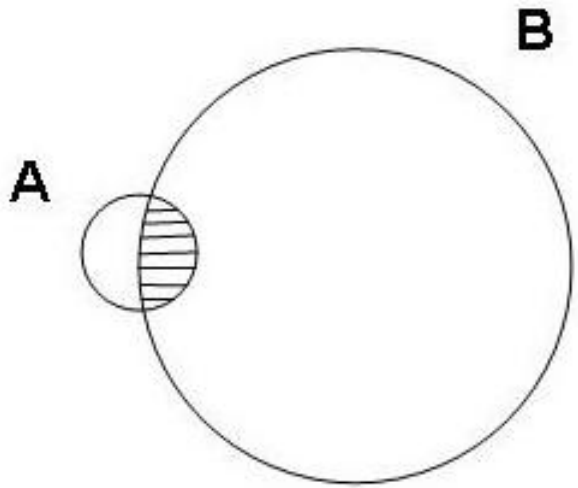
Bayesian Methods in Brain Imaging

Will Penny

First Technical Course, European Centre for Soft Computing, Mieres, Spain.
4th July 2011

Bayes rule

Given probabilities $p(A)$, $p(B)$, and the joint probability $p(A, B)$, we can write the conditional probabilities



$$p(B|A) = \frac{p(A,B)}{p(A)}$$

$$p(A|B) = \frac{p(A,B)}{p(B)}$$

Eliminating $p(A, B)$ gives Bayes rule

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

Bayes rule

In the context of model fitting, if we have data y and model parameters w, the terms in Bayes rule

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

are referred to as the prior, $p(w)$, the likelihood, $p(y|w)$, and the posterior, $p(w|y)$.

The probability $p(y)$ is a normalisation term and can be found by *marginalisation*. For continuously valued parameters

$$p(y) = \int p(y|w)p(w)dw$$

or for discrete parameters

$$p(y) = \sum_i p(y|w_i)p(w_i)$$

$p(y)$ is referred to as the marginal likelihood or model evidence.

Medical Decision Making

Johnson et al (2001) consider Bayesian inference in for Magnetic Resonance Angiography (MRA). An Aneurysm is a localized, blood-filled balloon-like bulge in the wall of a blood vessel. They commonly occur in arteries at the base of the brain. There are two tests:

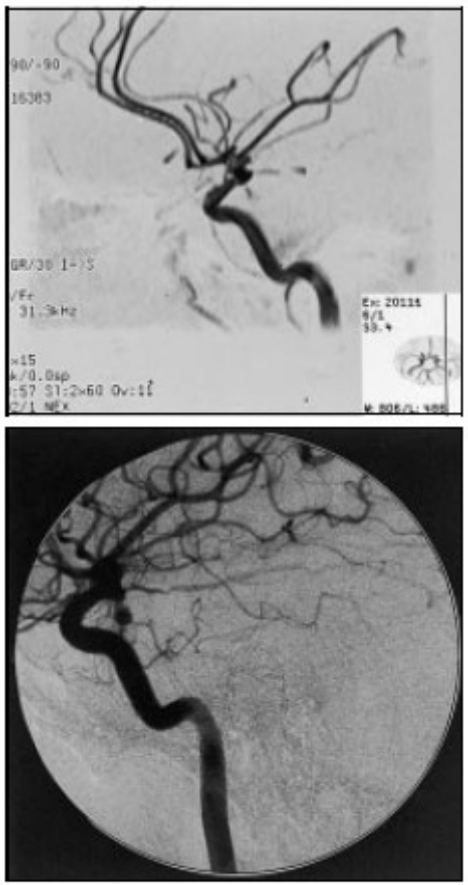


Fig 1 Case 1: magnetic resonance angiography with maximum intensity projection has normal appearance (top), whereas intra-arterial digital subtraction angiography (injection of left internal carotid artery) shows a large left posterior communicating artery aneurysm (bottom)

(1) MRA can miss sizable Intracranial Aneurysms (IA)'s but is non-invasive (top).

(2) Intra-Arterial Digital Subtraction Angiography (DSA) (bottom) is the gold standard method for detecting IA but is an *invasive* procedure requiring local injection of a contrast agent via a tube inserted into the relevant artery.

Medical Decision Making

Given patient 1's symptoms (oculomotor palsy), the prior probability of IA (prior to MRA) is believed to be 90%.

For IAs bigger than 6mm MRA has a sensitivity and specificity of 95% and 92%.

What then is the probability of IA given a *negative* MRA test result ?

Medical Decision Making

The probability of IA given a negative test can be found from Bayes rule

$$p(IA = 1|MRA = 0) = \frac{p(MRA = 0|IA = 1)p(IA = 1)}{p(MRA = 0|IA = 1)p(IA = 1) + p(MRA = 0|IA = 0)p(IA = 0)}$$

where $p(IA = 1)$ is the probability of IA prior to the MRA test. MRA test sensitivity and specificity are

$$p(MRA = 1|IA = 1)$$
$$p(MRA = 0|IA = 0)$$

We have $p(MRA = 0|IA = 1) = 1 - p(MRA = 1|IA = 1)$

Medical Decision Making

Negative test result

Prior (clinical) probability = 0.90

$$\text{Posterior probability} = \frac{(1-\text{sensitivity}) \times \text{prior probability}}{(1-\text{sensitivity}) \times \text{prior probability} + \text{specificity} \times (1-\text{prior probability})}$$

$$= \frac{(1-0.95) \times 0.90}{(1-0.95) \times 0.90 + 0.92 \times (1-0.90)}$$

Posterior probability = 0.3285

Positive test result

Prior (clinical) probability = 0.90

$$\text{Posterior probability} = \frac{\text{sensitivity} \times \text{prior probability}}{(\text{sensitivity} \times \text{prior probability}) + (1-\text{specificity}) \times (1-\text{prior probability})}$$

$$= \frac{0.95 \times 0.90}{(0.95 \times 0.90) + (1-0.92) \times (1-0.90)}$$

Posterior probability = 0.9907

Fig 3 Probability of a posterior communicating artery aneurysm given a negative or positive result from magnetic resonance angiography and a prior clinical probability of 90%. Sensitivity and specificity of angiography are 95% and 92% respectively. Probabilities are expressed between 0.0 (0%) and 1.0 (100%)

Medical Decision Making

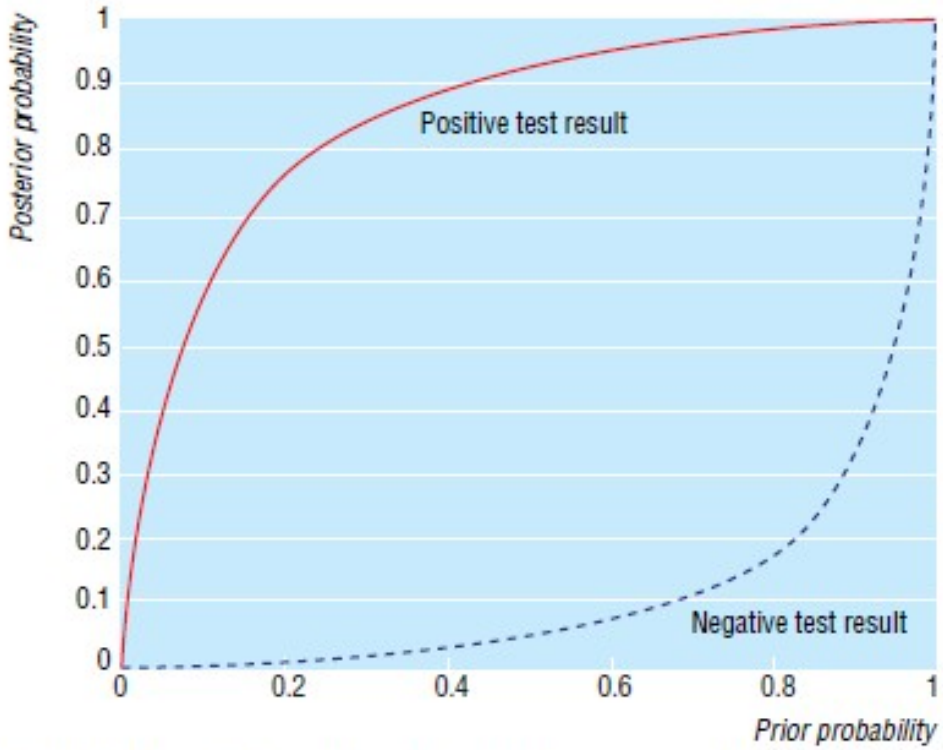
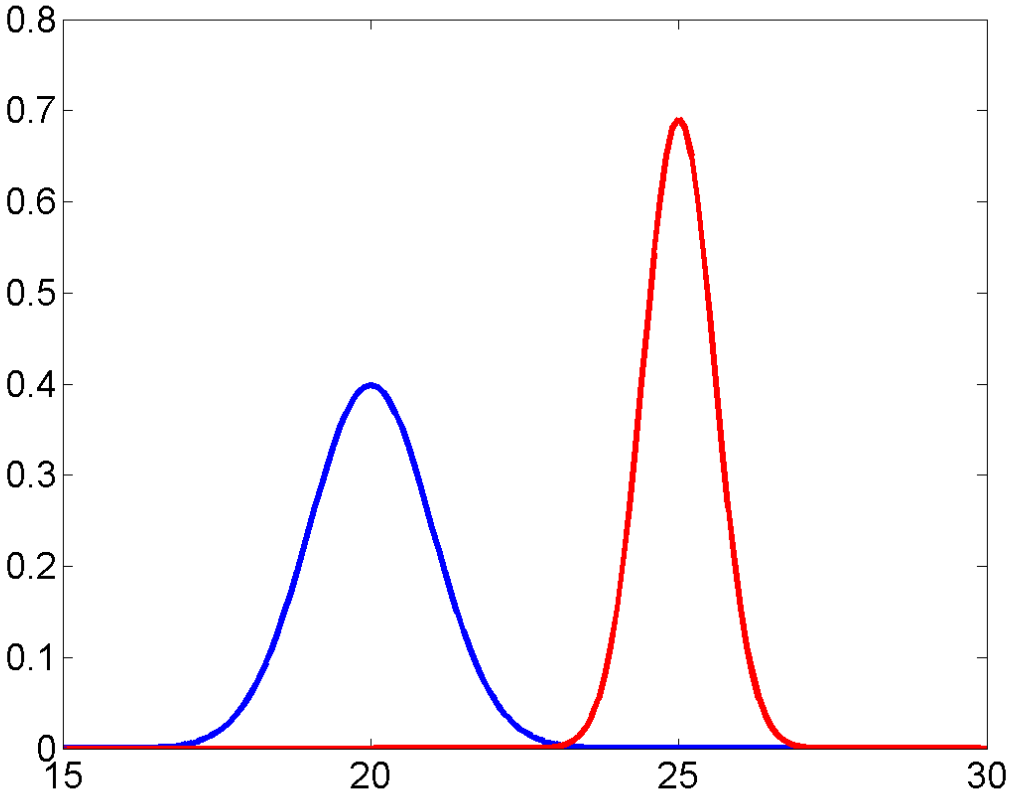


Fig 4 Influence of prior clinical probability on the probability of a disease after a negative or positive test result. Test sensitivity and specificity are 95% and 92% respectively

A negative MRA cannot therefore be used to exclude a diagnosis of IA. In both reported cases IA was initially excluded, until other symptoms developed or other tests also proved negative.

Optimal Data Fusion

For the prior (blue) we have $m_0 = 20$, $\lambda_0 = 1$ and for the likelihood (red) $m_D = 25$ and $\lambda_D = 3$.



Precision, λ, is inverse variance.

Bayes rule for Gaussians

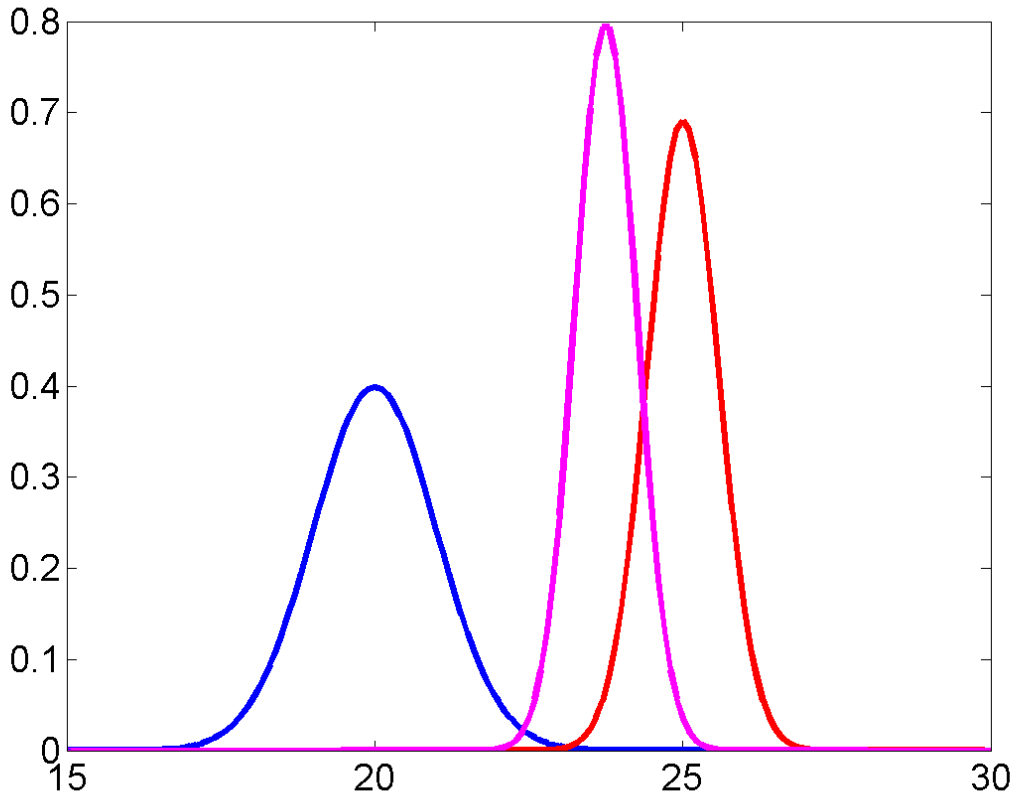
For a Gaussian prior with mean m_0 and precision λ_0, and a Gaussian likelihood with mean m_D and precision λ_D the posterior is Gaussian with

$$\begin{aligned} \lambda &= \lambda_0 + \lambda_D \\ m &= \frac{\lambda_0}{\lambda} m_0 + \frac{\lambda_D}{\lambda} m_D \end{aligned}$$

So, (1) precisions add and (2) the posterior mean is the sum of the prior and data means, but each weighted by their relative precision.

Bayes rule for Gaussians

For the prior (blue) $m_0 = 20$, $\lambda_0 = 1$ and the likelihood (red) $m_D = 25$ and $\lambda_D = 3$, the posterior (magenta) shows the posterior distribution with $m = 23.75$ and $\lambda = 4$.



The posterior is closer to the likelihood because the likelihood has higher precision.

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

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

where the Normal density is given by

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

Maximum Likelihood

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

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

We can compute the gradient with help from the Matrix Reference Manual

$$\frac{dL}{dw} = X^T C_y^{-1} y - X^T C_y^{-1} Xw$$

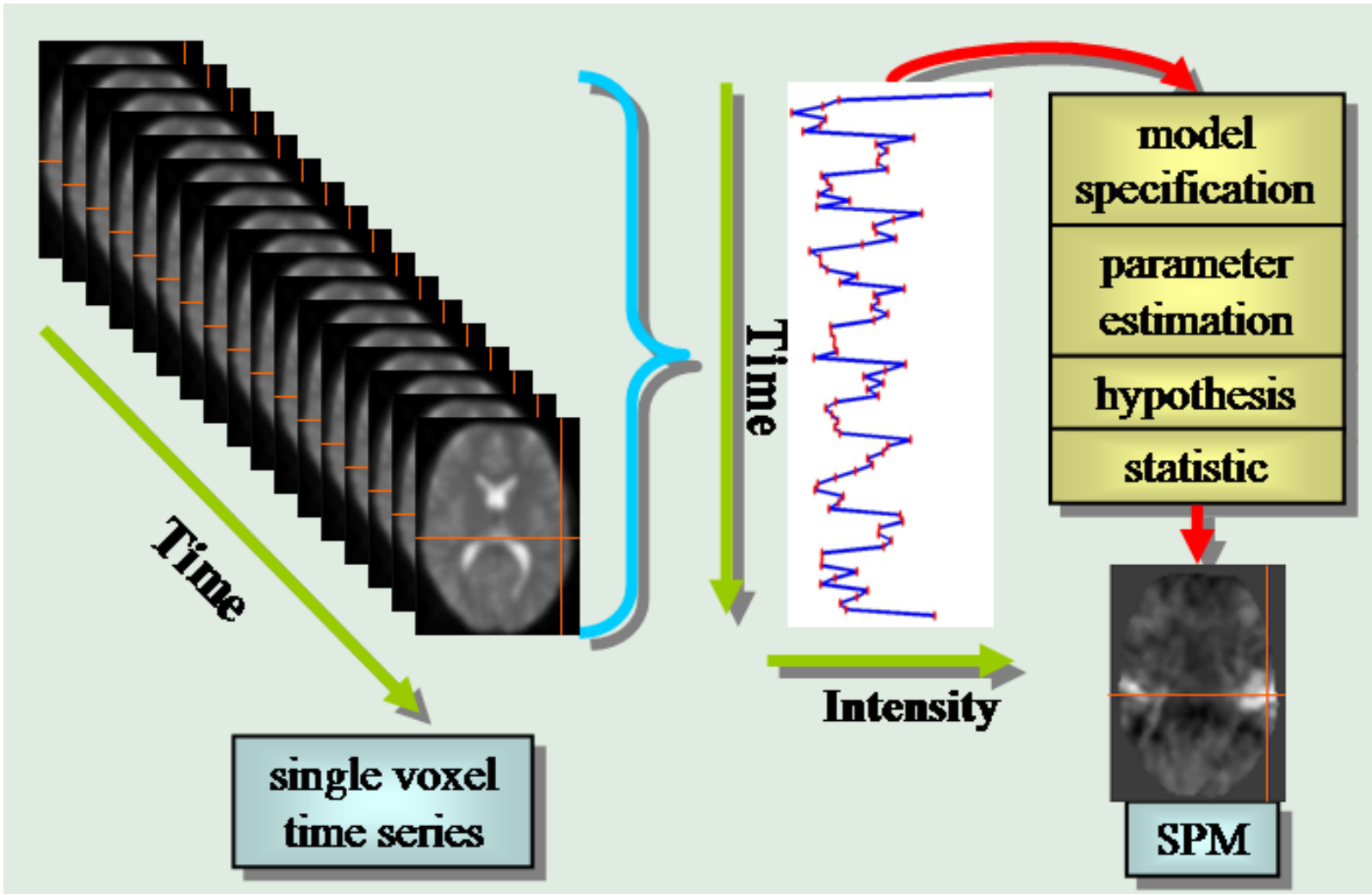
to zero. This leads to the solution

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

This is the Maximum Likelihood (ML) solution.

fMRI time series analysis

In software such as SPM or FSL brain mapping is implemented with the following method. At the *i*th voxel (volume element) we have time series y_i



fMRI time series analysis

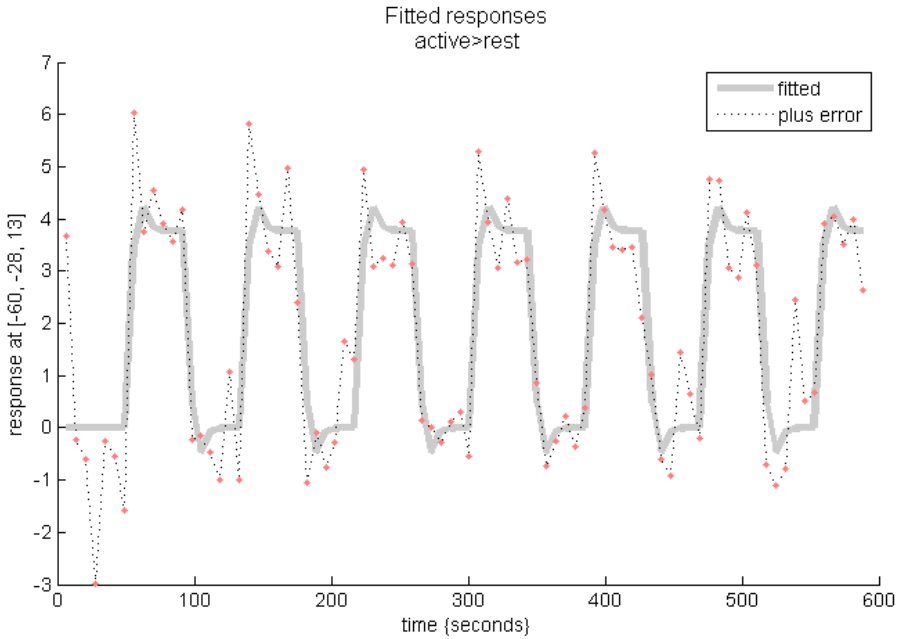
In a standard analysis linear models are fitted separately at each voxel

$$y_i = Xw_i + e_i$$

$C_y = \mathrm{Cov}(e_i)$ is the error covariance and then the regression coefficients are computed using Maximum Likelihood (ML) estimation

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

The fitted responses are then $\hat{y}_i = X\hat{w}_i$



Statistical Parametric Maps

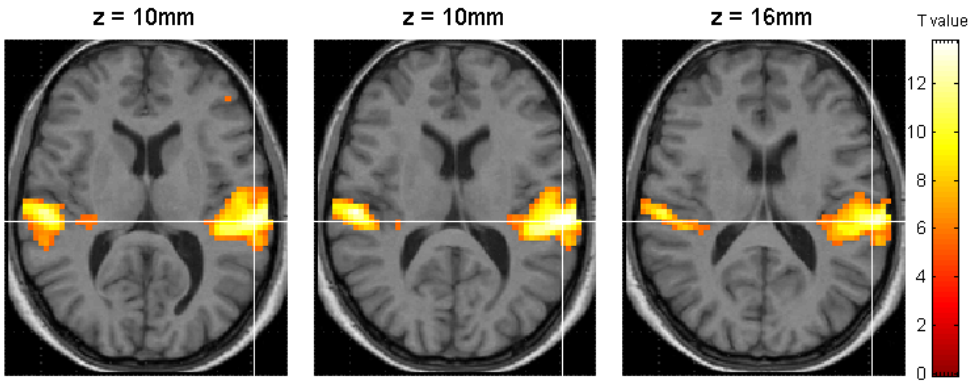
We can then test if effects are significantly non-zero.

The uncertainty in the ML estimates is given by

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

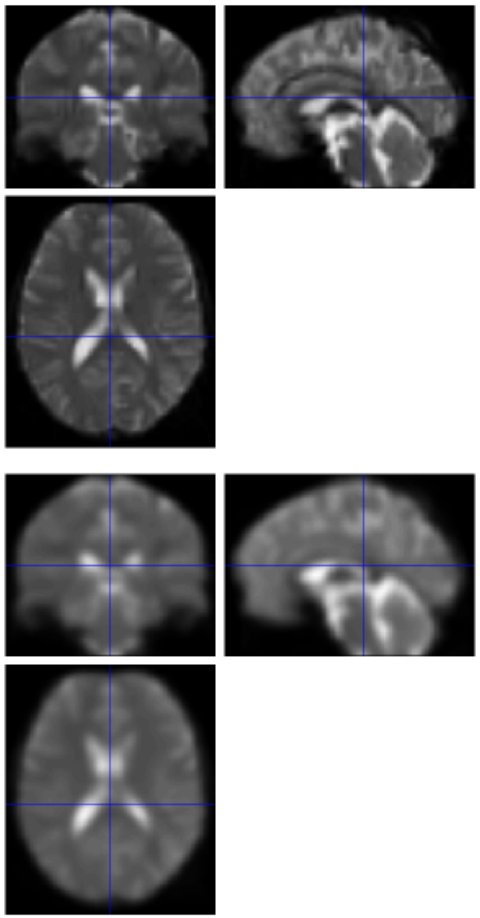
A t-score is then given by

$$t_i = \hat{w}_i(k) / \sqrt{S(k,k)}$$



Data smoothing

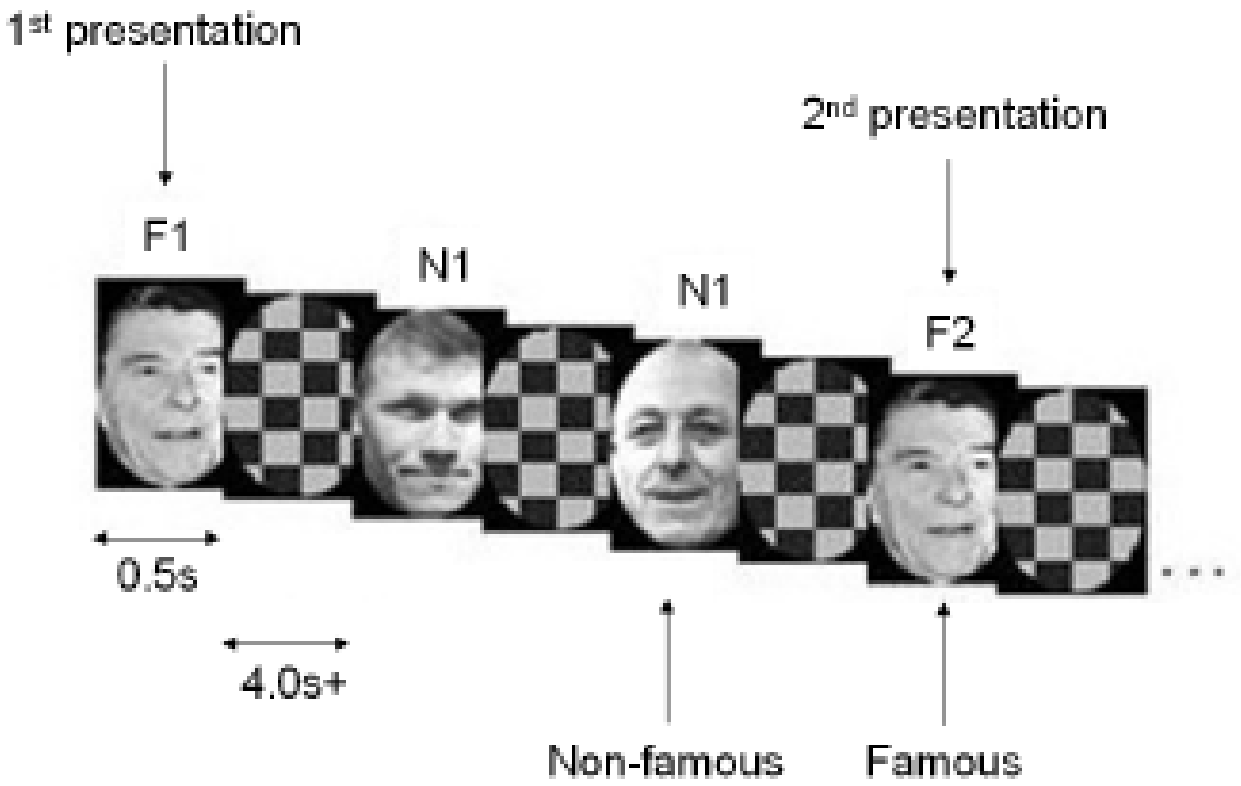
The standard approach in SPM and FSL smooths the data, y, before fitting time series models at each voxel. This increases the SNR.



Top 3 views show original fMRI images, bottom 3 views show data smoothed by a Gaussian kernel of width 6mm.

Multiple effects

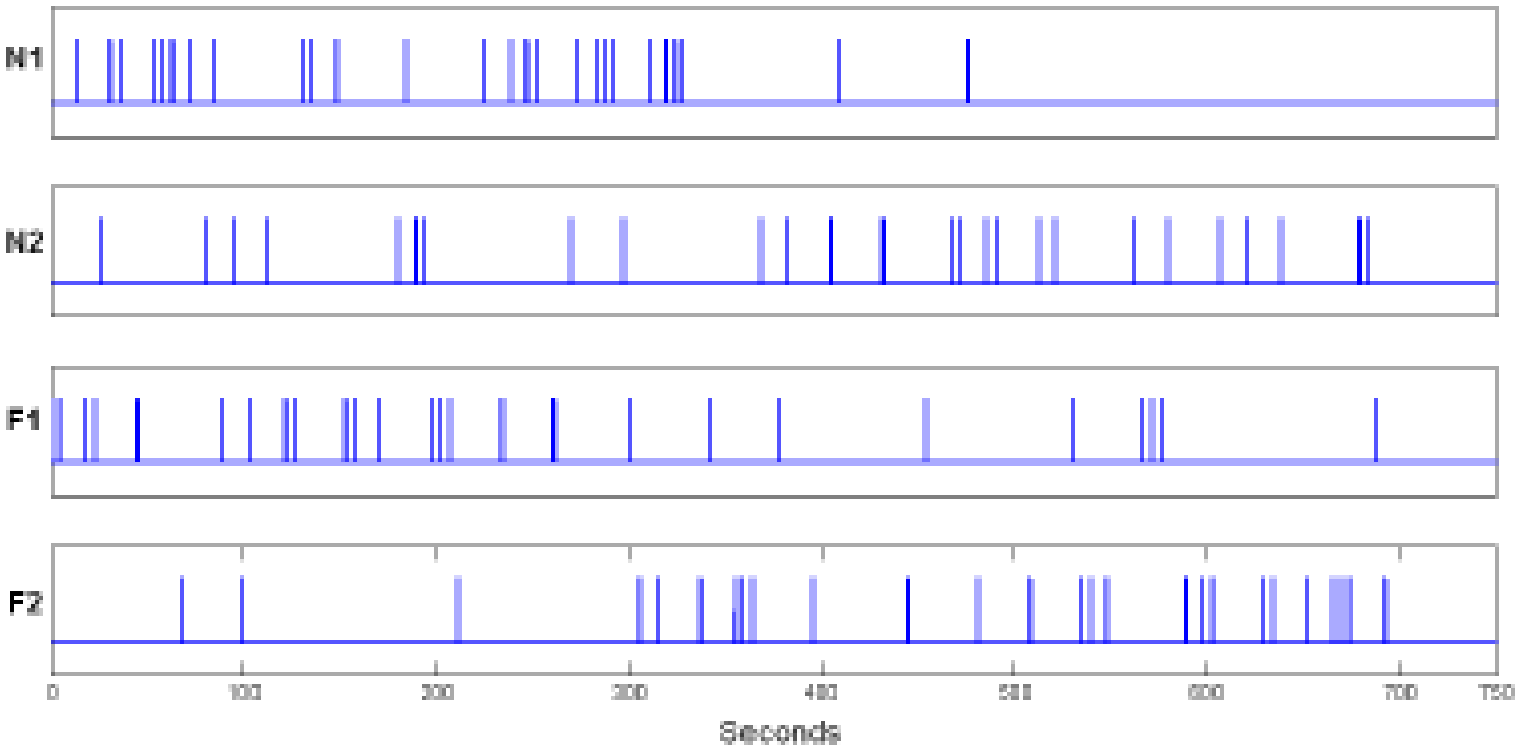
Generally, we are looking to test for multiple effects at each point in the brain.



Henson et al. (2002) was looking to find which brain regions show different responses for repeated and familiar stimuli in the context of face processing.

Multiple effects

Generally, we are looking to test for multiple effects at each point in the brain.

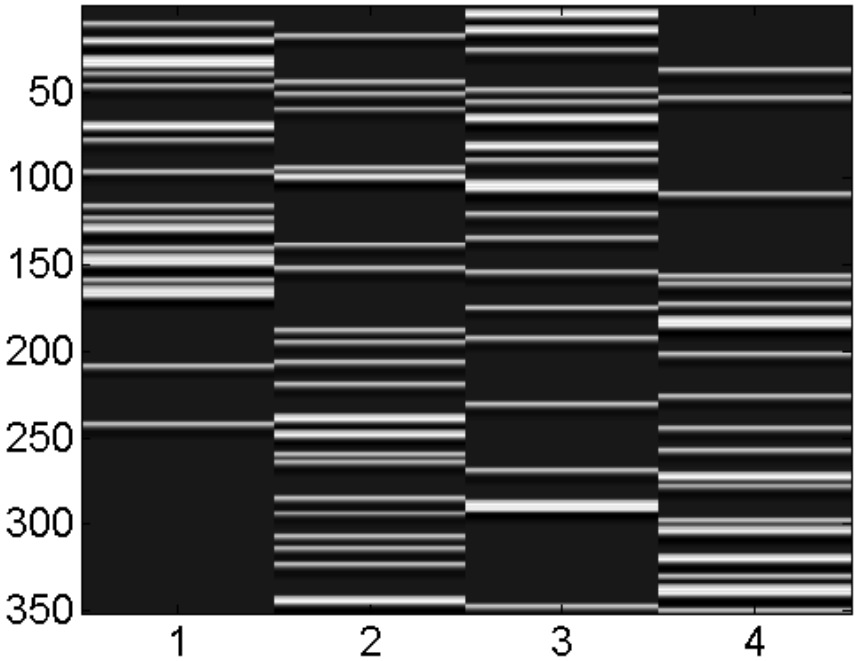


Multiple effects

Generally, we are looking to test for multiple effects at each point in the brain. At voxel i

$$y_i = Xw_i + e_i$$

where eg. X is as below and w_i is a 4-element vector.



Contrasts

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

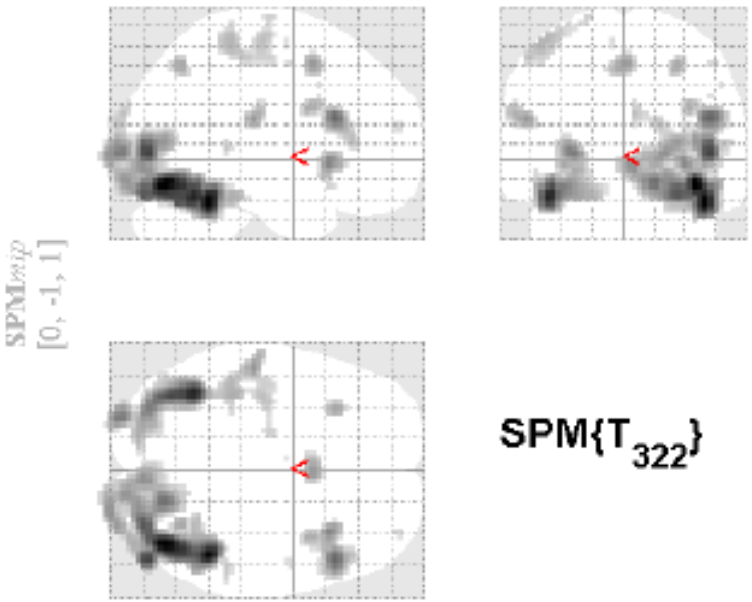
$$\mu_c = c^T \hat{w}_i$$

For example, to look at the average response to faces (regardless of type) $c = [1/4, 1/4, 1/4, 1/4]$.

The uncertainty in the effect is then

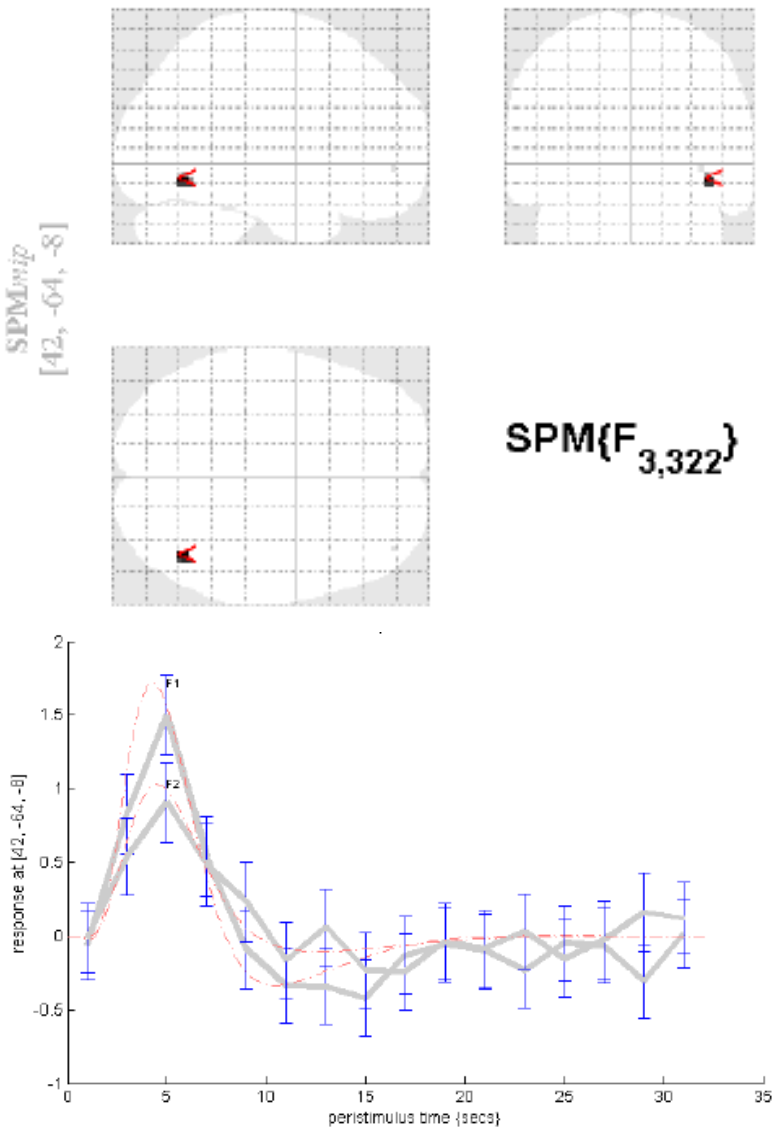
$$\sigma_c^2 = c^T S c$$

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



Contrasts

To look at the effect of repetition
$c = [1/2, -1/2, 1/2, -1/2]$.



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 $\text{Cov}[e_1] = C_y$ and $\text{Cov}[e_2] = C_w$.

$$\begin{aligned} p(y|w) &= N(y; Xw, C_y) \\ p(w) &= N(w; \mu_w, C_w) \end{aligned}$$

GLM posterior

The posterior density is (Bishop, 2006)

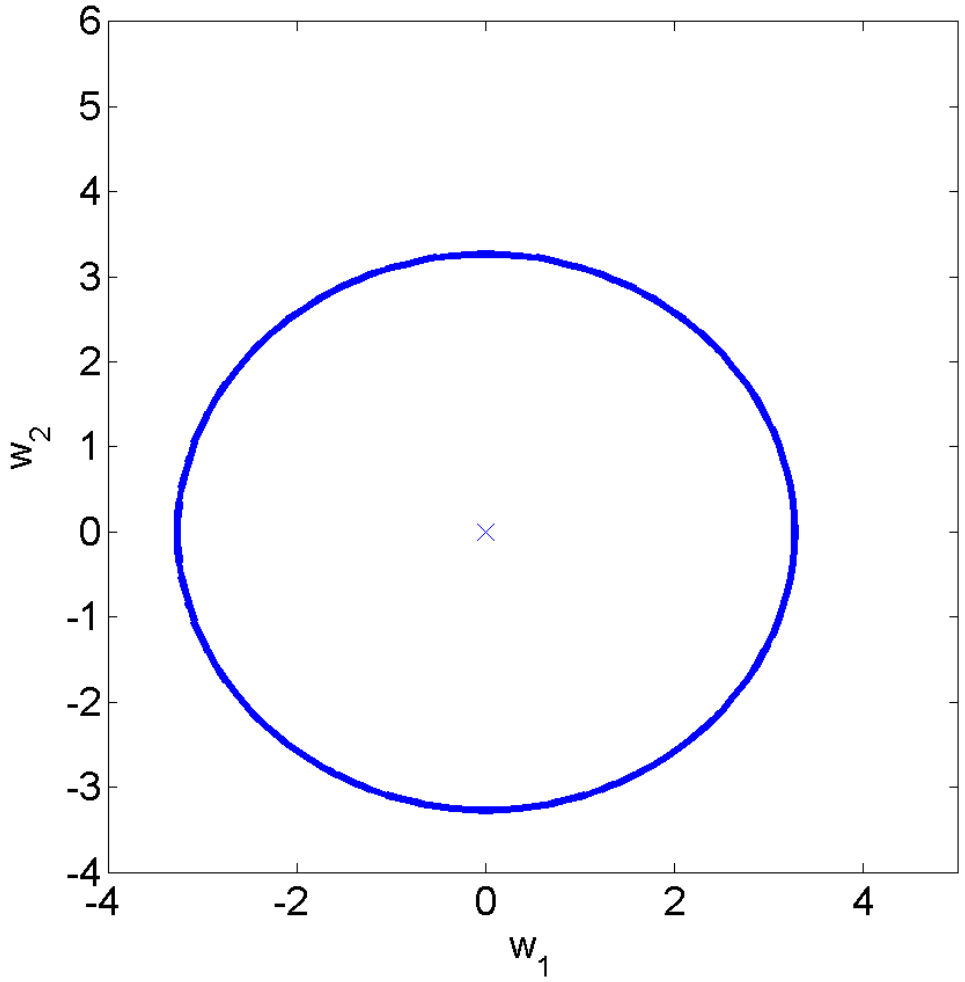
$$\begin{aligned} p(w|y) &= \mathrm{N}(w; m_w, S_w) \\ S_w^{-1} &= X^T C_y^{-1} X + C_w^{-1} \\ m_w &= S_w(X^T C_y^{-1} y + C_w^{-1} \mu_w) \end{aligned}$$

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.

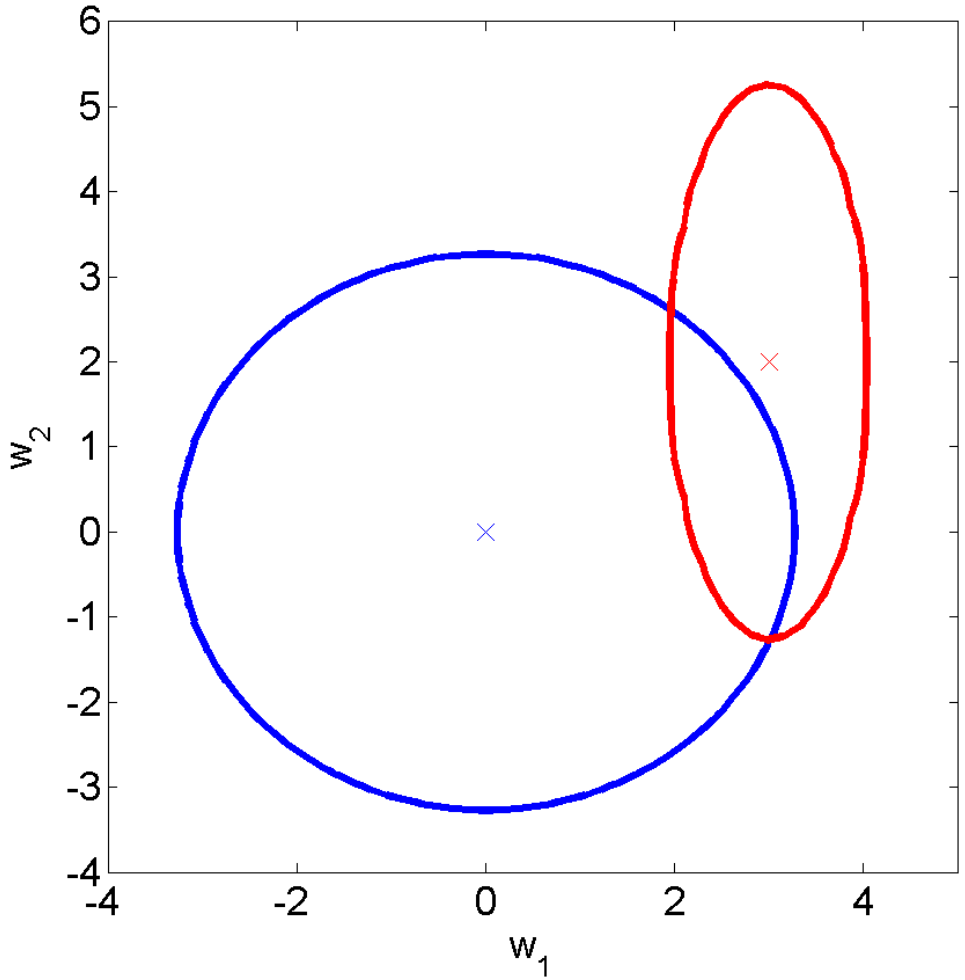
Bayesian GLM with two parameters

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



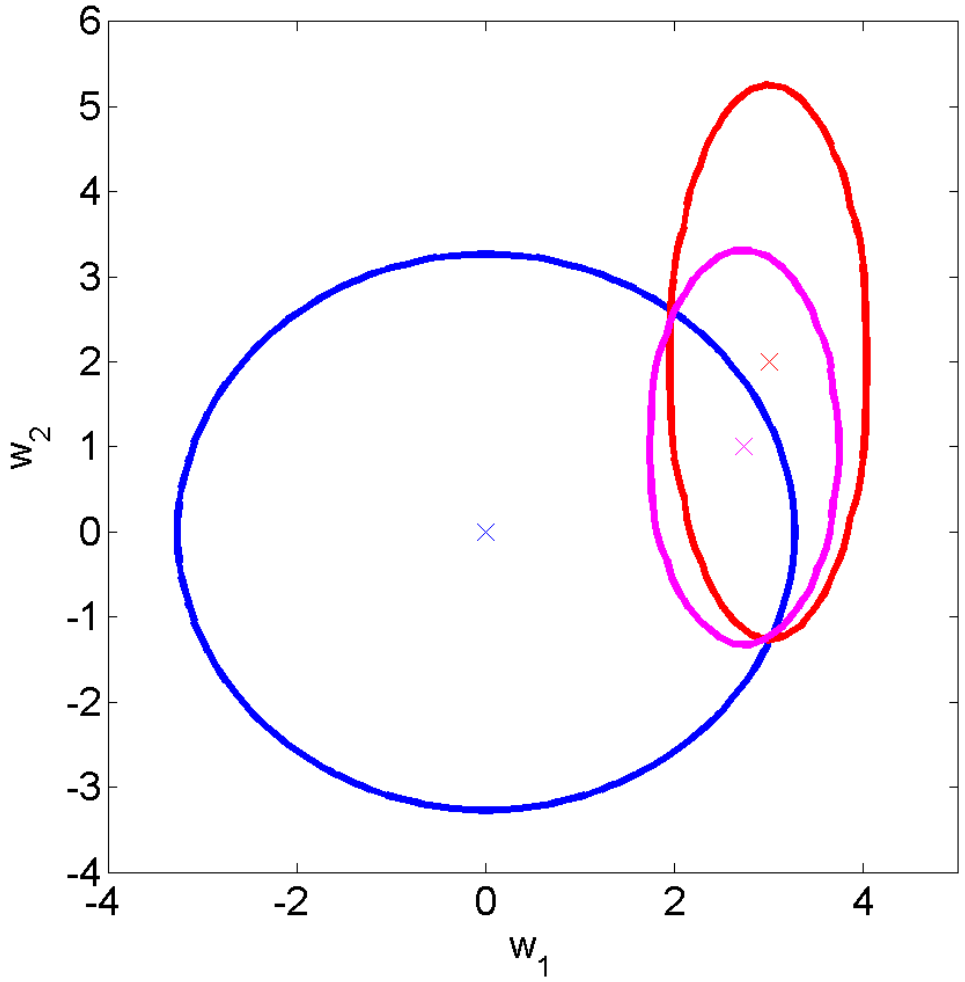
Bayesian GLM with two parameters

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



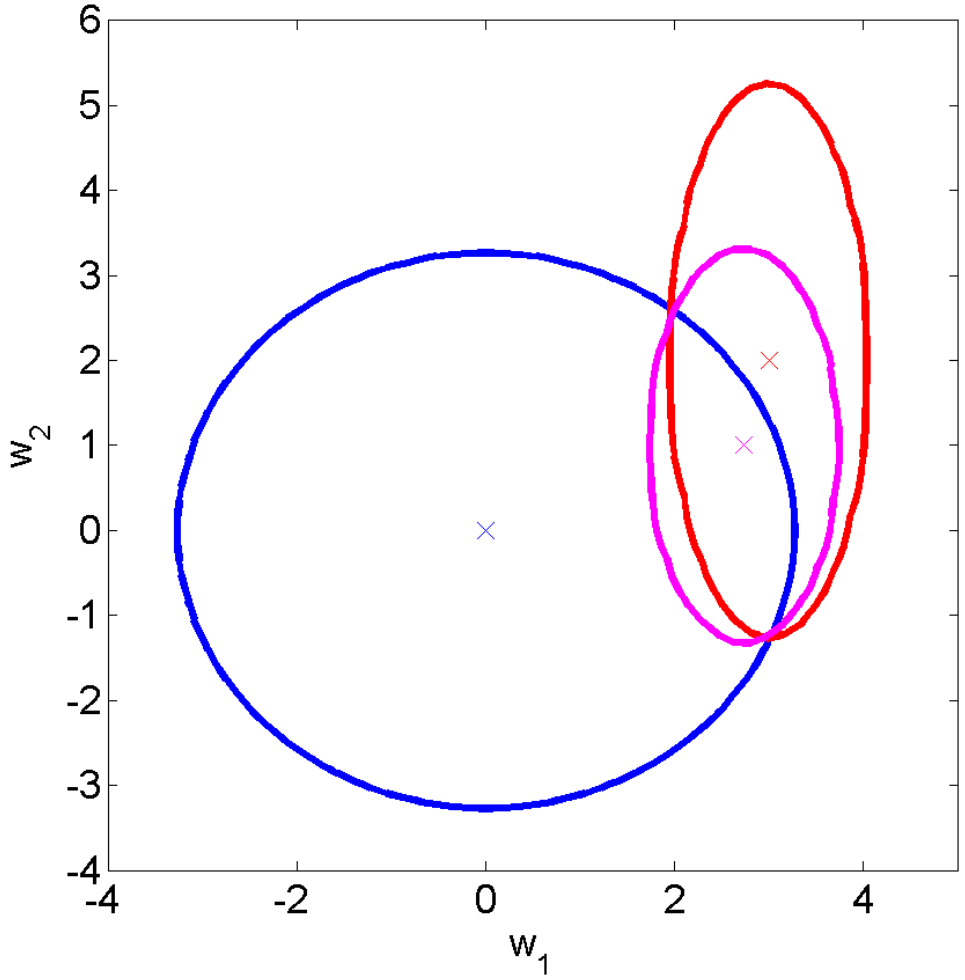
Bayesian GLM with two parameters

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



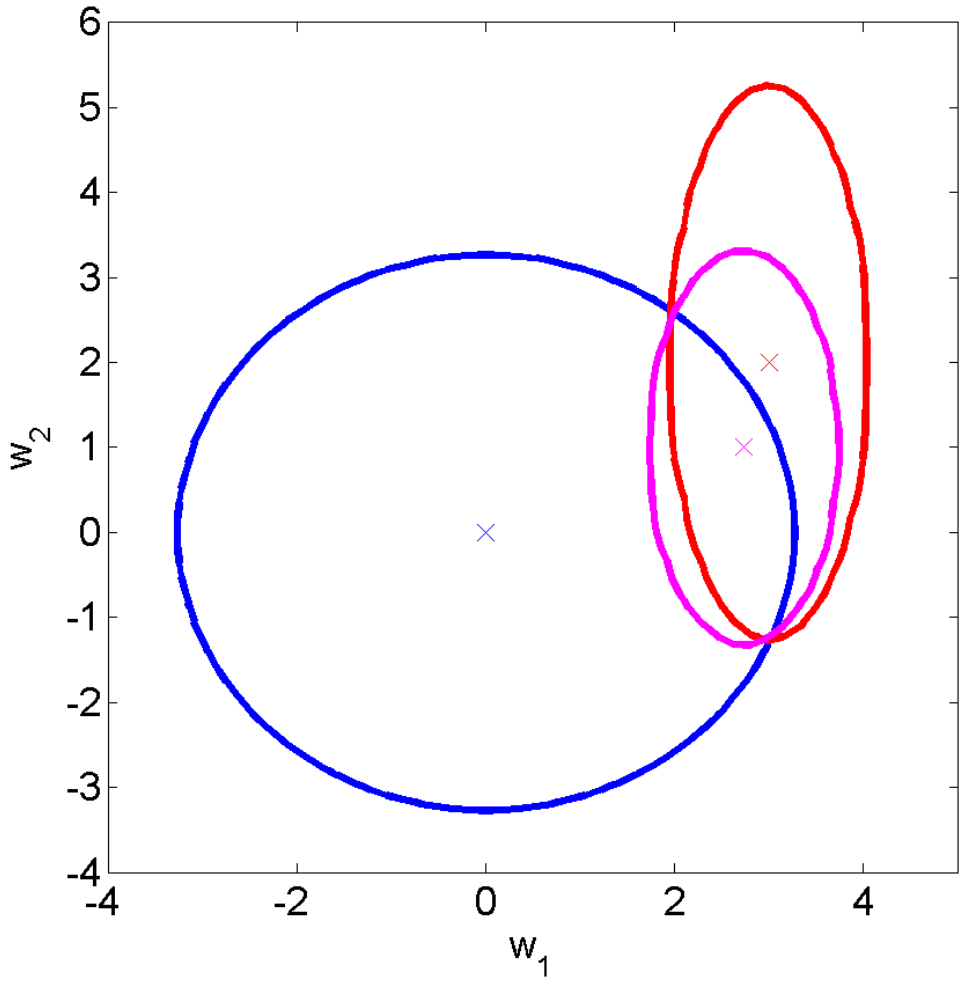
Bayesian GLM with two parameters

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



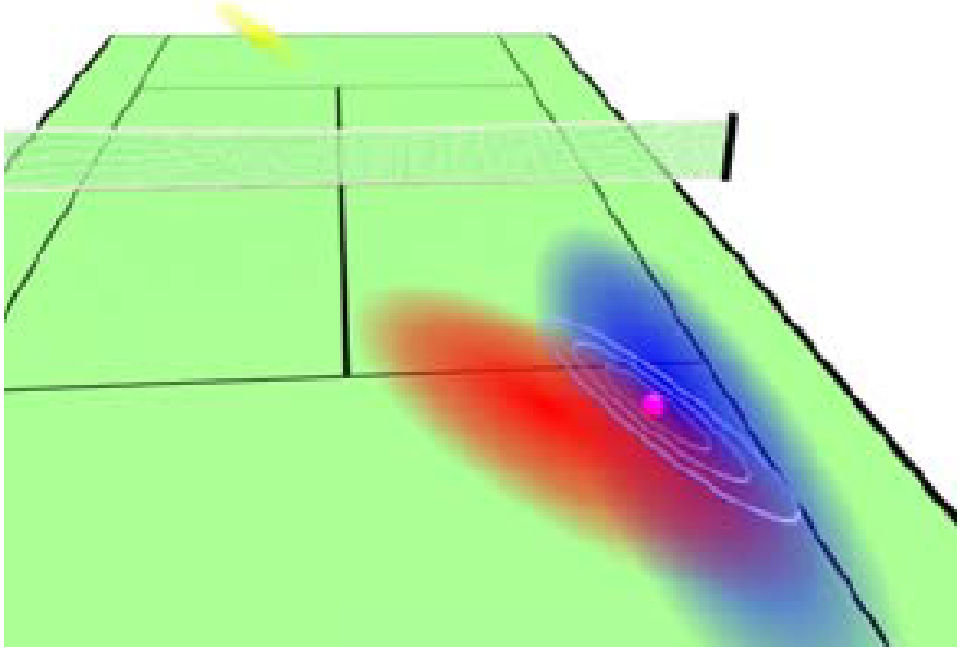
Shrinkage Prior

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



Tennis

From Wolpert and Ghahramani (2006)



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

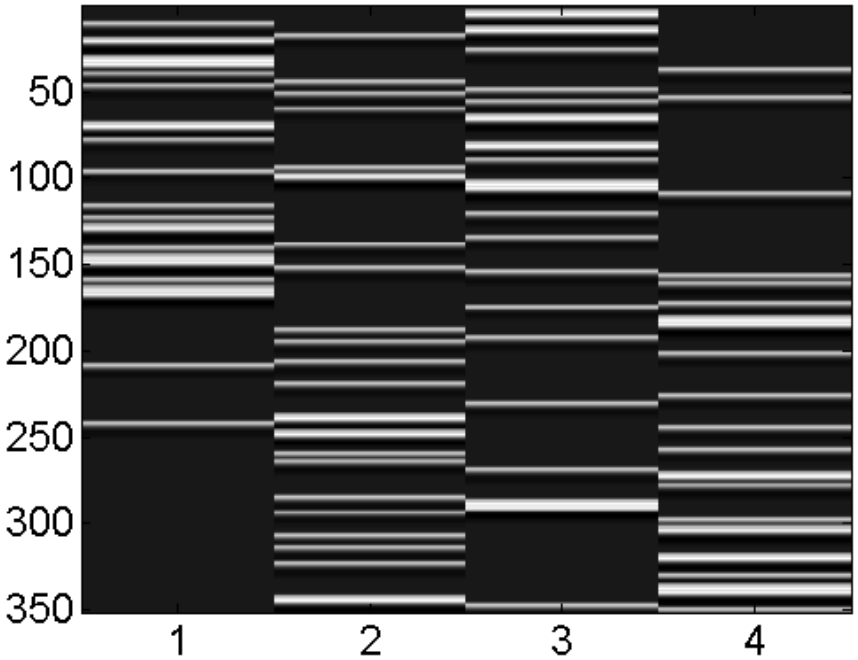
There is evidence that humans combine information in this optimal way (Doya et al 2006).

Time series models

Usually, we wish to test for multiple experimental effects. At the ith spatial position

$$y_i = Xw_i + e_i$$

where eg X is a $T \times K$ design matrix as below and w_i is a $K = 4$ element vector and y_i is a $T = 351$ element time series



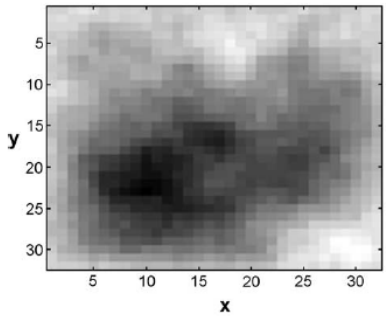
Time series model for images

We can write the model over all spatial positions $i = 1..V$ as

$$Y = XW + E$$

where Y is an $T \times V$ fMRI data matrix, X is a $V \times K$ design matrix, W is a $K \times V$ matrix of regression coefficients, and E is an $T \times V$ matrix of errors.

The ith column of Y is the time series at the ith voxel. The ith column of W, w_i, is the vector of regression coefficients at the ith voxel.



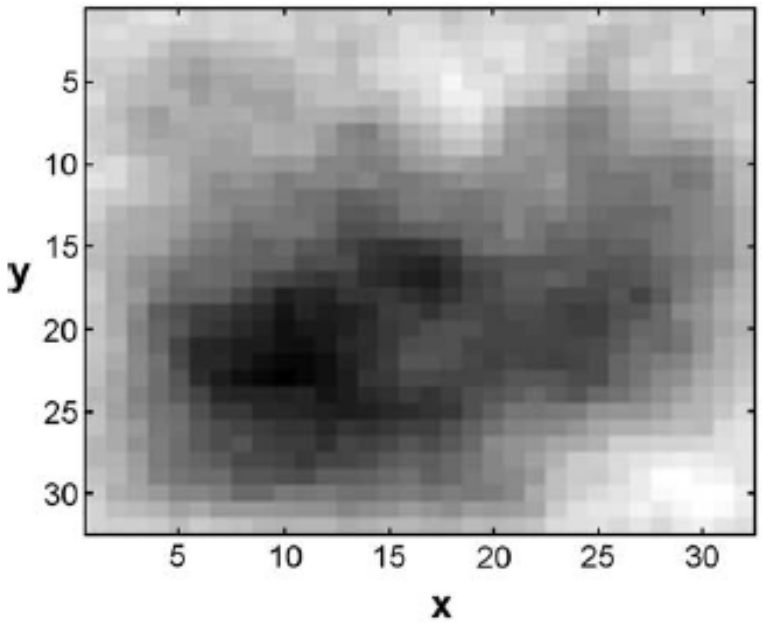
The kth *row* of W, w_k, is a V-element vector. It contains the image of regression coefficients for the kth effect we are testing for.

Spatial Prior

We can define a spatial prior over each of the $k = 1..K$ regression coefficient images

$$p(w_k|\alpha) = \mathrm{N}(w_k; 0, \alpha Q_w)$$

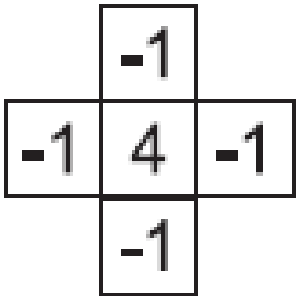
to capture our prior information that regression coefficients will be similar at nearby voxels, where the larger α produces smoother images.



The matrix Q_w can be set to reflect different assumptions abouth the types of smoothness.

Laplacian Prior

If we define a spatial kernel S with elements



Then $z_k = Sw_k$ will be a vector of local discrepancies between neighbouring voxels. The quantity

$$z_k^T z_k = w_k^T S^T S w_k$$

is then the sum of squared discrepancies. A shrinkage prior on z_k, that is one that encourages minimal discrepancy, is given by

$$p(w_k|\alpha_k) = \mathrm{N}(w_k; 0, \alpha_k Q_w)$$

where $Q_w^{-1} = S^T S$.

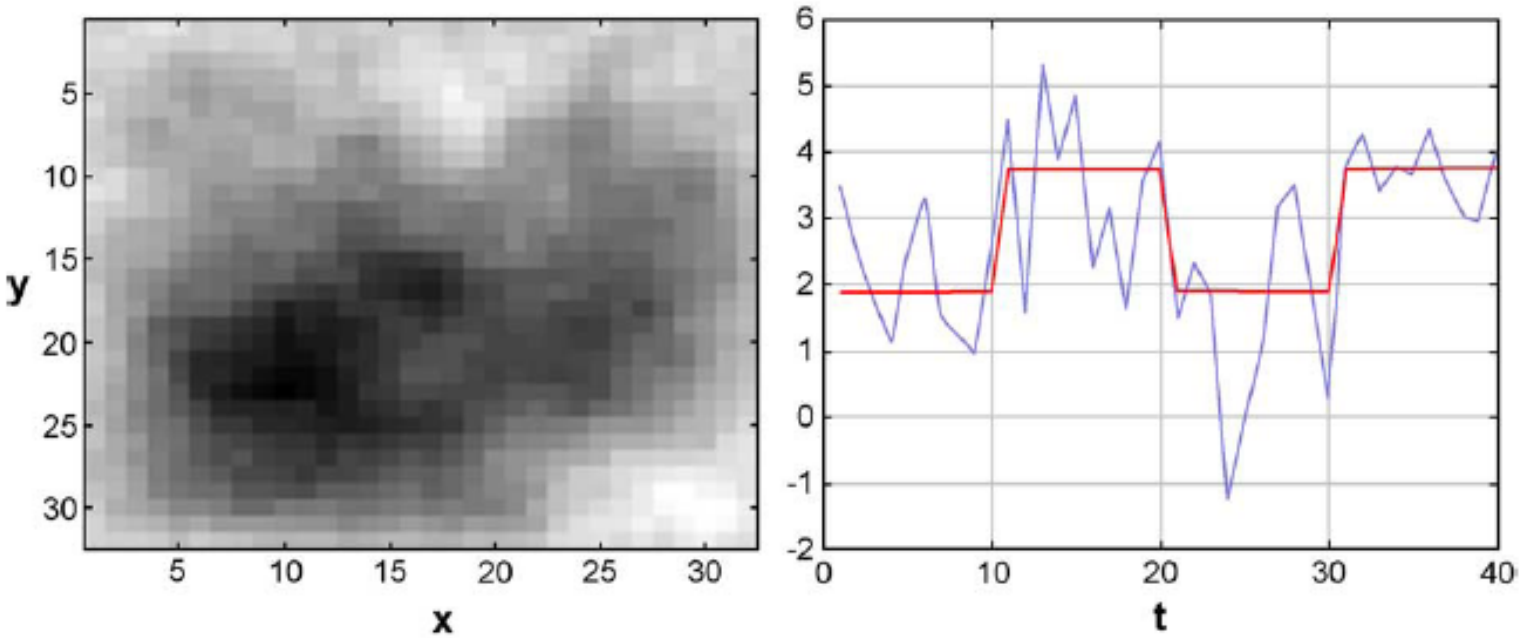
Spatio-Temporal Model

We define a spatio-temporal model

$$p(Y, W) = p(Y|W)p(W|\alpha)$$

where

$$p(W|\alpha) = \prod_{k=1}^{K} p(w_k|\alpha_k)$$



Different effects (eg first versus second presentation) can have different smoothnesses. For full generative model see Penny et al (2005).

Variational Bayes

Given fMRI data, the posterior distibution for the spatio-temporal model is Gaussian but the posterior covariance is of dimension $VK \times VK$ which is too large to handle.

Instead we use an approximate posterior that factorises over voxels

$$q(W|Y) = \prod_i q(w_i|Y)$$

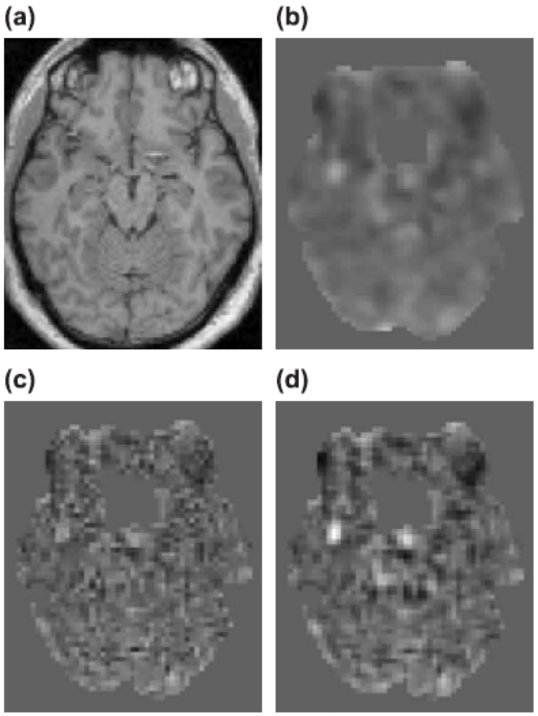
This approximate posterior can be fitted using the Variational Bayes (VB) framework (Bishop 2006, Penny et al. 2005).

The VB method can also be used to estimate the prior precisions α. This is a special case of Empirical Bayes (see MEG example later).

Results

- (a) Structural image (MRI)
- (b) ML applied to smooth data Y
- (c) Bayesian inference with Global shrinkage prior ($Q_w = I$)
- (d) Bayesian inference with Laplacian prior ($Q_w = S^T S$)

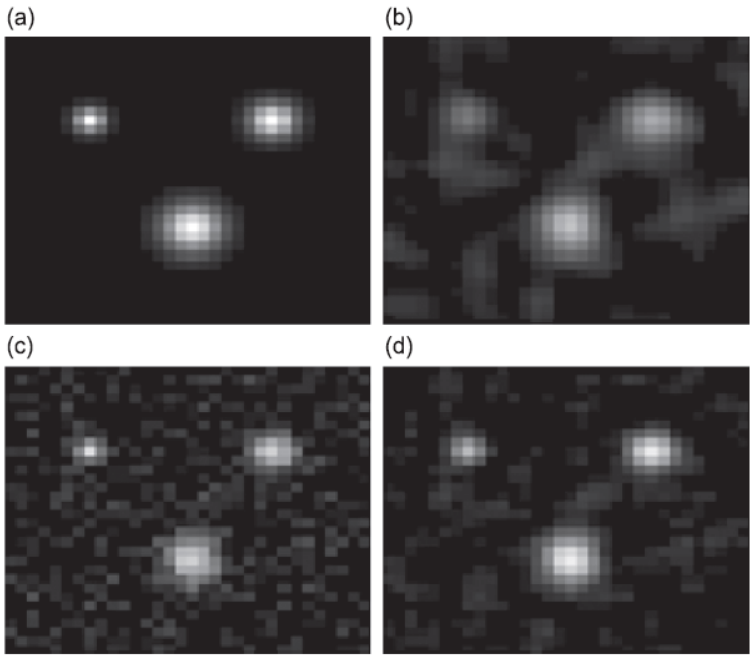
.



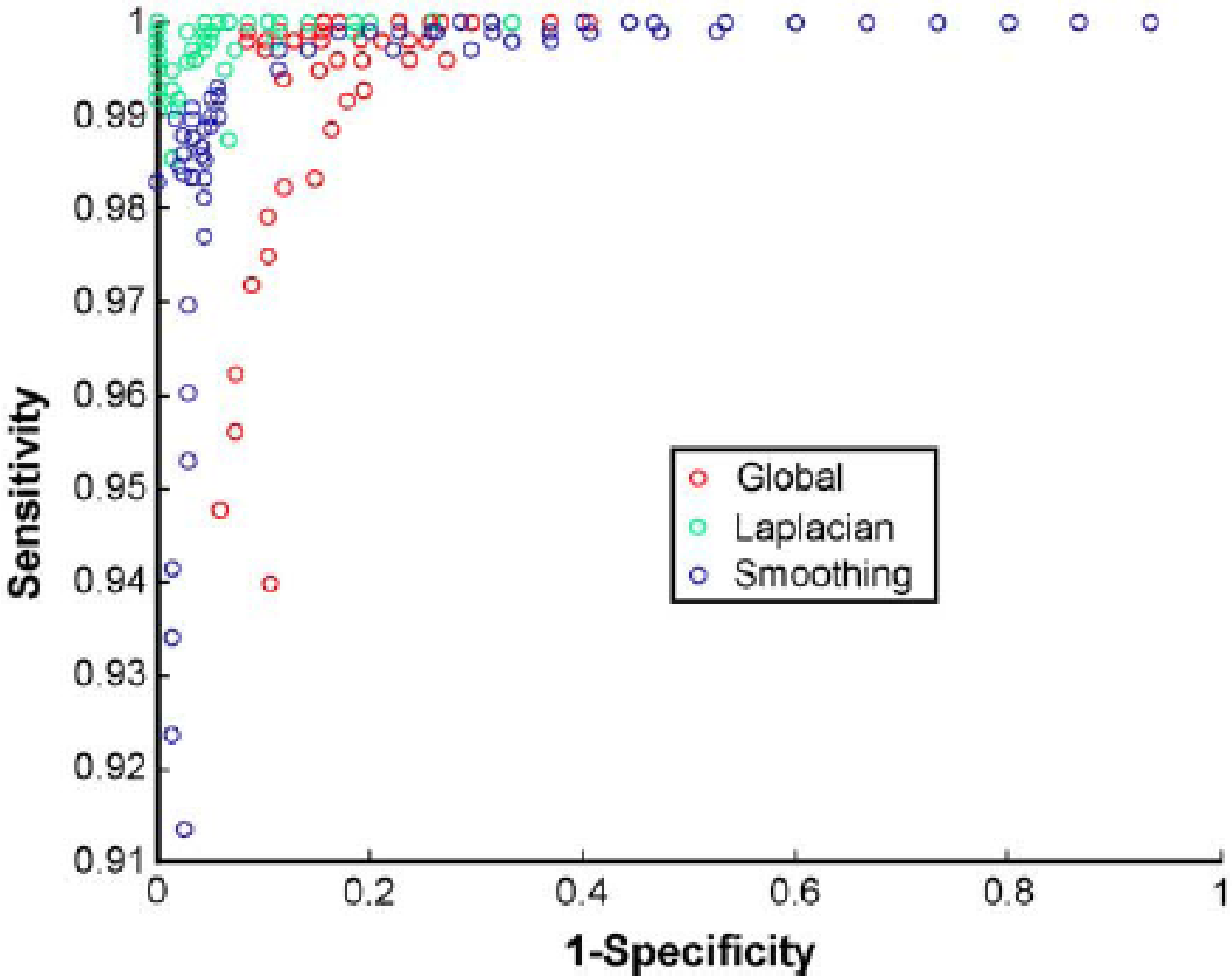
For (c) a Gaussian smoothing kernel of 8mm FWHM was used.

Synthetic data

- (a) True activations
- (b) ML applied to smooth data
- (c) Bayesian inference with Global shrinkage prior ($Q_w = I$)
- (d) Bayesian inference with Laplacian prior ($Q_w = S^T S$)



Receiver Operating Characteristic



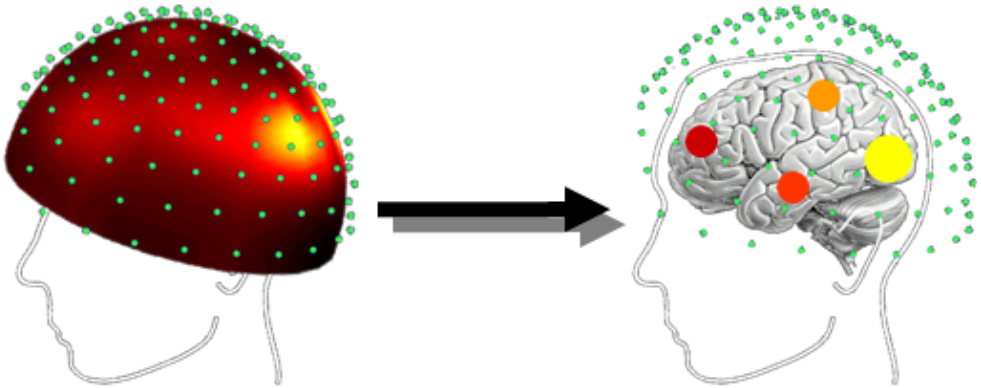
MEG Source Reconstruction

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

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

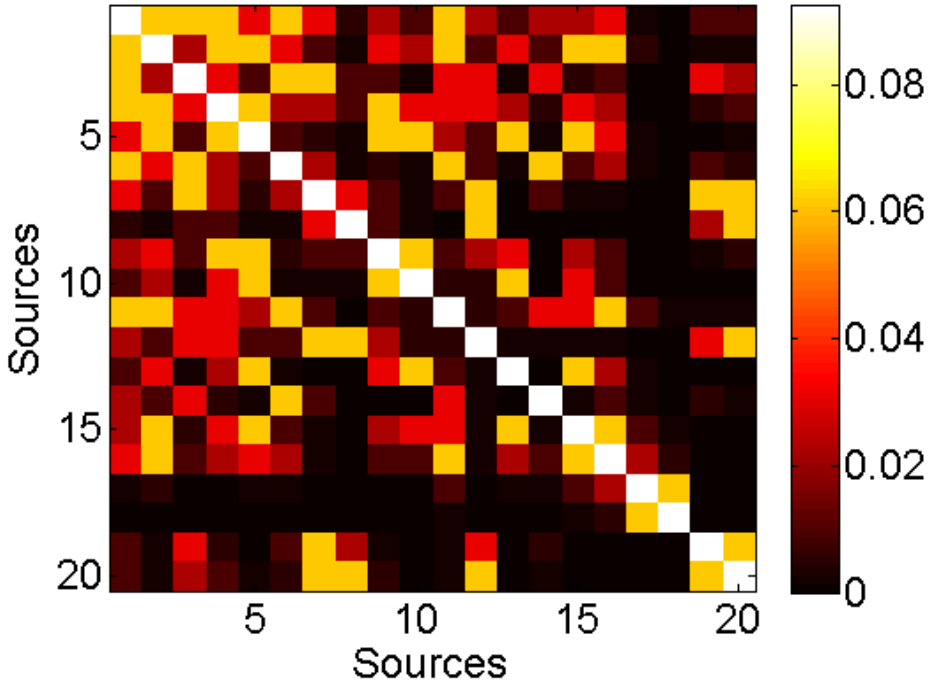
We let

$$\begin{aligned} C_y &= \lambda_y Q_1 \\ C_w &= \lambda_w Q_2 \end{aligned}$$

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

$$p(w|y) = \mathrm{N}(w; m_w, S_w)$$

where

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

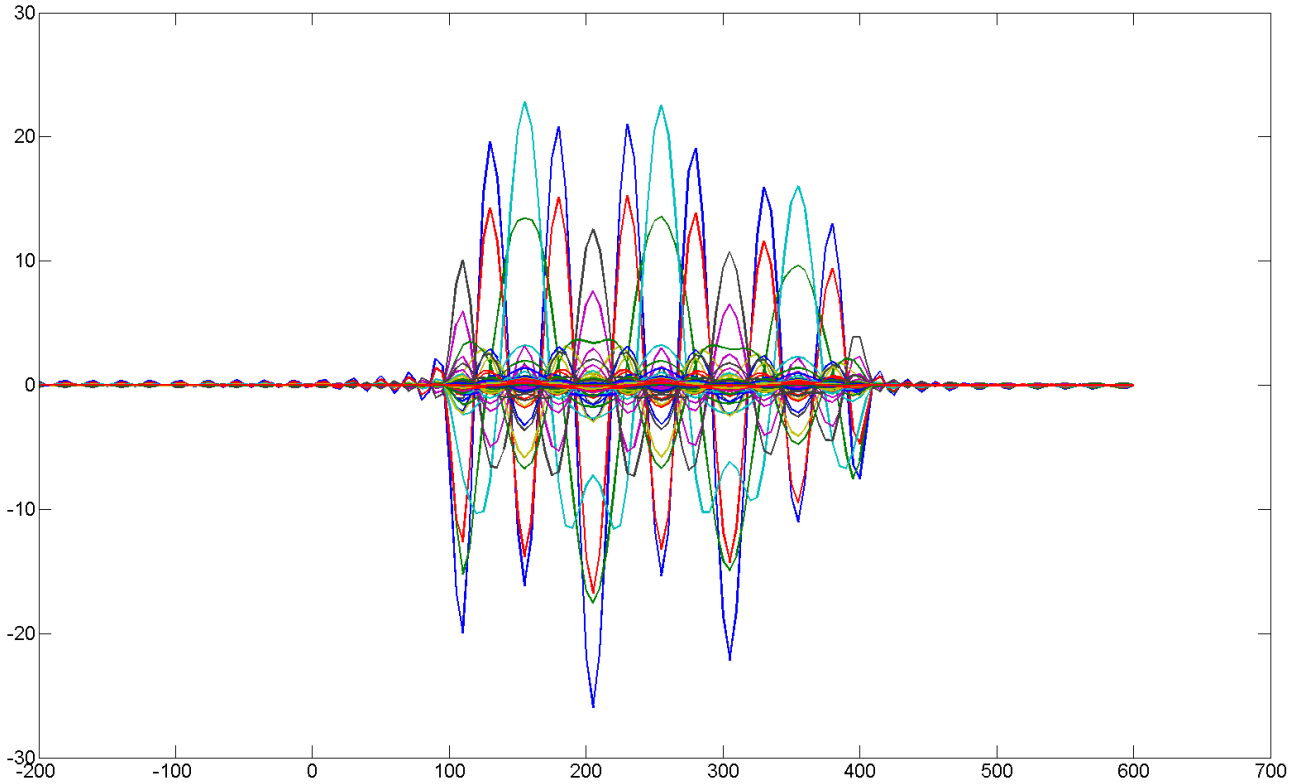
However, S_w is $p \times p$ with $p = 8192$ so cannot be inverted easily. But we can use the matrix inversion lemma, also known as the Woodbury identity (Bishop, 2006)

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

to ensure that only $d \times d$ matrices need inverting.

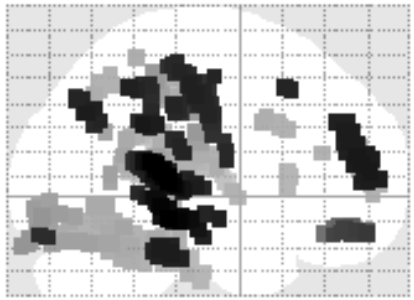
Simulation

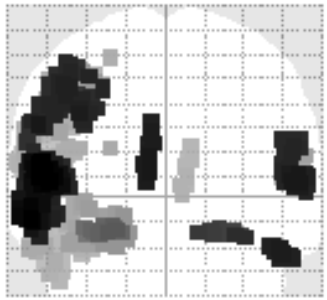
Two sinusoidal sources were placed in bilateral auditory cortex and produced this MEG data (Barnes, 2010), comprising $d = 274$ time series (butterfly plot)

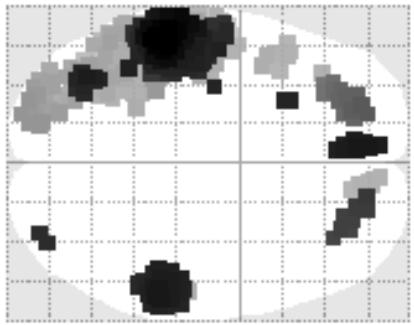


LORETA

We fix $\lambda_y = 1$. Here we set $\lambda_w = 0.01$.



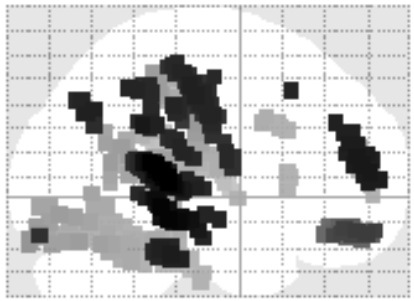


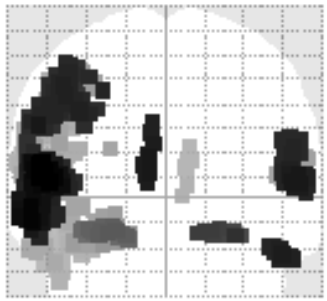


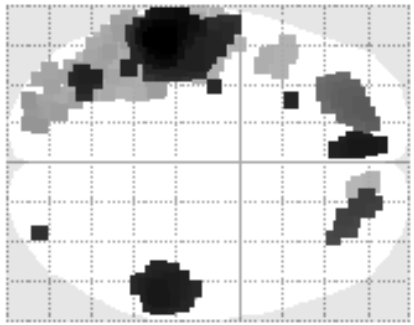
This shows the posterior mean activity for the 500 dipoles with the greatest power (over peristimulus time)

LORETA

We fix $\lambda_y = 1$. Here we set $\lambda_w = 0.1$.

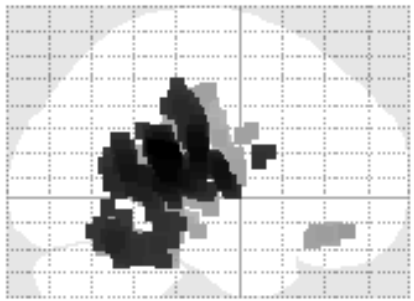


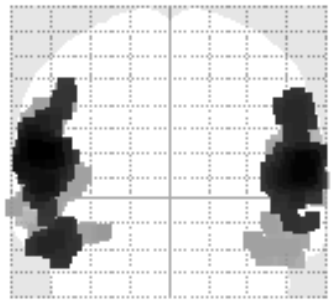


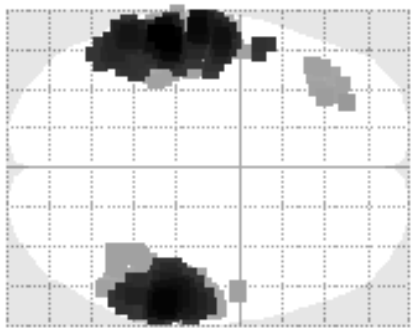


LORETA

We fix $\lambda_y = 1$. Here we set $\lambda_w = 1$.







Empirical Bayes

Hyperparameters, λ, can be estimated so as to maximise the model evidence. This forms the basis of Empirical Bayes.

The marginal likelihood or model evidence is given by

$$\begin{aligned} p(y|\lambda) &= \int p(y, w, \lambda) dw \\ &= \int p(y|w, \lambda) p(w|\lambda) dw \end{aligned}$$

The log model evidence is

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

For linear models this can be derived as in Bishop (2006) or as in my Maths for Brain Imaging notes.

In this formulation λ are not treated as random variables. There is no prior on them.

Empirical Bayes

We iterate between finding the parameters w and hyperparameters λ. For linear Gaussian models this corresponds to computing the posterior over w

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

and then setting λ to maximise the model evidence.

$$\hat{\lambda} = \arg\max_{\lambda} L(\lambda)$$

These two steps are then iterated and can be thought of as E and M steps in an EM optimisation algorithm.

Model Evidence

The model evidence is composed of sum squared precision weighted prediction errors and Occam factors

$$\begin{aligned} L(\lambda) &= -\frac{1}{2} e_y^T C_y^{-1} e_y - \frac{1}{2} \log |C_y| - \frac{d}{2} \log 2\pi \\ &- \frac{1}{2} e_w^T C_w^{-1} e_w - \frac{1}{2} \log \frac{|C_w|}{|S_w|} \end{aligned}$$

where λ is a vector of hyperparameters that parameterise the covariances $C_w = \lambda_w Q_w$ and $C_y = \lambda_y Q_y$. The prediction errors are the difference between what is expected and what is observed

$$\begin{aligned} e_y &= y - X m_w \\ e_w &= m_w - \mu_w \end{aligned}$$

Isotropic Covariances

For a Bayesian GLM

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

with isotropic covariances (eg minimum norm source reconstruction)

$$\begin{aligned} C_y &= \lambda_y I_N \\ C_w &= \lambda_w I_p \end{aligned}$$

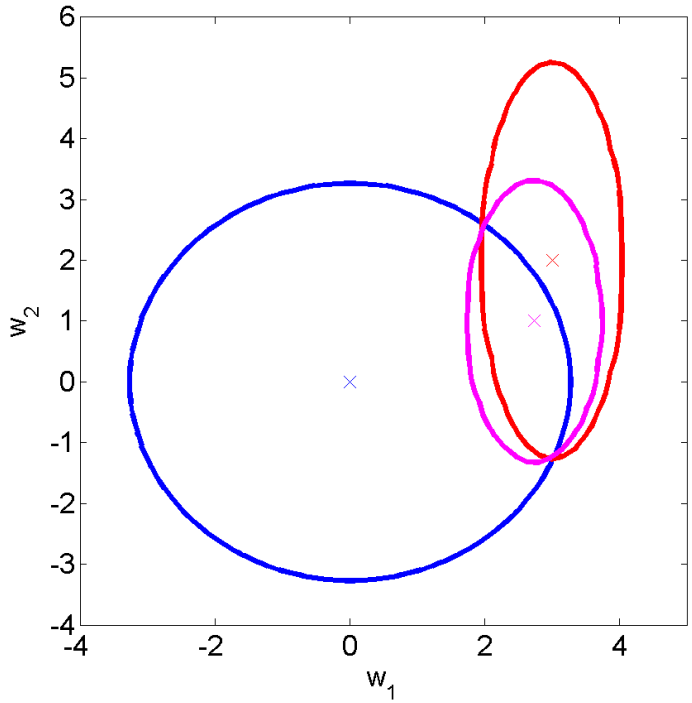
and d data points and p parameters. The equations for updating λ can be derived as shown in Chapter 10 of Bishop (2005).

Well-determined parameters

Define

$$\gamma = \sum_{j=1}^{p} \frac{\alpha_j}{\alpha_j + \hat{\lambda}_w}$$

where α_j are eigenvalues of the data precision term $X^T C_y^{-1} X$. If $\alpha_j >> \hat{\lambda}_w$ for all j then $\gamma = p$. Parameters have all been determined by the data. So γ is equivalent to number of well-determined parameters.



M-Step

Then

$$\frac{1}{\hat{\lambda}_w} = \frac{e_w^T e_w}{\gamma}$$
$$\frac{1}{\hat{\lambda}_y} = \frac{e_y^T e_y}{d - \gamma}$$

where the prediction errors are

$$e_y = y - X m_w$$
$$e_w = m_w - \mu_w$$

This effectively partitions the degrees of freedom in the data into those for estimating the prior and the likelihood.

Setting λ to maximise the *marginal* likelihood produces unbiased estimates of variances whereas ML estimation produces biased estimates.

Linear Covariances

For a Bayesian GLM

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

with covariances

$$\begin{aligned} C_y &= \sum_i \lambda_i Q_i \\ C_w &= \sum_{i'} \lambda_{i'} Q_{i'} \end{aligned}$$

where Q are known covariance basis functions. The M-step is

$$\hat{\lambda} = \arg\max_{\lambda} L(\lambda)$$

Gradient Ascent

This maximisation is effected by first computing the gradient and curvature of $L(\lambda)$ at the current parameter estimate, λ^{old}

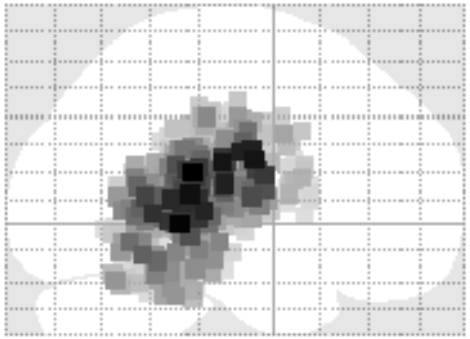
$$
\begin{aligned}
j_\lambda(i) &= \frac{dL(\lambda)}{d\lambda(i)} \\
H_\lambda(i,j) &= \frac{d^2L(\lambda)}{d\lambda(i)d\lambda(j)}
\end{aligned}
$$

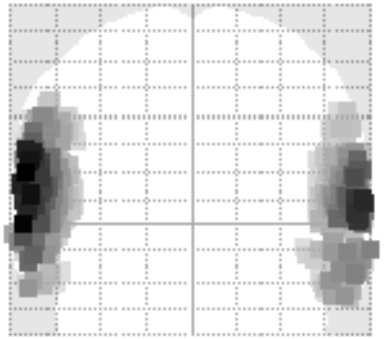
where i and j index the ith and jth parameters, j_λ is the gradient vector and H_λ is the curvature matrix. The new estimate is then given by

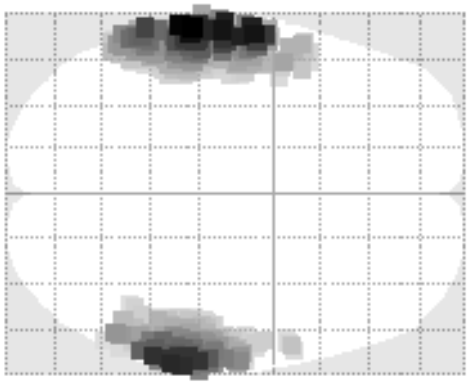
$$\lambda^{new} = \lambda^{old} - H_\lambda^{-1} j_\lambda$$

MEG Source Reconstruction

Hyperparameters set using Empirical Bayes.



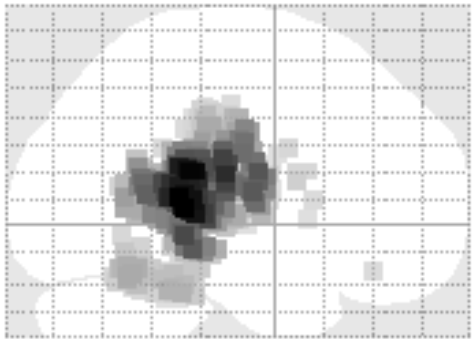


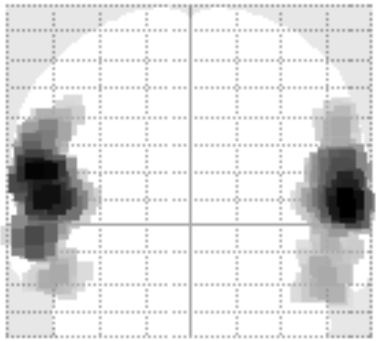


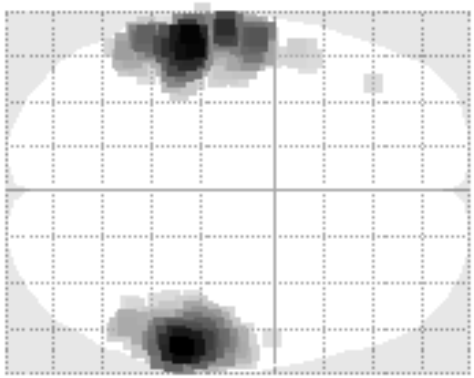
The *minimum norm* method, also implemented in SPM as the IID option.

Smoothness Priors

Hyperparameters set using Empirical Bayes.







This is similar to the LORETA method, implemented in SPM as the COH option.

References

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

C. Bishop (2006) Pattern Recognition and Machine Learning, Springer.

K. Doya et al (2006) Bayesian Brain. Probabilistic approaches to neural coding. MIT Press.

R. Henson (2003) Cerebral Cortex, 12: 178-186.

M. Johnson et al. (2001) BMJ 322, 1347-1349.

D. Mackay (2003) Information Theory, Inference and Learning Algorithms, Cambridge.

W. Penny et al. (2007) Bayesian fMRI time series analysis with spatial priors. Neuroimage 24, 350-362.

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