Classical and Bayesian Inference

Lecture Notes

Ahmet Ademoglu, *PhD*
Bogazici University
Institute of Biomedical Engineering

Some concepts and illustrations in this lecture are adapted from the textbook,
Statistical Parametric Mapping: The Analysis of Functional Brain Images, Editors: K. Friston, J. Ashburner, S. Kiebel, T. Nichols and W. Penny, *Academic Press*, 2006.

- The p value in classical inference pertains to the probability of getting the data under the null hypothesis.
- In Bayesian inference it is the probability that, given the data, the contrast exceeds a certain threshold γ.
- Classical inference has the same *specificity* everywhere because of a constant p value.
- To have the same effect in Bayesian inference, we either change the threshold for the effect size or the confidence about that effect.
- **Classical Approach**: One can have a test with uniform specificity.
- **Bayesian Approach**: One effect of uniform size with uniform confidence.

Hierarchical Linear Observation Models

$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}^{(1)}\theta^{(1)} + \epsilon^{(1)} \\
\theta^{(1)} &= \mathbf{X}^{(2)}\theta^{(2)} + \epsilon^{(2)} \\
&\vdots \\
\theta^{(n-1)} &= \mathbf{X}^{(n)}\theta^{(n)} + \epsilon^{(n)}
\end{aligned}
$$

$\epsilon^i \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_\epsilon^{(i)})$,
$\mathbf{y}$: response variable,
$\mathbf{X}^{(i)}$: design matrices,
$\theta^{(i)}$: parameter vector at level (i).

Non-hierarchical Form

$$\mathbf{y} = \epsilon^{(1)} + \mathbf{X}^{(1)}\epsilon^{(2)} + \ldots + \mathbf{X}^{(1)}\ldots\mathbf{X}^{(n-1)}\epsilon^{(n)} + \mathbf{X}^{(1)}\ldots\mathbf{X}^{(n)}\theta^{(n)}$$

$$E[\mathbf{y}\mathbf{y}^T] = \underbrace{\mathbf{C}_\epsilon^{(1)}}_{\textit{error}} + \ldots + \underbrace{\mathbf{X}^{(1)}\ldots\mathbf{X}^{(i-1)}\mathbf{C}_\epsilon^{(i)}\ldots\mathbf{X}^{(1)T}}_{\textit{ith level random effects}}$$

$$+ \ldots + \underbrace{\mathbf{X}^{(1)}\ldots\mathbf{X}^{(n)}\theta^{(n)}\theta^{(n)T}\ldots\mathbf{X}^{(1)T}}_{\textit{fixed effects}} \text{ where } \mathbf{C}_\epsilon^{(i)} = Cov\{\epsilon^{(i)}\}.$$

$$E[\mathbf{y}\mathbf{y}^T] = \tilde{\mathbf{C}}_\epsilon + \mathbf{X}^{(1)}\ldots\mathbf{X}^{(n)}\theta^{(n)}\theta^{(n)T}\ldots\mathbf{X}^{(1)T}$$

where $\tilde{\mathbf{C}}_\epsilon = \mathbf{C}_\epsilon^{(1)} + \cdots + \mathbf{X}^{(1)}\ldots\mathbf{X}^{(n-1)}\mathbf{C}_\epsilon^{(n)}\mathbf{X}^{(n-1)T}\ldots\mathbf{X}^{(1)T}$

Classical Perspective

$$\begin{aligned}
\mathbf{y} &= \tilde{\mathbf{X}}\theta^{(n)} + \tilde{\epsilon} \\
\tilde{\mathbf{X}} &= \mathbf{X}^{(1)}\mathbf{X}^{(2)}\ldots\mathbf{X}^{(n)} \\
\tilde{\epsilon} &= \epsilon^{(1)} + \mathbf{X}^{(1)}\epsilon^{(2)} + \ldots + \mathbf{X}^{(1)}\mathbf{X}^{(2)}\ldots\mathbf{X}^{(n-1)}\epsilon^{(n)}
\end{aligned}$$

Bayesian Perspective

$$\begin{aligned}
\mathbf{y} &= \mathbf{X}\theta + \epsilon^{(1)} \\
\mathbf{X} &= [\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(1)}\mathbf{X}^{(2)}\ldots\mathbf{X}^{(n-1)}, \mathbf{X}^{(1)}\mathbf{X}^{(2)}\ldots\mathbf{X}^{(n)}] \\
\theta^T &= [\epsilon^{(2)}\cdots\epsilon^{(n)}\ \theta^{(n)}]
\end{aligned}$$

Classical Perspective

The objective is to estimate $\theta^{(n)}$ and to make inference about "how large" they are based upon an estimate of their standard error.

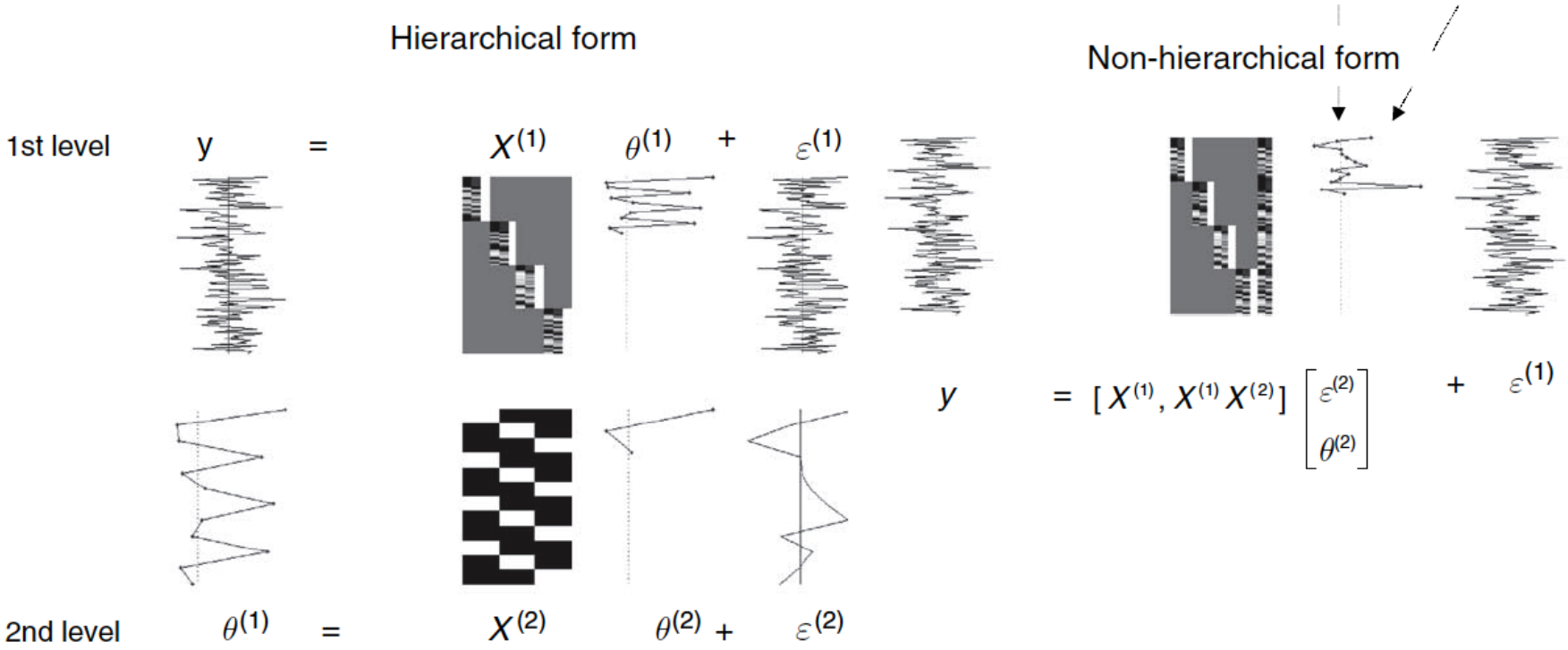
$$\begin{aligned} \eta_{ML} &= \mathbf{M}\mathbf{y} \\ \mathbf{M} &= (\tilde{\mathbf{X}}^T \mathbf{C}_{\tilde{\epsilon}}^{-1} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{C}_{\tilde{\epsilon}}^{-1} \end{aligned}$$

t-statistics $t = \mathbf{c}^T \eta_{ML} / \sqrt{\mathbf{c}^T Cov\{\eta_{ML}\}\mathbf{c}}$

$$Cov\{\eta_{ML}\} = \mathbf{M}\tilde{\mathbf{C}}_\epsilon \mathbf{M}^T = (\mathbf{X}^T \tilde{\mathbf{C}}_\epsilon^{-1} \mathbf{X})^{-1}$$

$\tilde{\mathbf{C}}_\epsilon = Cov[\tilde{\epsilon}]$
$= \mathbf{C}_\epsilon^{(1)} + \mathbf{X}^{(1)} \mathbf{C}_\epsilon^{(2)} \mathbf{X}^{(1)T} \ldots + \mathbf{X}^{(1)} \ldots \mathbf{X}^{(n-1)} \mathbf{C}_\epsilon^{(n)} \mathbf{X}^{(n-1)T} \ldots \mathbf{X}^{(1)T}$

We need to estimate covariance of the composite errors, from all levels, projected down the hierarchy onto the observation space *i.e.* $\tilde{\mathbf{C}}_\epsilon$.


Hierarchical form
Non-hierarchical form
1st level
y = X(1) θ(1) + ε(1)
y = [X(1), X(1)X(2)] [ε(2); θ(2)] + ε(1)
2nd level
θ(1) = X(2) θ(2) + ε(2)

Bayesian Perspective

By the nature of hierarchical model

$$\begin{aligned} \mathbf{y} &= \mathbf{X}^{(1)}\theta^{(1)} + \epsilon^{(1)} \\ \theta^{(1)} &= \mathbf{X}^{(2)}\theta^{(2)} + \epsilon^{(2)} \\ \theta^{(2)} &= \mathbf{X}^{(3)}\theta^{(3)} + \epsilon^{(3)} \\ &\vdots \\ \theta^{(n-1)} &= \mathbf{X}^{(n)}\theta^{(n)} + \epsilon^{(n)} \end{aligned}$$

If estimate the conditional mean $\eta^{(i)}_{\theta|y}$ and conditional covariance $\mathbf{C}^{(i)}_{\theta|y}$ at any level, we can draw inferences at that specific level.

$$\begin{aligned} E[\theta^{(i-1)}] &= \eta^{(i-1)}_{\theta} = \mathbf{X}^{(i)}\theta^{(i)} \\ Cov\{\theta^{(i-1)}\} &= \mathbf{C}^{(i-1)}_{\theta} = \mathbf{C}^{(i)}_{\epsilon} \end{aligned}$$

At the final level we can treat $\theta^{(n)}$ as i) unknown so that their priors are flat and they are treated as fixed effects, or ii) known so that there is nothing to make an inference at the final level.

Bayes Rule : $p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta)$

Priors for $p(\theta) \sim \mathcal{N}(\eta_\theta, \mathbf{C}_\theta)$

$$\eta_\theta^T = [0\,0\,\ldots\,\eta_\theta^{(n)}] \quad \mathbf{C}_\theta = \begin{bmatrix} \mathbf{C}_\epsilon^{(2)} & 0 & 0 \\ 0 & \mathbf{C}_\epsilon^{(n)} & 0 \\ 0 & 0 & \mathbf{C}_\theta^{(n)} \end{bmatrix} \quad \begin{array}{ll} \mathbf{C}_\theta^n = \infty & \text{unkown} \\ \mathbf{C}_\theta^n = 0 & \text{known} \end{array}$$

The likelihood and priors are

$$\begin{aligned} p(\mathbf{y}|\theta) &\propto \exp\left\{ -\tfrac{1}{2}(\mathbf{X}\theta - \mathbf{y})^T \mathbf{C}_\epsilon^{(1)-1}(\mathbf{X}\theta - \mathbf{y}) \right\} \\ p(\theta) &\propto \exp\left\{ -\tfrac{1}{2}(\theta - \eta_\theta)^T \mathbf{C}_\theta^{-1}(\theta - \eta_\theta) \right\} \end{aligned}$$

Using Bayes Rule we obtain

$$p(\theta|\mathbf{y}) \propto \exp\left\{ -\tfrac{1}{2}(\theta - \eta_{\theta|y})^T \mathbf{C}_{\theta|y}^{-1}(\theta - \eta_{\theta|y}) \right\}$$

where

$$\begin{aligned} \mathbf{C}_{\theta|y} &= (\mathbf{X}^T \mathbf{C}_\epsilon^{(1)-1}\mathbf{X} + \mathbf{C}_\theta^{-1})^{-1} \\ \eta_{\theta|y} &= \mathbf{C}_{\theta|y}(\mathbf{X}^T \mathbf{C}_\epsilon^{(1)-1}\mathbf{y} + \mathbf{C}_\theta^{-1}\eta_\theta) \end{aligned}$$

Empirical Bayesian Approach $\longrightarrow \mathbf{C}_\theta^{(n)} = \infty$ and $\mathbf{C}_\theta^{-1}\eta_\theta = 0$

No need to define η_θ since it never appears in estimation equations.

If the priors are flat $\mathbf{C}_\theta^{-1} = \mathbf{0} \longrightarrow$ Minimum Variance Estimator, referred to as *Gauss-Markov* estimator.
Furthermore, if $\mathbf{C}_\epsilon^{(1)} = \mathbf{I} \longrightarrow$ Ordinary Least Squares

The Augmented Form

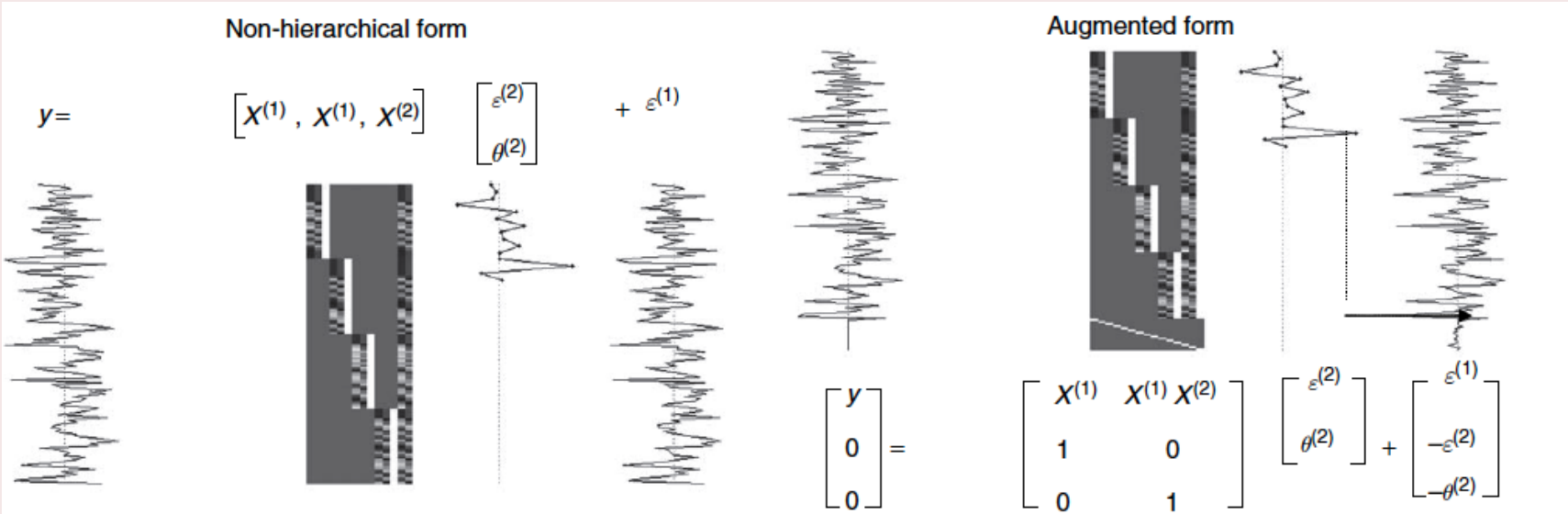
By defining $\overline{\mathbf{y}} = [\mathbf{y}\, \eta_\theta]^T$, $\overline{\mathbf{X}} = [\mathbf{X}\, \mathbf{I}]^T$ and $\mathbf{C}_\epsilon = \begin{bmatrix} \mathbf{C}_\epsilon^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_\theta \end{bmatrix}$

$$\begin{aligned} \mathbf{C}_{\theta|y} &= (\overline{\mathbf{X}}^T \mathbf{C}_\epsilon^{-1} \overline{\mathbf{X}})^{-1} \\ \eta^{\theta|y} &= \mathbf{C}_{\theta|y} (\overline{\mathbf{X}}^T \mathbf{C}_\epsilon^{-1} \overline{\mathbf{y}}) \end{aligned}$$

If the priors for the last level are flat, its prior expectation can be set to $\mathbf{0}$ *i.e.* $\theta^{(n)} \sim \mathcal{N}(\mathbf{0}, \infty)$. $E[\theta^{(i)}] = \mathbf{0}$ for the lower levels and $\mathbf{C}_\theta^{-1} = \mathbf{0}$. It only requires the estimation of $\mathbf{C}_\epsilon^{(1)} = \mathbf{C}_\epsilon$. The problem reduces to the ML solution in classical approach *i.e.*

$$\begin{aligned} \eta_{ML} &= \mathbf{M}\mathbf{y} \\ Cov\{\eta_{ML}\} &= \mathbf{M}\mathbf{C}_\epsilon\mathbf{M}^T = (\mathbf{X}^T\mathbf{C}_\epsilon^{-1}\mathbf{X})^{-1} \\ \mathbf{M} &= (\mathbf{X}^T\mathbf{C}_\epsilon^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{C}_\epsilon^{-1} \end{aligned}$$

which requires the estimation of covariance components $\mathbf{C}_\epsilon^{(i)} = \sum \lambda_i^{(i)} \mathbf{Q}_i^{(i)}$.

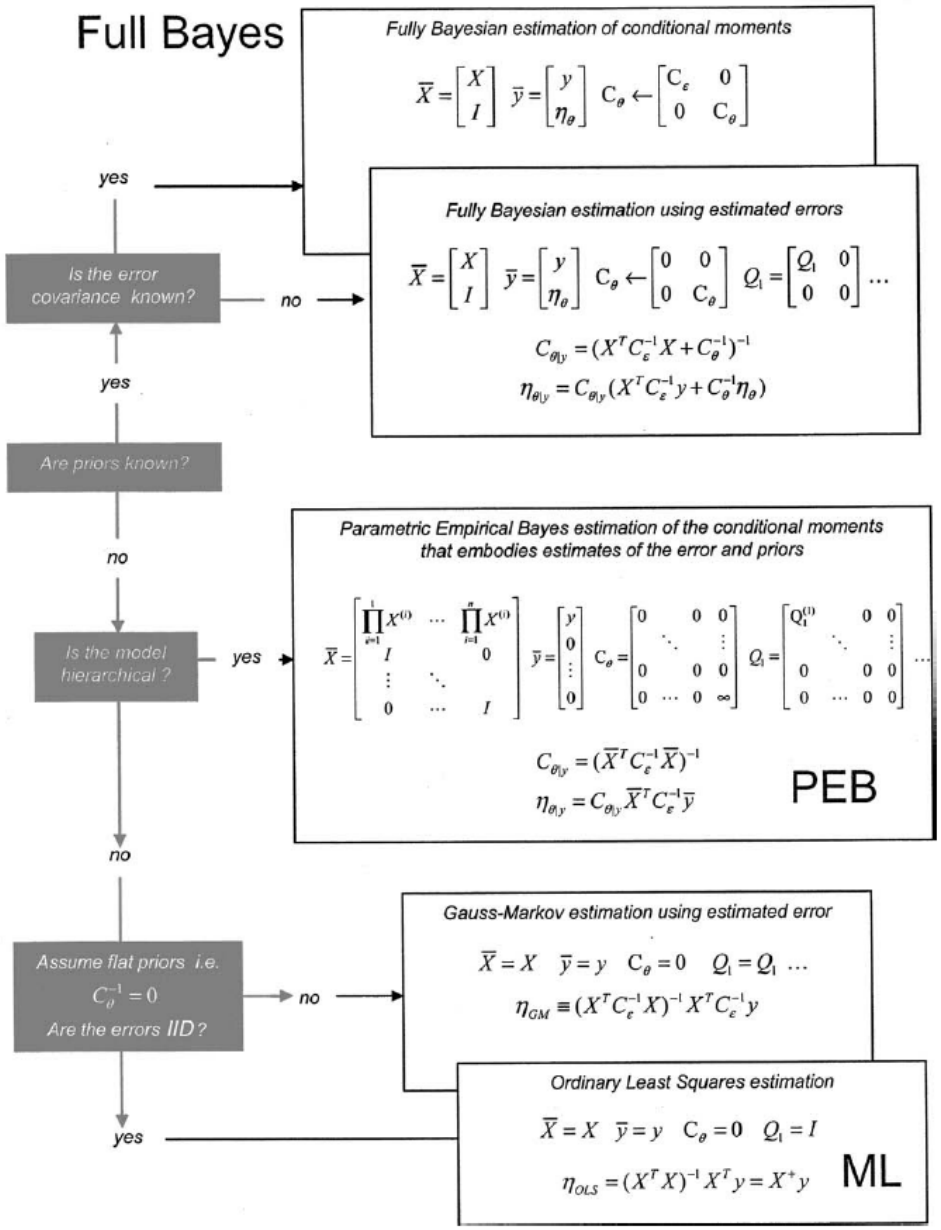


After determining the error covariance components $\mathbf{C}_\epsilon^{(i)}$ or $\mathbf{C}_\epsilon$ and the conditonal mean and covariances about the effects at any level

$$\eta_{\theta|y} = E[\theta|y] = [\eta_{\epsilon|y}^{(2)}] \ldots \eta_{\epsilon|y}^{(n)}\, \eta_{\theta|y}^{(n)}]^T \text{ and } \eta_{\theta|y}^{(i-1)} = E[\theta^{(i-1)}|\mathbf{y}] = \mathbf{X}^{(i)}\eta_{\theta|y}^{(i)} + \eta_{\epsilon|y}^{(i)}$$

The conditional mean and ML estimation are the same at the final level establishing the close connection between classical random effects analyses and hierarchical Bayesian Models.

$$t^{(i)} = \mathbf{c}^T \eta_{\theta|y}^{(i)} / \sqrt{\mathbf{c}^T \mathbf{C}_{\theta|y}^{(i)} \mathbf{c}}$$


Full Bayes
Fully Bayesian estimation of conditional moments
$\overline{X}=\begin{bmatrix}X\\I\end{bmatrix}\ \overline{y}=\begin{bmatrix}y\\\eta_\theta\end{bmatrix}\ C_\theta\leftarrow\begin{bmatrix}C_\varepsilon & 0\\0 & C_\theta\end{bmatrix}$
yes
Is the error covariance known?
no
Fully Bayesian estimation using estimated errors
$\overline{X}=\begin{bmatrix}X\\I\end{bmatrix}\ \overline{y}=\begin{bmatrix}y\\\eta_\theta\end{bmatrix}\ C_\theta\leftarrow\begin{bmatrix}0 & 0\\0 & C_\theta\end{bmatrix}\ Q_1=\begin{bmatrix}Q_1 & 0\\0 & 0\end{bmatrix}\cdots$
$C_{\theta|y}=(X^TC_\varepsilon^{-1}X+C_\theta^{-1})^{-1}$
$\eta_{\theta|y}=C_{\theta|y}(X^TC_\varepsilon^{-1}y+C_\theta^{-1}\eta_\theta)$
yes
Are priors known?
no
Is the model hierarchical ?
yes
Parametric Empirical Bayes estimation of the conditional moments that embodies estimates of the error and priors
$\overline{X}=\begin{bmatrix}\prod_{i=1}^{1}X^{(i)} & \cdots & \prod_{i=1}^{n}X^{(i)}\\I & & 0\\\vdots & \ddots & \\0 & \cdots & I\end{bmatrix}\ \overline{y}=\begin{bmatrix}y\\0\\\vdots\\0\end{bmatrix}\ C_\theta=\begin{bmatrix}0 & & 0 & 0\\ & \ddots & & \vdots\\0 & & 0 & 0\\0 & \cdots & 0 & \infty\end{bmatrix}\ Q_1=\begin{bmatrix}Q_1^{(1)} & & 0 & 0\\ & \ddots & & \vdots\\0 & & 0 & 0\\0 & \cdots & 0 & 0\end{bmatrix}\cdots$
$C_{\theta|y}=(\overline{X}^TC_\varepsilon^{-1}\overline{X})^{-1}$
$\eta_{\theta|y}=C_{\theta|y}\overline{X}^TC_\varepsilon^{-1}\overline{y}$
PEB
no
Assume flat priors i.e. $C_\theta^{-1}=0$ Are the errors IID?
no
Gauss-Markov estimation using estimated error
$\overline{X}=X\quad\overline{y}=y\quad C_\theta=0\quad Q_1=Q_1\ \cdots$
$\eta_{GM}\equiv(X^TC_\varepsilon^{-1}X)^{-1}X^TC_\varepsilon^{-1}y$
yes
Ordinary Least Squares estimation
$\overline{X}=X\quad\overline{y}=y\quad C_\theta=0\quad Q_1=I$
$\eta_{OLS}=(X^TX)^{-1}X^Ty=X^+y$
ML

PPM

Empirical Bayes estimates the variances of the prior distributions directly from the data using a hierarchical model where the parameters and hyperparameters ata any particular level are treated as priors for a lower level.

- 1st level : Experimental effect at any particular voxel
- 2nd level : The effects over voxels

Variation of a contrast over voxels serve as the prior variance of that contrast at any particular voxel.
For a general linear model
$\mathbf{y} = \mathbf{X}\theta + \epsilon$ with $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_\epsilon)$ and prior for $\theta \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_\theta)$
The posterior distribution is determined by

$$\begin{aligned} \mathbf{C}_{\theta|y} &= (\mathbf{X}^T \mathbf{C}_\epsilon^{(1)-1} \mathbf{X} + \mathbf{C}_\theta^{-1})^{-1} \\ \eta_{\theta|y} &= \mathbf{C}_{\theta|y} (\mathbf{X}^T \mathbf{C}_\epsilon^{(1)-1} \mathbf{y}) \end{aligned}$$

The posterior probability that a particular contrast exceeds a threshold γ is easily computed as $p = \Phi\left(\frac{\gamma - \mathbf{c}^T \eta_{\theta|y}}{\sqrt{\mathbf{c}^T \mathbf{C}_{\theta|y} \mathbf{c}}}\right)$ where $\Phi \sim \mathcal{N}(0, 1)$.
The image obtained from these posterior probabilities is called **PPM**.

Estimating the prior density with Empirical Bayes

The variation of a particular parameter over voxels can be used as the prior variance of that parameter at any particular voxel.

Estimating the prior density with Empirical Bayes

A 2-level hierarchical model

$$\begin{aligned} \mathbf{y} &= [\mathbf{X}_1\ \mathbf{X}_0]\begin{bmatrix}\theta_1\\ \theta_0\end{bmatrix} + \epsilon^{(1)} \\ \theta_1 &= \mathbf{0} + \epsilon^{(2)} \end{aligned}$$

θ_0 : Confounds as drifts, constant terms *etc.*

θ_1 : Random effects that we want estimate over the voxels.

This model assumes a voxel-wide prior distribution for the parameters θ_1 with zero mean and unknown covariance $\mathbf{C}_{\epsilon^{(2)}} = \sum_i \lambda_i \mathbf{Q}_i^{(2)}$.

By rearranging the above system,

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_0\theta_0 + \xi \\ \xi &= \mathbf{X}_1\epsilon^{(2)} + \epsilon^{(1)} \end{aligned} \qquad \mathbf{C}_\xi = E[\xi\xi^T] = \sum \lambda_k \mathbf{Q}_k$$

$\mathbf{Q} = \{\mathbf{X}_1\mathbf{Q}_1^{(2)}\mathbf{X}_1^T, \ldots, \mathbf{X}_1\mathbf{Q}_m^{(2)}\mathbf{X}_1^T, \mathbf{Q}_1^\epsilon, \ldots, \mathbf{Q}_l^\epsilon\}$: covariance structures

$\lambda = [\lambda_1^{(2)}, \ldots, \lambda_m^{(2)}, \lambda_1^\epsilon, \ldots, \lambda_l^\epsilon]^T$: hyperparameters for covariance structure

The global parameters are estimated by pooling all the voxels and using ReML

$$
\begin{array}{lcl}
\mathbf{C}_\xi & = & \sum \lambda_i \mathbf{Q}_k \\
\mathbf{C}_{\theta_0|y} & = & (\mathbf{X_0^T C_\xi^{-1} X}_0)^{-1} \\
\mathbf{P} & = & \mathbf{C}_\xi^{-1} - \mathbf{C}_\xi^{-1}\mathbf{X}_0\mathbf{C}_{\theta_0|y}\mathbf{X}_0^T\mathbf{C}_\xi^{-1} \\
\mathbf{g}_i & = & \frac{1}{2} Tr[\mathbf{P}^T\mathbf{Q}_i\mathbf{P}\frac{1}{n}\mathbf{Y}\mathbf{Y}^T - \frac{1}{2} Tr[\mathbf{P}\mathbf{Q}_i] \\
H_{ij} & = & \frac{1}{2} Tr[\mathbf{P}\mathbf{Q}_i\mathbf{P}\mathbf{Q}_j] \\
\lambda & \longleftarrow & \lambda + \mathbf{H}^{-1}\mathbf{g}
\end{array}
$$

$\mathbf{C}_{\theta_0}$ is assumed to be flat *i.e.* $\mathbf{C}_{\theta_0} = \infty$, since it is the prior covariance of fixed effects.

Therefore, the prior covariance of parameter vector $\theta = [\theta_1\ \theta_0]^T$

$$
\mathbf{C}_\theta = \begin{bmatrix}
\sum \lambda_i \mathbf{Q}_i^{(2)} & \cdots & & 0 \\
\vdots & & \infty & \\
 & & & \ddots \\
0 & & & \infty
\end{bmatrix}
$$

can be reconstructed using the $\lambda_i^{(2)}$ and $\mathbf{Q}_i^{(2)}$ and the flat priors for θ_0.

Once an empirical prior covariance is estimated for $\mathbf{C}_\theta$, voxel specific conditional mean $\eta_{\theta|y}$ and covariance $\mathbf{C}_{\theta|y}$ for θ can be determined using a general linear model;

$$\mathbf{y} = \mathbf{X}\theta + \epsilon$$

with $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_\epsilon)$ and prior for $\theta \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_\theta)$.
Now we have a precise estimate of $\mathbf{C}_\theta$ that we can determine $\mathbf{C}_\epsilon$ for each voxel by running the ReML and Bayesian estimation with voxel data.

Estimating the error covariance $\mathbf{C}_\epsilon$ with REML

$$
\begin{aligned}
\mathbf{C}_\epsilon &= \sum \lambda_i^\epsilon \mathbf{Q}_i^\epsilon \\
\mathbf{C}_{\theta|y} &= (\mathbf{X}^\mathbf{T}\mathbf{C}_\epsilon^{-1}\mathbf{X} + \mathbf{C}_\theta^{-1})^{-1} \\
\mathbf{P} &= \mathbf{C}_\epsilon^{-1} - \mathbf{C}_\epsilon^{-1}\mathbf{X}\mathbf{C}_{\theta|y}\mathbf{X}^T\mathbf{C}_\epsilon^{-1} \\
\mathbf{g}_i &= \frac{1}{2} Tr[\mathbf{P}^T\mathbf{Q}_i\mathbf{P}\mathbf{y}\mathbf{y}^T - \frac{1}{2} Tr[\mathbf{P}\mathbf{Q}_i] \\
H_{ij} &= \frac{1}{2} Tr[\mathbf{P}\mathbf{Q}_i\mathbf{P}\mathbf{Q}_j] \\
\lambda_i^\epsilon &\longleftarrow \lambda_i^\epsilon + \mathbf{H}^{-1}\mathbf{g}
\end{aligned}
$$

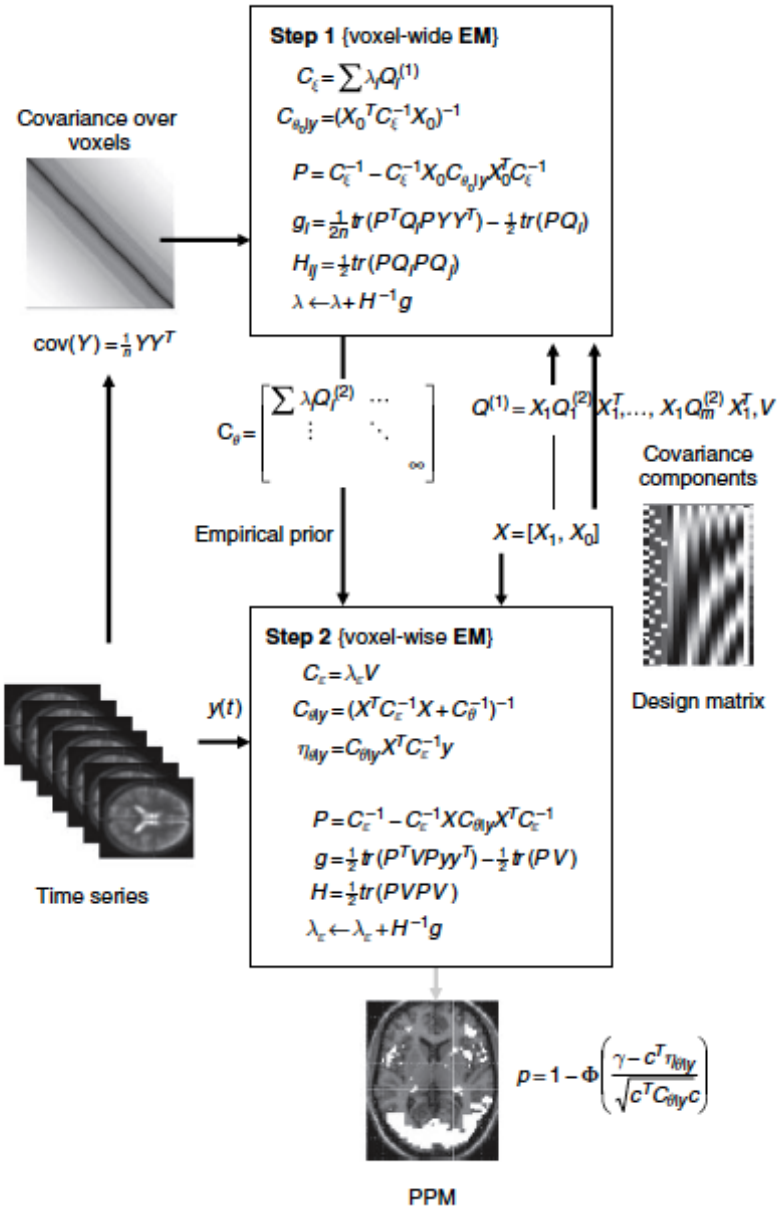
$\mathbf{C}_\epsilon = \sum_i \lambda_i^\epsilon \mathbf{Q}_i^\epsilon$ is the correlation matrix of the errors at any voxel.

The posterior distribution is determined by

$$
\begin{aligned}
\mathbf{C}_{\theta|y} &= (\mathbf{X}^T\mathbf{C}_\epsilon^{-1}\mathbf{X} + \mathbf{C}_\theta^{-1})^{-1} \\
\eta_{\theta|y} &= \mathbf{C}_{\theta|y}(\mathbf{X}^T\mathbf{C}_\epsilon^{-1}\mathbf{y})
\end{aligned}
$$

The posterior probability that a particular contrast exceeds a threshold γ is easily computed as $p = \Phi\left(\frac{\gamma - \mathbf{c}^T\eta_{\theta|y}}{\sqrt{\mathbf{c}^T\mathbf{C}_{\theta|y}\mathbf{c}}}\right)$ where $\Phi \sim \mathcal{N}(0,1)$.

The image obtained from these posterior probabilities is called **PPM**.



PPM SCHEME

If the error covariance $\mathbf{C}_\epsilon$ is modeled by multiple hyperparameters $\lambda_1^\epsilon, \ldots, \lambda_l^\epsilon$ then the covariance structure for any particular voxel $\mathbf{V} = \sum \lambda_i^\epsilon \mathbf{Q}_i^\epsilon$ is normalized ensuring that the voxel-specific error hyperparameters are estimated with high precision by a single hyperparameter.

Posterior Probability Map (PPM)

Once the posterior mean and covariances of parameters θ are known, the posterior probability that a particular effect specified by a contrast weight vector $\mathbf{C}$ exceeds some threshold γ is easily computed

$$p = 1 - \Phi\left(\frac{\gamma - \mathbf{C}^T \eta_{\theta|y}}{\sqrt{\mathbf{C}^T \mathbf{C}_{\theta|y} \mathbf{C}}}\right)$$

Φ is the cumulative density function of the unit normal distribution.

- **PPM** inference is about an effect, or activation, being greater than some specified size that has some meaning in relation to underlying neurophysiology.

  This contrasts with classical inference, in which the inference is about the effect being significantly different from zero.

- **PPM** is eschews the multiple-comparison problem. Classical inference requires an adjustment or correction to the p values which means that classical inference becomes less sensitive or powerful with large search volumes.

- We choose some fixed threshold γ, for all voxels in a classical SPM which ensures that the resulting inference has the same specificity everywhere.
- To emulate this uniform specificity, when thresholding a PPM, we would have to keep the thershold $\gamma + u\mathbf{C}_{\theta|y}$ constant, u being $\Phi^{-1}(1-p)$. The critical thing here is that if the prior covariance or observation error changes from voxel to voxel then either γ or u must change to maintain the same specificity.
- This means that the nature of the inference changes fundamentally, either in terms of the size of the inferred activation γ or the confidence about that effect u.
- Therefore, one can either have a test with uniform specificity (the classical approach) or one can infer an effect of uniform size with uniform confidence (the Bayesian approach).

- In summary, classical inference uses a criterion that renders the specificity fixed. However, this is at the price that the size of the effect, subtending the inferred activation, will change from voxel to voxel or brain region to brain region. By explicitly framing the inference in terms of the posterior probability, Bayesian inference sacrifices a constant specificity to ensure the inference is about the same thing at every voxel.
- In regions with high prior variability the classical threshold is relaxed to ensure type II errors are avoided.
- In this context the classical specificity represents the lower bound for Bayesian inference.
- In other words, Bayesian inference is generally much more specific than classical inference with equivalence when the prior variance becomes very large.

A Self Study Problem : A 2 level Model

$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}^{(1)}\theta^{(1)} + \epsilon^{(1)}, \quad \epsilon_1 \sim \mathcal{N}(\mathbf{0}, \Sigma),\ \Sigma = \lambda_1^{(1)}\mathbf{I}_t + \lambda_2^{(1)}\mathbf{C} \\
\theta^{(1)} &= \mathbf{X}^{(2)}\theta^{(2)} + \epsilon^{(2)}, \quad \epsilon_2 \sim \mathcal{N}(\mathbf{0}, diag([\lambda_1^{(2)} \dots \lambda_p^{(2)}]))
\end{aligned}
$$

$$
\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_s \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^{(1)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbf{X}_s^{(1)} \end{bmatrix} \begin{bmatrix} \theta_1^{(1)} \\ \vdots \\ \theta_s^{(1)} \end{bmatrix} + \epsilon^{(1)}
$$

$$
\mathbf{Q}_1^{(1)} = \begin{bmatrix} \mathbf{I_t} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \dots, \mathbf{Q}_s^{(1)} = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{I}_t \end{bmatrix}, \ \mathbf{I}_t\text{: White Covariance}
$$

$$
\mathbf{Q}_{s+1}^{(1)} = \begin{bmatrix} \mathbf{C} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \dots, \mathbf{Q}_{2s}^{(1)} = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{C} \end{bmatrix}, \ \mathbf{C}\text{: Colored Covariance}
$$

$$
\mathbf{X}^{(2)} = \mathbf{1}_s \otimes \mathbf{I}_p
$$

$$
\mathbf{Q}_1^{(2)} = \mathbf{I}_s \otimes \begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \dots, \mathbf{Q}_p^{(2)} = \mathbf{I}_s \otimes \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}
$$

A Self Study Problem: A two level model with data generated for 12 subjects.

- Level 1 Colored Noise Model $\mathbf{y} = \mathbf{X}^{(1)}\theta^{(1)} + \epsilon^{(1)}$ with $\mathbf{Q}_1^{(1)} = \mathbf{I}$ and $\mathbf{Q}_2^{(2)} = \mathbf{K}\mathbf{K}^T$,
  $k_{ij} = \begin{cases} a^{j-i} & i > j \\ 0 & i \leq j \end{cases}$
  a is the AR coefficient of colored noise model with hyperparameters $\lambda^{(1)} = [0.5 - 0.1]$.
  $\epsilon^{(1)}(n) = a\epsilon^{(1)}(n-1) + w(n)$, and $w(n) \sim \mathcal{N}(0, \sigma^2\mathbf{I})$
- Level 1 Parameters
  Event related pulse trains with SOA between [2.5 - 5.5] sec with a mean jitter of 4 sec.
  Two hemodynamic functions with 0 and 3 sec delays for early and late response with sampe rate 0.1 sec.
  $\mathbf{X}^{(1)}$: 2 hemodynamic functions convolved with stimulus train + DC offset
- Level 2 Parameters
  $\theta^{(1)} = \mathbf{X}^{(2)}\theta^{(2)} + \epsilon^{(2)}$ $\theta^{(2)} = [0.5\ 0\ 0]^T$ with variances $\lambda^{(2)} = [0.02\ 0.006\ 0]$.
  $\mathbf{X}^{(2)} = \mathbf{1}_{12} \otimes \mathbf{I}_3$

Exercise: ReML

Write the ReML code for 2-level Hierarchical model using the above $\mathbf{X}^{(i)}$ and $\mathbf{Q}^{(i)}$ to estimate the θ for each subject and resconstruct their HRF responses by $\mathbf{X}^{(1)}(1:2,:)\theta(1:2)$. Compare the estimated θ with their real values.

```
%%%% 2 level simulation data generation:  Ahmet Ademoglu, April 1, 2020%%%% 2 level simulation data genera
Rt = 0.1; % hrf sampling time (sec)
Ns = 12; % Number of subjects
lambda_2 = sqrt([ 0.02 0.006 0]); % level 2 hyperparameters
v = randn(Ns,3); % level 2 noise
v = (v- ones(Ns,1)*mean(v))./ (ones(Ns,1)*std(v)); % z-transform
v = v.*(ones(Ns,1)*lambda_2); v= v'; % variance scaling
theta_2 = [0.5 0 0]'; % level 2 parameters
X2 = kron(ones(Ns,1),eye(length(theta_2),length(theta_2))); % level 2 design matrix
theta_1 = X2*theta_2 + v(:); % level 1 parameters
% 1st level signal generation
SOA =rand(90,1)*3+2.5; % random event related stimuli in [2.5-5.5] with 4sec mean
S = round(10*cumsum(SOA)); % locating the time with 0.1 sec resol
P = zeros(max(S),1); P(S) = 1; % inserting pulse to event locations
[H,p]=spm_hrf(Rt); % hrf function with sample rate 0.1 sec
p(6) = 3; Hd = spm_hrf(Rt,p); % 3 sec delayed hrf function with sample rate 0.1 sec
H = H / max(H); Hd = Hd / max(Hd); % normalize peak
x1=[ conv(P,H) conv(P,Hd) ]; % convolve events with hrfs: columns of x1 are design matrix regressors
x1 = x1(100:20:end-200,:); % downsample 0.1 sec data to 2 sec data
N = size(x1,1); x1 = 1*[x1 ones(N,1)]; % add dc regressor
a = 1/exp(1); lambda_1 = [ 0.5 -0.1]; % AR(1) parameters and Hyperparameters for level 1
% Diagonal Noise Component (white) and colored Noise component
Q1 = diag(ones(N,1)); Q2 = convmtx(a.^([0:N-1])',N);
Q2 = Q2(1:N,1:N) - eye(N,N); % removing diagonal  ones
Q2 = Q2 *Q2'; Q2= Q2 - diag(diag(Q2)); % forming the color covariance structure as KK' in SPM
Ce = Q1*lambda_1(1) + Q2*lambda_1(2) ; % Error Covariance Matrix
F = chol(Ce); % Cholesky decomposition of Ce=F' * F for color filtering
w = randn(N,Ns); % generate white noise
w = (w-ones(N,1)*mean(w))./(ones(N,1)*std(w)); % correct for zero mean & normalize for unit variance
e = F*w; % Filtering the white noise to obtain colored error
y = kron(eye(Ns,Ns),x1)*theta_1; % generate level 1 noiseless data
Y = y + e(:);  Y = reshape(Y,size(Y,1)/Ns,Ns); % Signal + Noise
```

Generation of design matrices $\mathbf{X}^{(i)}$ and covariance structures $\mathbf{Q}^{(i)}$

```
% Generation of design matrices Xi covariance structures Qi
X1 = sparse(kron(eye(Ns),x1)); % 1st level design matrix
% Generation of 1st level covariance structures
for i=1:Ns,
s = zeros(Ns,Ns); s(i,i)=1;
q1{i} = full(kron( sparse(s),speye(N,N)));
q1{i+Ns} = full(kron( sparse(s),sparse(Q2)));
end;
P= (size(x1,2));
X2= kron(ones(Ns,1),eye(P)); % 2nd level design matrix
% Generation of 2nd level covariance structures
s = zeros(P,1); s(1)=1;
for i=1:P,
q2{i} = kron( eye(Ns) , diag(s) );
s=circshift(s,1);
end;
```