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

Overview
• Variance-Covariance Matrix
• What is (and isn’t) 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
Weight of Swedish men

\[ \mu=180\text{cm}, \sigma=14\text{cm}\ (\sigma^2=200) \]
\[ \mu=80\text{kg}, \sigma=14\text{kg}\ (\sigma^2=200) \]

Each completely characterised by \( \mu \) (mean) and \( \sigma^2 \) (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 \( (p(l,w)) \).

\[ \mu = \begin{bmatrix} 180 \\ 80 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 200 & 100 \\ 100 & 200 \end{bmatrix} \quad p(l,w|\mu,\Sigma) \]

What is (and isn’t) sphericity?

Sphericity \( \leftrightarrow \text{iid} \leftrightarrow N(\mu,\Sigma=\sigma^2I) \)

\[ \Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \]

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

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"

The rain in Bergen continued

Y
µ
\( \hat{\epsilon} \)
Residual error for 1900
Residual error for Dec 31
Residual error for Dec 30 and Dec 31

The rain in Bergen concluded

\( \hat{\epsilon} \hat{\epsilon}' \) S
\( \hat{\epsilon} \hat{\epsilon}' \) S
\( \hat{\epsilon} \hat{\epsilon}' \) S
Estimate based on 10 years
Estimate based on 50 years
Estimate based on 100 years
True \( \Sigma \)
Why is non-sphericity a problem?

![Graphs showing marginal and conditional distributions with an example of c.f. Blonde hair and blue eyes.]

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

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

```
Σ = \begin{bmatrix}
  200 & 100 \\
  100 & 200
\end{bmatrix}
```

\[\varepsilon = 0.069\]

We thought we had 100*365=36500 points. It was 2516

More Greenhouse-Geisser

\[\varepsilon = 0.107 \rightarrow \text{df} = 8.60\]

\[\varepsilon = 0.473 \rightarrow \text{df} = 37.8\]

\[\varepsilon = 0.999 \rightarrow \text{df} = 79.9\]

How was it solved in SPM99?

- Remember, if we know \(\Sigma\) we can correct \(\text{df}\).

Hope SPM2 is a bit more clever than that.

Same underlying model (AR)

A matrix inverse \(K^{-1}\) undoes what \(K\) did.

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?

\[
E\{zz^T\} = \sigma^2 I
\]

\[
\Sigma = E\{ee^T\} = E\{Kzz^T\} = \sigma^2 KK^T
\]

I.e. \(K\) is the matrix root of \(\Sigma\), so all we need to do is estimate it.

Remember how we estimated \(\Sigma\) for the rain in Bergen?

\[
\hat{\Sigma} = \hat{\Phi} \hat{\Phi}^T
\]

That’s pretty much what SPM2 does too.

Matrix model…

\[
Y = X \beta + \varepsilon
\]

\(\hat{\beta}\) estimate parameters by least-squares

Restricted Maximum Likelihood

\[
\lambda \Phi + \lambda_0 \Omega
\]

Maximum Likelihood

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

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

a.k.a \(p = 0.396\)

\[
\mu = 5, \sigma^2 = 1
\]

Maximum Likelihood

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

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

a.k.a \(p = 0.322\)

\[
\mu = 5, \sigma^2 = 1
\]
Maximum Likelihood
• If we have a model and know it’s parameters we can calculate the likelihood (sort of) of any data point.

\[ p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \]

a.k.a \[ \mu=5, \sigma^2=1 \]

Etc

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^{-10}

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

You have your data Calculate your likelihoods

But, does that really make us any happier? (Yeah!)
• Let us say we have a more complicated model

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

(Rather typical first level fMRI model)

\[ \Sigma(\lambda) = \lambda_1 + \lambda_2 \]

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

Maximum Likelihood
• And we can calculate the likelihood of the entire data vector.

\[ \prod_{i=1}^{n} p(y_i|\mu_i, \sigma_i^2) \]

\[ \mu=5, \sigma^2=4 \]

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)!

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?

Or seen slightly differently
• Data (20 points) drawn from an \( N(5,1) \) distribution.

Likelihood as function of \( \mu \) and \( \sigma^2 \)
\( \mu \) and \( \sigma^2 \) at the location of the peak is the ML-estimate

Unbiased estimate
ML-estimate

\( \sigma^2 \) max as function of \( \mu \)

N.B. location of max for \( \sigma^2 \) depends on estimate of \( \mu \)

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

Non-sphericity for 2nd level models

- Errors are independent but not identical
- Errors are not independent and not identical

…random effects

…we know this. Bummer!

Hur man än vänder sig är rumpan bak

True variance-covariance matrix

\[ \Sigma = \mathbb{E}\{ee^T\} = \mathbb{E}\{\hat{e}\hat{e}^T\} + \text{X Cov}(\beta)\text{X}^T \]

This is what we want

This is what we observe

This we can calculate if...

ReML/EM

Multi-subject analysis...?
Example 1  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?

Example 2  U. Noppeney et al.

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

<table>
<thead>
<tr>
<th>motion</th>
<th>sound</th>
<th>visual</th>
<th>action</th>
</tr>
</thead>
<tbody>
<tr>
<td>“jump”</td>
<td>“click”</td>
<td>“pink”</td>
<td>“turn”</td>
</tr>
</tbody>
</table>

• 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?