Classical (frequentist) inference

Methods & models for fMRI data analysis

24 October 2017

Klaas Enno Stephan

With many thanks for slides & images to:

FIL Methods group

Translational Neuromodeling Unit

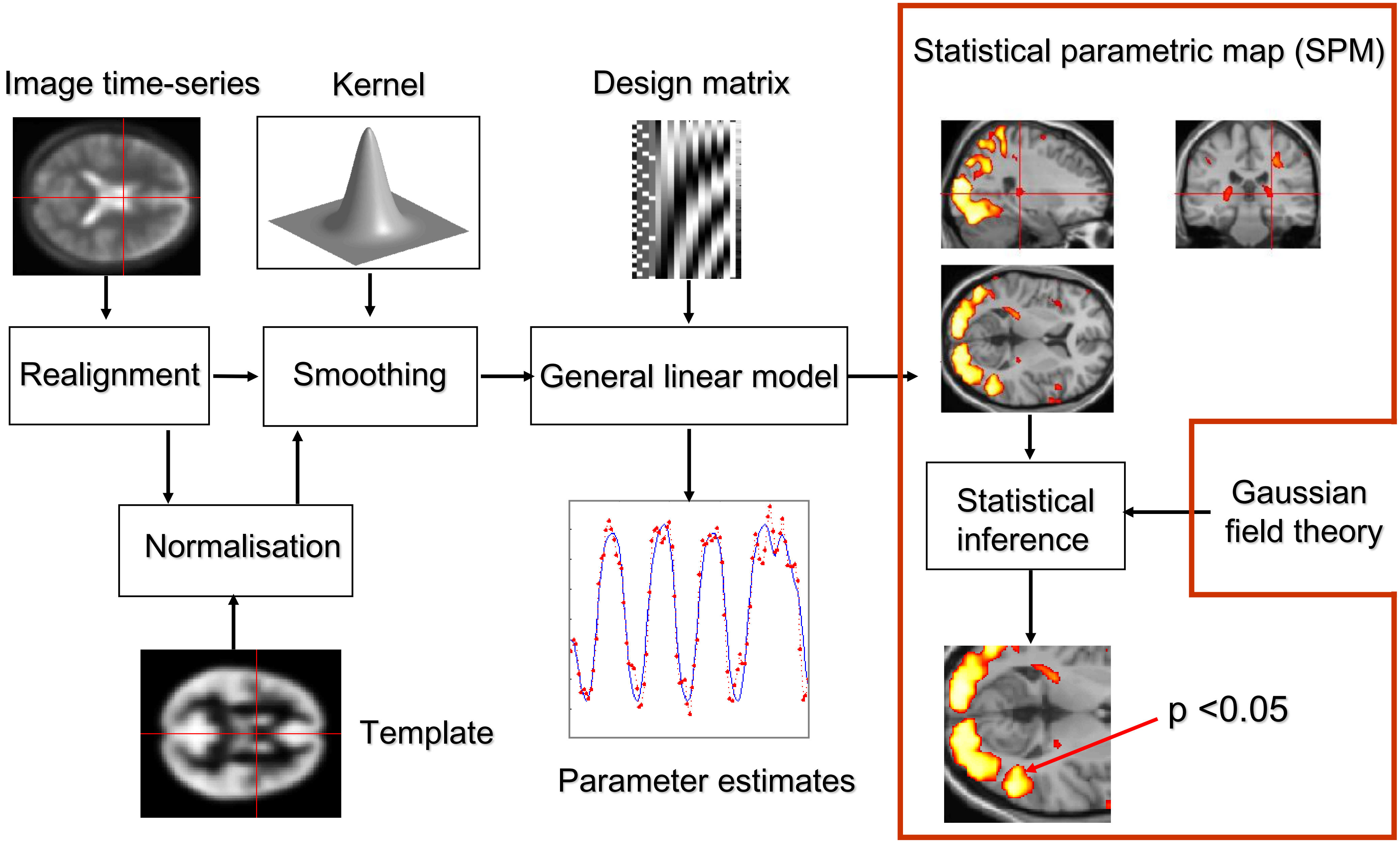
Universität Zürich[UZH]

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

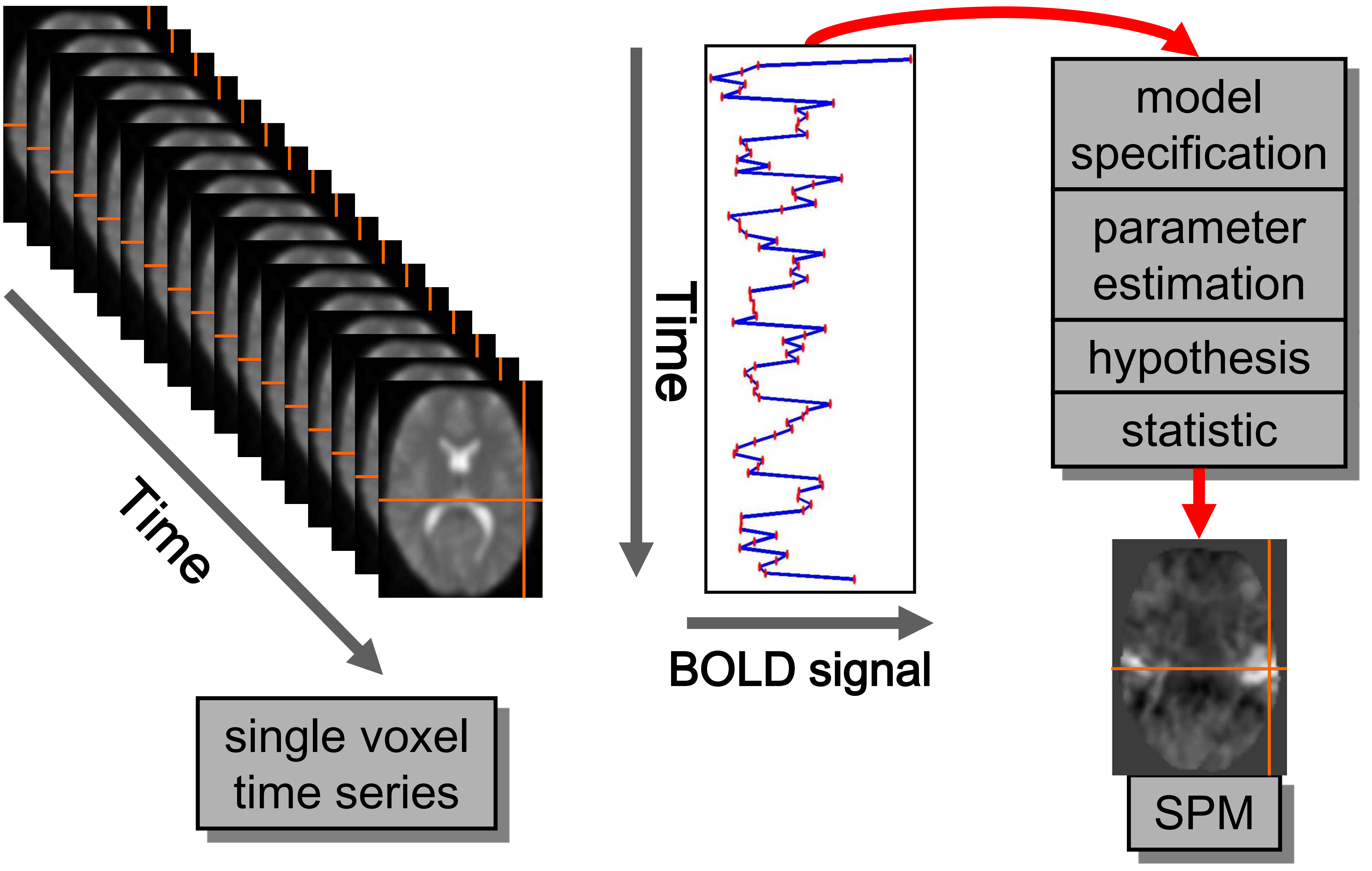
Overview

- A recap of model specification and parameter estimation
- Hypothesis testing
- Contrasts and estimability
    - *T*-tests
    - *F*-tests
- Design orthogonality

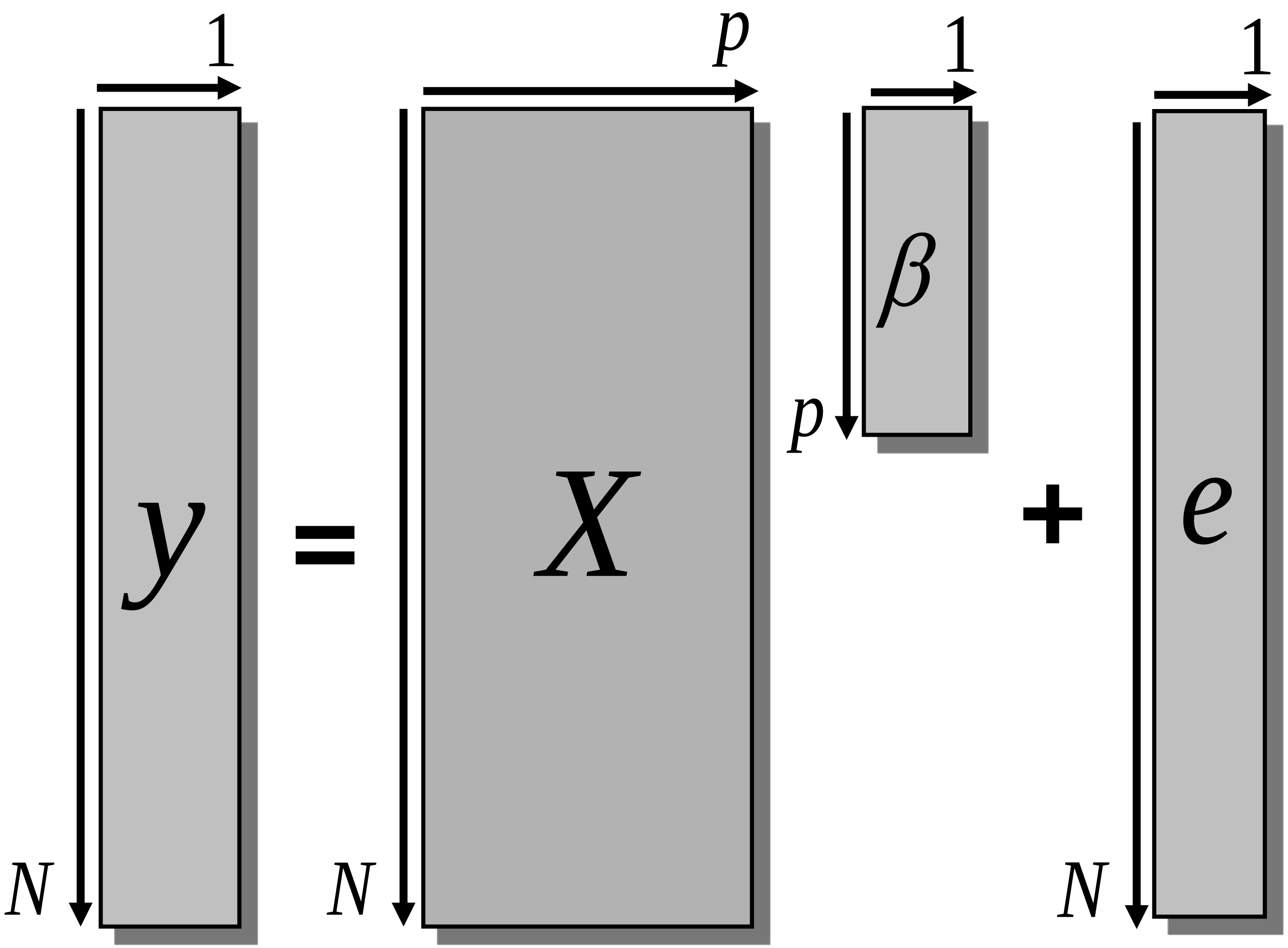
Overview of SPM



Voxel-wise time series analysis



Mass-univariate analysis: voxel-wise GLM



$$y = X\beta + e$$

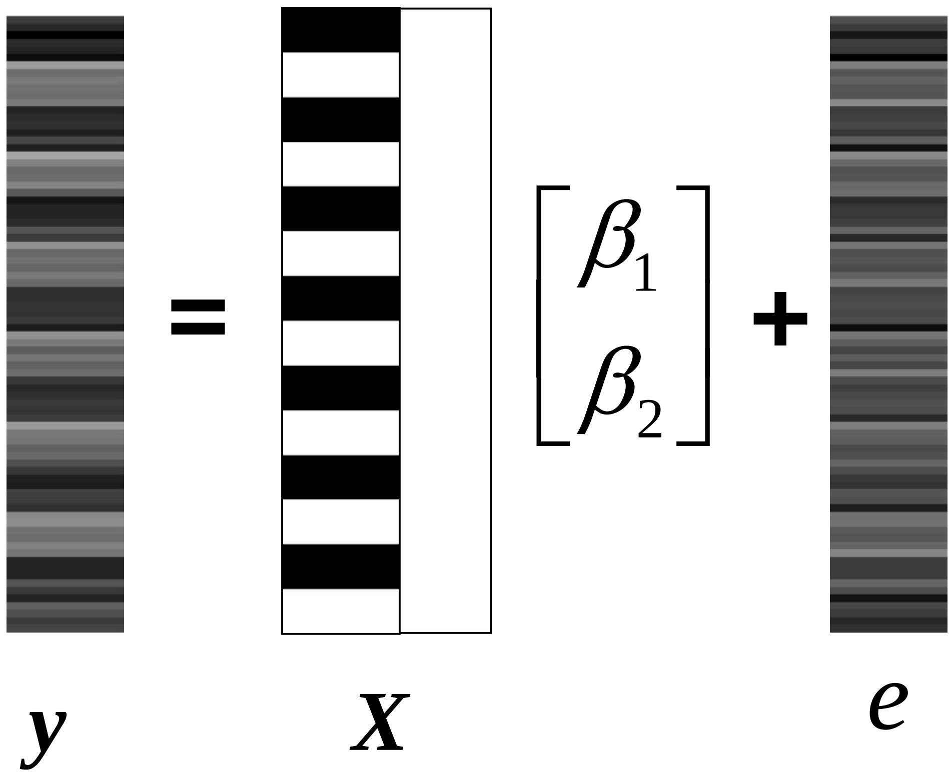
$$e \sim N(0, \sigma^2 I)$$

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.

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):

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

OLS parameter estimation

The Ordinary Least Squares (OLS) estimators are: $\hat{\beta} = (X^T X)^{-1} X^T y$

These estimators minimise $\sum e_t^2 = e^T e$. They are found solving either

$$\frac{\partial\left(\sum e_t^2\right)}{\partial \hat{\beta}_t} = 0$$

or

$$X^T e = 0$$

Under i.i.d. assumptions, the OLS estimates correspond to ML estimates:

$$e \sim N(0, \sigma^2 I) \longrightarrow Y \sim N(X\beta, \sigma^2 I)$$

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

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

NB: precision of our estimates depends on design matrix!

Maximum likelihood (ML) estimation

probability density function (θ fixed!)

$$y \mapsto f(y \mid \theta)$$

likelihood function (y fixed!)

$$\theta \mapsto f(y \mid \theta)$$

$$\theta \mapsto L(\theta \mid y)$$

$$L(\theta \mid y) = f(y \mid \theta)$$

ML estimator

$$\hat{\theta} = \arg\max_{\theta} L(\theta \mid y)$$

For $cov(e)=\sigma^2 I$, the ML estimator is equivalent to the OLS estimator:

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

OLS

For $cov(e)=\sigma^2 V$, the ML estimator is equivalent to a weighted least squares (WLS) estimate (with $W=V^{-1/2}$):

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

WLS

Bonus material: t-statistic based on ML estimates in SPM

$$Wy = WX\beta + We$$

$$\hat{\beta} = (WX)^{+} Wy$$

$$c = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$$

$$t = \frac{c^T \hat{\beta}}{\hat{std}(c^T \hat{\beta})}$$

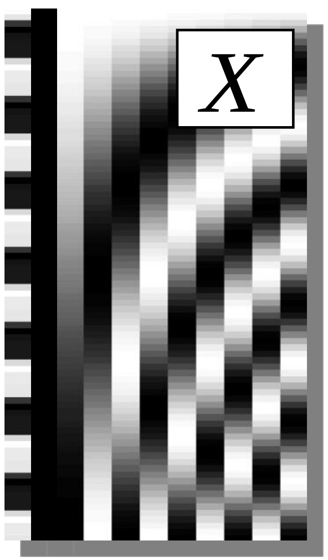
$$\hat{std}(c^T \hat{\beta}) = \sqrt{\hat{\sigma}^2 c^T (WX)^{+} (WX)^{+T} c}$$

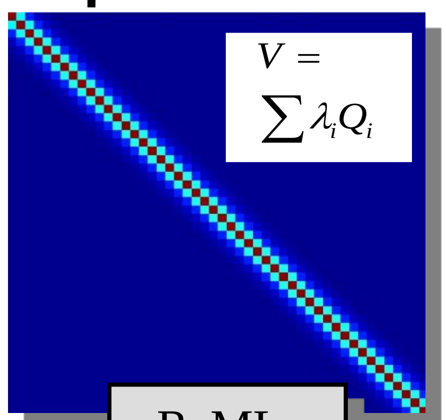
$$W = V^{-1/2}$$

$$\sigma^2 V = Cov(e)$$

$$\hat{\sigma}^2 = \frac{\sum \left(Wy - WX\hat{\beta}\right)^2}{tr(R)}$$

$$R = I - WX(WX)^{+}$$





ReML-estimates

For brevity:

$$(WX)^{+} = (X^T WX)^{-1} X^T$$

Statistic

- A **statistic** is the result of applying a mathematical function to a **sample** (set of data).

- More formally, a **statistic** is a function of a sample where the function itself is independent of the sample's distribution.
  (The term is used both for the function and for the value of the function on a given sample.)

- A statistic is distinct from an unknown statistical **parameter**, which is a population property and can only be estimated approximately from a sample.

- A statistic used to estimate a parameter is called an **estimator**.
  For example, the sample mean is a statistic and an estimator for the population mean, which is a parameter.

Hypothesis testing

To test an hypothesis, we construct a "test statistic".

- **"Null hypothesis" H_0 = "there is no effect"** $\Rightarrow$ $c^T\beta = 0$

  This is what we want to disprove.

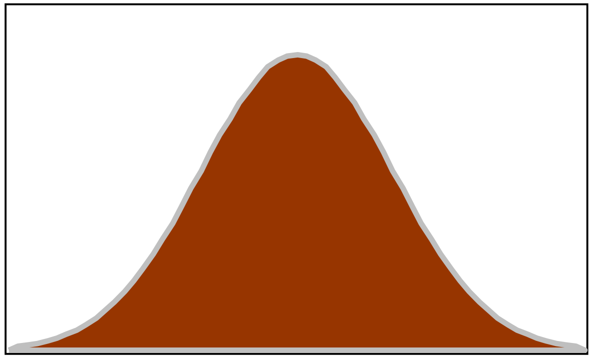
  ⇨ The "alternative hypothesis" H_1 represents the outcome of interest.

- **The test statistic T**

  The test statistic summarises the evidence for H_0.

  Typically, the test statistic is small in magnitude when H_0 is true and large when H_0 is false.

  ⇨ We need to know the distribution of T under the null hypothesis.

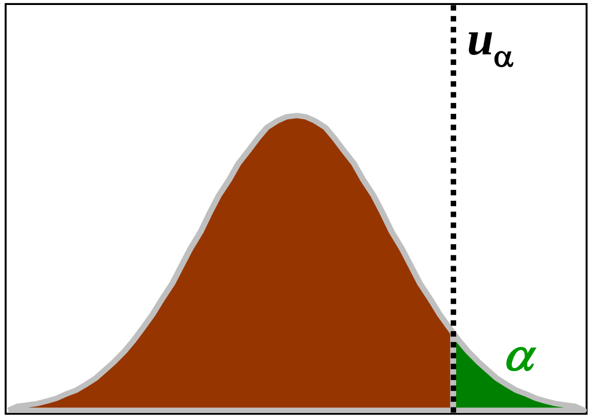


Null Distribution of T

Hypothesis testing

- **Type I Error α:**
  Acceptable *false positive rate* α.
  Threshold u_α controls the false positive rate

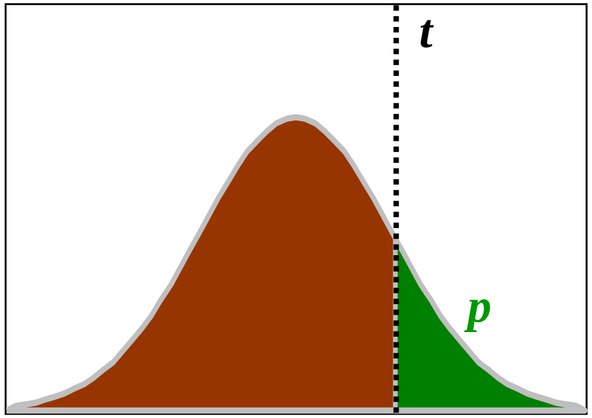
$$\alpha = p(T > u_\alpha \mid H_0)$$



Null Distribution of T

- **Observation of test statistic t, a realisation of T:**
  A *p*-value summarises evidence against H_0.
  This is the probability of observing t, or a more extreme value, under the null hypothesis:

$$p(T \geq t \mid H_0)$$



Null Distribution of T

- **The conclusion about the hypothesis:**
  We reject H_0 in favour of H_1 if $t > u_\alpha$

Types of error

Actual condition

Test result

	H_0 true	H_0 false
Reject H_0	**False positive (FP)** **Type I error α**	**True positive (TP)**
Failure to reject H_0	**True negative (TN)**	**False negative (FN)** **Type II error β**

specificity: 1-α
= TN / (TN + FP)
= proportion of actual negatives which are correctly identified

sensitivity (power): 1-β
= TP / (TP + FN)
= proportion of actual positives which are correctly identified

One cannot accept the null hypothesis (one can only fail to reject it)

Absence of evidence is not evidence of absence!

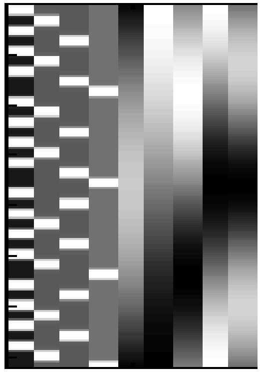
If we do not reject H_0, then all can we say is that there is not enough evidence in the data to reject H_0. This does not mean that we can accept H_0.

What does this mean for neuroimaging results based on classical statistics?

A failure to find an “activation” in a particular area does not mean we can conclude that this area is not involved in the process of interest.

Contrasts

- We are usually not interested in the whole β vector.
- A contrast $c^T\beta$ selects a specific effect of interest:
  ⇨ a contrast vector c is a vector of length p
  ⇨ $c^T\beta$ is a linear combination of regression coefficients β



$c^T = [1\ 0\ 0\ 0\ 0\ \ldots]$

$c^T\beta = \mathbf{1}\beta_1 + \mathbf{0}\beta_2 + \mathbf{0}\beta_3 + \mathbf{0}\beta_4 + \mathbf{0}\beta_5 + \ldots$

$c^T = [0\ {-1}\ 1\ 0\ 0\ \ldots]$

$c^T\beta = \mathbf{0}\beta_1 + \mathbf{-1}\beta_2 + \mathbf{1}\beta_3 + \mathbf{0}\beta_4 + \mathbf{0}\beta_5 + \ldots$

- Under i.i.d assumptions:

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

NB: the precision of our estimates depends on design matrix and the chosen contrast !

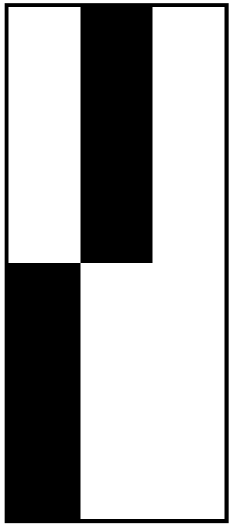
Bonus material: Estimability of parameters

- If X is not of **full rank** then different parameters can give identical predictions, i.e. $X\beta_1 = X\beta_2$ with $\beta_1 \neq \beta_2$.
- The parameters are therefore ‘non-unique’, ‘non-identifiable’ or ‘**non-estimable**’.
- For such models, X^TX is not invertible so we must resort to generalised inverses (SPM uses the Moore-Penrose **pseudo-inverse**).
- This gives a parameter vector that has the smallest norm of all possible solutions.
- However, even when parameters are non-estimable, certain contrasts may well be!

One-way ANOVA

(unpaired two-sample *t*-test)

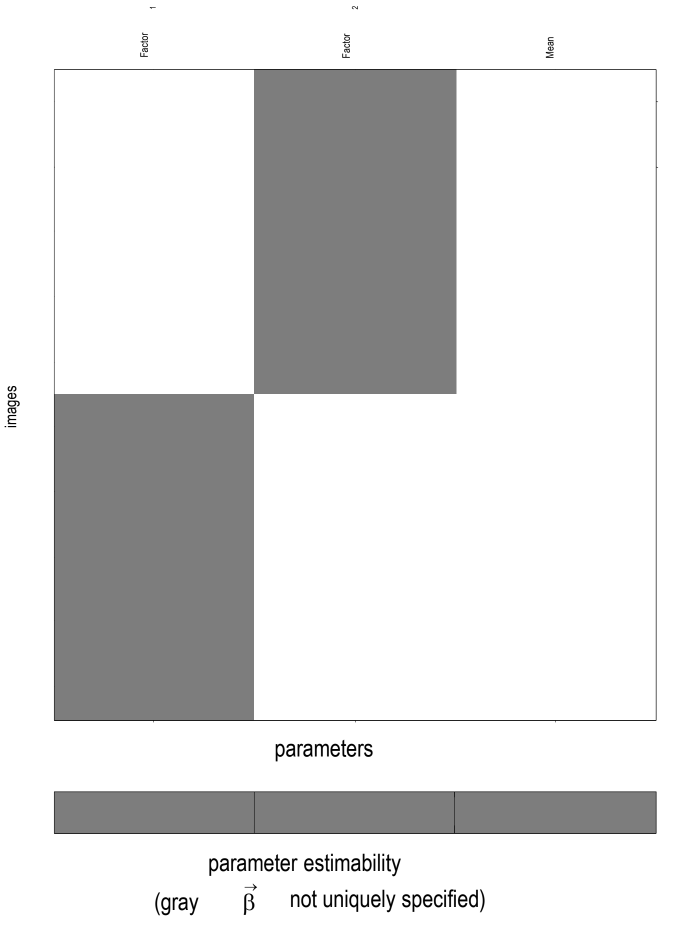
$$\begin{matrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{matrix}$$



Rank(X)=2

Bonus material: Estimability of parameters

- If X is not of **full rank** then different parameters can give identical predictions, i.e. $X\beta_1 = X\beta_2$ with $\beta_1 \neq \beta_2$.
- The parameters are therefore 'non-unique', 'non-identifiable' or '**non-estimable**'.
- For such models, $X^T X$ is not invertible so we must resort to generalised inverses (SPM uses the Moore-Penrose **pseudo-inverse**).
- This gives a parameter vector that has the smallest norm of all possible solutions.
- However, even when parameters are non-estimable, certain contrasts may well be!



Bonus material: Estimability of contrasts

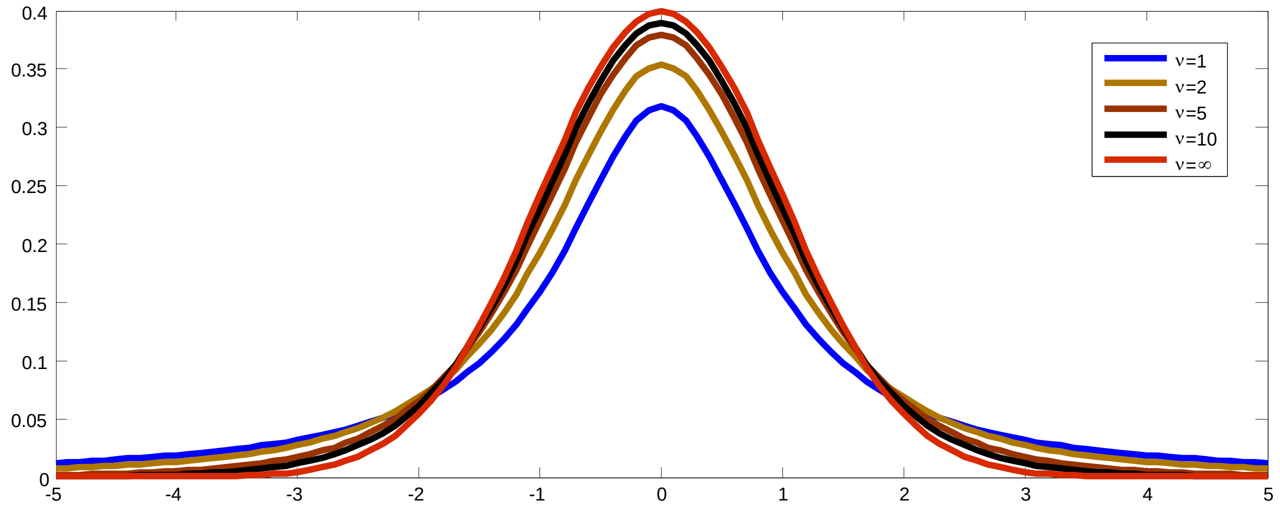
- Linear dependency: there is one contrast vector q for which $Xq = 0$.
- Thus: $y = X\beta+Xq+e = X(\beta+q)+e$
- So if we test $c^T\beta$ we also test $c^T(\beta+q)$, thus an estimable contrast has to satisfy $c^Tq = 0$.

- In the above ANOVA example (unpaired t-test), any contrast that is orthogonal to [1 1 -1] is estimable:

[1 0 0], [0 1 0], [0 0 1] are not estimable.
[1 0 1], [0 1 1], [1 -1 0], [0.5 0.5 1] are estimable.

Student's t-distribution

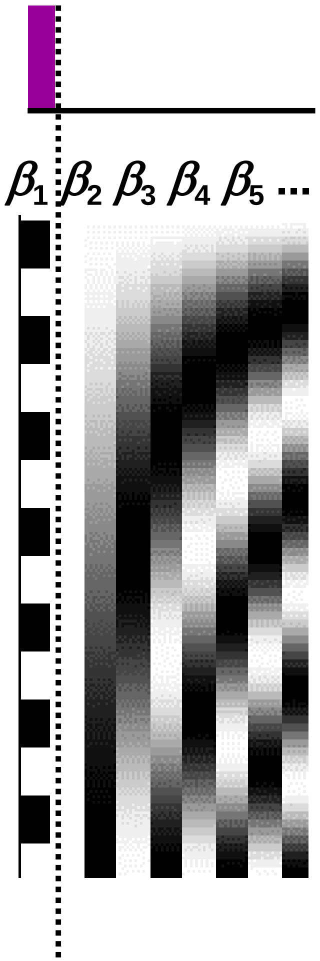
- first described by William Sealy Gosset, a statistician at the Guinness brewery at Dublin
- t-statistic is a signal-to-noise measure: t = effect / standard deviation
- t-distribution is an approximation to the normal distribution for small samples
- t-contrasts are simply linear combinations of the betas
  $\Rightarrow$ the t-statistic does not depend on the scaling of the regressors or on the scaling of the contrast
- Unilateral test: $H_0 : c^T \beta = 0$ vs. $H_1 : c^T \beta > 0$



Probability density function of Student's t distribution

t-contrasts – SPM{t}

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



Question: box-car amplitude > 0 ?

=

$H_1 = c^T\beta > 0$?

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

Test statistic:

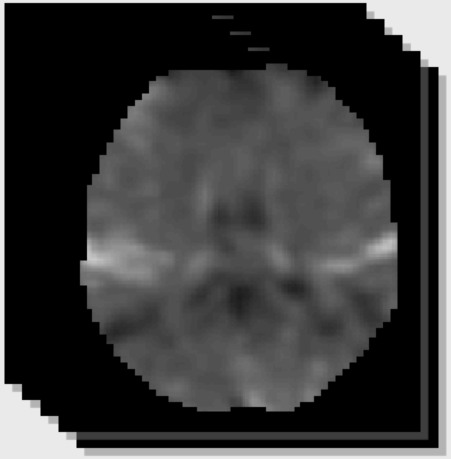
$$t = \frac{\textit{contrast}\text{ of estimated parameters}}{\sqrt{\text{variance estimate}}}$$

$p(y \mid c^T\hat{\beta} = 0)$

$$t = \frac{c^T\hat{\beta}}{s\hat{t}d(c^T\hat{\beta})} = \frac{c^T\hat{\beta}}{\sqrt{\hat{\sigma}^2 c^T\left(X^T X\right)^{-1} c}} \sim t_{N-p}$$

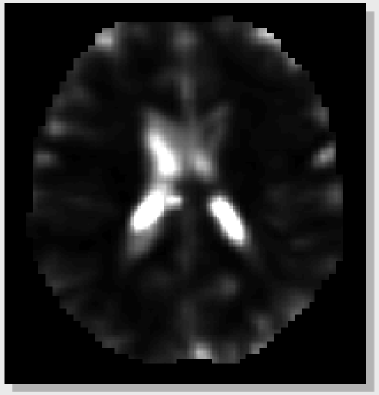
t-contrasts in SPM

For a given contrast *c*:



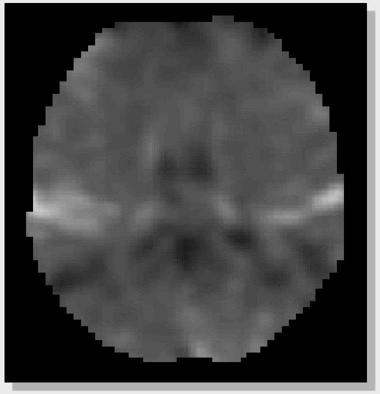
beta_???? images

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



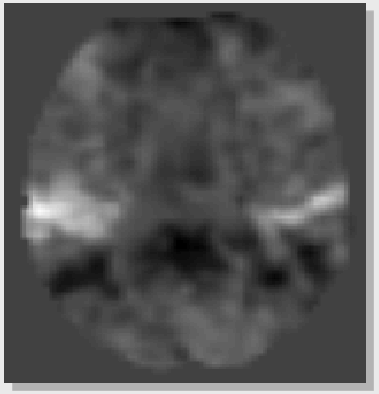
ResMS image

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



con_???? image

$$c^T \hat{\beta}$$



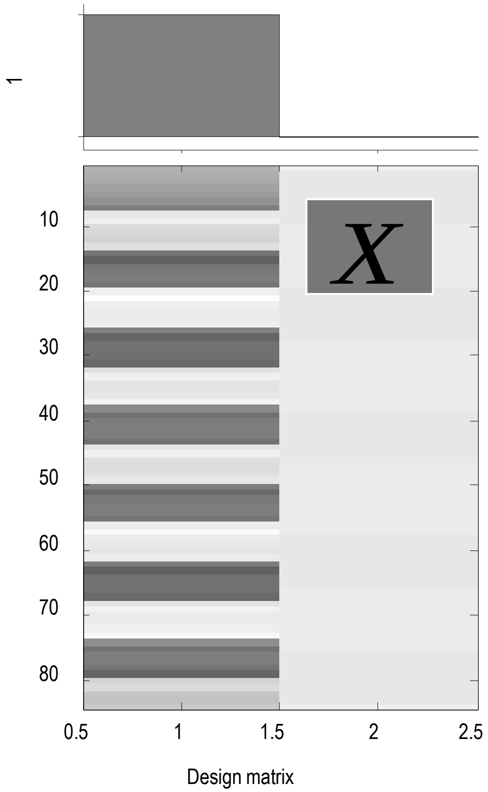
spmT_???? image

SPM{*t*}

t-contrast: a simple example

Passive word listening versus rest

$c^T = [\ 1 \quad 0\ ]$

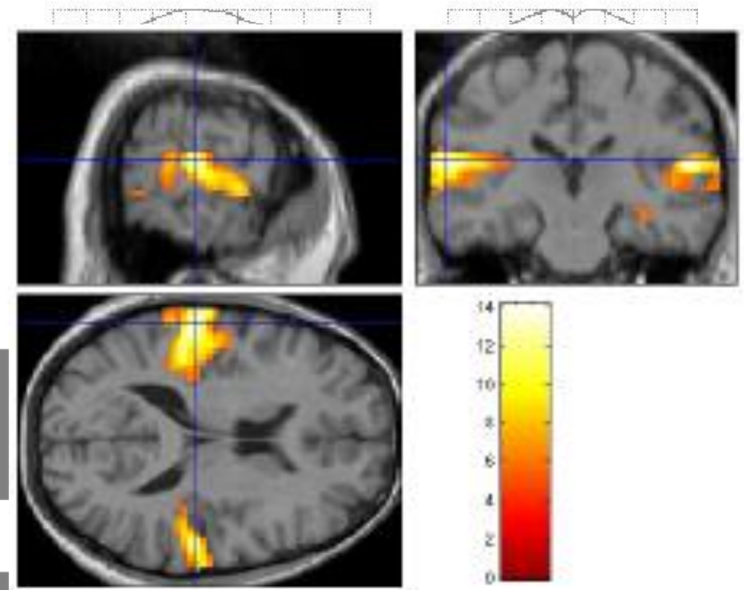


Q: activation during listening ?

Null hypothesis: $\beta_1 = 0$

$$t = \frac{c^T \hat{\beta}}{Std(c^T \hat{\beta})}$$

$$p(y \mid c^T \hat{\beta} = 0)$$



SPMresults:
Height threshold T = 3.2057 {p<0.001}

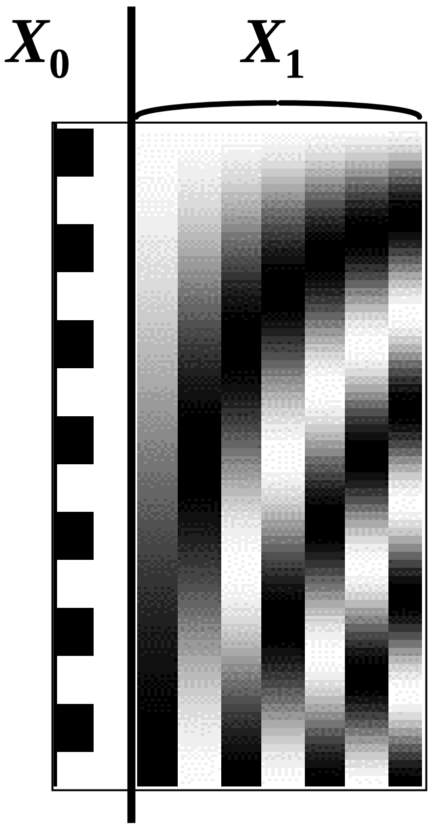
Statistics: ***p-values adjusted for search volume***

set-level		cluster-level			voxel-level					mm mm mm		
p	c	$p_{corrected}$	k_E	$p_{uncorrected}$	$p_{FWE-corr}$	$p_{FDR-corr}$	T	$(Z_{\equiv})$	$p_{uncorrected}$			
0.000	10	0.000	520	0.000	0.000	0.000	13.94	Inf	0.000	-63	-27	15
					0.000	0.000	12.04	Inf	0.000	-48	-33	12
					0.000	0.000	11.82	Inf	0.000	-66	-21	6
		0.000	426	0.000	0.000	0.000	13.72	Inf	0.000	57	-21	12
					0.000	0.000	12.29	Inf	0.000	63	-12	-3
					0.000	0.000	9.89	7.83	0.000	57	-39	6
		0.000	35	0.000	0.000	0.000	7.39	6.36	0.000	36	-30	-15
		0.000	9	0.000	0.000	0.000	6.84	5.99	0.000	51	0	48
		0.002	3	0.024	0.001	0.000	6.36	5.65	0.000	-63	-54	-3
		0.000	8	0.001	0.001	0.000	6.19	5.53	0.000	-30	-33	-18
		0.000	9	0.000	0.003	0.000	5.96	5.36	0.000	36	-27	9
		0.005	2	0.058	0.004	0.000	5.84	5.27	0.000	-45	42	9
		0.015	1	0.166	0.022	0.000	5.44	4.97	0.000	48	27	24
		0.015	1	0.166	0.036	0.000	5.32	4.87	0.000	36	-27	42

F-test: the extra-sum-of-squares principle

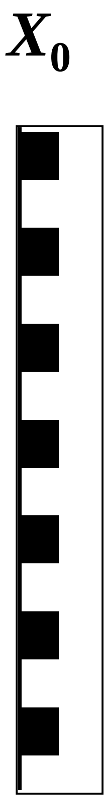
Model comparison: Full vs. reduced model

Null Hypothesis H_0: True model is X_0 (reduced model)



$\longrightarrow$ **RSS** $\sum \hat{e}^2_{full}$

Full model ($X_0 + X_1$)?



$\longrightarrow$ **RSS_0** $\sum \hat{e}^2_{reduced}$

Or reduced model?

F-statistic: ratio of unexplained variance under X_0 and total unexplained variance under the full model

$$F \propto \frac{RSS_0 - RSS}{RSS}$$

$$F \propto \frac{ESS}{RSS} \sim F_{\nu_1,\nu_2}$$

$\nu_1 = \mathrm{rank}(X) - \mathrm{rank}(X_0)$
$\nu_2 = N - \mathrm{rank}(X)$

F-test: multidimensional contrasts – SPM{F}

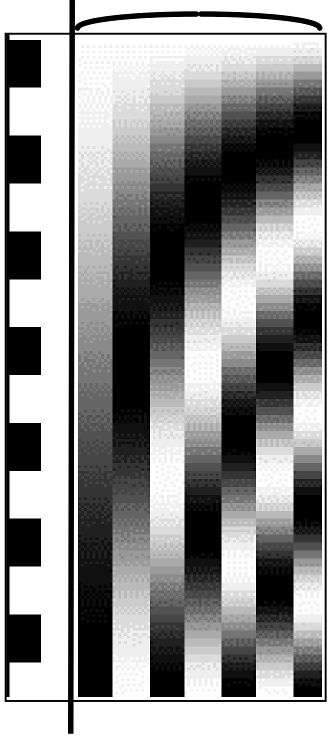
Tests multiple linear hypotheses:

H_0: True model is X_0

H_0: $\beta_3 = \beta_4 = \ldots = \beta_9 = 0$

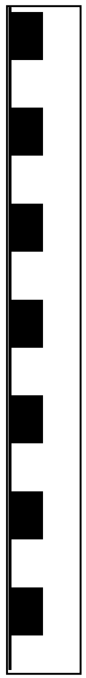
test H_0: $c^T\beta = 0$?

X_0 $\quad X_1$ $(\beta_{3\text{-}9})$



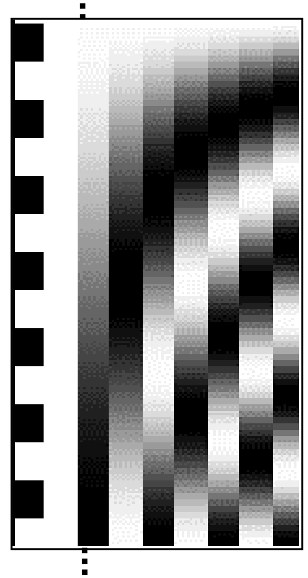
Full model?

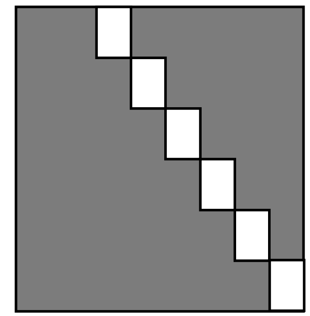
X_0

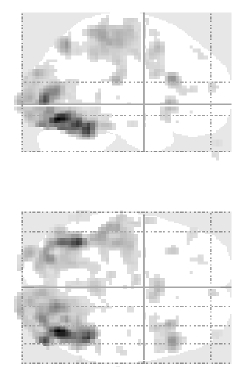


Reduced model?

$$c^T = \begin{matrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{matrix}$$







SPM{$F_{6,322}$}

F-test: a few remarks

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

- Hypotheses:

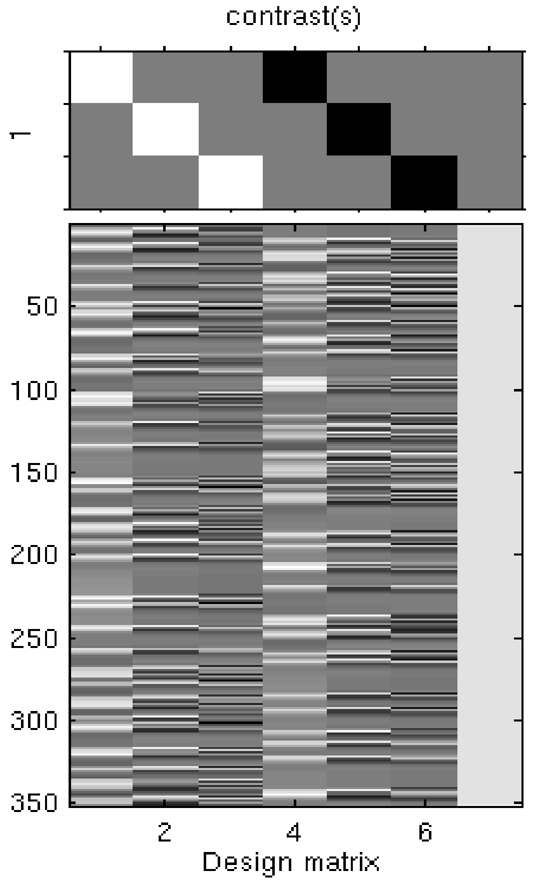
$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Null hypothesis H_0: $\beta_1 = \beta_2 = ... = \beta_p = 0$

Alternative hypothesis H_1: At least one $\beta_k \neq 0$

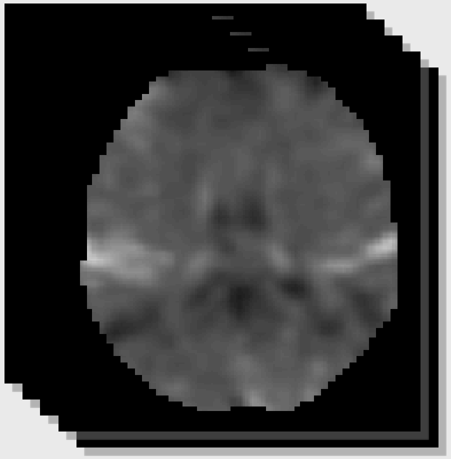
- F-tests are not directional:
  When testing a uni-dimensional contrast with an *F*-test, for example $\beta_1 - \beta_2$, the result will be the same as testing $\beta_2 - \beta_1$.

Bonus material: Differential F-contrasts



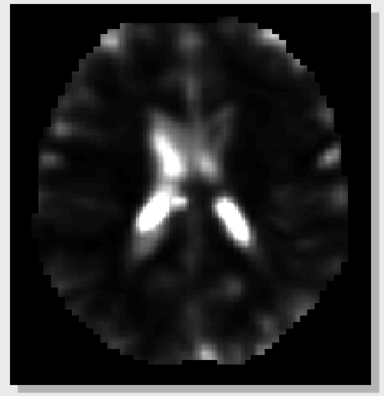
- equivalent to testing for effects that can be explained as a linear combination of the 3 differences
- useful when using informed basis functions and testing for overall shape differences in the HRF between two conditions

F-contrast in SPM



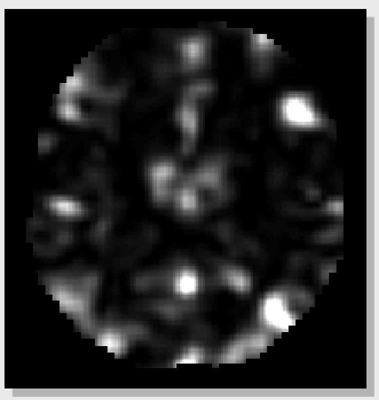
beta_???? images

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



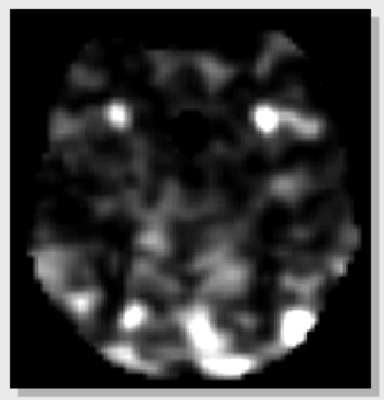
ResMS image

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



ess_???? images

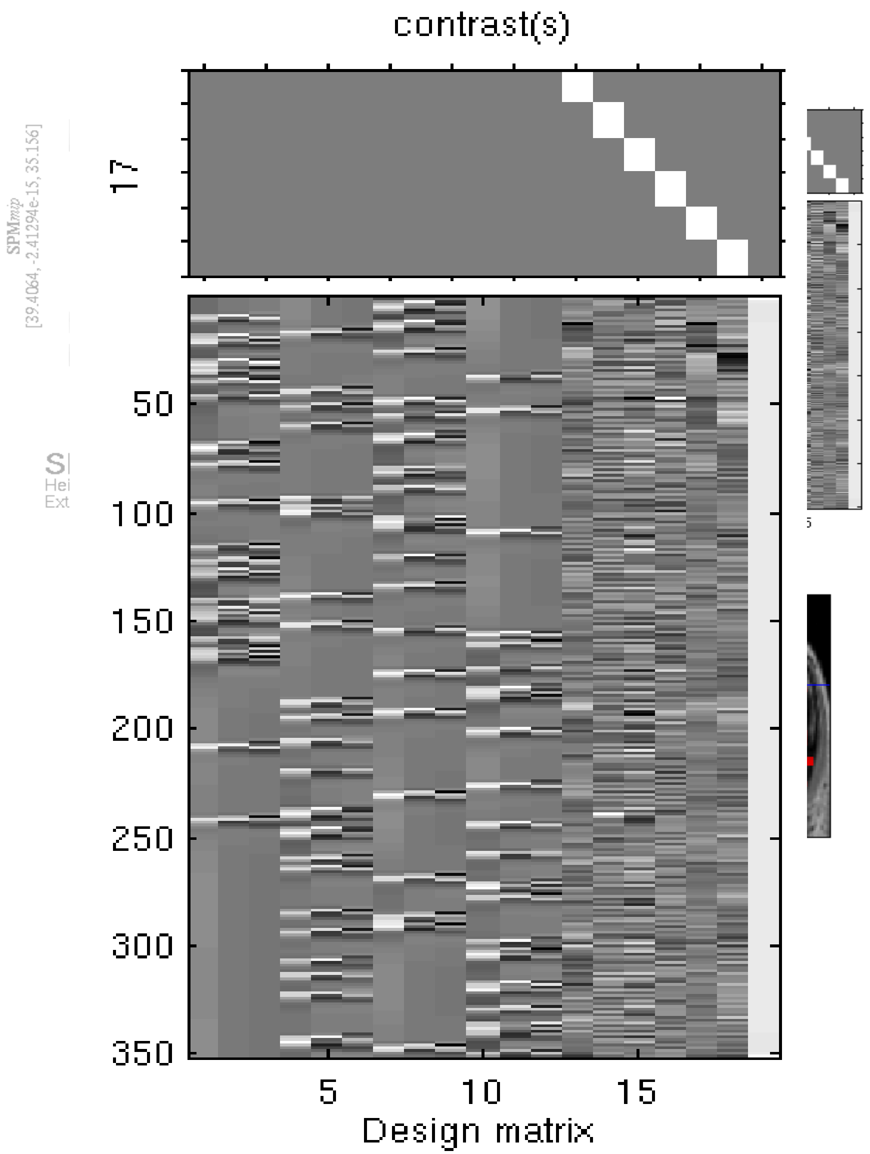
$$(RSS_0 \text{ - } RSS)$$



spmF_???? images

SPM{F}

F-test example: movement related effects

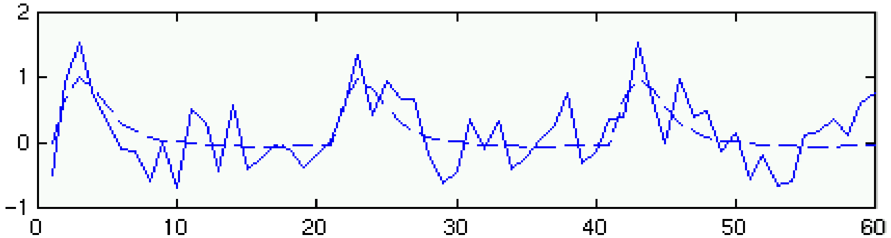


To assess movement-related activation:

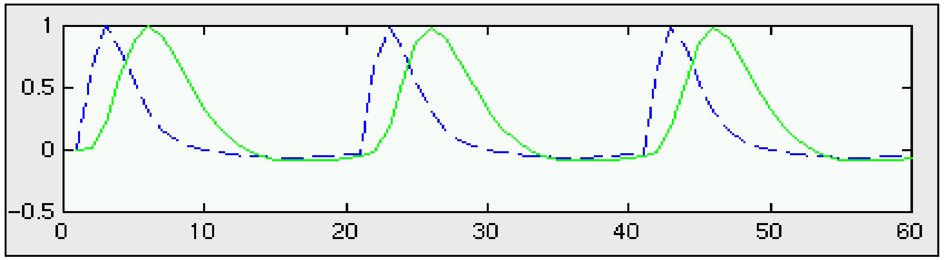
There is a lot of residual movement-related artifact in the data (despite spatial realignment), which tends to be concentrated near the boundaries of tissue types.

By including the realignment parameters in our design matrix, we can “regress out” linear components of subject movement, reducing the residual error, and hence improve our statistics for the effects of interest.

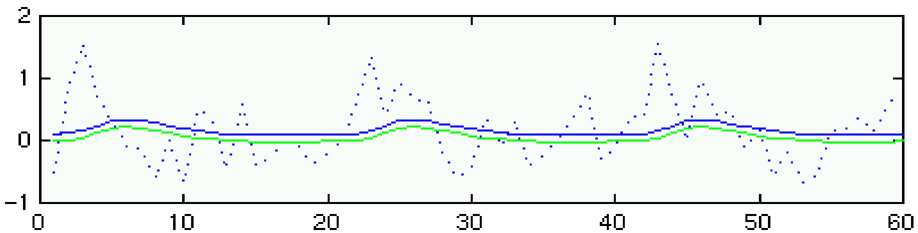
Example: a suboptimal model



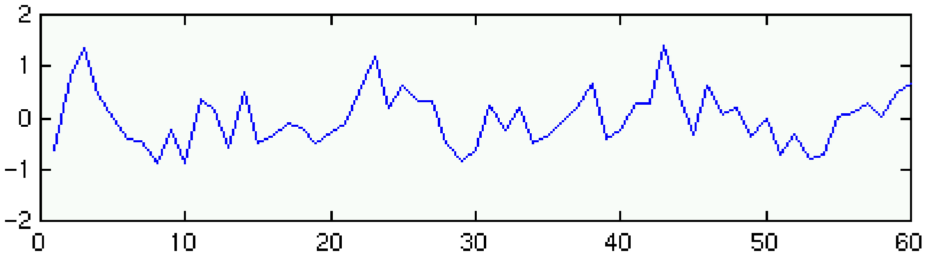
True signal (--) and observed signal



Model (green, peak at 6sec)
TRUE signal (blue, peak at 3sec)



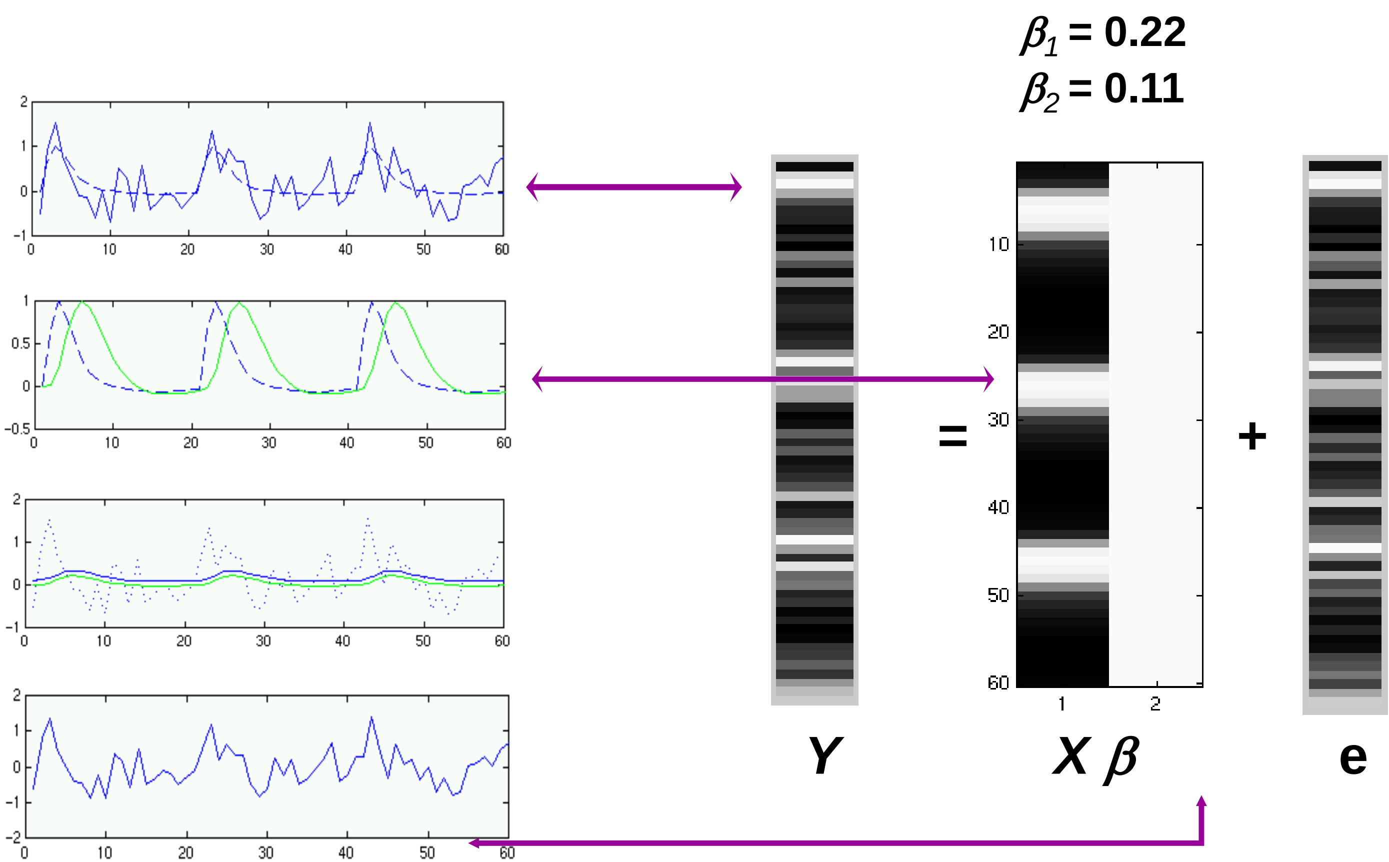
Fitting (β_1 = 0.2, β_2 (const.) = 0.11);
(here: blue solid line = total fit)



Residuals (still contain some signal)

⇨ Test for the green regressor not significant

Example: a suboptimal model

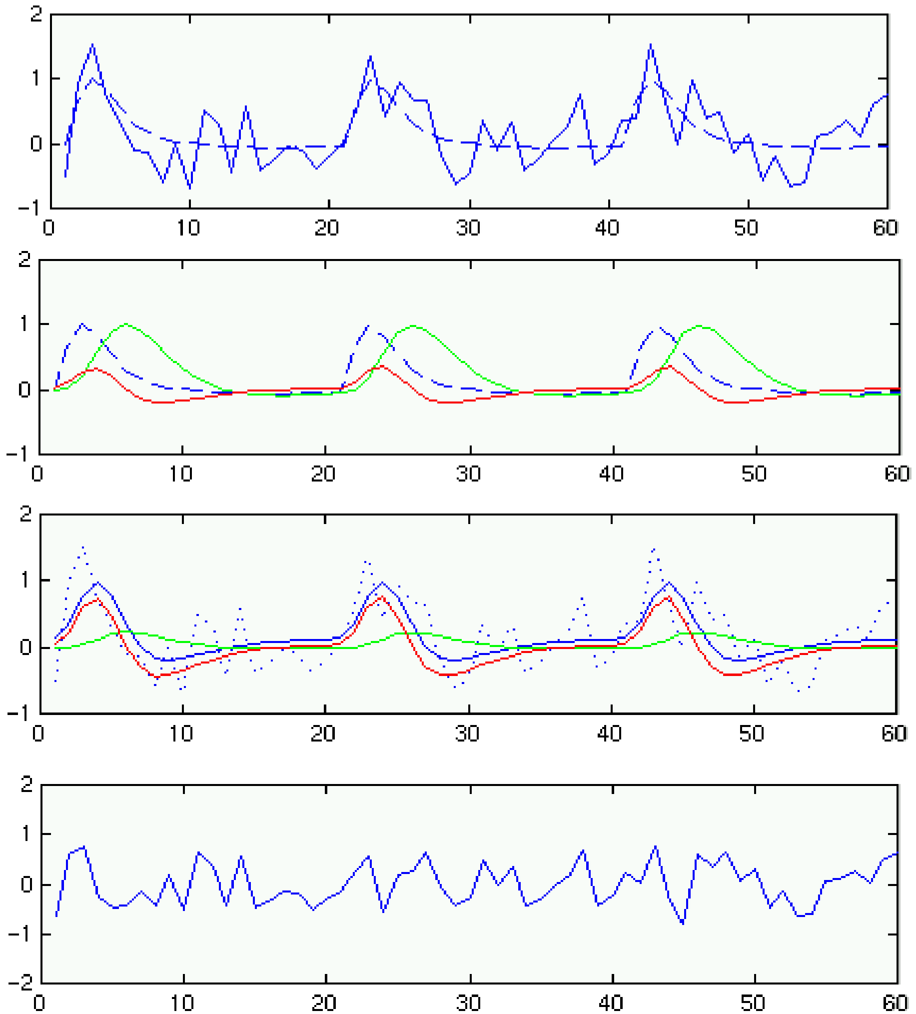


Residual Var. = 0.3

$p(Y|\ \mathbf{b}_1 = 0)$ ⇨
p-value = 0.1
(*t*-test)

$p(Y|\ \mathbf{b}_1 = 0)$ ⇨
p-value = 0.2
(*F*-test)

A better model



True signal + observed signal

Model (green and red)
and true signal (blue ---)
Red regressor: temporal derivative of the green regressor

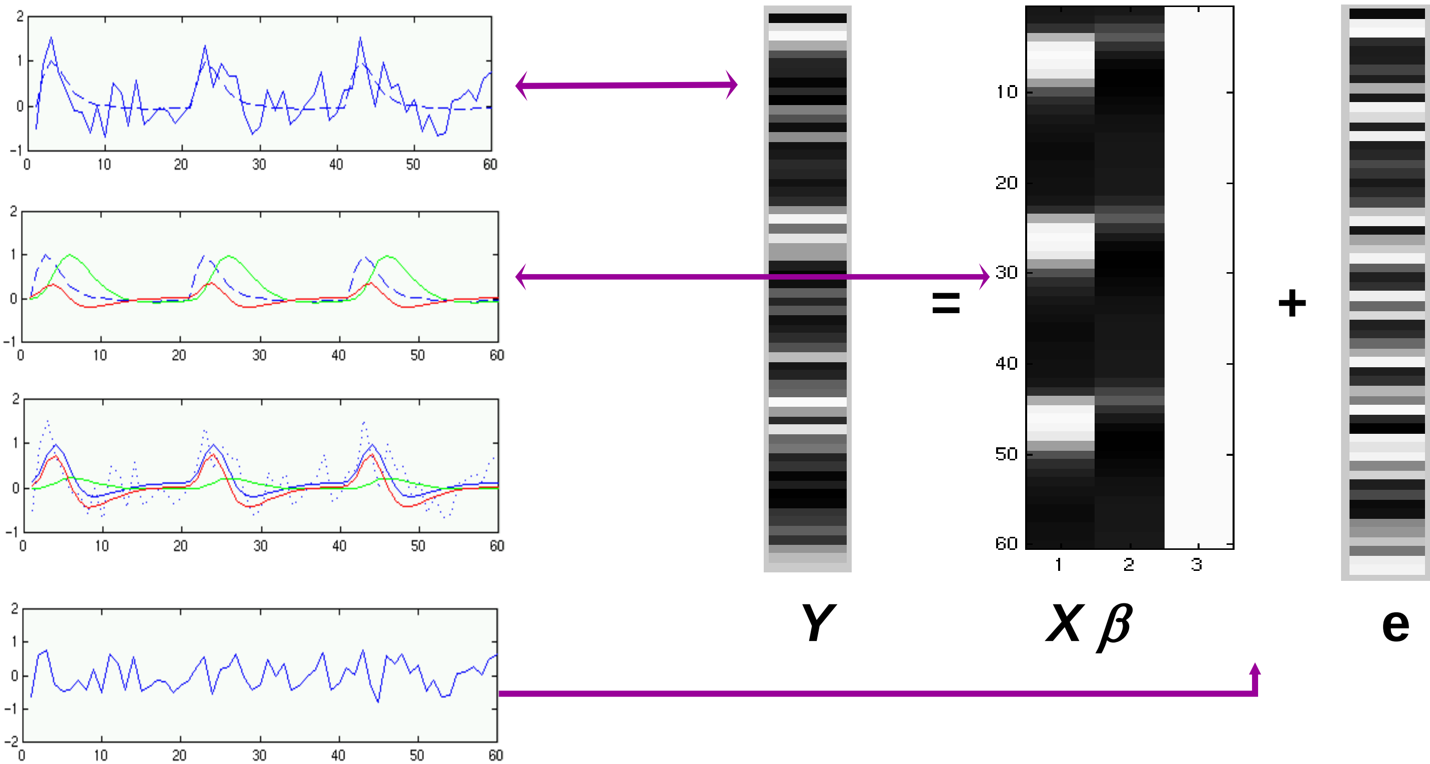
Total fit (blue)
and partial fit (green & red)
Adjusted and fitted signal

Residuals (less variance & structure)

- ⇨ *t*-test of the green regressor almost significant
- ⇨ *F*-test very significant
- ⇨ *t*-test of the red regressor very significant

A better model

β_1 = 0.22
β_2 = 2.15
β_3 = 0.11

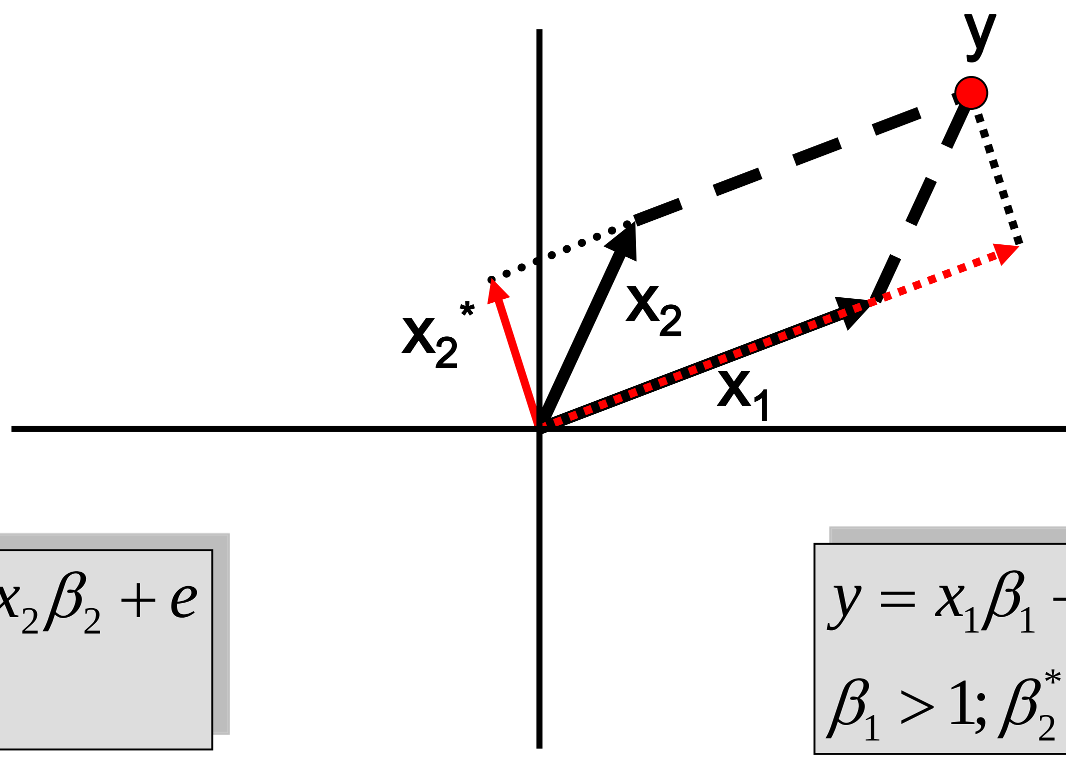


Residual Var. = 0.2

$p(Y|\, \mathbf{b}_1 = 0) \Rightarrow$
p-value = 0.07
(*t*-test)

$p(Y|\, \mathbf{b}_1 = 0, \mathbf{b}_2 = 0) \Rightarrow$
p-value = 0.000001
(*F*-test)

Recap from previous lecture: Correlation among regressors



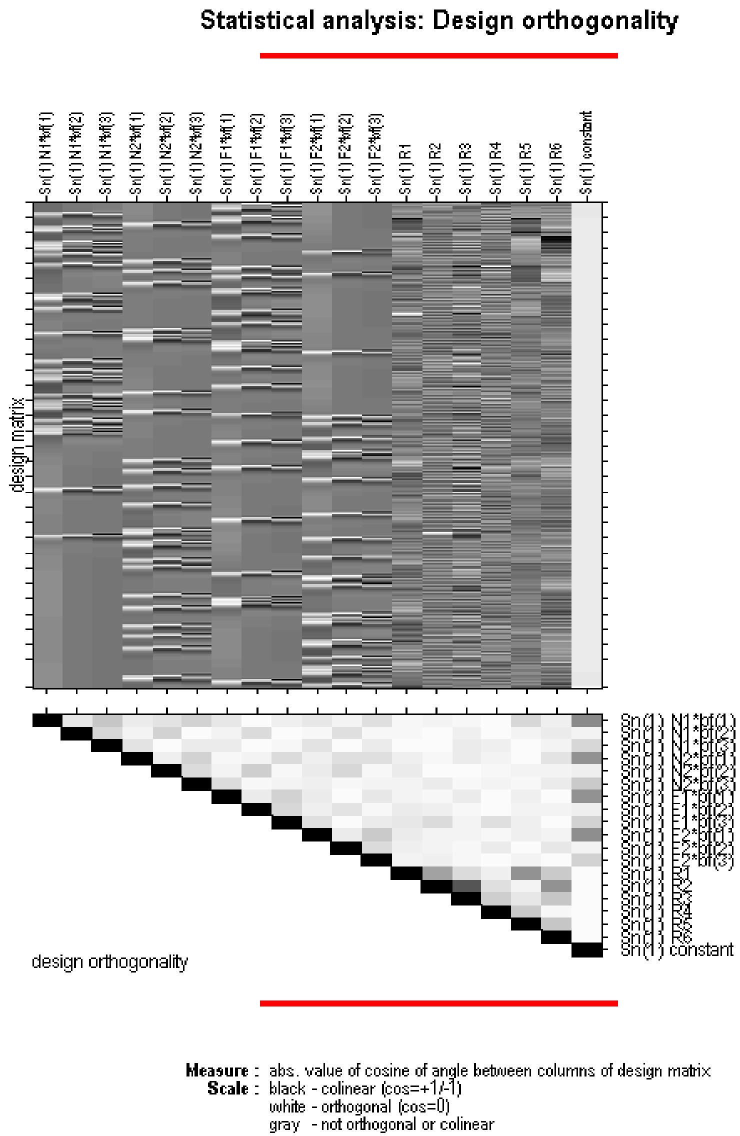
$$y = x_1\beta_1 + x_2\beta_2 + e$$
$$\beta_1 = \beta_2 = 1$$

$$y = x_1\beta_1 + x_2^*\beta_2^* + e$$
$$\beta_1 > 1; \beta_2^* = 1$$

Correlated regressors = explained variance is shared between regressors

When x_2 is orthogonalized with regard to x_1, only the parameter estimate for x_1 changes, not that for x_2!

Design orthogonality



- 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.

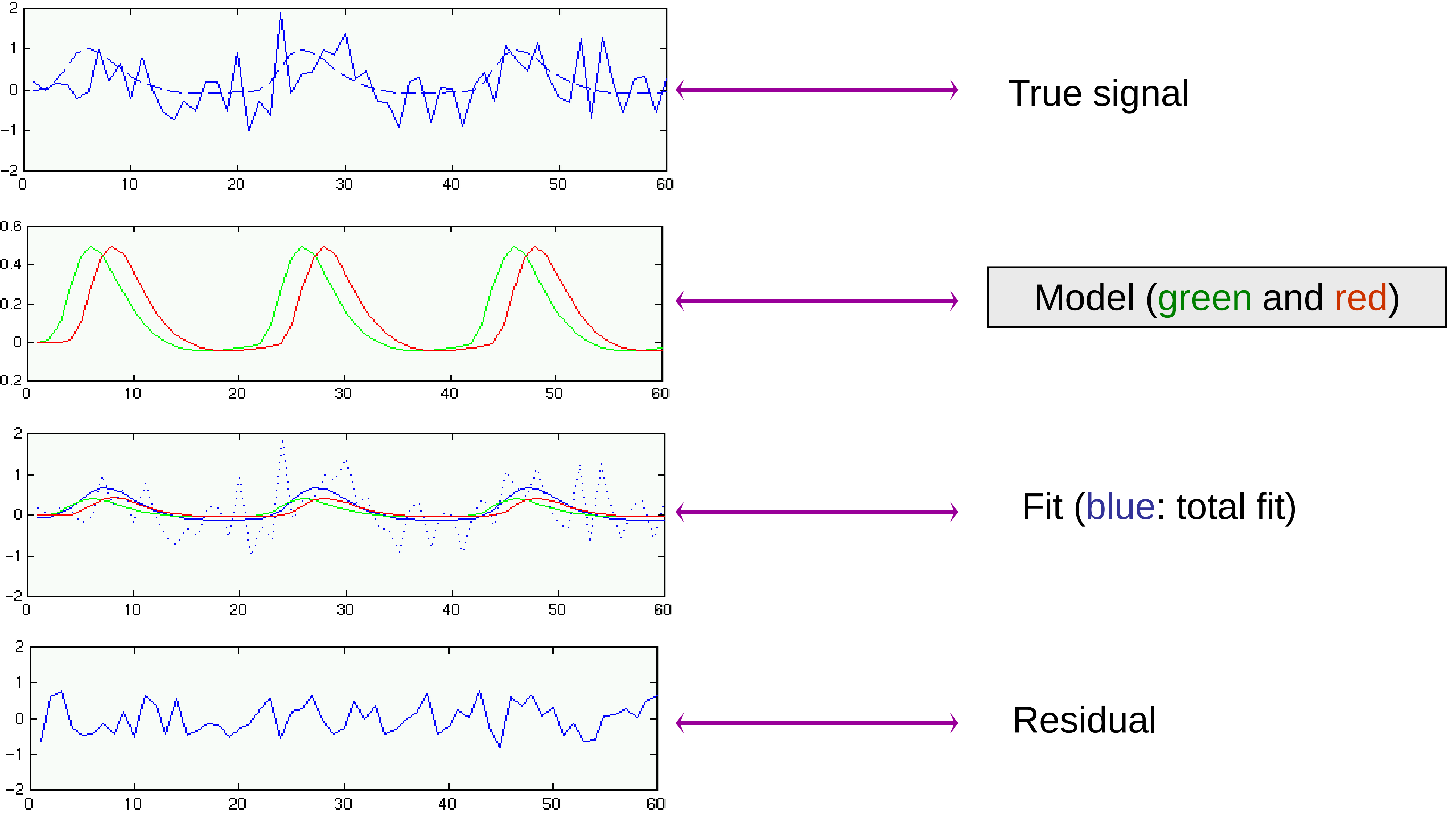
- The cosine of the angle between two vectors *a* and *b* is obtained by:

$$\cos\alpha = \frac{ab}{|a||b|}$$

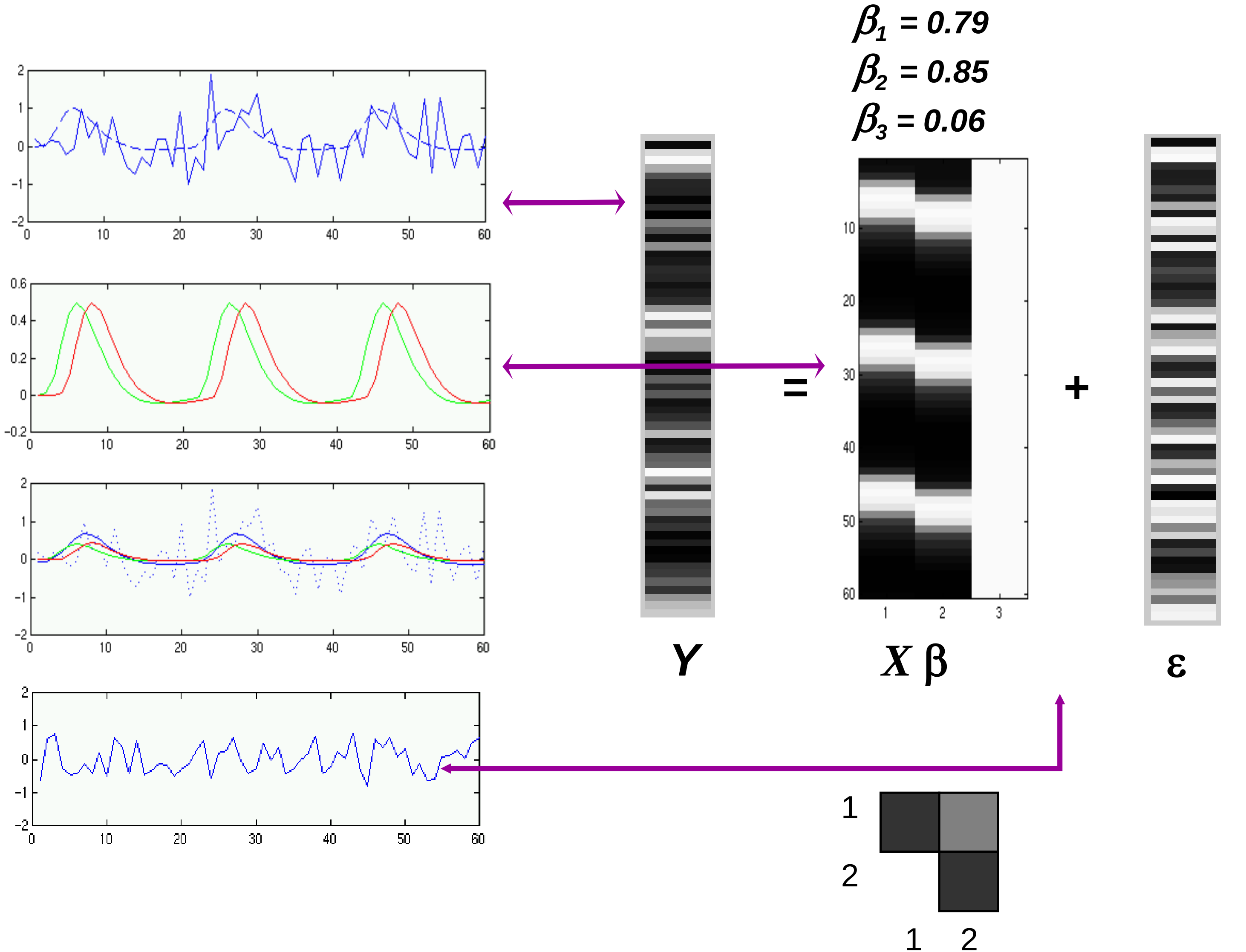
- For **zero-mean vectors**, the cosine of the angle between the vectors is the same as the **correlation** between the two variates:

$$\cos\alpha = corr_{a,b}$$

Correlated regressors



Correlated regressors



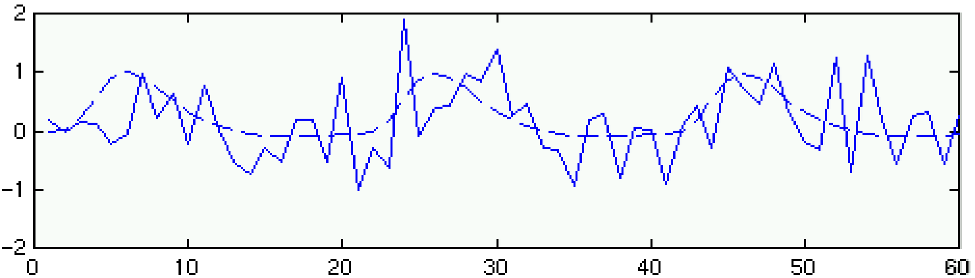
Residual var. = 0.3

$p(Y|\, \mathbf{b}_1 = 0) \Rightarrow$
p-value = 0.08
(*t*-test)

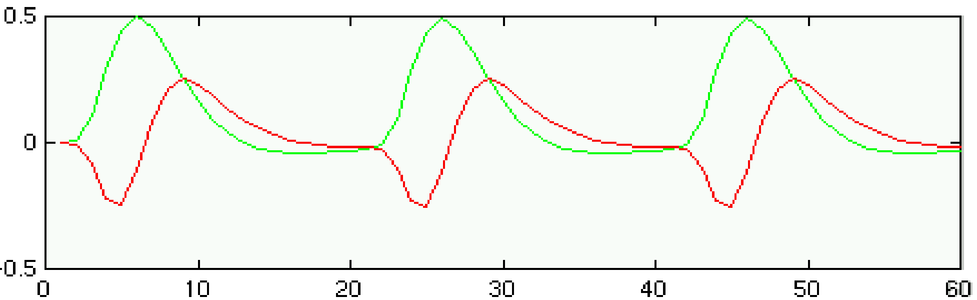
$P(Y|\, \mathbf{b}_2 = 0) \Rightarrow$
p-value = 0.07
(*t*-test)

$p(Y|\, \mathbf{b}_1 = 0, \mathbf{b}_2 = 0) \Rightarrow$
p-value = 0.002
(*F*-test)

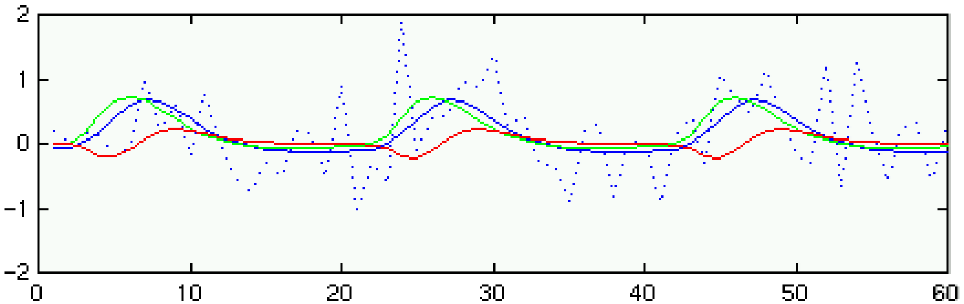
After orthogonalisation



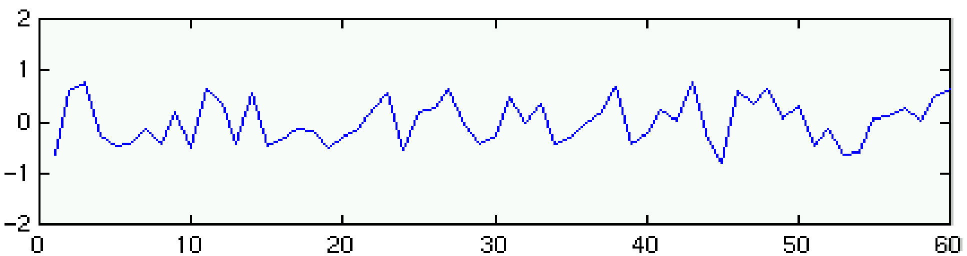
True signal



Model (green and red)
red regressor has been
orthogonalised with respect to the green one
⇔ remove everything that correlates with the green regressor



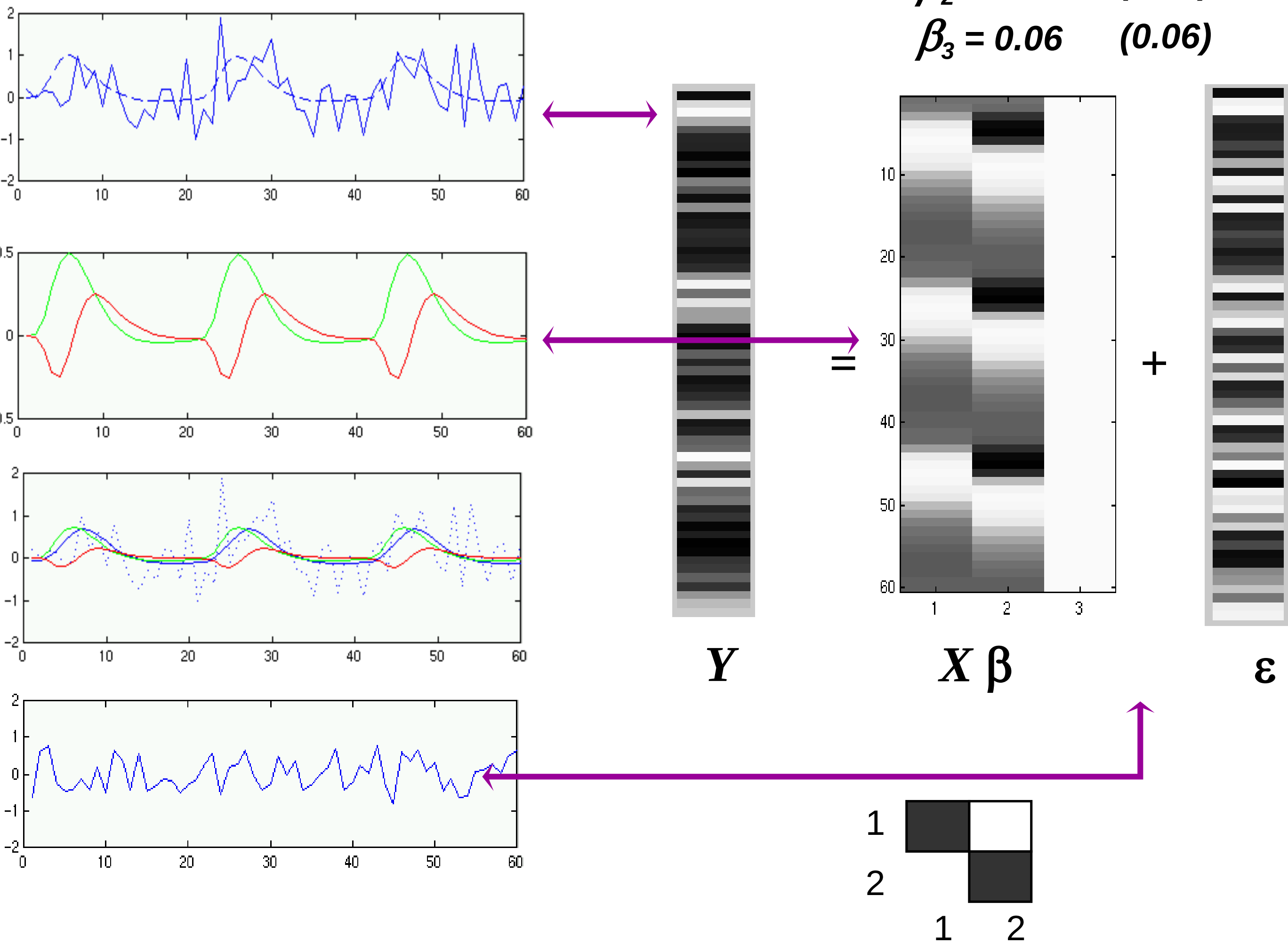
Fit (does not change)



Residuals (do not change)

After orthogonalisation

β_1 = 1.47 (0.79)
β_2 = 0.85 (0.85)
β_3 = 0.06 (0.06)



Residual var. = 0.3

$p(Y|\, \mathbf{b}_1 = 0)$
p-value = 0.0003
(t-test)

does change

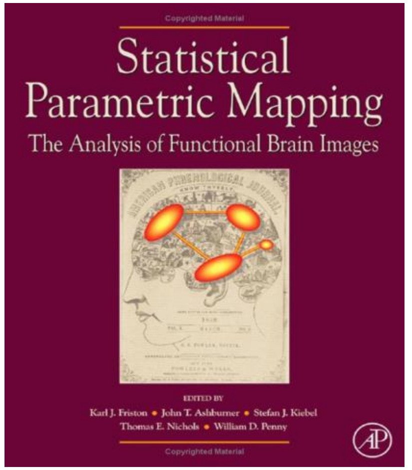
$p(Y|\, \mathbf{b}_2 = 0)$
p-value = 0.07
(t-test)

does not change

$p(Y|\, \mathbf{b}_1 = 0, \mathbf{b}_2 = 0)$
p-value = 0.002
(F-test)

does not change

Bibliography



- Friston KJ et al. (2007) *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Elsevier.

- Christensen R (1996) *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer.

- Friston KJ et al. (1995) Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping* 2: 189-210.

Thank you