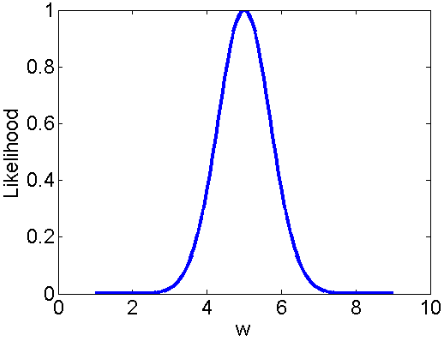
Bayesian Inference for Linear Models

Bayesian Course
Wellcome Trust Centre for Neuroimaging at UCL
Feb 2013.

Maximum Likelihood

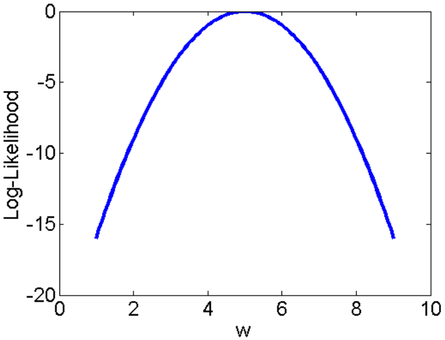
Set parameter(s) w to maximise the likelihood, $p(y|w)$.



Maximum Likelihood

Set parameter(s) w to maximise the log-likelihood

$$L = \log p(y|w)$$



The gradient $dL/dw = 0$ at the maximum.

General Linear Model

The General Linear Model (GLM) is given by

$$y = Xw + e$$

where y are data, X is a design matrix, and e are zero mean Gaussian errors with covariance V. The above equation implicitly defines the likelihood function

$$p(y|w) = \mathrm{N}(y; Xw, V)$$

where the Normal density is given by

$$\mathrm{N}(x; \mu, V) = \frac{1}{(2\pi)^{N/2}|V|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T V^{-1}(x-\mu)\right)$$

Maximum Likelihood

If we know V then we can estimate w by maximising the likelihood or equivalently the log-likelihood

$$L = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|V| - \frac{1}{2}(y - Xw)^T V^{-1}(y - Xw)$$

We can compute the gradient with help from the Matrix Reference Manual or Wikipedia's Matrix Calculus page:

$$\frac{dx^T A}{dx} = A, \frac{dAx}{dx} = A^T$$

and

$$\frac{dx^T Ax}{dx} = Ax + A^T x = 2Ax$$

if A is symmetric.

Maximum Likelihood

If we know V then we can estimate w by maximising the likelihood or equivalently the log-likelihood

$$L = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|V| - \frac{1}{2}(y - Xw)^T V^{-1}(y - Xw)$$

We can compute the gradient with help from the Matrix Reference Manual or Wikipedia's Matrix Calculus page:

$$\frac{dL}{dw} = X^T V^{-1} y - X^T V^{-1} X w$$

and set it to zero. This leads to the 'normal equations' and the solution

$$\hat{w}_{ML} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$$

This is often referred to as Weighted Least Squares (WLS), $\hat{w}_{ML} = \hat{w}_{WLS}$.

fMRI analysis

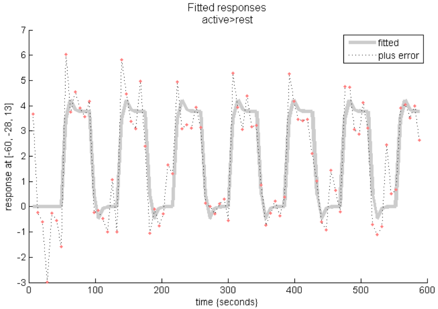
For fMRI time series analysis we have a linear model at each voxel i

$$y_i = Xw_i + e_i$$

$V_i = \text{Cov}(e_i)$ is estimated first (see later) and then the regression coefficients are computed using Maximum Likelihood (ML) estimation.

$$\hat{w}_i = (X^T V_i^{-1} X)^{-1} X^T V_i^{-1} y_i$$

The fitted responses are then $\hat{y}_i = X\hat{w}_i$ (SPM Manual)



fMRI analysis

The uncertainty in the ML estimates is given by

$$S = (X^T V_i^{-1} X)^{-1}$$

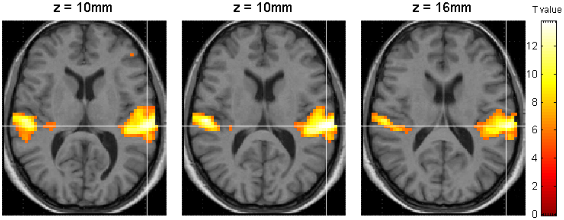
Contrast vectors c can then be used to test for specific effects

$$\mu_c = c^T \hat{w}_i$$

The uncertainty in the effect is then

$$\sigma_c^2 = c^T S c$$

and a t-score is then given by $t = \mu_c / \sigma_c$



Least Squares

For isotropic error covariance $V = \lambda I$, the normal equations are

$$\frac{dL}{dw} = \lambda X^T y - \lambda X^T X w$$

This leads to the Ordinary Least Squares (OLS) solution $\hat{w}_{ML} = \hat{w}_{OLS}$,

$$\hat{w}_{OLS} = (X^T X)^{-1} X^T y$$

Bayesian GLM

A Bayesian GLM is defined as

$$\begin{aligned} y &= Xw + e_1 \\ w &= \mu_w + e_2 \end{aligned}$$

where the errors are zero mean Gaussian with covariances $\mathrm{Cov}[e_1] = C_y$ and $\mathrm{Cov}[e_2] = C_w$.

$$\begin{aligned} p(y|w) &\propto \exp\left(-\frac{1}{2}(y - Xw)^T C_y^{-1}(y - Xw)\right) \\ p(w) &\propto \exp\left(-\frac{1}{2}(w - \mu_w)^T C_w^{-1}(w - \mu_w)\right) \end{aligned}$$

Bayesian GLM

The posterior distribution is then

$$p(w|y) \propto p(y|w)p(w)$$

Taking logs and keeping only those terms that depend on w gives

$$\begin{aligned}
\log p(w|y) &= -\frac{1}{2}(y - Xw)^T C_y^{-1}(y - Xw) \\
&- \frac{1}{2}(w - \mu_w)^T C_w^{-1}(w - \mu_w) + .. \\
&= -\frac{1}{2} w^T (X^T C_y^{-1} X + C_w^{-1}) w \\
&+ w^T (X^T C_y^{-1} y + C_w^{-1} \mu_w) + ..
\end{aligned}$$

Bayesian GLM

If $p(x) = \mathrm{N}(x; m, S)$ then

$$p(x) \propto \exp\left(-\frac{1}{2}(x-m)^T S^{-1}(x-m)\right)$$

Taking logs of the Gaussian density $p(x)$ and keeping only those terms that depend on x gives

$$\log p(x) = -\frac{1}{2}x^T S^{-1} x + x^T S^{-1} m + ..$$

For our posterior we have

$$\begin{aligned}\log p(w|y) &= -\frac{1}{2}w^T(X^T C_y^{-1} X + C_w^{-1})w \\ &+ w^T(X^T C_y^{-1} y + C_w^{-1}\mu_w) + ..\end{aligned}$$

Equating terms gives

$$\begin{aligned}p(w|y) &= \mathrm{N}(w; m_w, S_w) \\ S_w^{-1} &= X^T C_y^{-1} X + C_w^{-1} \\ m_w &= S_w(X^T C_y^{-1} y + C_w^{-1}\mu_w)\end{aligned}$$

GLM posterior

The posterior density is

$$\begin{aligned} p(w|y) &= \mathrm{N}(w; m_w, S_w) \\ S_w^{-1} &= X^T C_y^{-1} X + C_w^{-1} \\ m_w &= S_w(X^T C_y^{-1} y + C_w^{-1} \mu_w) \end{aligned}$$

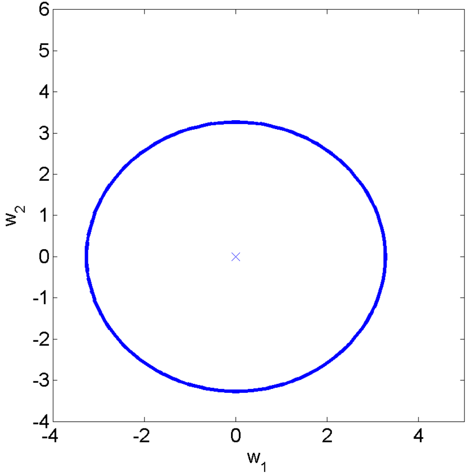
The posterior precision is the sum of the prior precision and the data precision.

The posterior mean is a relative precision weighted combination of the data mean and the prior mean.

If $\mu_w = 0$ we have a *shrinkage prior*.

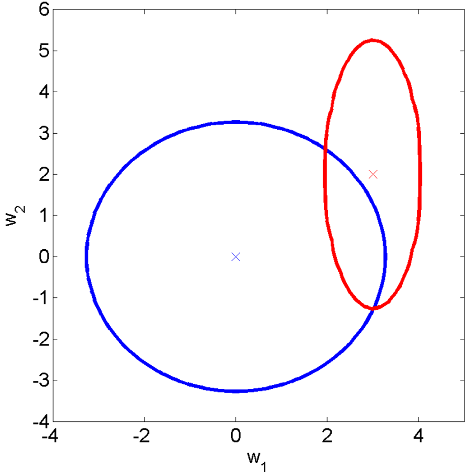
Bayesian GLM with two parameters

The prior has mean $\mu_w = [0, 0]^T$ (cross) and precision $C_w^{-1} = \text{diag}([1, 1])$.



Bayesian GLM with two parameters

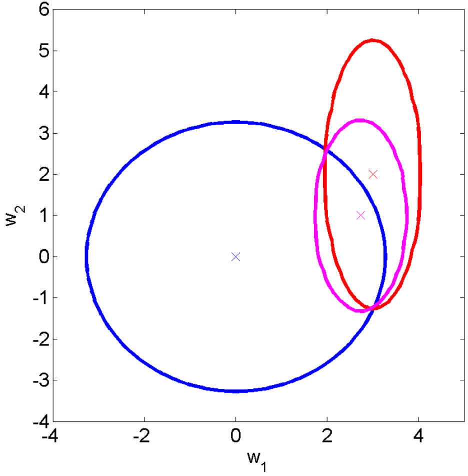
The likelihood has mean $X^T y = [3, 2]^T$ (circle) and precision $(X^T C_y^{-1} X)^{-1} = \text{diag}([10, 1])$.



Bayesian Inference for Linear Models
Maximum Likelihood
Linear Models
fMRI analysis
Bayesian Linear Models
fMRI example
Augmented Form
MAP Learning
MEG Source Reconstruction
References

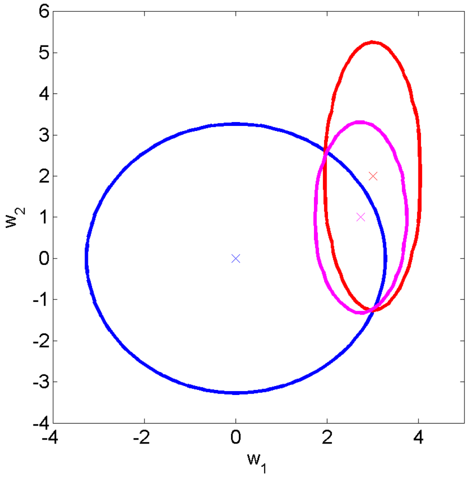
Bayesian GLM with two parameters

The posterior has mean $m = [2.73, 1]^T$ (cross) and precision $S_w^{-1} = \text{diag}([11, 2])$.



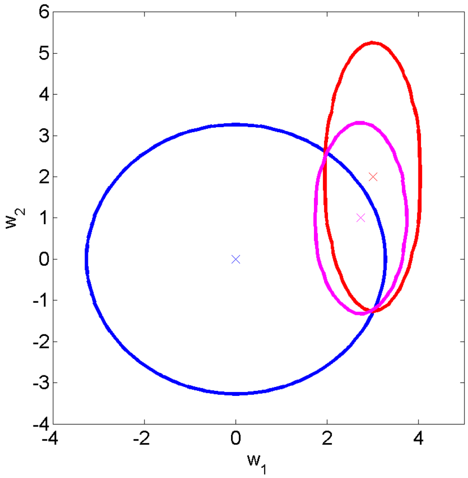
Bayesian GLM with two parameters

In this example, the measurements are more informative about w_1 than w_2. This is reflected in the posterior distribution.



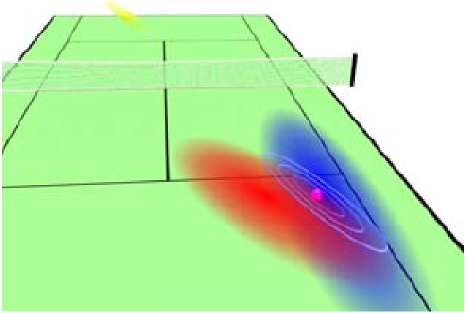
Shrinkage Prior

If $\mu_w = 0$ we have a *shrinkage prior*.



Tennis

From Wolpert and Ghahramani (2006)



$$
\begin{aligned}
p(w|y) &= \mathrm{N}(w; m_w, S_w) \\
S_w^{-1} &= X^T C_y^{-1} X + C_w^{-1} \\
m_w &= S_w(X^T C_y^{-1} y + C_w^{-1} \mu_w)
\end{aligned}
$$

fMRI example - contrast

Given a contrast, C, testing for effect $s = C^T w$ the posterior distribution of the effect is

$$p(s|Y) = \mathrm{N}(s; m_s, C_s)$$

where

$$m_s = C^T m_w$$

and

$$C_s = C^T S_W C$$

For example $C^T = [1 - 1]$ to look for difference between two conditions.

fMRI example - Effect Size

Define an effect size threshold, s_t

Define the following classifications

Positively Activated

$$p_{posact} = p(s > s_t | y)$$

Negatively Activated

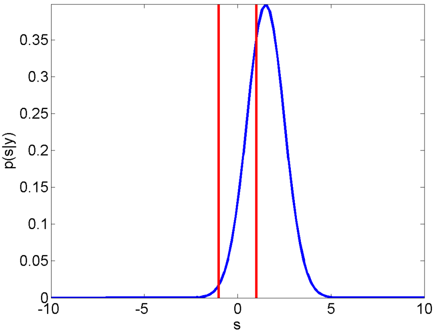
$$p_{negact} = p(s < -s_t | y)$$

Not Activated

$$p_{notact} = p(-s_t \leq s \leq s_t | y)$$

fMRI example

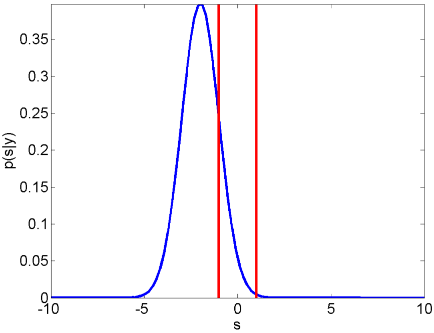
Weakly positively activated voxel



$p_{posact} = 0.69$, $p_{negact} = 0.01$, $p_{notact} = 0.3$

fMRI example

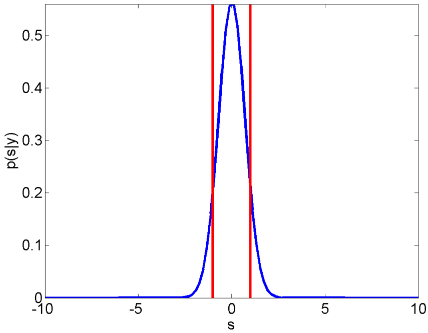
Weakly negatively activated voxel



$$p_{posact} = 0.00,\ p_{negact} = 0.84,\ p_{notact} = 0.16$$

fMRI example

Weakly not activated voxel



$p_{posact} = 0.16$, $p_{negact} = 0.16$, $p_{notact} = 0.68$

fMRI example - Odds Ratios

Define Activated = positive or negative

$$p_{act} = p_{posact} + p_{negact}$$

Activated odds

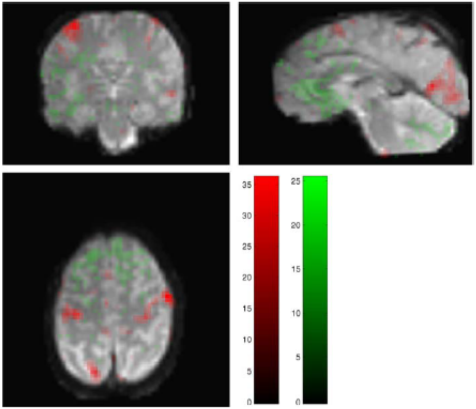
$$R_{act} = \frac{p_{act}}{1 - p_{act}}$$

Deactivated odds

$$R_{notact} = \frac{p_{notact}}{1 - p_{notact}}$$

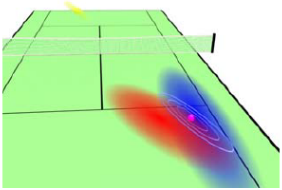
fMRI example - Finger Tapping

log R_{act} (Red) and log R_{notact} (Green)



Finger tapping task by Joerg Magergurth et al. (ISMRM, 2013). Only plot voxels for which log $R \geq 10$.

Augmented Form



The posterior over w

$$
\begin{aligned}
S_w^{-1} &= X^T C_y^{-1} X + C_w^{-1} \\
m_w &= S_w (X^T C_y^{-1} y + C_w^{-1} \mu_w)
\end{aligned}
$$

can also be written in a more compact form.

Augmented Form

This compact form is

$$
\begin{aligned}
S_w^{-1} &= \bar{X}^T V^{-1} \bar{X} \\
m_w &= S_w (\bar{X}^T V^{-1} \bar{y})
\end{aligned}
$$

where

$$
\begin{aligned}
\bar{X} &= \begin{bmatrix} X \\ I_p \end{bmatrix} \\
V &= \begin{bmatrix} C_y & 0 \\ 0 & C_w \end{bmatrix} \\
\bar{y} &= \begin{bmatrix} y \\ \mu_w \end{bmatrix}
\end{aligned}
$$

where we've augmented the data matrix with prior expectations; $\bar{y}$ is $(d+p) \times 1$ and $\bar{X}$ is $(d+p) \times p$.

Augmented Form

Estimation in a Bayesian GLM is therefore equivalent to Maximum Likelihood estimation (ie. for IID covariances this is the same as Weighted Least Squares) with *augmented* data.

$$m_w = (\bar{X}^T V^{-1} \bar{X})^{-1} \bar{X}^T V^{-1} \bar{y}$$

Prior beliefs can be thought of as extra data points.

MAP Learning

The posterior density is given by Bayes rule

$$p(w|y) = \frac{p(y|w)p(w)}{p(y)}$$

The Maximum A Posterior (MAP) estimate is given by

$$\hat{w} = \arg\max_{w} p(w|y)$$

Because the maxima of $\log[x]$ is the same as the maximum of x we can also write

$$\hat{w} = \arg\max_{w} L(y, w)$$

where

$$L = \log[p(y|w)p(w)]$$

is the joint log likelihood. For Linear Gaussian models MAP parameters are equivalent to the posterior mean.

MAP Learning and regularised least squares

With

$$
\begin{aligned}
p(y|w) &= \mathrm{N}(y; Xw, \lambda_1^{-1} I) \\
p(w) &= \mathrm{N}(w; 0, \lambda_2^{-1} C_w)
\end{aligned}
$$

we have

$$
\begin{aligned}
L(y,w) &= \log[p(y|w)p(w)] \\
&= -\frac{\lambda_1}{2}(y - Xw)^T (y - Xw) - \frac{\lambda_2}{2} w^T C_w^{-1} w
\end{aligned}
$$

So we are trying to minimise

$$
(y - Xw)^T (y - Xw) + \frac{\lambda_2}{\lambda_1} w^T C_w^{-1} w
$$

a data-dependent error term and a regularisation term

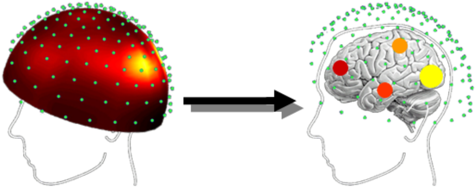
MEG Source Reconstruction

MEG Source Reconstruction is achieved through inversion of the linear model

$$y = Xw + e$$

$$(d \times 1) = (d \times p)(p \times 1) + (d \times 1)$$

for MEG data, y with d sensors and p potential sources, w, lying perpendicular to the cortical surface. The lead field matrix is specified by X. For our example we have $d = 274$ and $p = 8192$.



The above equation is for a single time point.

Generative Models

Likelihood

$$p(y|w) = \mathrm{N}(y; Xw, C_y)$$

Prior

$$p(w) = \mathrm{N}(w; 0, C_w)$$

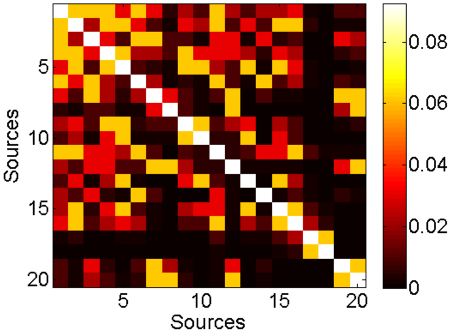
We let

$$\begin{aligned} C_y &= \lambda_1 Q_1 \\ C_w &= \lambda_2 Q_2 \end{aligned}$$

For shrinkage priors $Q_2 = I_p$, MAP estimation results in the minimum norm method of source reconstruction. This is implemented in SPM as the 'IID' option

Smoothness Priors

For smoothness priors $Q_2 = KK^T$ corresponding to the operation of a Gaussian smoothing kernel, MAP estimation results something similar to the Low Resolution Tomography (LORETA) method.



This is implemented in SPM as the ‘COH’ option. Note, these are not location priors.

Posterior Density

From earlier we have

$$\begin{aligned} S_w^{-1} &= X^T C_y^{-1} X + C_w^{-1} \\ m_w &= S_w X^T C_y^{-1} y \end{aligned}$$

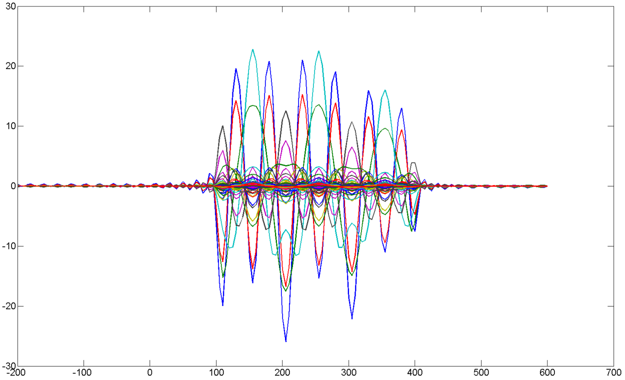
However, S_w is $p \times p$ with $p = 8192$ so cannot be inverted easily. But we can use the matrix inversion lemma, also known as the Woodbury identity (Bishop, 2006)

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

to ensure that only $d \times d$ matrices need inverting. See 'Bayesian MEG' notes on website.

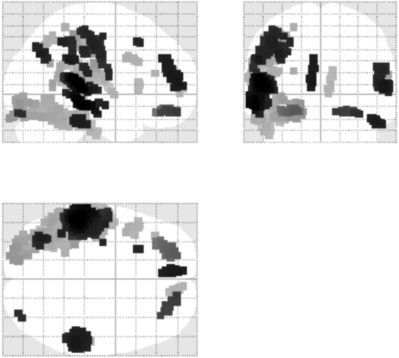
Simulation

Two sinusoidal sources were placed in bilateral auditory cortex and produced this MEG data (Barnes, 2010), comprising $d = 274$ time series (butterfly plot)



LORETA

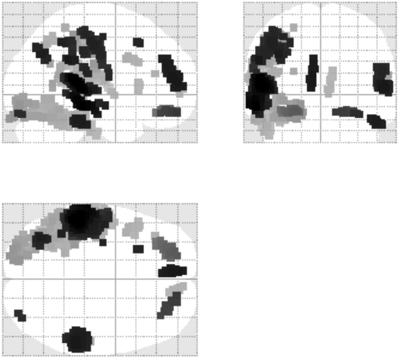
We fix $\lambda_1 = 1$. Here we set $\lambda_2 = 0.01$.



This shows the posterior mean activity for the 500 dipoles with the greatest power (over peristimulus time)

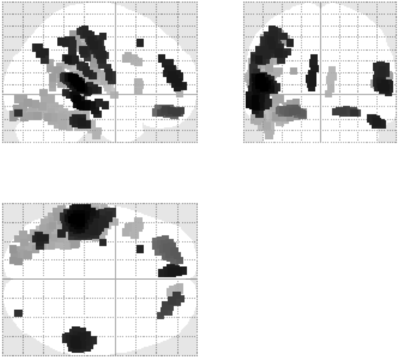
LORETA

We fix $\lambda_1 = 1$. Here we set $\lambda_2 = 0.01$.



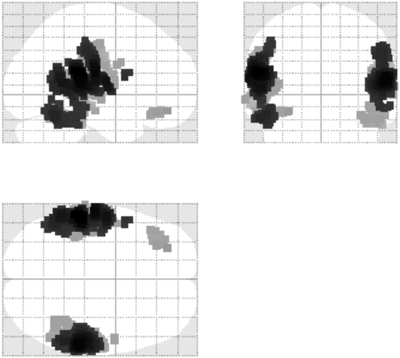
LORETA

We fix $\lambda_1 = 1$. Here we set $\lambda_2 = 0.1$.



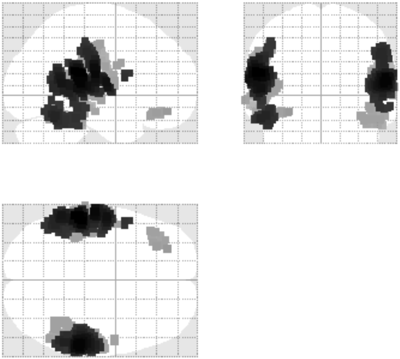
LORETA

We fix $\lambda_1 = 1$. Here we set $\lambda_2 = 1$.



LORETA

We fix $\lambda_1 = 1$. Here we set $\lambda_2 = 1$.



Use Empirical Bayes to optimise λ_2 or multiple hyperparameters.

References

G. Barnes (2010) MEG Source Localisation, SPM Manual, Chapter 35

C. Bishop (1995) Neural Networks for Pattern Recognition. OUP.

W. Penny, J Kilner and F.Blankenburg (2007) Neuroimage 36, 661-671.

SPM Manual. http://www.fil.ion.ucl.ac.uk/spm/doc/

D. Wolpert and Z. Ghahramani (2004) In Gregory RL (ed) Oxford Companion to the Mind, OUP.