Maths for Brain Imaging: Lecture 9

W.D. Penny
Wellcome Department of Imaging Neuroscience,
University College, London WC1N 3BG.

December 6, 2006

1 Contents

- Laplace approximation
- Kullback-Liebler divergence
- Variational Bayes
- Application: Single subject fMRI with GLM-AR models
- Expectation Maximisation
- Mixture models
- Application: Identifying degenerate systems

2 Laplace approximation

Laplace's method approximates the integral of a function $\int f(\theta)d\theta$ by fitting a Gaussian at the maximum $\hat{\theta}$ of $f(\theta)$, and computing the volume of the Gaussian. The covariance of the Gaussian is determined by the Hessian matrix of $\log f(\theta)$ at the maximum point $\hat{\theta}$ [3].

The term 'Laplace approximation' is used for the method of approximating a posterior distribution with a Gaussian centered at the Maximum a Posterior (MAP) estimate. This is the application of Laplace's method with $f(\theta) = p(Y|\theta)p(\theta)$.

3 Kullback-Liebler divergence

For densities $q(\theta)$ and $p(\theta)$ the Relative Entropy or Kullback-Liebler (KL) divergence from q to p is [2]

$$KL[q||p] = \int q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta \qquad (1)$$

The KL-divergence satisfies the Gibb's inequality [4]

$$KL[q||p] \geq 0 \tag{2}$$

with equality only if $q = p$. In general $KL[q||p] \neq KL[p||q]$, so KL is not a distance measure.

4 Variational Bayes

Given a probabilistic model of some data, the log of the 'evidence' or 'marginal likelihood' can be written as

$$\begin{aligned} \log p(Y) &= \int q(\theta) \log p(Y) d\theta \\ &= \int q(\theta) \log \frac{p(Y,\theta)}{p(\theta|Y)} d\theta \\ &= \int q(\theta) \log \left[\frac{p(Y,\theta) q(\theta)}{q(\theta) p(\theta|Y)} \right] d\theta \\ &= F + KL(q(\theta)||p(\theta|Y)) \end{aligned} \tag{3}$$

where $q(\theta)$ is considered, for the moment, as an arbitrary density. We have

$$F = \int q(\theta) \log \frac{p(Y, \theta)}{q(\theta)} d\theta, \tag{4}$$

which in statistical physics is known as the *negative* variational free energy. The second term in equation 3 is the KL-divergence between the density $q(\theta)$ and the true posterior $p(\theta|Y)$. Equation 3 is the fundamental equation of the VB-framework and is shown graphically in Figure 1. Because KL is always positive, due to the Gibbs inequality, F provides a lower bound on the model evidence. Moreover, because KL is zero when two densities are the same, F will become equal to the model evidence when $q(\theta)$ is equal to the true posterior. For this reason $q(\theta)$ can be viewed as an *approximate posterior*.

4.1 Example

The solid lines in Figure 2 show a posterior distribution p which is a Gaussian mixture density comprising two

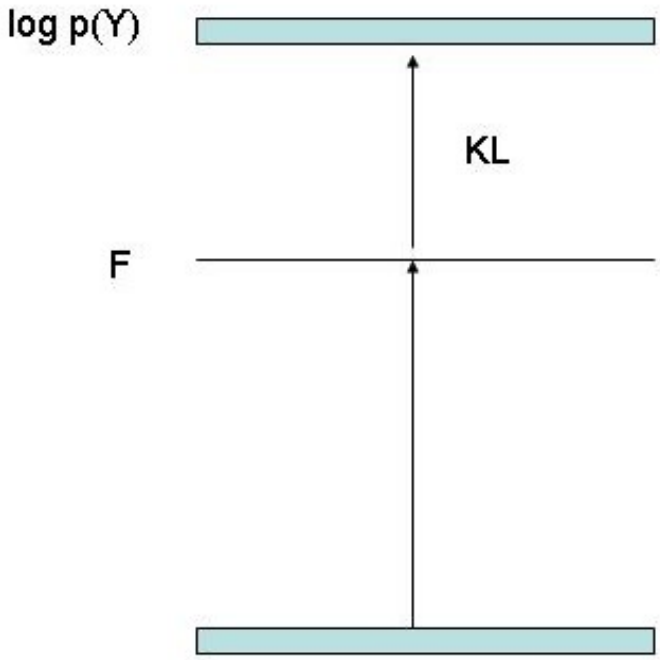


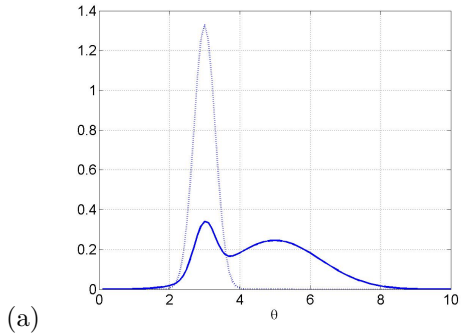
Figure 1: *The negative variational free energy, F, provides a lower bound on the log-evidence of the model with equality when the approximate posterior equals the true posterior.*

modes. The first contains the Maximimum A Posteriori (MAP) value and the second contains the majority of the probability mass.

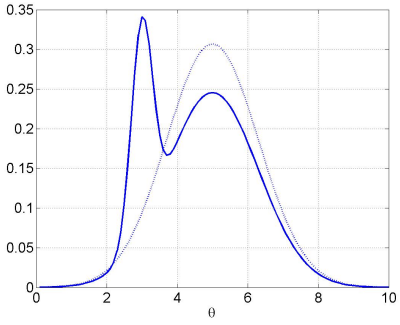
The Laplace approximation to p is therefore given by a Gaussian centred around the first, MAP mode. This is shown in Figure 2(a).

Figure 2(b) shows a Laplace approximation to the second mode, which could arise if MAP estimation found a local, rather than a global, maximum. Finally, Figure 2(c) shows the minimum KL-divergence approximation, assuming that q is a Gaussian. This is a fixed-form VB approximation, as we have fixed the form of the approximating density (ie. q is a Gaussian). This VB solution corresponds to a density q which is moment matched to p.

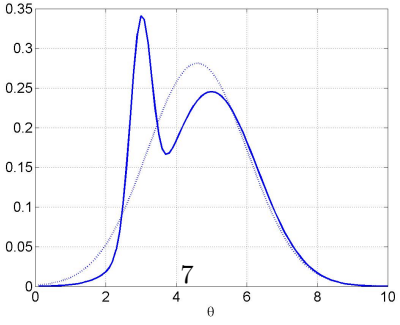
Figure 3 plots $KL[q||p]$ as a function of the mean and standard deviation of q, showing a minimum around the moment-matched values. These KL values were computed by discretising p and q and approximating equa-



(a)



(b)



(c)

Figure 2: *Probability densities $p(\theta)$ (solid lines) and $q(\theta)$ (dashed lines) for a Gaussian mixture $p(\theta) = 0.2 \times \mathsf{N}(m_1, \sigma_1^2) + 0.8 \times \mathsf{N}(m_2, \sigma_2^2)$ with $m_1 = 3, m_2 = 5$,$\sigma_1 = 0.3, \sigma_2 = 1.3$, and a single Gaussian $q(\theta) = \mathsf{N}(\mu, \sigma^2)$ with (a) $\mu = \mu_1, \sigma = \sigma_1$ which fits the first mode, (b) $\mu = \mu_2, \sigma = \sigma_2$ which fits the second mode and (c) $\mu = 4.6, \sigma = 1.4$ which is moment-matched to $p(\theta)$.*

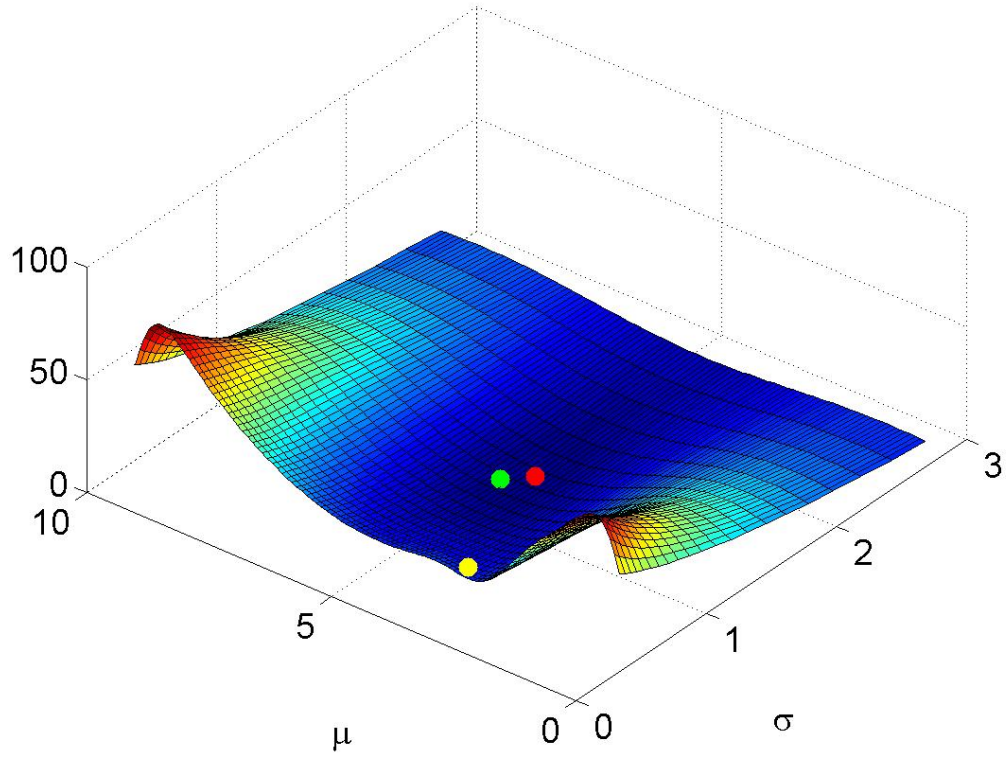


Figure 3: *KL-divergence, $KL(q||p)$ for p as defined in Figure 2 and q being a Gaussian with mean μ and standard deviation σ. The KL-divergences of the approximations in Figure 2 are (a) 11.73 for the first mode (yellow ball), (b) 0.93 for the second mode (green ball)*[8]*and (c) 0.71 for the moment-matched solution (red ball).*

tion 1 by a discrete sum. The MAP mode, maximum mass mode and moment-matched solutions have $KL[q||p]$ values of 11.7, 0.93 and 0.71 respectively. This shows that low KL is achieved when q captures most of the probability mass of p and, minimum KL when q is moment-matched to p. The figure also shows that, for reasonable values of the mean and standard deviation, there are no local minima. This is to be contrasted with the posterior distribution itself which has two maxima, one local and one global.

This example provides a good motivation for VB. But in higher dimensions due to (i) nature of KL and (ii) factorisations (see later) VB is not so optimal. See Minka [5] and Mackay [4] for further details.

4.2 Nonlinear functions of parameters

Capturing probability mass is particularly important if one is interested in nonlinear functions of parameter values, such as model predictions. Figures 4 and 5 show

histograms of model predictions for squared and logistic-map functions indicating that VB predictions are qualitatively better than those from the Laplace approximation.

Often in Bayesian inference, one quotes posterior exceedance probabilities. For the squared function, Laplace says 5% of samples are above $g = 12.2$. But in the true density, 71% of samples are. For the logisitic function 62% are above Laplace's 5% point. The percentage of samples above VB's 5% points are 5.1% for the squared function and 4.2% for the logistic-map function. So for this example, Laplace can tell you the posterior exceedance probability is 5% when, in reality it is an order of magnitude greater. This is not the case for VB.

4.3 Factorised Approximations

To obtain a practical learning algorithm we must also ensure that the integrals in F are tractable. One generic procedure for attaining this goal is to assume that the

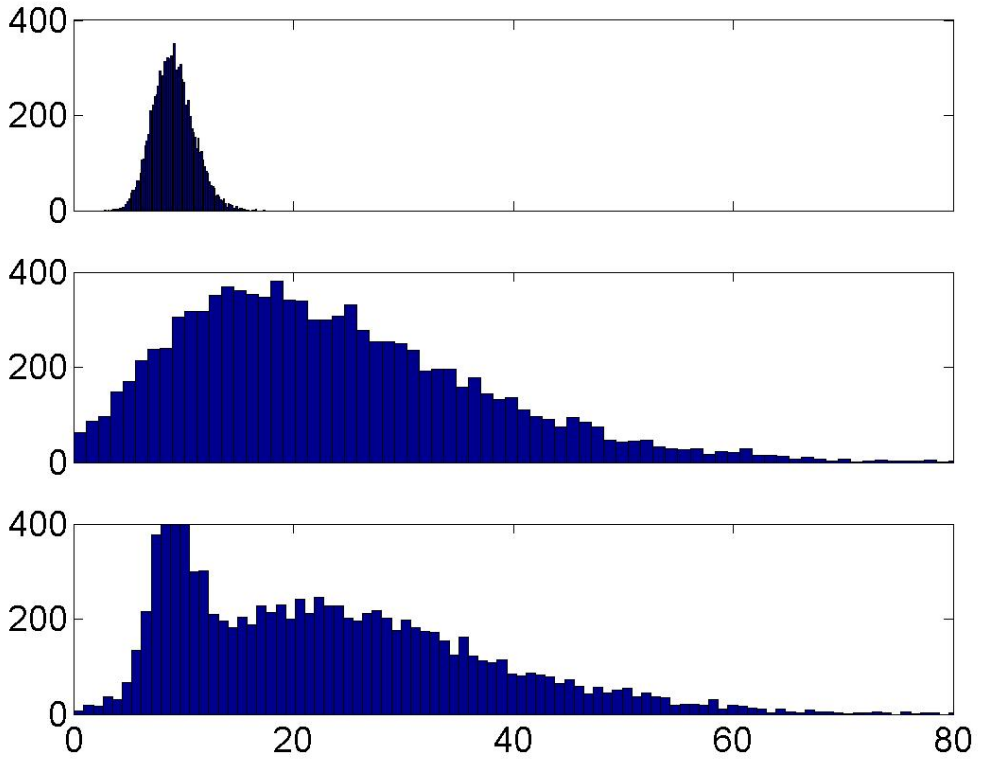


Figure 4: *Histograms of 10,000 samples drawn from $g(\theta)$ where the distribution over θ is from the Laplace approximation (top), VB approximation (middle) and true distribution, p, (bottom) for $g(\theta) = \theta^2$.*

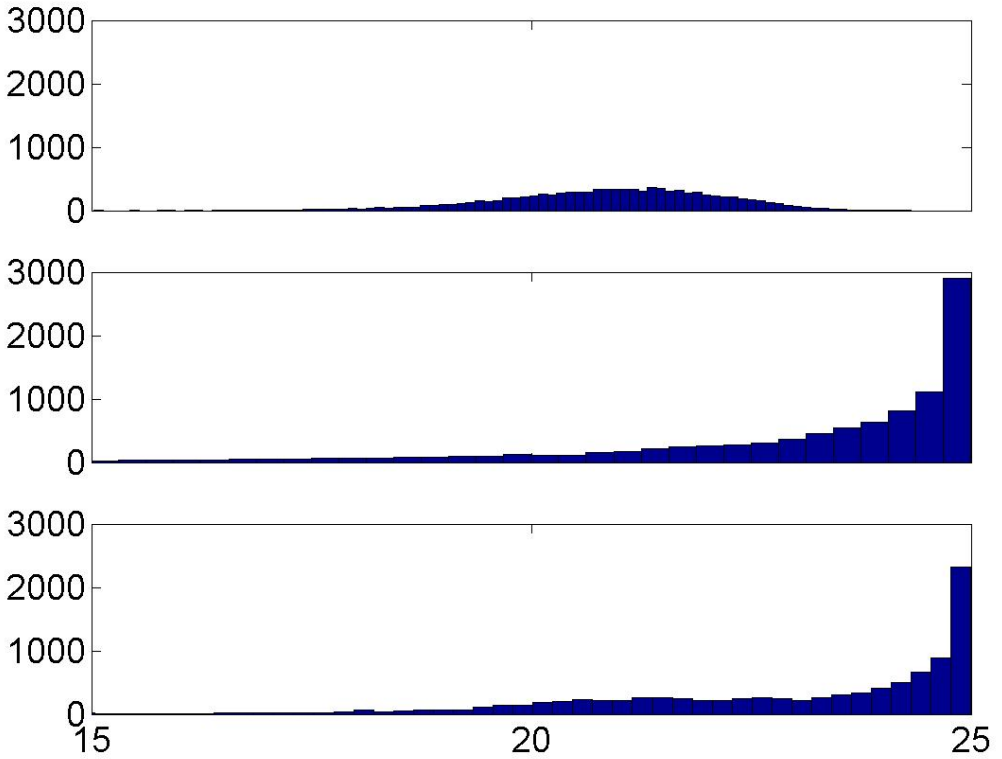


Figure 5: *Histograms of 10,000 samples drawn from $g(\theta)$ where the distribution over θ is from the Laplace approximation (top), VB approximation (middle) and true distribution, p, (bottom) for $g(\theta) = \theta * (10 - \theta)$. This is akin to a logistic map function encountered in dynamical systems [6].*

approximating density factorizes over groups of parameters. In physics, this is known as the mean field approximation. Thus, we consider:

$$q(\theta) = \prod_i q(\theta_i) \tag{5}$$

where θ_i is the ith group of parameters. We can also write this as

$$q(\theta) = q(\theta_i)q(\theta_{\setminus i}) \tag{6}$$

where $\theta_{\setminus i}$ denotes all parameters *not* in the ith group. The distributions $q(\theta_i)$ which maximise F can then be shown to be

$$q(\theta_i) = \frac{\exp[I(\theta_i)]}{Z} \tag{7}$$

where Z is the normalisation factor needed to make $q(\theta_i)$ a valid probability distribution and

$$I(\theta_i) = \int q(\theta_{\setminus i}) \log p(Y, \theta) d\theta_{\setminus i} \tag{8}$$

For proof see [8].

4.4 Model Inference

As we have seen earlier, the negative free energy, F, is a lower bound on the model evidence. If this bound is tight then F can be used as a surrogate for the model evidence and so allow for Bayesian model selection and averaging. Earlier, the negative free energy was written

$$F(m) = \int q(\theta|m) \log \frac{p(Y,\theta|m)}{q(\theta|m)} d\theta \tag{9}$$

By using $p(Y,\theta|m) = p(Y|\theta,m)p(\theta|m)$ we can express it as the sum of two terms

$$F(m) = \int q(\theta|m) \log p(Y|\theta,m) d\theta - KL[q(\theta|m)||p(\theta|m)] \tag{10}$$

where the first term is the average likelihood of the data and the second term is the KL between the approximating posterior and the *prior*.

4.5 KL for Gaussians

The KL divergence for Normal densities $q(x) = \mathsf{N}(\mu_q, \Sigma_q)$ and $p(x) = \mathsf{N}(\mu_p, \Sigma_p)$ is

$$\begin{aligned} KL_N(\mu_q, \Sigma_q; \mu_p, \Sigma_p) &= 0.5 \log \frac{|\Sigma_p|}{|\Sigma_q|} + 0.5 Tr(\Sigma_p^{-1}\Sigma_q) \\ &+ 0.5(\mu_q - \mu_p)^T \Sigma_p^{-1}(\mu_q - \mu_p) - \frac{d}{2} \end{aligned} \tag{11}$$

where $|\Sigma_p|$ denotes the determinant of the matrix Σ_p. The KL will tend to increase with the dimension of x.

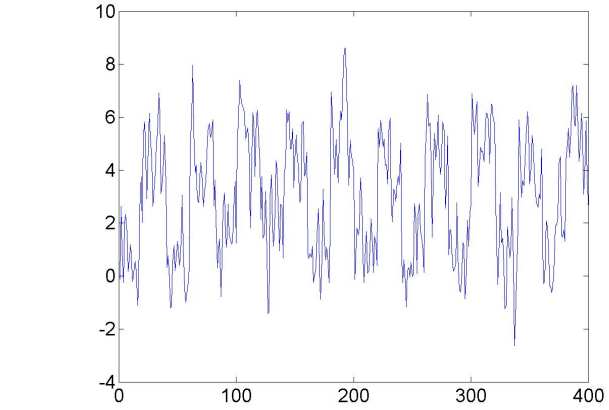
5 Single-subject fMRI: GLM-AR models

We generated data from a GLM-AR model having two regression coefficients and three autoregressive coefficients

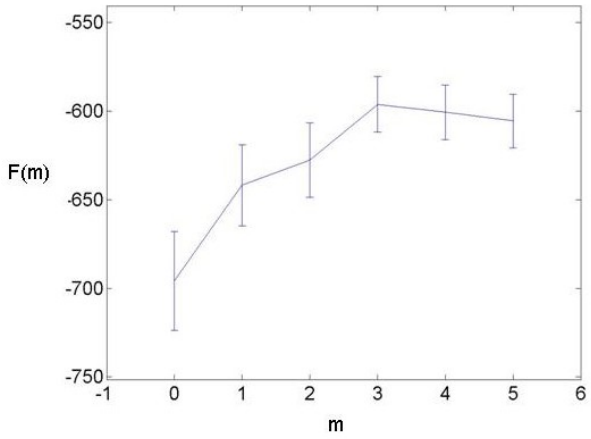
$$y_t = x_t w + e_t \tag{12}$$

$$e_t = \sum_{j=1}^{m} a_j e_{t-j} + z_t \tag{13}$$

where x_t is a two-element row vector, the first element flipping between a '-1' and '1' with a period of 40 scans (ie. 20 -1's followed by 20 1's) and the second element being '1' for all t. The two corresponding entries in w reflect the size of the activation, $w_1 = 2$, and the mean signal level, $w_2 = 3$. We used an AR(3) model for the errors with parameters $a_1 = 0.8$, $a_2 = -0.6$ and $a_3 = 0.4$. See [9] for further details.



(a)



(b)

Figure 6: The figures show (a) an example time series from a GLM-AR model with AR model order $m = 3$ and (b) a plot of the average negative free energy $F(m)$, with error bars, versus m. This shows that $F(m)$ picks out the correct model order.

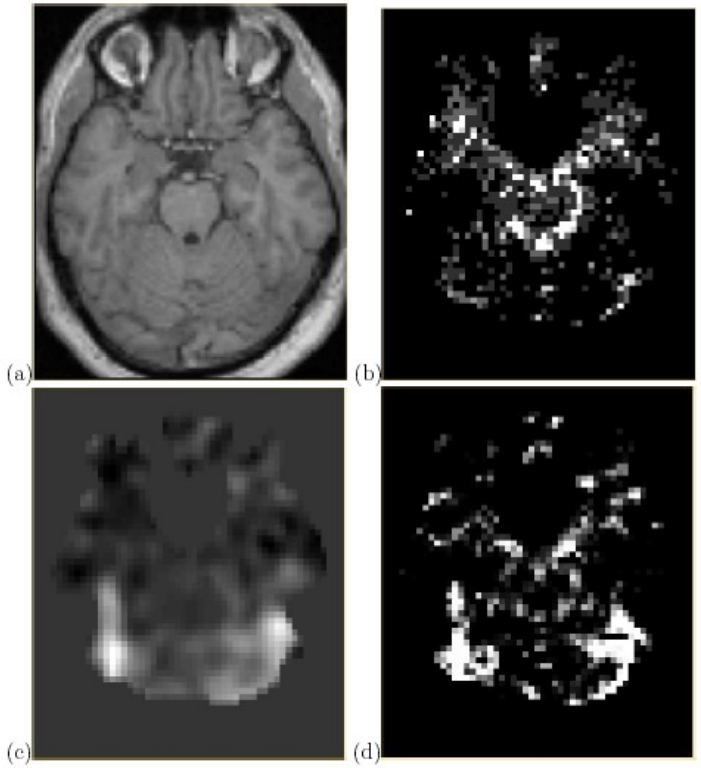


Figure 7: Face data: plot (b) shows argmax $F(m)$ as a function of voxel with $m = 0$ in black and $m = 3$ in white.

6 Mixture models

6.1 EM for mixture models

In this context EM is a maximum-likelihood algorithm for models with observed variables Y and hidden variables H. Hidden variable denotes which Gaussian is used to generate a data point. Select Gaussian k with probability k. That Gaussian has parameters μ_k and Σ_k.

Now, repeat 'VB derivation' but with eveything conditioned on parameters $\beta = \{\mu_k, \Sigma_k, \pi_k\}$ and replace θ with H. This gives

$$\log p(Y|\beta) = F_{EM} + KL[q(H)||p(H|Y,\beta)] \tag{14}$$

where

$$F_{EM} = \int q(H) \log \frac{p(H,Y|\beta)}{q(H)} dH \tag{15}$$

This gives rise to the following algorithm.

- E-Step: Set $q(H) = p(H|Y,\beta)$. This sets the KL

term to zero. This can be done by letting

$$q(h_n) = p(h_n|y_n, \beta) \quad (16)$$

$$= \frac{p(y_n|h_n, \beta)p(h_n|\beta)}{p(y_n|\beta)} \quad (17)$$

for all data points n. This is just Bayes rule. Write $\gamma_n^k = q(h_n = k)$, the responsibilies ie. the probability that data point n was generated from the kth Gaussian.

- M-step: Now, as $KL = 0$, $F_{EM} = \log p(Y|\beta)$, so we can maximise the likelihood wrt. β by maximising F_{EM} wrt. β. We have

$$F_{EM} = \sum_k \sum_n \gamma_k^n \log p(y_n|h_n = k)p(h_n = k) \quad (18)$$

$$= \sum_k \sum_n \gamma_k^n \log p(y_n|h_n = k) + \sum_k \sum_n \gamma_k^n p(h_n = k)$$

Setting the derivatives $dF_{EM}/d\beta$ to zero gives the

following updates

$$\begin{aligned} \mu_k &= \frac{\sum_n \gamma_n^k y_n}{\sum_n \gamma_n^k} \\ \Sigma_k &= \frac{\sum_n \gamma_n^k (y_n - \mu_k)(y_n - \mu_k)^T}{\sum_n \gamma_n^k} \\ \pi_k &= \frac{\sum_n \gamma_n^k}{N} \end{aligned} \tag{19}$$

See netlab demo `demgmm1.m`.

6.2 VB for mixture models

Allows for priors on model parameters eg. means of Gaussians. Provides approximation to model evidence based on the negative free energy. See Attias [1] and tech report `vbgmm.ps` for details.

6.3 Cross-modal priming fMRI

Mixture models have been applied to an analysis of intersubject variability in fMRI data. Model comparisons

based on VB identified two overlapping degenerate neuronal systems in subjects performing a crossmodal priming task [7].

SVD was applied to contrast images from 17 subjects and the first 5 spatial modes were used. A cluster analysis was then implemented in this 5-dimensional space.

Due to the problem of local maxima the cluster analysis was run 10,000 times. On 9,308 the evidence (as approximated using $F(m)$) for the 2-cluster model was higher. This was also the case if 2, 3 or 4 spatial modes were used.

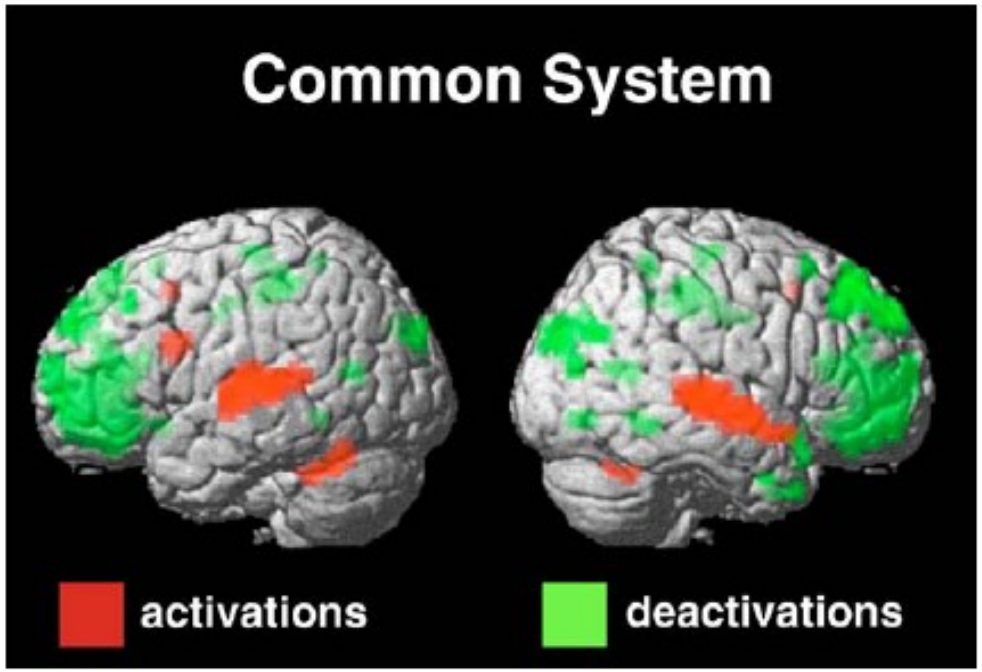


Fig. 1. Activations (red) and deactivations (green) for semantic decisions on crossmodal compound trials (for all normal subjects) relative to fixation are rendered on an averaged normalized brain. Height threshold: $P < 0.05$ corrected. Extent threshold > 3 voxels.

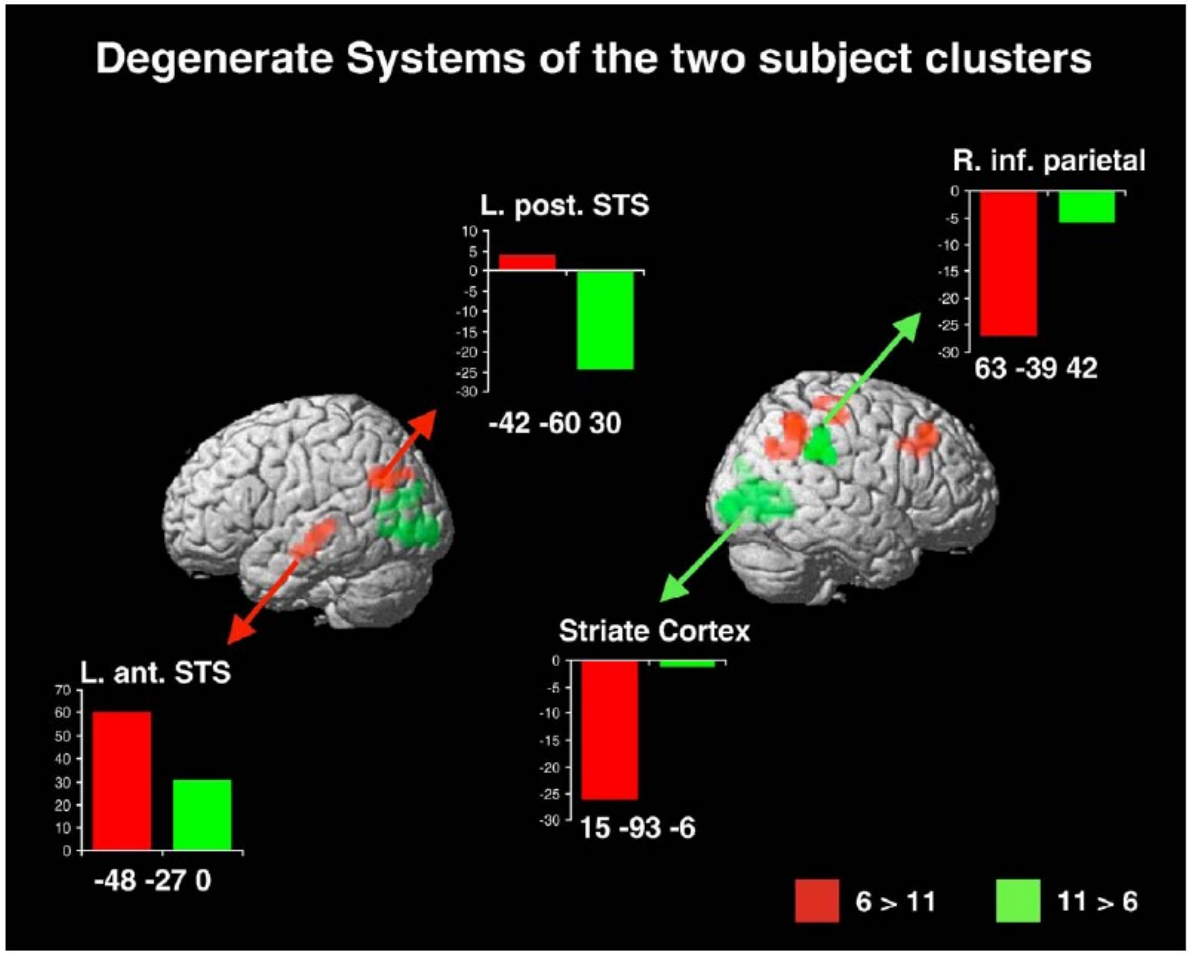


Fig. 2. Semantic decisions on crossmodal compound trials. Differential activation across groups is rendered on an averaged normalized brain. Height threshold: $P < 0.01$ uncorrected. Extent threshold > 50 voxels. Red = 6 > 11 subjects. Green = 11 > 6 subjects. Parameter estimates for 6 subject cluster (red) and 11 subject cluster (green) during semantic decisions on crossmodal stimuli. The bar graphs represent the size of the effect in adimensional units (corresponding to percent whole brain mean).

References

[1] H. Attias. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999.

[2] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[3] D. J. C. MacKay. Choice of basis for Laplace approximations. *Machine Learning*, 33:77–86, 1998.

[4] D.J.C Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, 2003.

[5] T. Minka. Divergence measures and message passing. Technical report, Microsoft Research Ltd, Cambridge, UK, 2005. MSR-TR-2005-173.

[6] T. Mullin. *The Nature of Chaos*. Oxford Science Publications, 1993.

[7] U. Noppeney, W. D. Penny, C. J. Price, G. Flandin, and K. J. Friston. Identification of degenerate neuronal systems based on intersubject variability. *Neuroimage*, 30:885–890, 2006.

[8] W. Penny, S. Kiebel, and K. Friston. Variational bayes. In K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny,

editors, *Statistical Parametric Mapping: The analysis of functional brain images*. Elsevier, London, 2006.

[9] W.D. Penny, S.J. Kiebel, and K.J. Friston. Variational Bayesian Inference for fMRI time series. *NeuroImage*, 19(3):727–741, 2003.