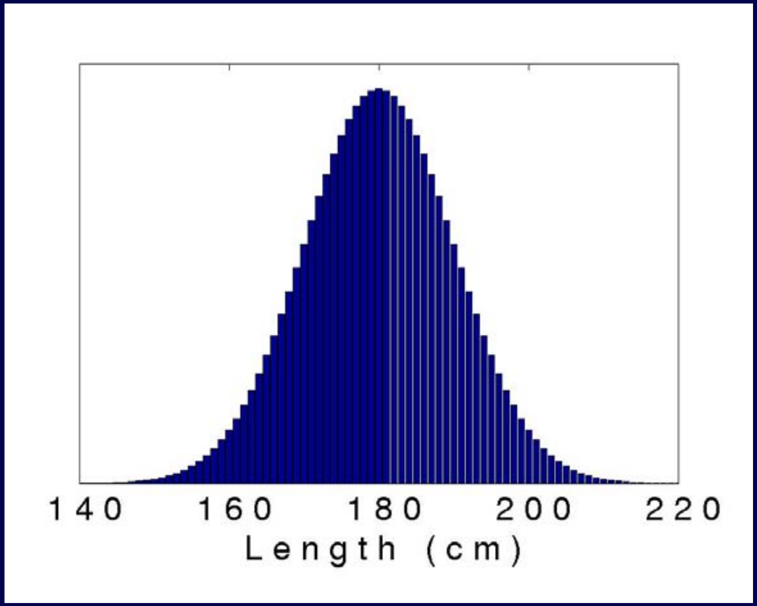
Variance Component Estimation
a.k.a.
Non-Sphericity Correction

Overview

- Variance-Covariance Matrix
- What is (and <u>isn't</u>) sphericity?
- Why is non-sphericity a problem?
- How do proper statisticians solve it?
- How did SPM99 solve it.
- How does SPM2 solve it?
- What is all the fuss?
- Some 2nd level examples.

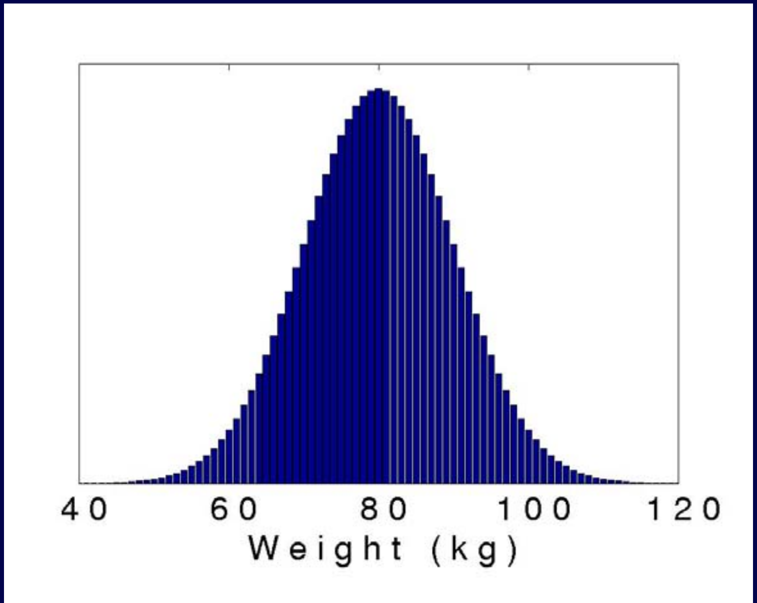
Variance-Covariance matrix

Length of Swedish men



μ=180cm, σ=14cm (σ^2=200)

Weight of Swedish men



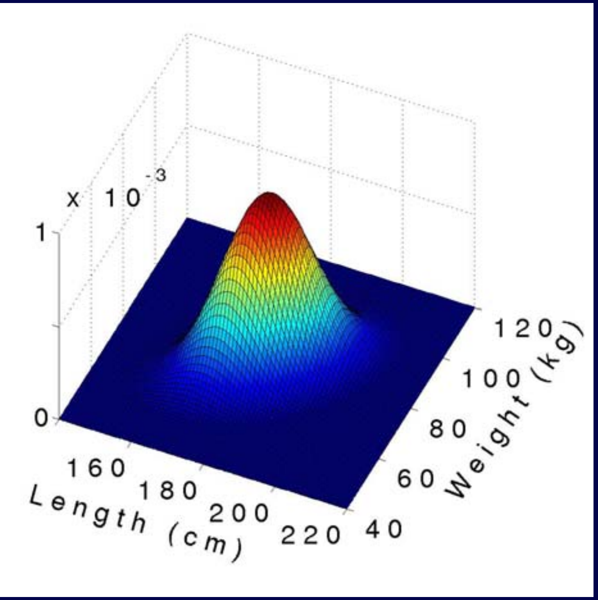
μ=80kg, σ=14kg (σ^2=200)

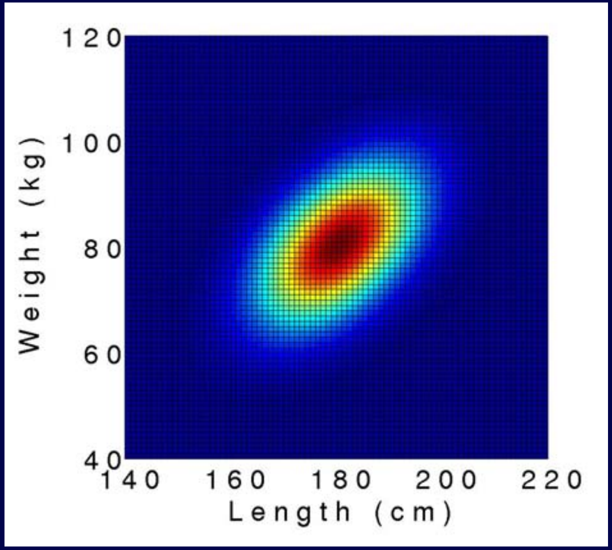
Each completely characterised by μ (mean) and σ^2 (variance),

i.e. we can calculate $p(l|\mu,\sigma^2)$ for any l

Variance-Covariance matrix

- Now let us view length and weight as a 2-dimensional stochastic variable (p(l,w)).





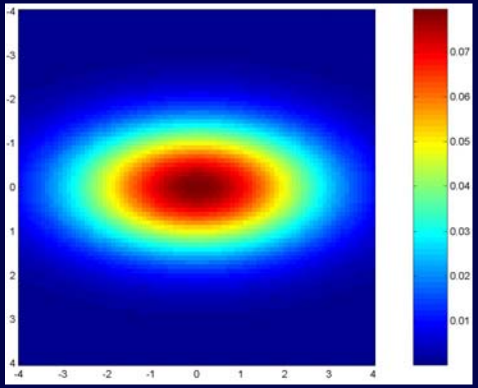
$$\boldsymbol{\mu} = \begin{bmatrix} 180 \\ 80 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 200 & 100 \\ 100 & 200 \end{bmatrix} \qquad p(l,w|\boldsymbol{\mu},\boldsymbol{\Sigma})$$

What is (and isn't) sphericity?

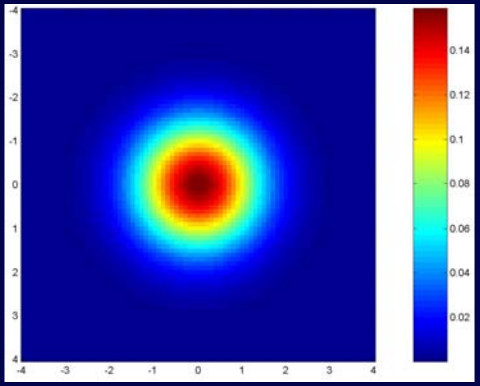
Sphericity $\leftrightarrow$ *iid* $\leftrightarrow$ $N(\boldsymbol{\mu},\boldsymbol{\Sigma}=\sigma^2\mathbf{I})$

$$\Downarrow$$

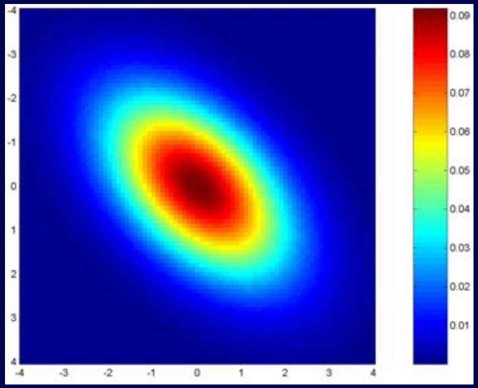
$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$$



$$Cov(\varepsilon) = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$



$$Cov(\varepsilon) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$Cov(\varepsilon) = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

Variance quiz

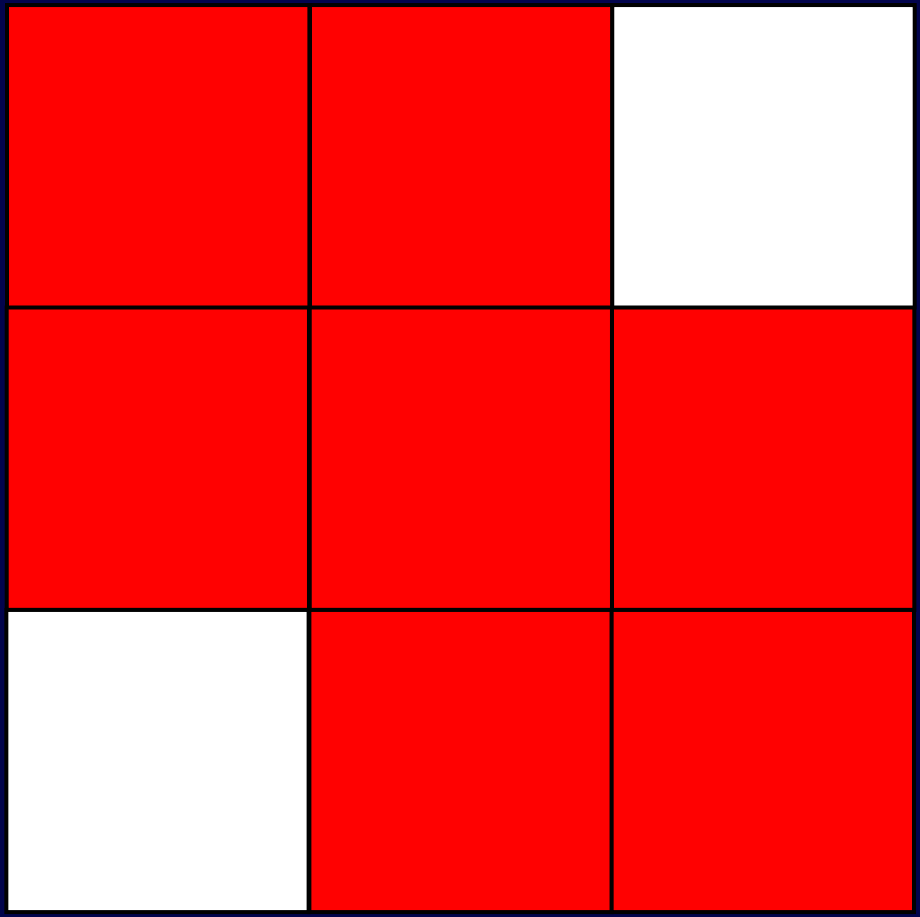
Height			
Weight			
# hours watching telly per day			

Variance quiz

Height

Weight

hours watching telly per day



Variance quiz

Height				
Weight				
# hours watching telly per day				
Shoe size				

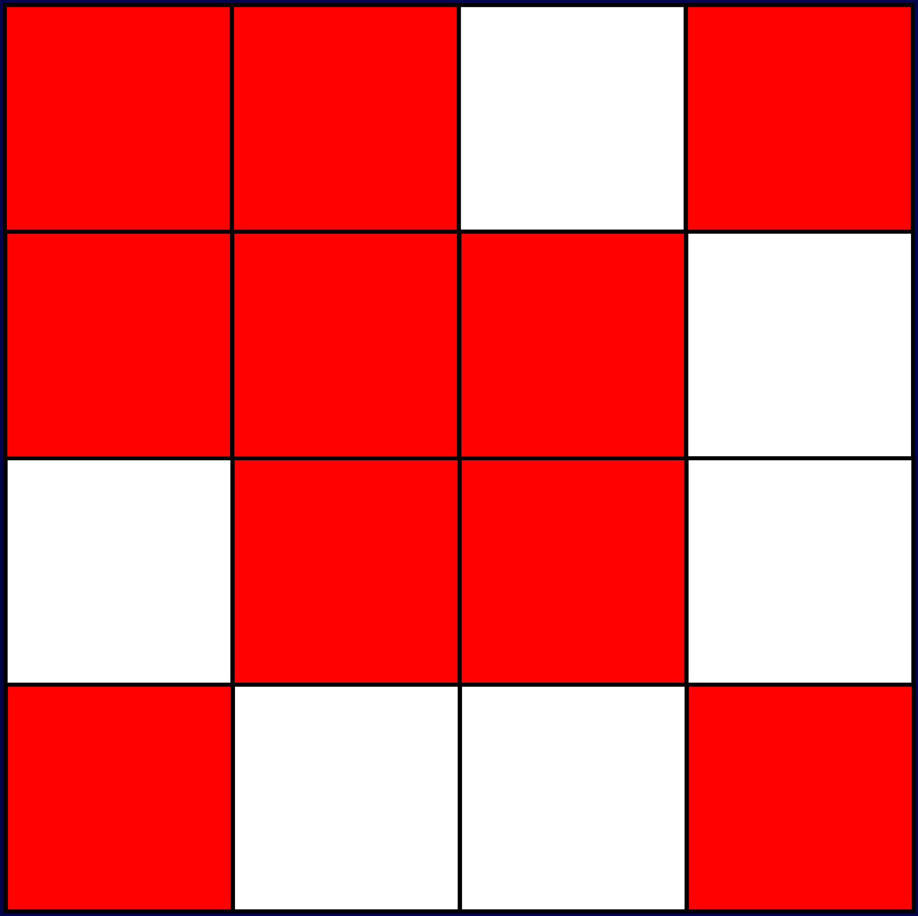
Variance quiz

Height

Weight

hours watching telly per day

Shoe size



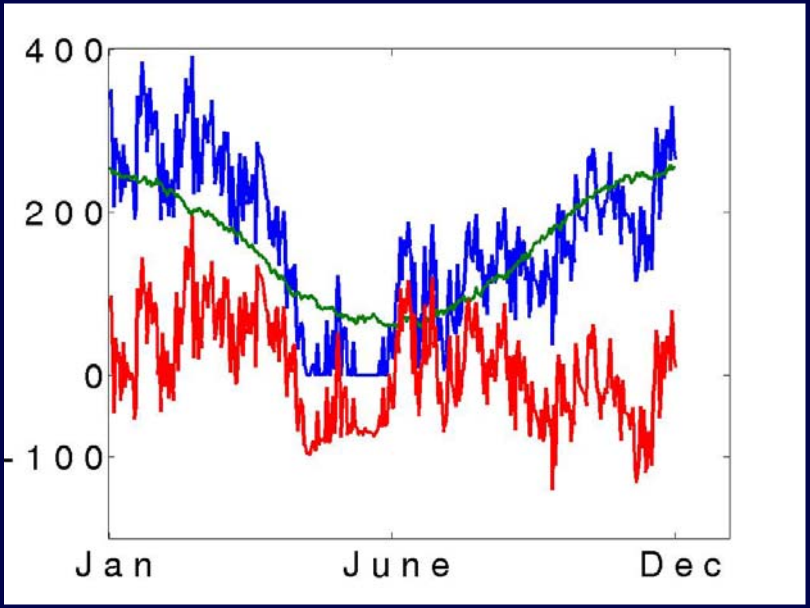
Example:

”The rain in Norway stays mainly in Bergen”

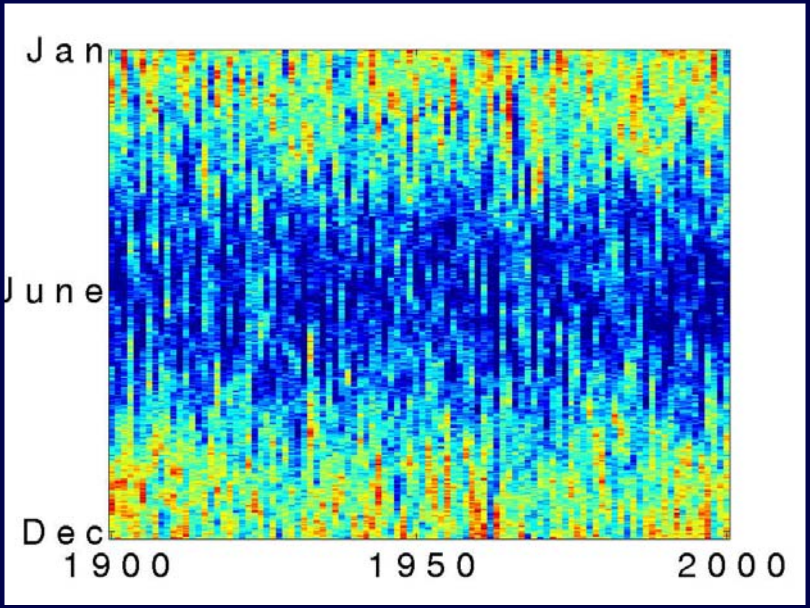
or

”A hundred years of gloominess”

Daily rainfall for 1950

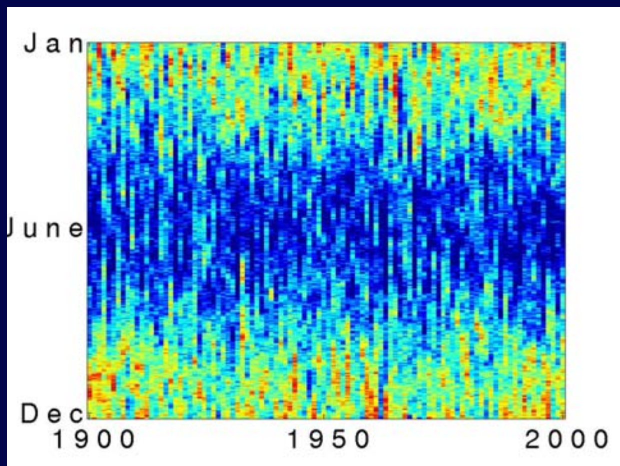


Daily rainfall for 20th century



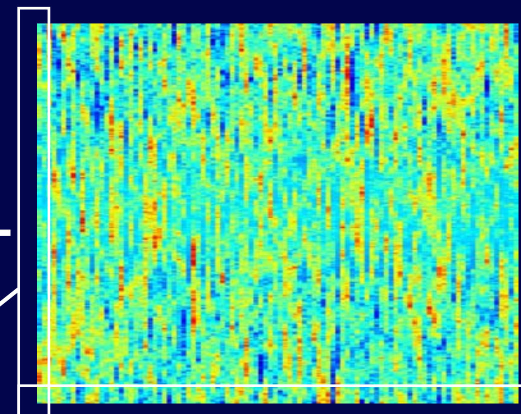
The rain in Bergen continued

The rain in Bergen



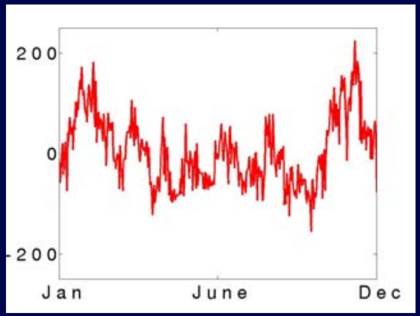
$=$

$+$

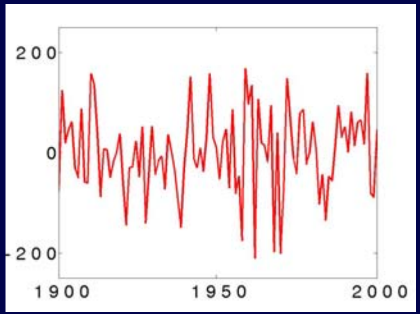


$\mathbf{Y}$ $\boldsymbol{\mu}$ $\hat{\mathbf{E}}$

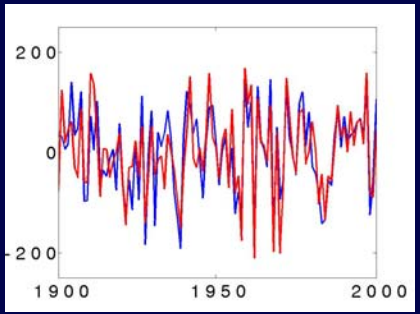
Residual error for 1900



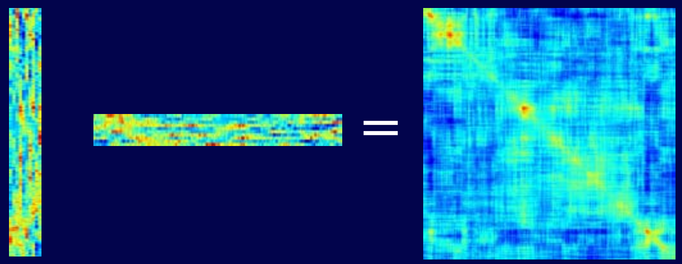
Residual error for Dec 31



Residual error for Dec 30 and Dec 31

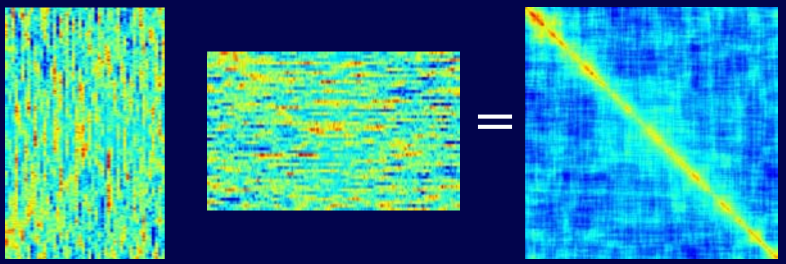


The rain in Bergen concluded



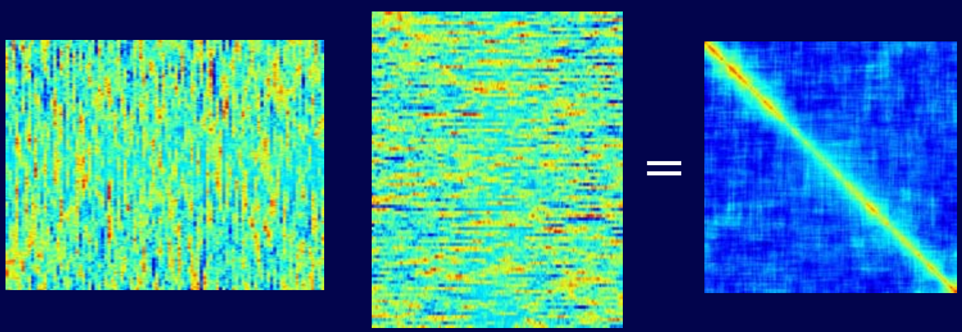
$\hat{\mathbf{E}}$ $\hat{\mathbf{E}}^T$ $\mathbf{S}$

Estimate based on 10 years



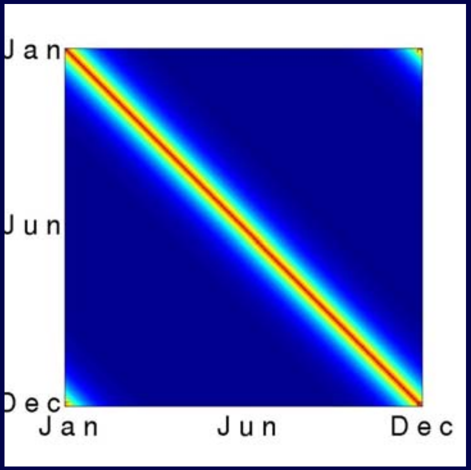
$\hat{\mathbf{E}}$ $\hat{\mathbf{E}}^T$ $\mathbf{S}$

Estimate based on 50 years



$\hat{\mathbf{E}}$ $\hat{\mathbf{E}}^T$ $\mathbf{S}$

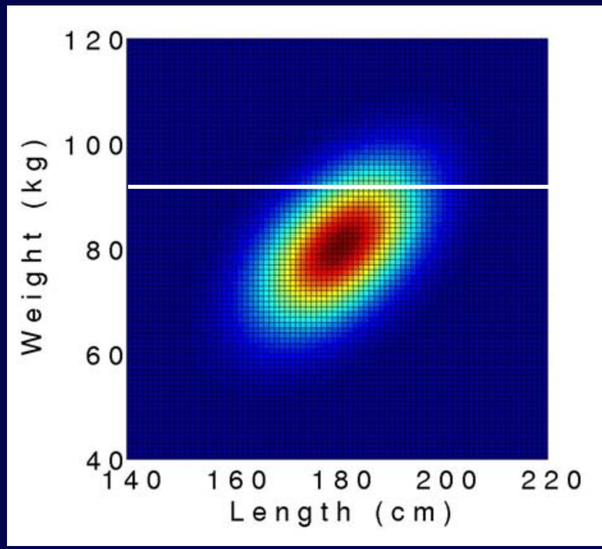
Estimate based on 100 years



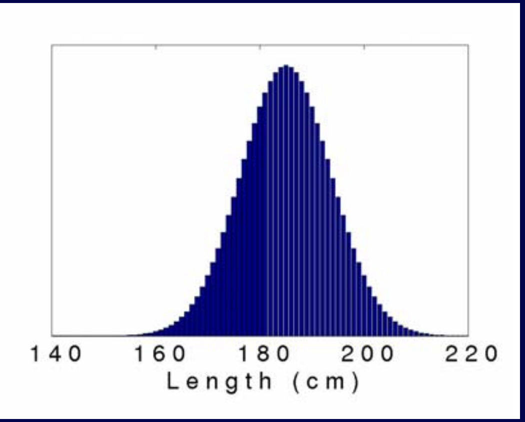
True Σ

Why is non-sphericity a problem?

$p(l,w)$



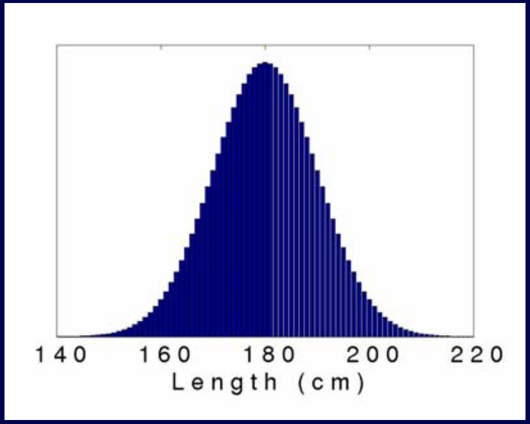
Conditional



$p(l|w=90kg)$

Marginal

$p(l)$

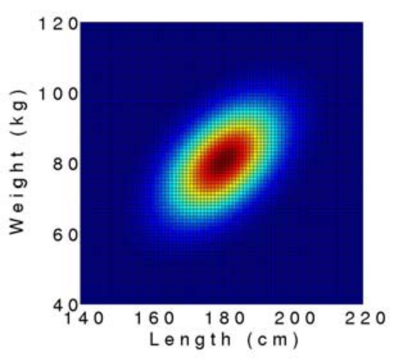


c.f. Blonde hair and blue eyes

How do ”proper” statisticians solve it? (they cheat)

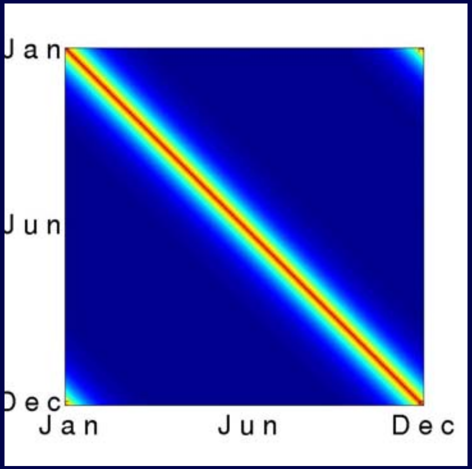
- Greenhouse-Geisser (Satterthwaite) correction.
- Correction factor $(n\text{-}1)^{-1} \leq \varepsilon \leq 1$

Remember?



$$\Sigma = \begin{bmatrix} 200 & 100 \\ 100 & 200 \end{bmatrix} \quad \varepsilon{=}0.8$$

$\Sigma =$

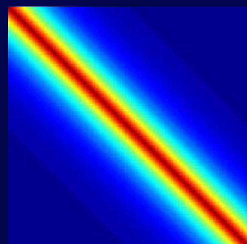
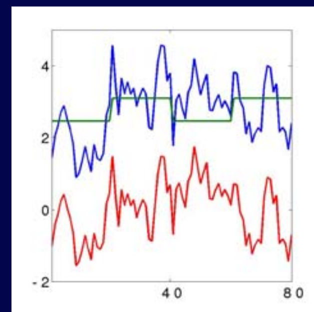


$\varepsilon{=}0.069$

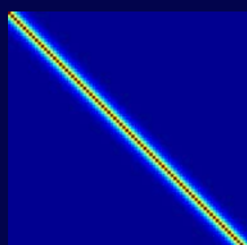
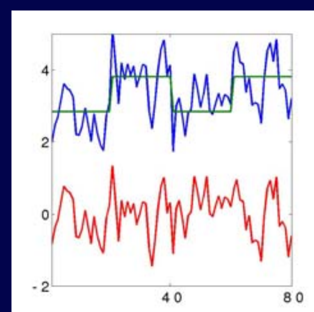
We **thought** we had 100*365=36500 points.

It **was** 2516

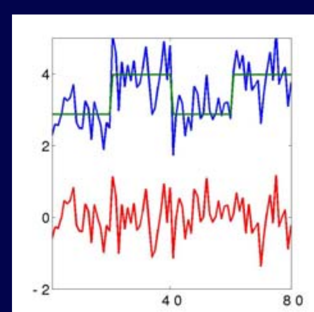
More Greenhouse-Geisser

$\Sigma=$  $\Rightarrow$ 

$\varepsilon=0.107 \rightarrow df=8.60$

$\Sigma=$  $\Rightarrow$ 

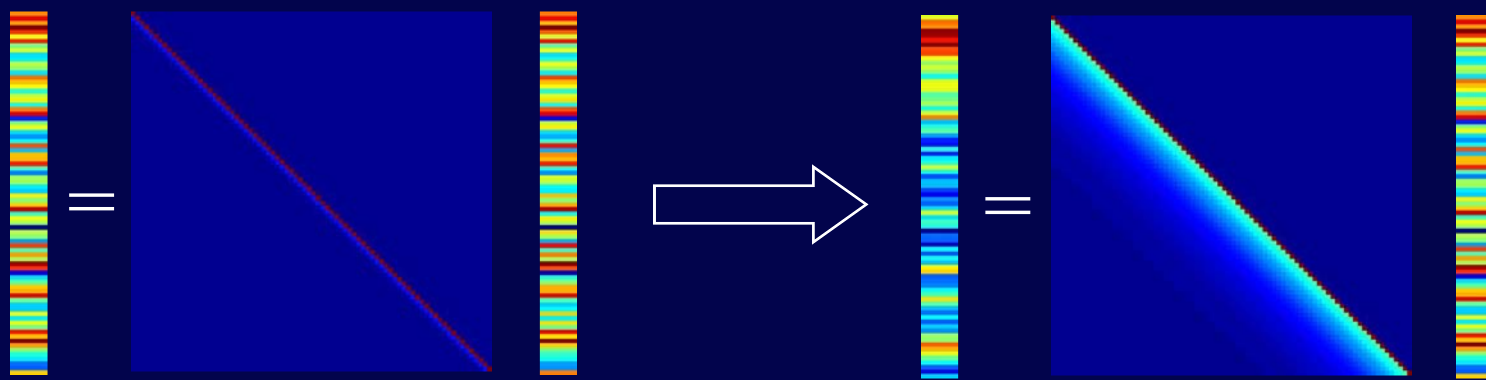
$\varepsilon=0.473 \rightarrow df=37.8$

$\Sigma=$  $\Rightarrow$ 

$\varepsilon=0.999 \rightarrow df=79.9$

How was it solved in SPM99?

- Remember, If we know **Σ** we can correct *df*.



$\mathbf{e}$ $\mathbf{K}$ $\mathbf{z}$ $\mathbf{e}_{pc}$ $\mathbf{S}$ $\mathbf{e}$

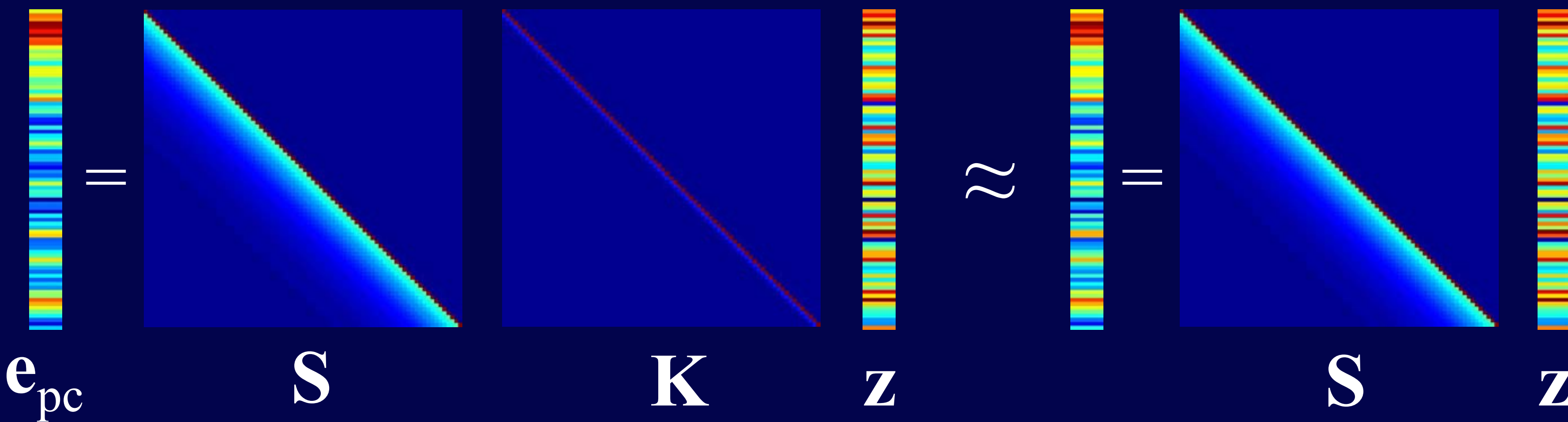
Unknown smoothing filter

So we smooth some more

Observed error

Unobservable uncorrelated error

Why on earth would we do that??

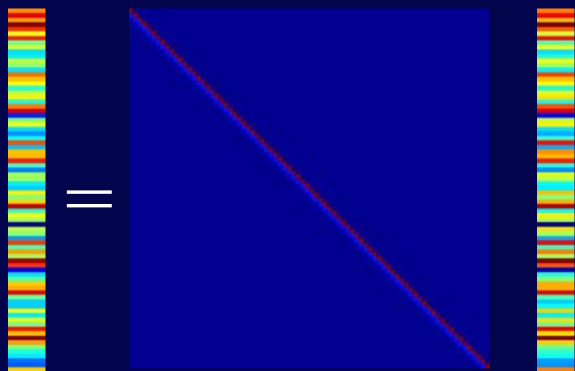


Because the effects of S makes K inconsequential. I.e. we can do a Greenhouse-Geisser based on (the known) **K**.

We ”precolour” with **K**

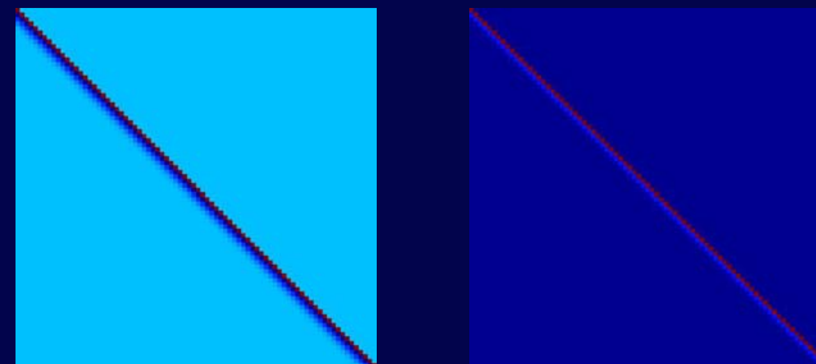
Hope SPM2 is a bit more clever than that.

Same underlying model (AR)

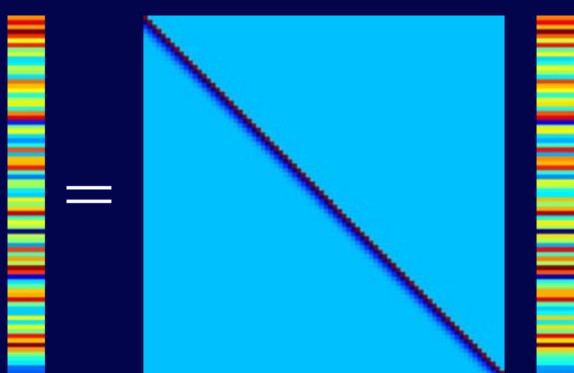


e **K** **z**

A matrix inverse $\mathbf{K}^{-1}$ undoes what **K** did



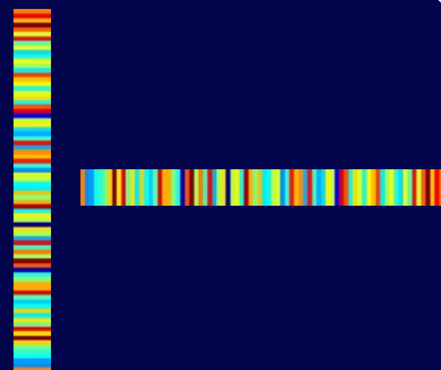
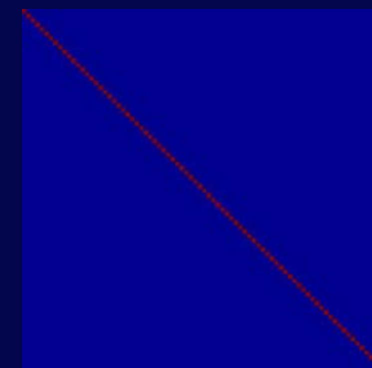
$\mathbf{K}^{-1}$ **K**

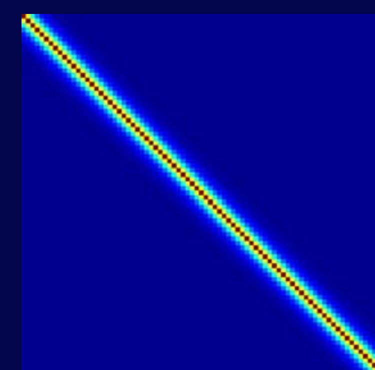


$\hat{\mathbf{z}}$ $\hat{\mathbf{K}}^{-1}$ **e**

SPM2 tries to estimate the matrix K^{-1}, that undoes what K did. If we can find that we can ”pre-whiten” the data, i.e. make them uncorrelated.

Well, how on earth can we do that?

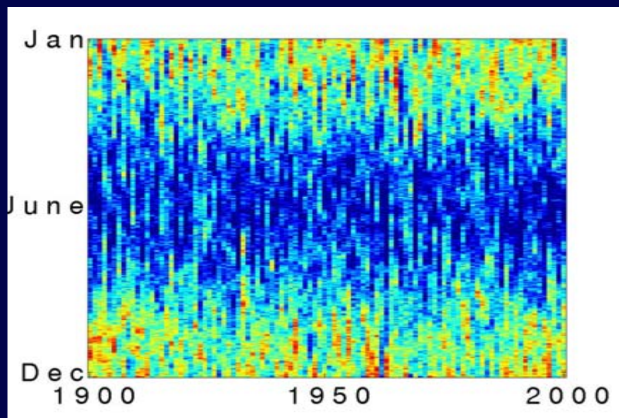
$E\{\mathbf{z}\mathbf{z}^T\} = E\{$  $\} = \sigma^2\mathbf{I} =$ 

$\Sigma = E\{\mathbf{e}\mathbf{e}^T\} = E\{\mathbf{K}\mathbf{z}\mathbf{z}^T\mathbf{K}^T\} = \sigma^2\mathbf{K}\mathbf{K}^T =$ 

I.e. K is the matrix root of Σ, so all we need to do is estimate it.

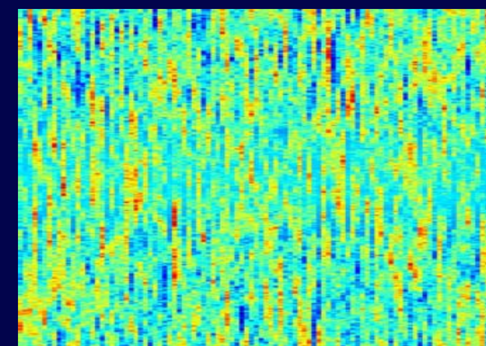
Remember how we estimated Σ for the rain in Bergen?

The rain in Bergen



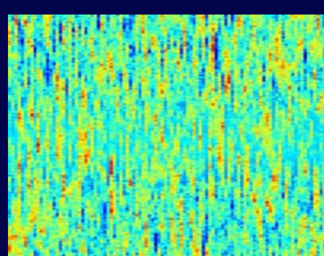
=

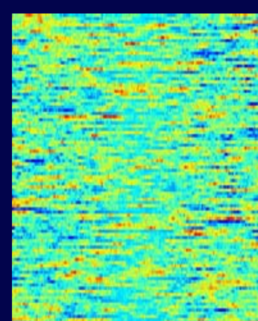
+



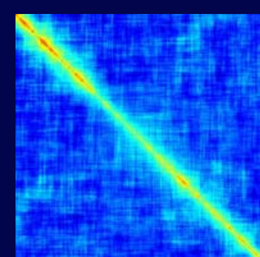
Y **μ** **Ê**

$\hat{\Sigma} =$





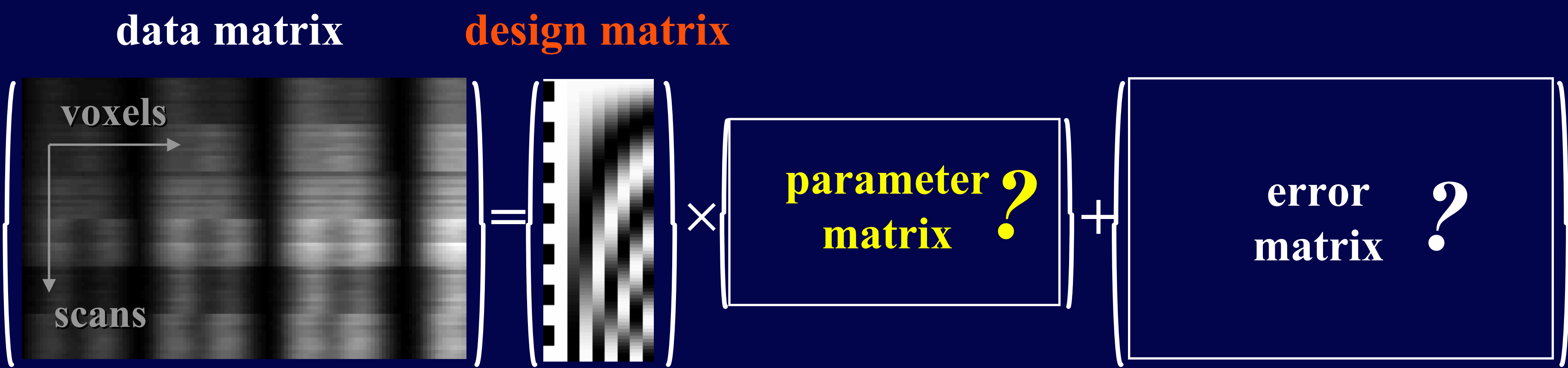
=



Ê $\hat{\mathbf{E}}^{\mathbf{T}}$ **S**

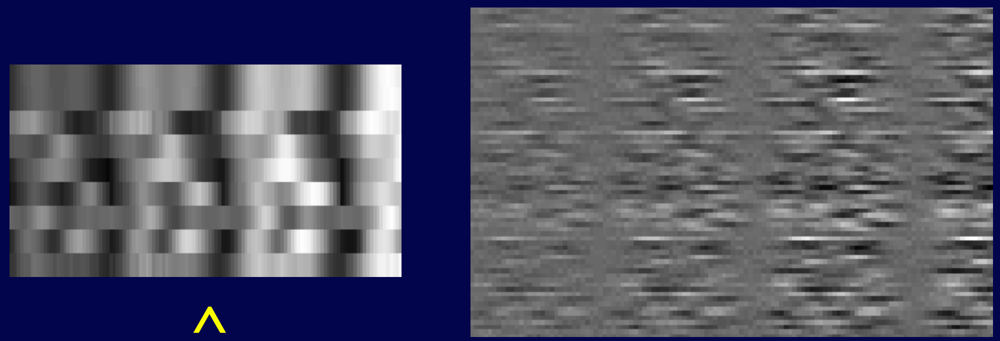
That’s pretty much what SPM2 does too.

Matrix model…



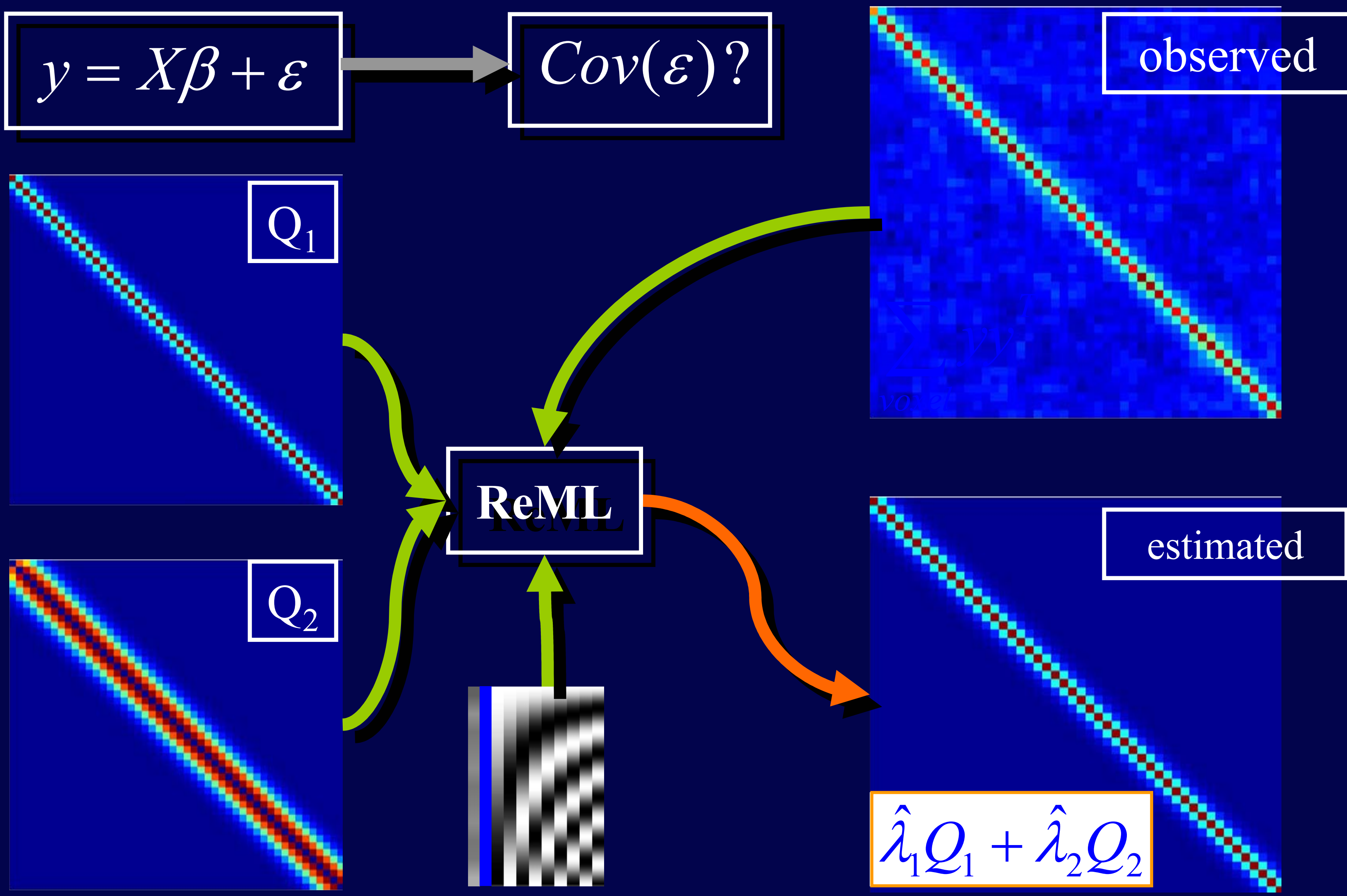
$$Y = X\beta + \varepsilon$$

➲ estimate **parameters** by least-squares ➲



$\hat{\beta}$

Restricted Maximum Likelihood



Maximum Likelihood

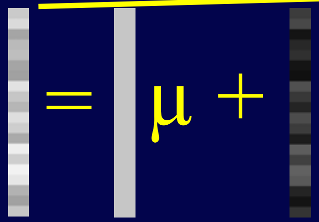
- If we have a model and know it's parameters we can calculate the likelihood (sort of) of any data point.

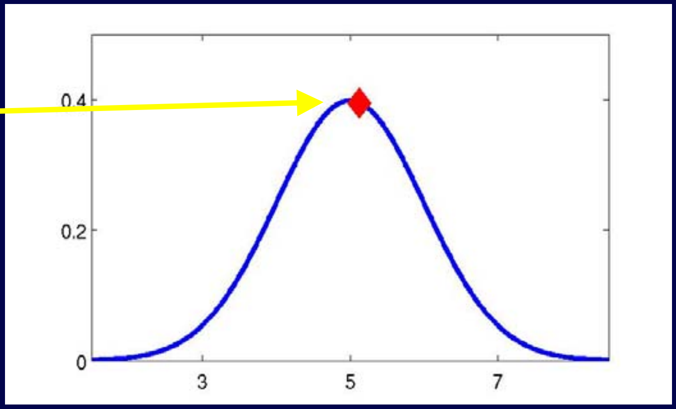
$$y_i = \mu + e_i \quad e \sim N(0, \sigma^2) \quad \Rightarrow \quad p(y_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i - \mu)^2}{\sigma^2}}$$

a.k.a

μ=5, σ^2=1

p=0.396





Maximum Likelihood

- If we have a model and know it's parameters we can calculate the likelihood (sort of) of any data point.

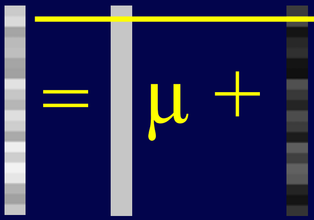
$$y_i = \mu + e_i \quad e \sim N(0, \sigma^2)$$

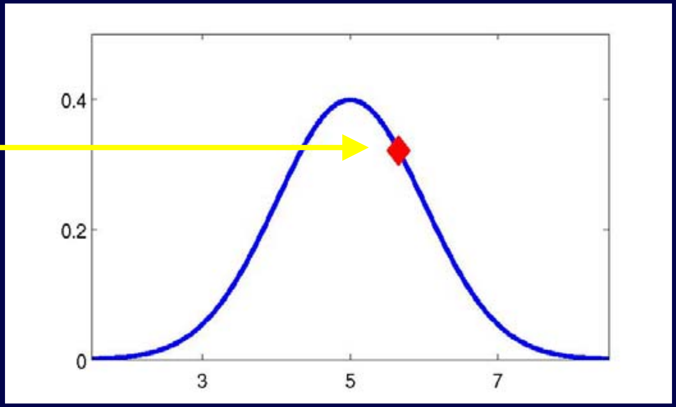
$$p(y_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i - \mu)^2}{\sigma^2}}$$

a.k.a

μ=5, σ²=1

p=0.322



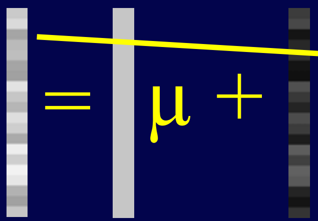


Maximum Likelihood

- If we have a model and know it's parameters we can calculate the likelihood (sort of) of any data point.

$$y_i = \mu + e_i \quad e \sim N(0, \sigma^2) \Rightarrow p(y_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i - \mu)^2}{\sigma^2}}$$

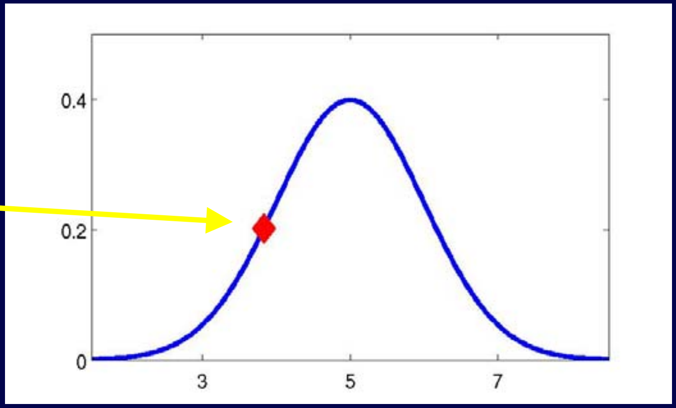
a.k.a



p=0.202

Etc

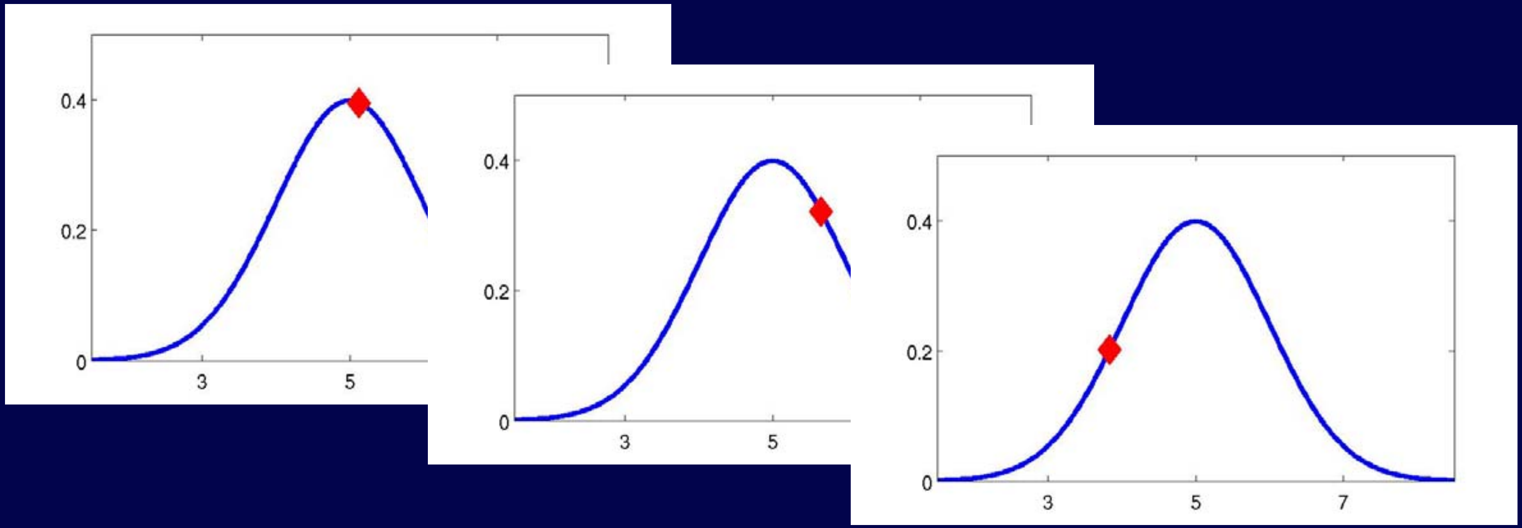
μ=5, σ^2=1



Maximum Likelihood

- And we can calculate the likelihood of the entire data vector.

$$p(\mathbf{y}|\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2}}$$



…

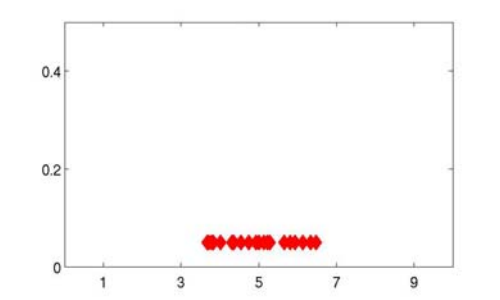
p=0.396 * 0.322 * 0.202 * …

But, does that really make us any happier?

- In reality we don’t know the parameters of our model. They are what we want to estimate.

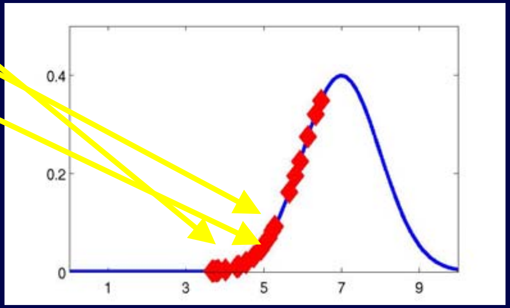
Not brilliant!

$p=0.069*0.162*0.003*\ ...\ =1.86*10^{-30}$



You have your data

”Guess” values for the parameters, here $\mu=7$ and $\sigma^2=1$

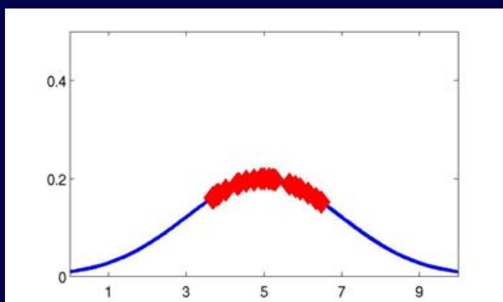


Calculate your likelihoods

But, does that really make us any happier?

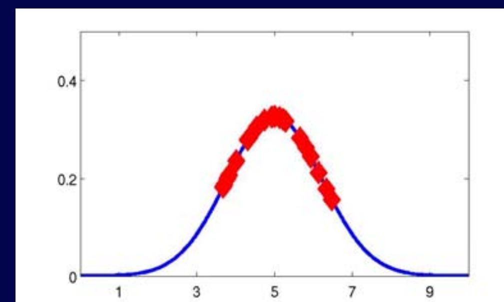
- So, let us try some other values for the parameters.

$\mu=5$, $\sigma^2=4$



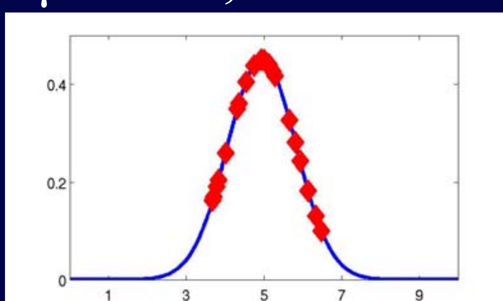
$p=1.38*10^{-15}$
Not bad!

$\mu=5$, $\sigma^2=1.5$



$p=9.41*10^{-13}$
Wow!

$\mu=4.95$, $\sigma^2=0.79$



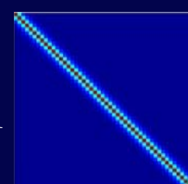
$p=5.28*10^{-12}$
And we have a winner (an ML estimate)!

And, that is actually how simple it is (promise)!

But, does that really make us any happier? (Yeah!)

- Let us say we have a more complicated model

e.g. $p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\Sigma}(\boldsymbol{\lambda})) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}$

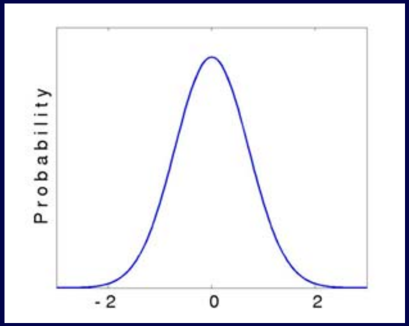
where $\boldsymbol{\Sigma}(\boldsymbol{\lambda}) = \lambda_1$  $+ \lambda_2$ 

(Rather typical first level fMRI model)

- We still have our data (**y**)
- We can still calculate the likelihood for each choice of $\boldsymbol{\beta}=[\beta_1\ \beta_2\ ...]$ and $\lambda=[\lambda_1\ \lambda_2]$.
- And, of course, we can still chose those that maximise the likelihood.

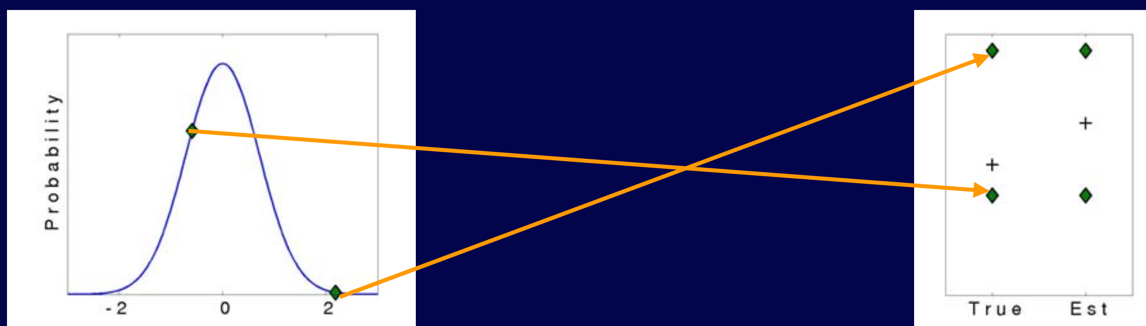
What is all the fuss then?

- Did you ever wonder about the (n-1) when estimating sample variance?



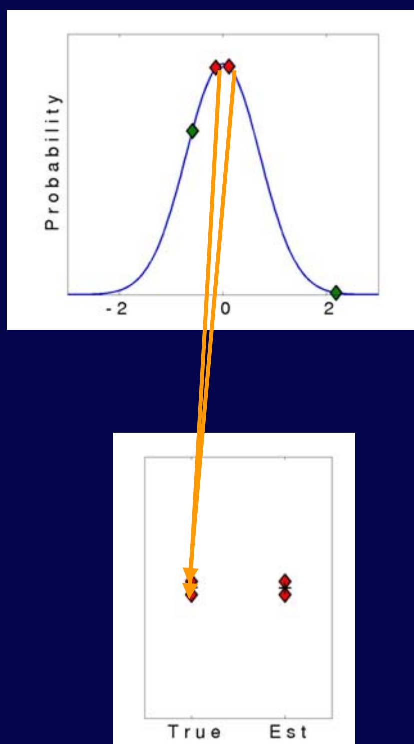
What is all the fuss then?

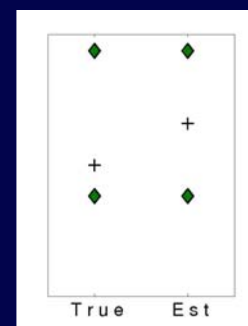
- Did you ever wonder about the (n-1) when estimating sample variance?



What is all the fuss then?

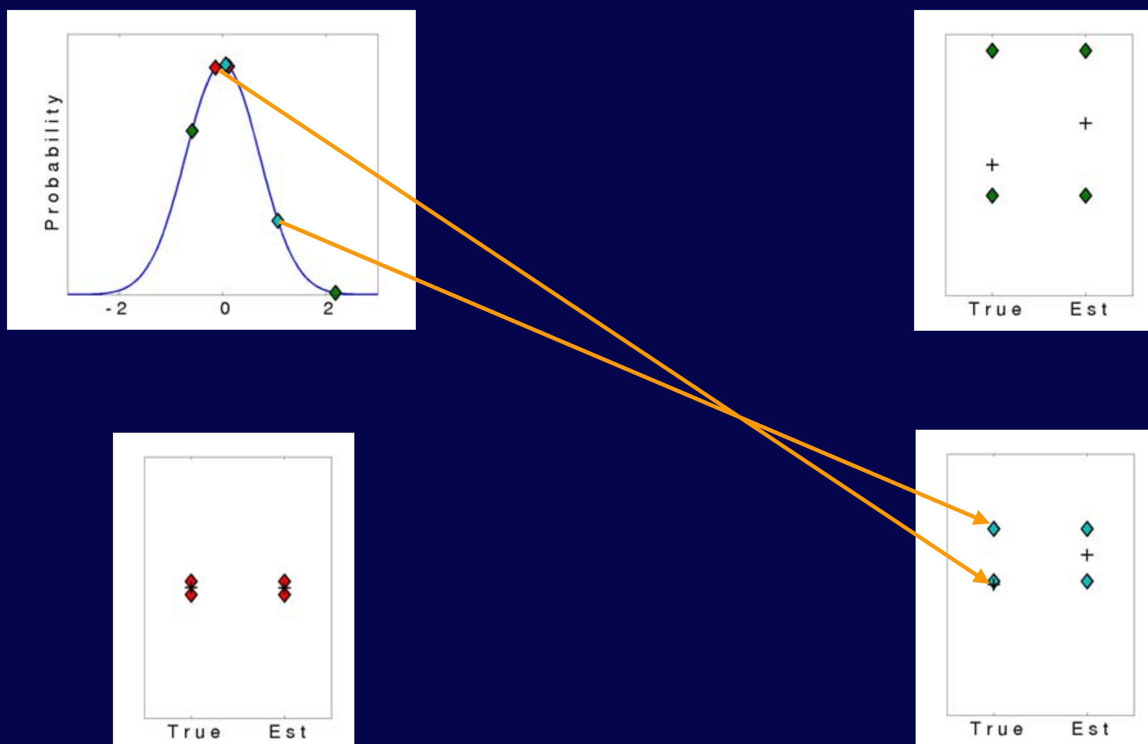
- Did you ever wonder about the (n-1) when estimating sample variance?





What is all the fuss then?

- Did you ever wonder about the (n-1) when estimating sample variance?

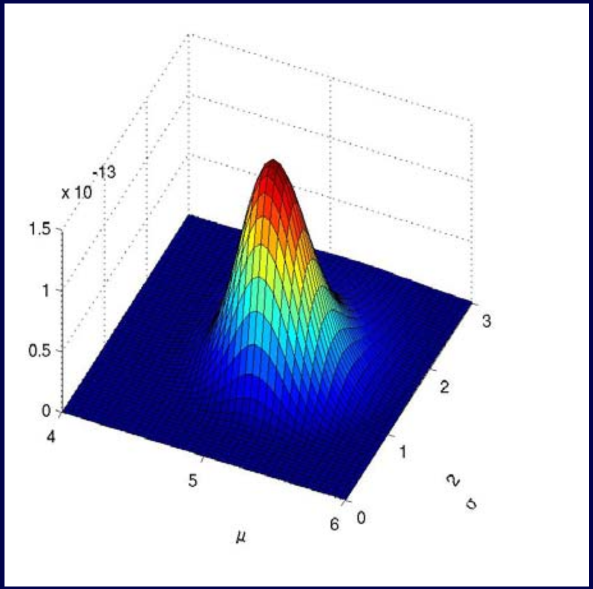


etc…

Or seen slightly differently

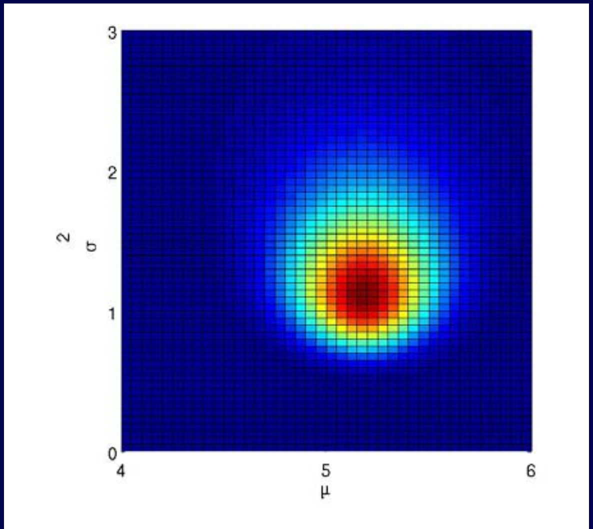
- Data (20 points) drawn from an $N(5,1)$ distribution.

Likelihood as function of μ and σ^2



μ and σ^2 at the location of the peak is the ML-estimate

And seen as an image

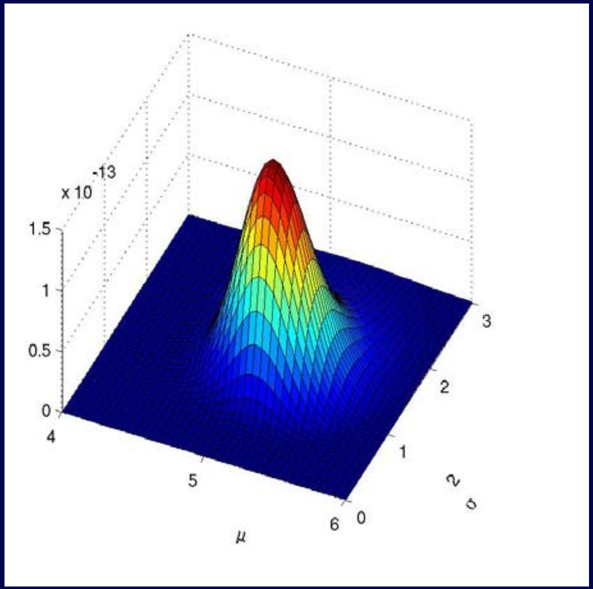


N.B. location of max for σ^2 depends on estimate of μ

Or seen slightly differently

- Data (20 points) drawn from an $N(5,1)$ distribution.

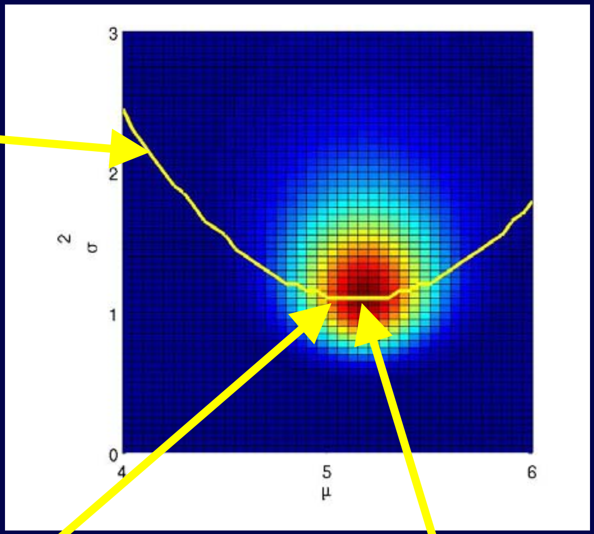
Likelihood as function of μ and σ^2



μ and σ^2 at the location of the peak is the ML-estimate

And seen as an image

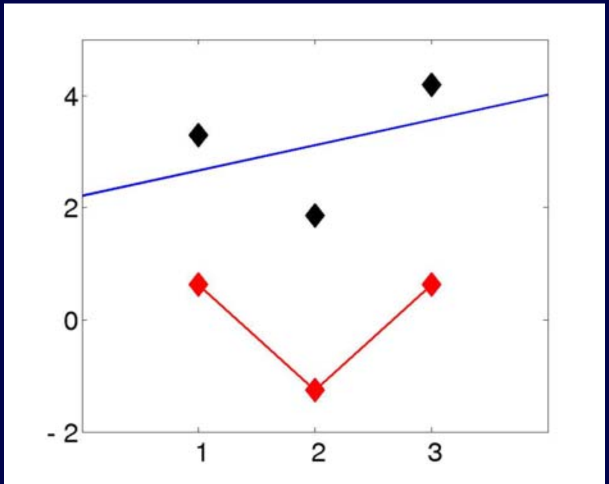
σ² max as function of μ

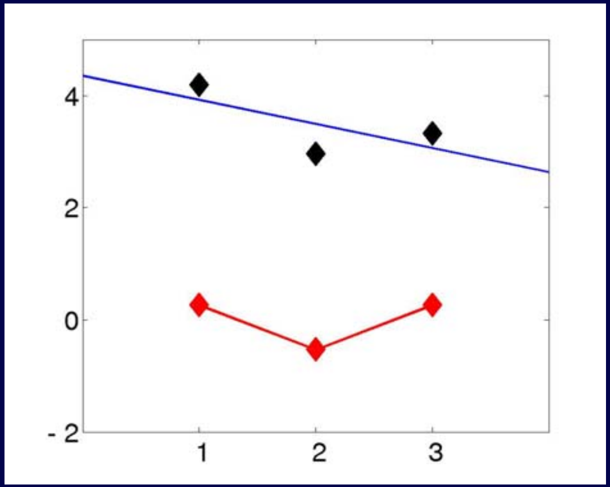


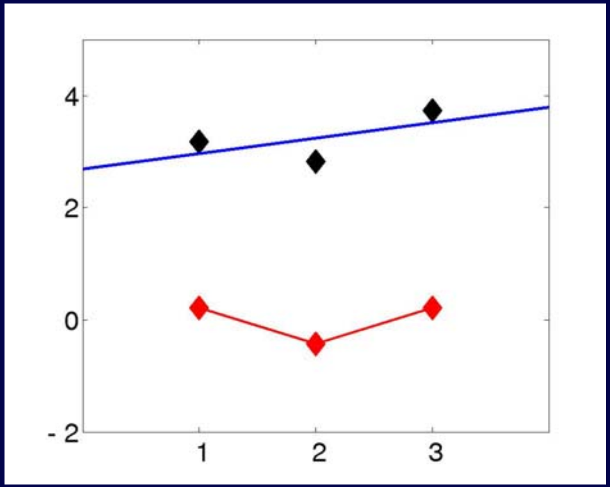
Unbiased estimate

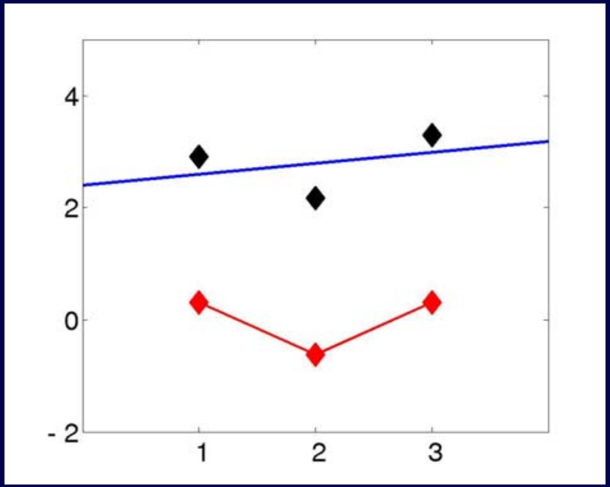
ML-estimate

And the same for estimating serial correlations (c.f. Durbin-Watson)









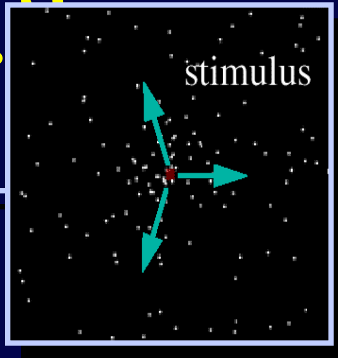
Hur man än vänder sig är rumpan bak

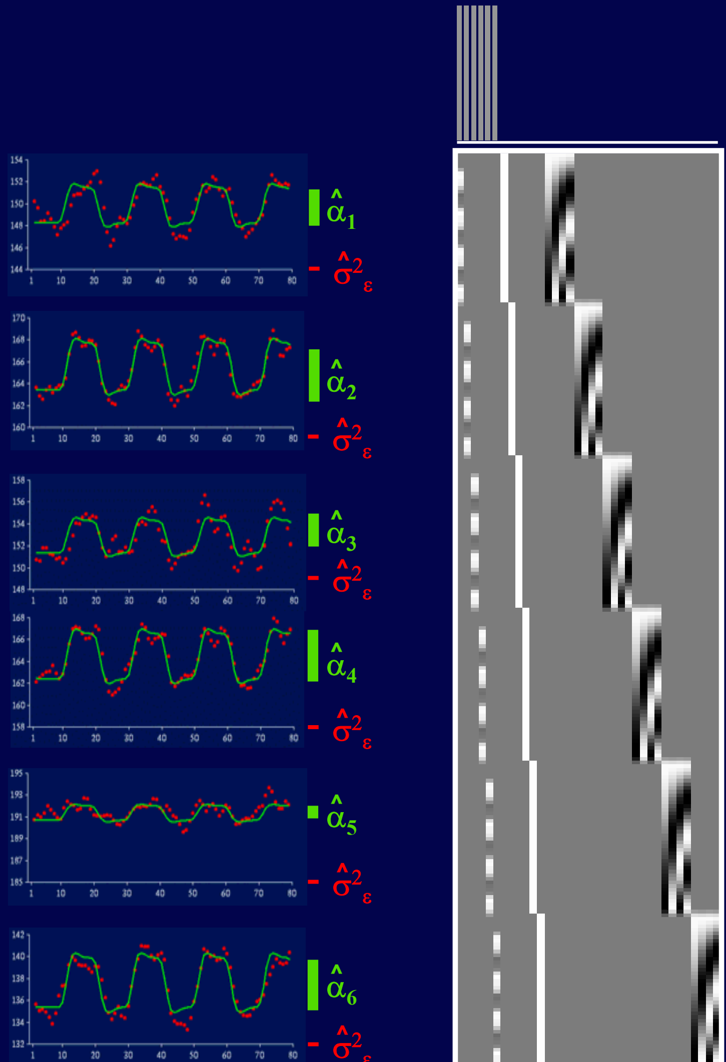
$$\overbrace{\underbrace{\Sigma = E\{\mathbf{e}\mathbf{e}^{T}\}}_{\text{This is what we want}}}^{\text{True variance-covariance matrix}} \overbrace{\underbrace{= E\{\hat{\mathbf{e}}\hat{\mathbf{e}}^{T}\}}_{\text{This is what we observe}}}^{\text{Sample variance-covariance matrix}} + \overbrace{\underbrace{XCov(\beta)X^{T}}_{\text{This we can calculate if…}}}^{\text{Effects of error in parameter estimates}}$$

…we know this. Bummer!

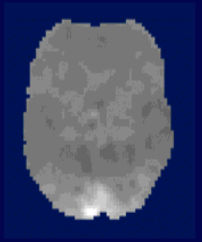
ReML/EM

Multi-subject analysis…?



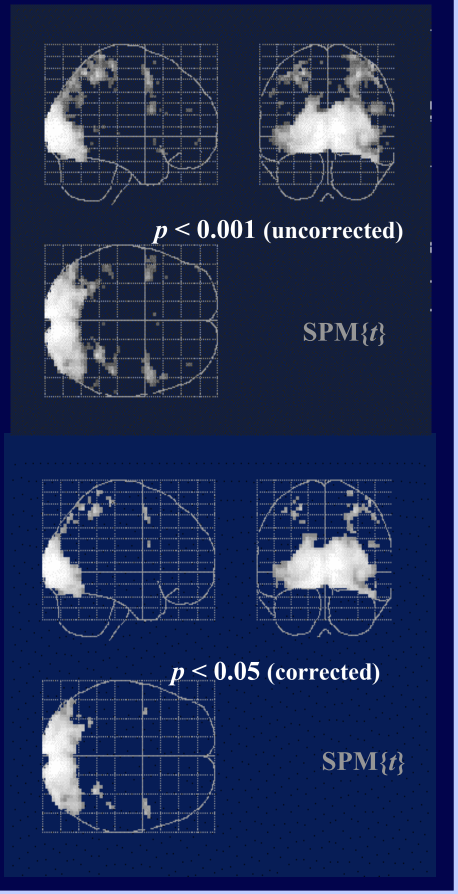


estimated mean activation image

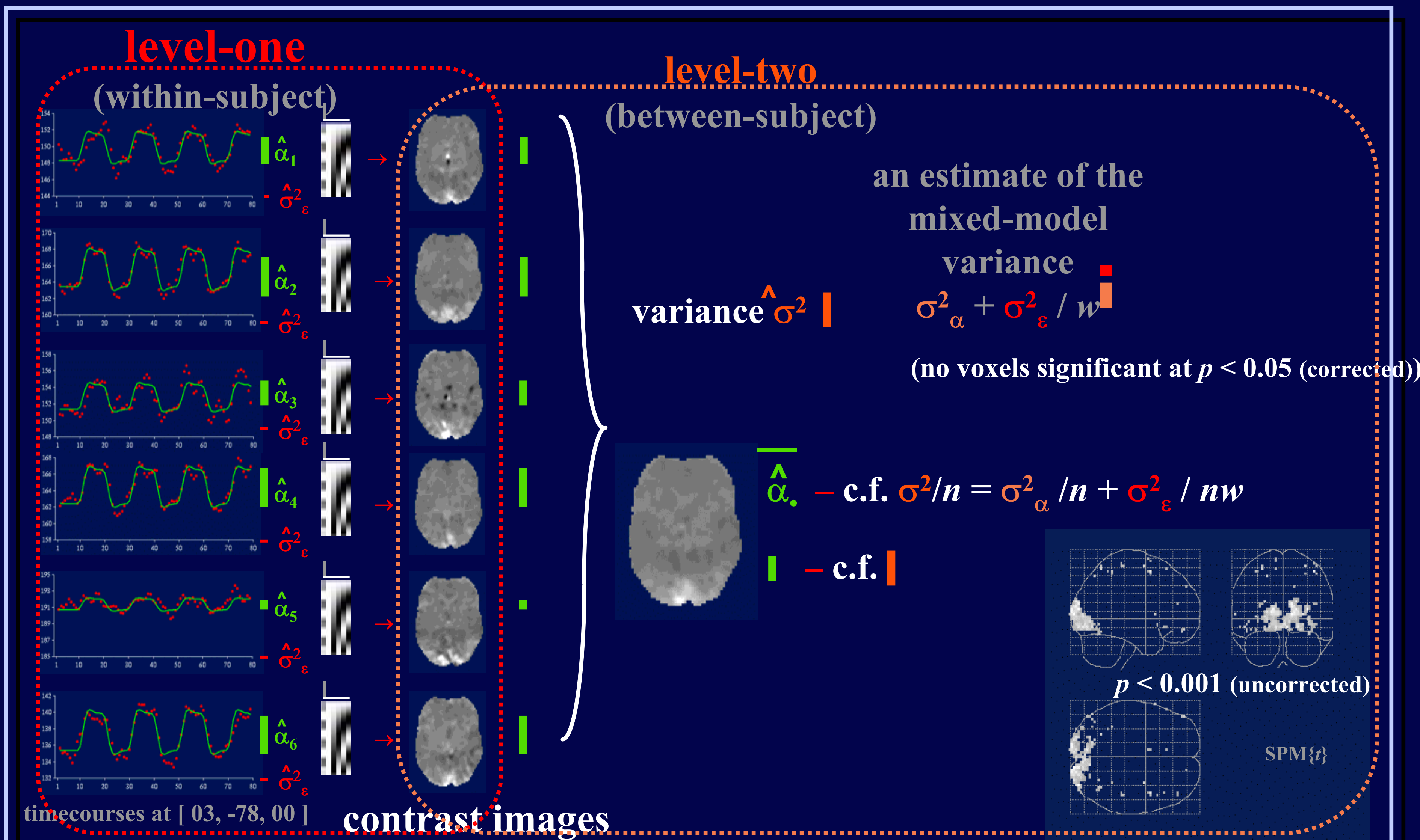


$\overline{\hat{\alpha}_\bullet}$ – c.f. σ^2_ε / *nw*

▌ – c.f. -



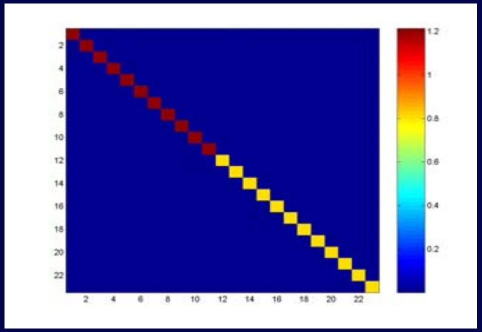
…random effects



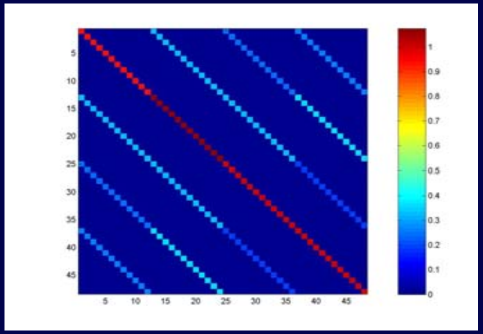
Non-sphericity for 2nd level models

Error Covariance

– Errors are independent but not identical



– Errors are not independent and not identical



Non-Sphericity

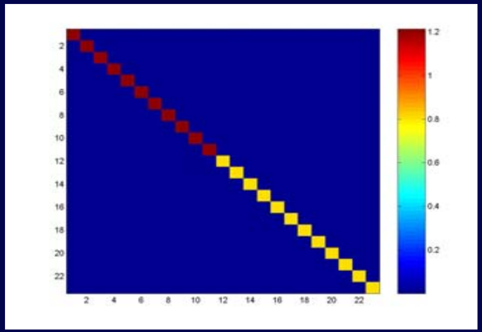
Error can be Independent but Non-Identical when…

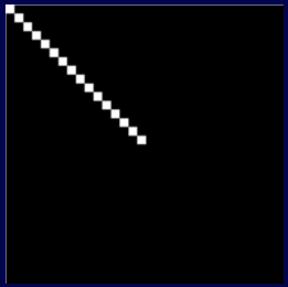
1) One parameter but from different groups

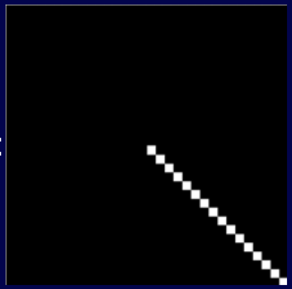
e.g. patients and control groups

2) One parameter but design matrices differ across subjects

e.g. subsequent memory effect

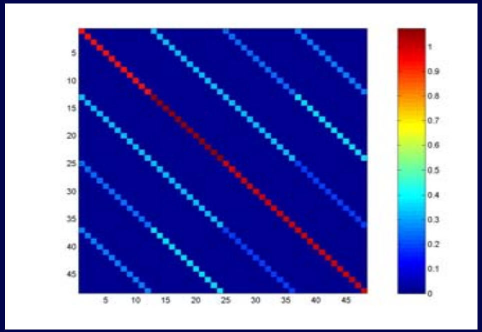


$Q_1 =$ 

$Q_2 =$ 

Non-Sphericity

Error can be Non-Independent and Non-Identical when…



1) Several parameters per subject

e.g. Repeated Measurement design

2) Conjunction over several parameters

e.g. Common brain activity for different cognitive processes

3) Complete characterization of the hemodynamic response

e.g. F-test combining HRF, temporal derivative and dispersion regressors

Example I

U. Noppeney et al.

Stimuli: Auditory Presentation (SOA = 4 secs) of (i) words and (ii) words spoken backwards

Subjects: (i) 12 control subjects
(ii) 11 blind subjects

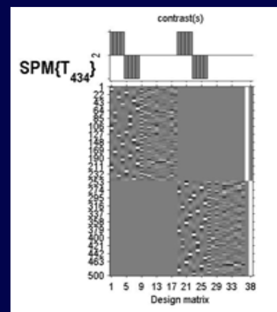
Scanning: fMRI, 250 scans per subject, block design

Q. What are the regions that activate for real words relative to reverse words in ***both*** blind and control groups?

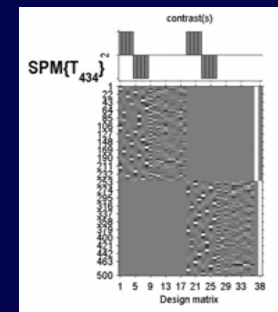
Independent but Non-Identical Error

1st Level

Controls

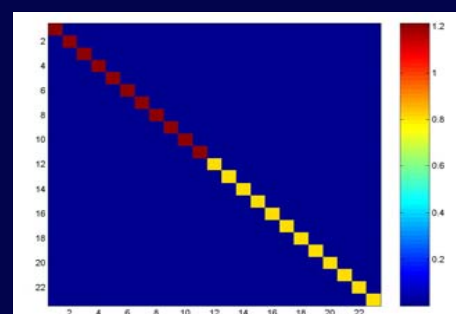


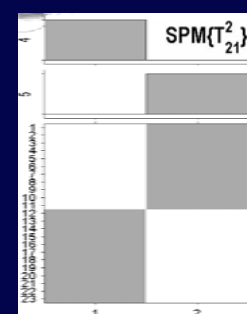
Blinds



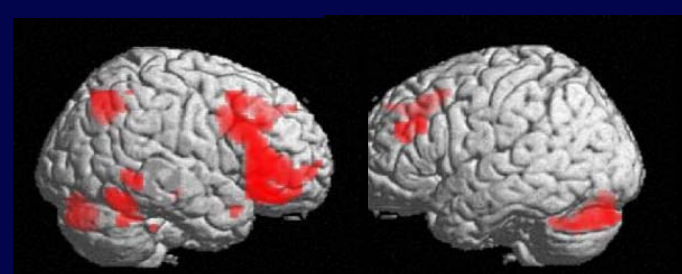
2nd Level

Controls and Blinds





Conjunction between the 2 groups



Example 2

U. Noppeney et al.

Stimuli: Auditory Presentation (SOA = 4 secs) of words

motion	sound	visual	action
"jump"	"click"	"pink"	"turn"

- Subjects: (i) 12 control subjects

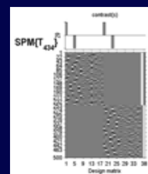
- Scanning: fMRI, 250 scans per subject, block design

Q. What regions are affected by the semantic content of the words ?

Non-Independent and Non-Identical Error

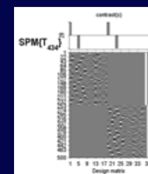
1st Leve

motion



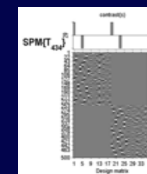
$\stackrel{?}{=}$

sound



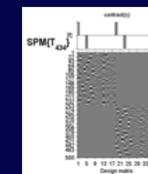
$\stackrel{?}{=}$

visual

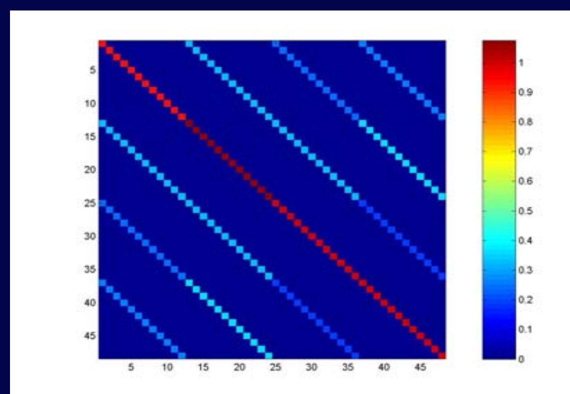


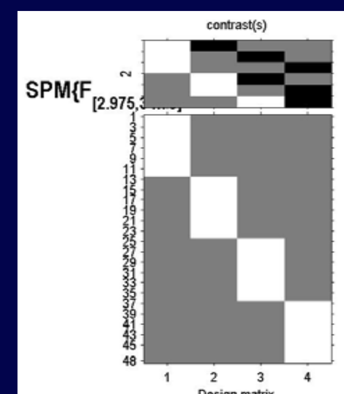
$\stackrel{?}{=}$

action



2nd Level





F-test