Chapter 11: Hierarchical Models

W. Penny and R. Henson

May 17, 2006

Introduction

Hierarchical models are central to many current analyses of functional imaging data including random effects analysis (Chapter 12), EEG source localization (Chapter 28 to 30) and spatiotemporal models of imaging data (Chapters 25 and 26 and [Friston et al. 2002b]). These hierarchical models posit linear relations among variables with error terms that are Gaussian. The General Linear Model (GLM), which to date has been so central to the analysis of functional imaging data, is a special case of these hierarchical models consisting of just a single layer.

Model fitting and statistical inference for hierarchical models can be implemented using a Parametric Empirical Bayes (PEB) algorithm described in Chapter 24 and in [Friston et al. 2002a]. The algorithm is sufficiently general to accomodate multiple hierarchical levels and allows for the error covariances to take on arbitrary form. This generality is particularly appealing as it renders the method applicable to a wide variety of modelling scenarios. Because of this generality, however, and the complexity of scenarios in which the method is applied, readers wishing to learn about PEB for the first time are advised to read this Chapter first. Chapter 24 then goes on to discuss the more general case. It also shows that the variance components that are estimated using PEB, can also be estimated using an algorithm from classical statistics called Restricted Maximum Likelihood (ReML).

In this Chapter we provide an introduction to hierarchical models and focus on some relatively simple examples. This Chapter covers the relevant mathematics and numerical examples are presented in the following Chapter. Each model and PEB algorithm we present is a special case of that described in [Friston et al. 2002a]. Whilst there are a number of tutorials on hierarchical modelling [Lee 1997, Carlin and Louis 2000] what we describe here has been tailored for functional imaging applications. We also note that a tutorial on hierarchical models is, to our minds, also a tutorial on Bayesian inference, as higher levels act as priors for parameters in lower levels. Readers are therefore encouraged to also consult background texts on Bayesian inference, such as [Gelman 1995].

This Chapter focusses on two-level models and shows how one computes the posterior distributions over the first- and second-level parameters. These are derived, initially, for completely general designs and error covariance matrices. We then consider two special cases; (i) models with equal error variances and (ii) separable models. We assume initially that the covariance components are known, and then in the section on PEB, we show how they can be estimated. A numerical example is then given showing PEB in action. The Chapter then describes how Bayesian inference can be implemented for hierarchical models with arbitrary probability distributions (eg. non-Gaussian), using the belief propagation algorithm. We close with a discussion.

In what follows, the notation $\mathsf{N}(m, \Sigma)$ denotes a uni/multivariate normal distribution with mean m and variance/covariance Σ and lower-case p's denote probability densities. Upper case letters denote matrices, lower case denote column vectors and x^T denotes the transpose of x. We will also make extensive use of the normal density ie. if $p(x) = \mathsf{N}(m, \Sigma)$ then

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

We also use $\mathsf{Var}[]$ to denote variance, $\otimes$ to denote the Kronecker product and X^+ to denote the pseudo-inverse.

Two-level models

We consider two-level linear Gaussian models of the form

$$\begin{aligned} y &= Xw + e \\ w &= M\mu + z \end{aligned} \tag{2}$$

where the errors are zero mean Gaussian with covariances $\mathsf{Cov}[e] = C$ and $\mathsf{Cov}[z] = P$. The model is shown graphically in Figure 1. The column vectors y and w have K and N entries respectively. The vectors w and μ are the first- and second-level parameters and X and M are the first- and second-level design matrices. Models of this form have been used in functional imaging. For example, in random effects analysis the second level models describe the variation of subject effect sizes about a population effect size, μ. In Bayesian inference with shrinkage priors, the second-level models variation of effect-size over voxels around a whole-brain mean effect size of $\mu = 0$ (ie. for a given cognitive challenge the response of a voxel chosen at random is, on average, zero). See, for example, [Friston et al. 2002b].

The aim of Bayesian inference is to make inferences about w and μ based on the posterior distributions $p(w|y)$ and $p(\mu|y)$. These can be derived as follows. We first note that the above equations specify the likelihood and prior probability distributions

$$\begin{aligned} p(y|w) &\propto \exp\left(-\tfrac{1}{2}(y-Xw)^T C^{-1}(y-Xw)\right) \\ p(w) &\propto \exp\left(-\tfrac{1}{2}(w-M\mu)^T P^{-1}(w-M\mu)\right) \end{aligned} \tag{3}$$

The posterior distribution is then

$$p(w|y) \propto p(y|w)p(w) \tag{4}$$

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

$$\log p(w|y) = -\frac{1}{2}(y-Xw)^T C^{-1}(y-Xw) \tag{5}$$

$$
\begin{aligned}
- & \frac{1}{2}(w - M\mu)^T P^{-1}(w - M\mu) + .. \\
= & -\frac{1}{2}w^T(X^T C^{-1} X + P^{-1})w + w^T(X^T C^{-1} y + P^{-1} M\mu) + ..
\end{aligned}
$$

Taking logs of the Gaussian density $p(x)$ in equation 1 and keeping only those terms that depend on x gives

$$\log p(x) = -\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}m + .. \tag{6}$$

Comparing equation 5 with terms in the above equation shows that

$$
\begin{aligned}
p(w|y) &= \mathsf{N}(m, \Sigma) \qquad (7) \\
\Sigma^{-1} &= X^T C^{-1} X + P^{-1} \\
m &= \Sigma(X^T C^{-1} y + P^{-1} M\mu)
\end{aligned}
$$

The posterior distribution over the second-level coefficient is given by Bayes' rule as

$$p(\mu|y) = \frac{p(y|\mu)p(\mu)}{p(y)} \tag{8}$$

However, because we do not have a prior $p(\mu)$ this posterior distribution becomes identical to the likelihood term, $p(y|\mu)$, which can be found by eliminating the first-level parameters from our two equations ie. by substituting the second level equation into the first giving

$$y = XM\mu + Xz + e \tag{9}$$

which can be written as

$$y = \tilde{X}\mu + \tilde{e} \tag{10}$$

where $\tilde{X} = XM$ and $\tilde{e} = Xz + e$. The solution to equation 10 then gives

$$
\begin{aligned}
p(\mu|y) &= \mathsf{N}(\hat{\mu}, \Sigma_\mu) \qquad (11) \\
\hat{\mu} &= (\tilde{X}^T\tilde{C}^{-1}\tilde{X})^{-1}\tilde{X}^T\tilde{C}^{-1}y \\
\Sigma_\mu &= (\tilde{X}^T\tilde{C}^{-1}\tilde{X})^{-1}
\end{aligned}
$$

where the covariance term

$$
\begin{aligned}
\tilde{C} &= \mathsf{Cov}[\tilde{e}] \qquad (12) \\
&= XPX^T + C
\end{aligned}
$$

We have now achieved our first goal, the posterior distributions of first- and second-level parameters being expressed in terms of the data, design and error-covariance matrices. We now consider the special cases of sensor fusion, equal variance models and separable models.

Sensor Fusion

The first special case is the univariate model

$$
\begin{aligned}
y &= w + e \qquad (13) \\
w &= \mu + z
\end{aligned}
$$

with a single scalar data point, y, and variances $C = 1/\beta$, $P = 1/\alpha$ specified in terms of the data precision β and the prior precision α (the 'precision' is the inverse variance). Plugging these values into equation 7 gives

$$\begin{aligned} p(w|y) &= \mathsf{N}(m, \lambda^{-1}) \\ \lambda &= \beta + \alpha \\ m &= \frac{\beta}{\lambda} y + \frac{\alpha}{\lambda} \mu \end{aligned} \tag{14}$$

Despite its simplicity this model possesses two important features of Bayesian learning in linear-Gaussian models. The first is that 'precisions add' - the posterior precision is the sum of the data precision and the prior precision. The second is that the posterior mean is the sum of the data mean and the prior mean, each weighted by their relative precisions. A numerical example is shown in Figure 2.

Equal variance

This special case is a two-level multivariate model as in equation 2 but with isotropic covariances at both the first and second levels. We have $C = \beta^{-1} I_K$ and $P = \alpha^{-1} I_N$. This means that observations are independent and have the same error variance. This is an example of the errors being Independent and Identically Distribution (IID), where in this case the distribution is a zero-mean Gaussian having a particular variance. In this Chapter we will also use the term 'sphericity' for any model with IID errors. Models without IID errors will have 'non-sphericity' (as an aside we note that IID is not actually a requirement of 'sphericity' and readers looking for a precise definition are referred to [Winer et al. 1991] and to Chapter 10).

On a further point of terminology, the unknown vectors w and μ will be referred to as 'parameters' whereas variables related to error covariances will be called 'hyperparameters'. The variables α and β are therefore hyperparameters. The posterior distribution over first level parameters is given by

$$\begin{aligned} p(w|y) &= \mathsf{N}(\hat{w}, \hat{\Sigma}) \\ \hat{\Sigma} &= (\beta X^T X + \alpha I_N)^{-1} \\ \hat{w} &= \hat{\Sigma} \left(\beta X^T y + \alpha M \mu \right) \end{aligned} \tag{15}$$

Note that if $\alpha = 0$ we recover the Maximum Likelihood estimate

$$\hat{w}_{ML} = (X^T X)^{-1} X^T y \tag{16}$$

This is the familiar Ordinary Least Squares (OLS) estimate used in the GLM [Holmes et al. 1997]. The posterior distribution over the second level parameters is given by equation 11 with

$$\tilde{C} = \beta^{-1} I_K + \alpha^{-1} X X^T \tag{17}$$

Separable model

We now consider 'separable models' which can be used, for example, for random effects analysis. Figure 3 shows the corresponding generative model. In these models, the first-level splits into N separate sub-models. For each sub-model, i, there are n_i observations. These form the n_i-element vector y_i giving information about the parameter w_i via the design vector x_i. For fMRI analysis these design vectors comprise stimulus functions eg. boxcars or delta functions, convolved with an assumed hemodynamic response. The overall first-level design matrix X then has a block-diagonal form $X = \mathsf{blkdiag}(x_1, .., x_i, .., x_N)$ and the covariance is given by $C = \mathsf{diag}[\beta_1 1_{n_1}^T, .., \beta_i 1_{n_i}^T, .., \beta_N 1_{n_N}^T]$ where 1_n is a column vector of 1's with n entries. For example, for $N = 3$ groups with $n_1 = 2$, $n_2 = 3$ and $n_3 = 2$ observations in each group

$$X = \begin{bmatrix} x_1(1) & 0 & 0 \\ x_1(2) & 0 & 0 \\ 0 & x_2(1) & 0 \\ 0 & x_2(2) & 0 \\ 0 & x_2(3) & 0 \\ 0 & 0 & x_3(1) \\ 0 & 0 & x_3(2) \end{bmatrix} \tag{18}$$

and $C^{-1} = \mathsf{diag}[\beta_1, \beta_1, \beta_2, \beta_2, \beta_2, \beta_3, \beta_3]$. The covariance at the second level is $P = \alpha^{-1} I_N$, as before, and we also assume that the second level design matrix is a column of 1's, $M = 1_N$. The posterior distribution over first level parameters is found by substituting X and C into equation 7. This gives a distribution which factorises over the different first level coefficients such that

$$\begin{aligned} p(w|y) &= \prod_{i=1}^{N} p(w_i|y) \qquad (19) \\ p(w_i|y) &= N(\hat{w}_i, \hat{\Sigma}_{ii}) \\ \hat{\Sigma}_{ii}^{-1} &= \beta_i x_i^T x_i + \alpha \\ \hat{w}_i &= \hat{\Sigma}_{ii} \beta_i x_i^T y_i + \hat{\Sigma}_{ii} \alpha \mu \end{aligned}$$

The posterior distribution over second level parameters is, from equation 11, given by

$$\begin{aligned} p(\mu|y) &= \mathsf{N}(\hat{\mu}, \sigma_\mu^2) \qquad (20) \\ \sigma_\mu^2 &= \frac{1}{\sum_{i=1}^{N} x_i^T (\alpha^{-1} x_i x_i^T + \beta_i^{-1})^{-1} x_i} \\ \hat{\mu} &= \sigma_\mu^2 \sum_{i=1}^{N} x_i^T (\alpha^{-1} x_i x_i^T + \beta_i^{-1})^{-1} y_i \end{aligned}$$

We note that in the absence of any second level variability, ie. $\alpha \to \infty$, the estimate $\hat{\mu}$ reduces to the mean of the first level coefficients weighted by their precision

$$\hat{\mu} = \frac{\sum_i \beta_i x_i^T y_i}{\sum_i \beta_i x_i^T x_i} \tag{21}$$

Parametric Empirical Bayes

In the previous section we have shown how to compute the posterior distributions $p(w|y)$ and $p(\mu|y)$. As can be seen from equations 7 and 11, however, these equations depend on covariances P and C. In this section we show how covariance components can be estimated for the special cases of equal variance models and separable models.

In [Friston et al. 2002a] the covariances are decomposed using using

$$\begin{aligned} C &= \sum_j \lambda_j^1 Q_j^1 \\ P &= \sum_j \lambda_j^2 Q_j^2 \end{aligned} \tag{22}$$

where Q_j^1 and Q_j^2 are basis functions that are specified by the modeller depending on the application in mind. For example, for analysis of fMRI data from a single subject two basis functions are used, the first relating to error variance and the second relating to temporal autocorrelation [Friston et al. 2002b]. The hyperparameters $\lambda = [\{\lambda_j^1\}, \{\lambda_j^2\}]$ are unknown but can be estimated using the PEB algorithm described in [Friston et al. 2002a]. Variants of this algorithm are known as the *evidence framework* [Mackay 1992] or *Maximum Likelihood II (ML-II)* [Berger 1985]. The PEB algorithm is also referred to as simply *Empirical Bayes* but we use the term PEB to differentiate it from the Nonparametric Empirical Bayes methods described in [Carlin and Louis 2000]. The hyperparameters are set so as to maximise the evidence (also known as the marginal likelihood)

$$p(y|\lambda) = \int p(y|w,\lambda) p(w|\lambda) dw \tag{23}$$

This is the likelihood of the data after we have integrated out the first-level parameters. For the two multivariate special cases described above, by substituting in our expressions for the prior and likelihood, integrating, taking logs and then setting the derivatives to zero, we can derive a set of update rules for the hyperparameters. These derivations are provided in the following two sections.

Equal variance

For the equal variance model the objective function is

$$p(y|\alpha,\beta) = \int p(y|w,\beta) p(w|\alpha) dw \tag{24}$$

Substituting in expressions for the likelihood and prior gives

$$p(y|\alpha,\beta) \;=\; \left(\frac{\beta}{2\pi}\right)^{K/2} \left(\frac{\alpha}{2\pi}\right)^{N/2} \int \exp\left(-\frac{\beta}{2} e(w)^T e(w) - \frac{\alpha}{2} z(w)^T z(w)\right) dw$$

where $e(w) = y - Xw$ and $z(w) = w - M\mu$. By re-arranging the terms in the exponent (and keeping all of them, unlike before) where we were only interested in w-dependent terms) the integral can be written as

$$\begin{aligned} I \;=\; & \left[\int \exp\left(-\frac{1}{2}(w-\hat{w})^T \hat{\Sigma}^{-1} (w-\hat{w})\right) dw\right] \\ & \cdot \left[\exp\left(-\frac{\beta}{2} e(\hat{w})^T e(\hat{w}) - \frac{\alpha}{2} z(\hat{w})^T z(\hat{w})\right)\right] \end{aligned} \tag{25}$$

where the second term is not dependent on w. The first factor is then simply given by the normalising constant of the multivariate Gaussian density

$$(2\pi)^{N/2}|\hat{\Sigma}|^{1/2} \tag{26}$$

Hence,

$$p(y|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{K/2} \alpha^{N/2}|\hat{\Sigma}|^{1/2} \exp\left(-\frac{\beta}{2}e(\hat{w})^T e(\hat{w}) - \frac{\alpha}{2}z(\hat{w})^T z(\hat{w})\right)$$

where $|\hat{\Sigma}|$ denotes the determinant of $\hat{\Sigma}$. Taking logs gives the 'log-evidence'

$$F = \frac{K}{2}\log\frac{\beta}{2\pi} + \frac{N}{2}\log\alpha + \frac{1}{2}\log|\hat{\Sigma}| - \frac{\beta}{2}e(\hat{w})^T e(\hat{w}) - \frac{\alpha}{2}z(\hat{w})^T z(\hat{w}) \tag{27}$$

To find equations for updating the hyperparameters we must differentiate F with respect to α and β and set the derivative to zero. The only possibly problematic term is the log-determinant but this can be differentiated by first noting that the inverse covariance is given by

$$\hat{\Sigma}^{-1} = \beta X^T X + \alpha I_N \tag{28}$$

If λ_j are the eigenvalues of the first term then the eigenvalues of $\hat{\Sigma}^{-1}$ are $\lambda_j + \alpha$. Hence,

$$\begin{aligned}
|\hat{\Sigma}^{-1}| &= \prod_j (\lambda_j + \alpha) \qquad (29)\\
|\hat{\Sigma}| &= \frac{1}{\prod_j (\lambda_j + \alpha)}\\
\log|\hat{\Sigma}| &= -\sum_j \log(\lambda_j + \alpha)\\
\frac{\partial}{\partial\alpha}\log|\hat{\Sigma}| &= -\sum_j \frac{1}{\lambda_j + \alpha}
\end{aligned}$$

Setting the derivative $\partial F/\partial\alpha$ to zero then gives

$$\begin{aligned}
\alpha z(\hat{w})^T z(\hat{w}) &= N - \sum_j \frac{\alpha}{\lambda_j + \alpha} \qquad (30)\\
&= \sum_j \frac{\lambda_j + \alpha}{\lambda_j + \alpha} - \sum_j \frac{\alpha}{\lambda_j + \alpha}\\
&= \sum_j \frac{\lambda_j}{\lambda_j + \alpha}
\end{aligned}$$

This is an implicit equation in α which leads to the following update rule. We first define the quantity γ which is computed from the 'old' value of α

$$\gamma = \sum_{j=1}^{N} \frac{\lambda_j}{\lambda_j + \alpha} \tag{31}$$

and then let

$$\frac{1}{\alpha} = \frac{z(\hat{w})^T z(\hat{w})}{\gamma} \tag{32}$$

The update for β is derived by first noting that the eigenvalues λ_j are linearly dependent on β. Hence

$$\frac{\partial \lambda_j}{\partial \beta} = \frac{\lambda_j}{\beta} \tag{33}$$

The derivative of the log-determinant is then given by

$$\frac{\partial}{\partial \beta} \log |\hat{\Sigma}^{-1}| \quad = \quad \frac{1}{\beta} \sum_j \frac{\lambda_j}{\lambda_j + \alpha} \tag{34}$$

which leads to the update

$$\frac{1}{\beta} = \frac{e(\hat{w})^T e(\hat{w})}{K - \gamma} \tag{35}$$

The PEB algorithm consists of iterating the update rules in equations 31, 32, 35 and the posterior estimates in equation 15, until convergence.

The update rules in equations 31, 32 and 35 can be interpreted as follows. For every j for which $\lambda_j >> \alpha$, the quantity γ increases by 1. As α is the prior precision and λ_j is the data precision (of the jth 'eigencoefficient') γ therefore measures the number of parameters that are determined by the data. Given K data points, the quantity $K - \gamma$ therefore corresponds to the number of degrees of freedom in the data set. The variances α^{-1} and β^{-1} are then updated based on the sum of squares divided by the appropriate degrees of freedom.

Separable models

For separable models the objective function is

$$p(y|\alpha, \{\beta_i\}) = \int p(y|w, \{\beta_i\}) p(w|\alpha) dw \tag{36}$$

Because the second-level here is the same as for the equal variance case, so is the update for alpha. The updates for β_i are derived in a similar manner as before but we also make use of the fact that the first-level posterior distribution factorises (see equation 19). This decouples the updates for each β_i and results in the following PEB algorithm

$$\begin{aligned}
\hat{e}_i &= y_i - \hat{w}_i x_i && (37) \\
\hat{z}_i &= \hat{w}_i - \hat{\mu} \\
\lambda_i &= \beta_i x_i^T x_i \\
\gamma_i &= \frac{\lambda_i}{\lambda_i + \alpha} \\
\gamma &= \sum_i \gamma_i \\
\beta_i &= (n_i - \gamma_i)/\hat{e}_i^T \hat{e}_i \\
\alpha &= \gamma / \hat{z}^T \hat{z}
\end{aligned}$$

$$\begin{aligned}\hat{w}_i &= (\beta_i x_i^T y_i + \alpha\mu)/(\lambda_i + \alpha) \\ d_i &= (\alpha^{-1} x_i x_i^T + \beta_i^{-1} I_{n_i})^{-1} \\ \sigma_\mu^2 &= 1/(\sum_i x_i^T d_i x_i) \\ \hat{\mu} &= \sigma_\mu^2 \sum_i x_i^T d_i y_i\end{aligned}$$

Initial values for $\hat{w}_i$ and β_i are set using OLS, $\hat{\mu}$ is initially set to the mean of $\hat{w}_i$ and α is initially set to 0. The equations are then iterated until convergence (in our examples in Chapter 12 we never required more than ten iterations). Whilst the above updates may seem somewhat complex, they can perhaps be better understood in terms of messages passing among nodes in a hierarchical network. This is shown in Figure 4 for the 'prediction' and 'prediction error' variables.

The PEB algorithms we have described show how Bayesian inference can take place when the variance components are unknown (in the previous section, we assumed the variance components were known). An application of this PEB algorithm to random effects analysis is provided in the next Chapter. We now provide a brief numerical example demonstrating the iterations with PEB updates.

Numerical example

This numerical example caricatures the use of PEB for estimating effect sizes from functional imaging data described in Chapter 23. The approach uses a 'global shrinkage prior' which embodies a prior belief that across the brain (i) the average effect is zero, $\mu = 0$ and (ii) the variability of responses follows a Gaussian distribution with precision α. Mathematically, we can write $p(w_i) = \mathsf{N}(0, \alpha^{-1})$. Figure 5(a) shows effect sizes generated from this prior for a $N = 20$-voxel brain and $\alpha = 1$.

Chapter 23 allows for multiple effects to be expressed at each voxel and for PET/fMRI data to be related to effect sizes using the full flexibility of General Linear Models (GLMs). Here, we just assume that data at each voxel are normally distributed about the effect size at that voxel. That is, $p(y_i|w_i) = \mathsf{N}(w_i, \beta_i^{-1})$. Figure 5(b) shows $n_i = 10$ data points at each voxel generated from this likelihood. We have allowed the observation noise precision β_i to be different at each voxel. Voxels 2, 15 and 18, for example, have noisier data than others.

Effect sizes were then estimated from this data using Maximum-Likelihood (ML) and PEB. ML estimates are shown in Figure 5(c) and (d). These are simply computed as the mean value observed at each voxel. PEB was implemented using the updates in equation 37 with $\mu = 0$ and $x_i = 1_{n_i}$ and initialised with $\alpha = 0$ and β_i and $\hat{w}_i$ set to ML-estimated values.

Equation 37 was then iterated, resulting in effect size estimates shown in Figure 6 before iterations one, three, five and seven. These estimates seem rather stable after only two or three iterations. Only the effects at voxels 5 and 15 seem markedly changed between iterations three and seven. The corresponding estimates of α were 0, 0.82, 0.91 and 0.95, showing convergence to the true prior response precision value of 1.

It is well known that PEB provides estimates that are, on average, more accurate than ML. Here, we quantify this using, σ_s, the standard deviation across voxels of the difference between the true and estimated effects. For ML, $\sigma_s = 0.71$ and for PEB, $\sigma_s = 0.34$. That PEB estimates are twice as accurate on average can be seen by comparing Figures 6(a) and (d). Of course, PEB is only better 'on average'. It does better at most voxels at the expense of being worse at a minority, for example, voxel 2. This trade-off is discussed further in Chapter 22.

PEB can do better than ML because it uses more information. Here, the information that effects have a mean of zero across the brain and follow a Gaussian variability profile. This shows the power of Bayesian estimation, which combines prior information with data in an optimal way. In this example, a key parameter in this trade off is the parameter γ_i which is computed as in equation 37. This quantity is the ratio of the data precision to the posterior precision. A value of 1 indicates that the estimated effect is determined solely by the data, as in ML. A value of 0 indicates the estimate is determined solely by the prior. For most voxels in our data set we have $\gamma_i \approx 0.9$, but for the noisy voxels 2, 15 and 18, we have $\gamma_i \approx 0.5$. PEB thus relies more on prior information where data is unreliable.

PEB will only do better than ML if the prior is chosen appropriately. For functional imaging data, we will never know what the 'true prior' is, just as we'll never know what the 'true model' is. But some priors and models are better than others, and there is a formal method for deciding between them. This is 'Bayesian model selection' and is described in Chapter 35.

Finally, we note that the prior used here does not use spatial information ie. there is no notion that voxel 5 is 'next to' voxel 6. It turns out that for functional imaging data, spatial information is important. In Chapter 25 we describe Bayesian fMRI inference with spatial priors. Bayesian model selection shows that models with spatial priors are preferred to those without [Penny et al. 2006].

Belief propagation

This Chapter has focussed on the special case of two-level models and Gaussian distributions. It is worthwhile noting that the general solution to inference in tree-structured hierarchical models, which holds for all distributions, is provided by the 'sum-product' or 'belief propagation' algorithm [Pearl 1988, Jordan and Weiss 2002]. This is a message passing algorithm which aims to deliver the marginal distributions [1] at each point in the hierarchy. It does this by propagating evidence up the hierarchy and marginal distributions down. If the downward messages are passed after the upward messages have reached the top, then this is equivalent to propagating the posterior beliefs down the hierarchy. This is shown schematically in Figure 7.

This general solution is important as it impacts on non-Gaussian and/or nonlinear hierarchical models. Of particular relevance are the models of inference in cortical hierarchies [Friston 2003] referred to in later Chapters of the book. In these models evidence flows up the hierarchy, in the form of prediction errors,

[1] The probability distribution over a set of variables is known as the joint distribution. The distribution over a subset is known as the marginal distribution.

and marginal distributions flow down, in the form of predictions. Completion of the downward pass explains late components of event related potentials which are correlated with eg. extra-classical receptive field effects [Friston 2003]. This general solution also motivates a data analysis approach known as Bayesian Model Averaging (BMA), described further in Chapter 35, where eg. x_3 in Figure 7 embodies assumptions about model structure. The downward pass of belief propagation then renders our final inferences independent of these assumptions. See Chapter 16 of [Mackay 2003] and [Ghahramani 1998] for further discussion of these issues.

Discussion

We have described Bayesian inference for some particular two-level linear-Gaussian hierarchical models. A key feature of Bayesian inference in this context is that the posterior distributions are Gaussian with precisions that are the sum of the data and prior precisions. The posterior means are the sum of the data and prior means, but each weighted according to their relative precision. With zero prior precision, two-level models reduce to a single-level model (ie. a GLM) and Bayesian inference reduces to the familiar maximum-likelihood estimation scheme. With non-zero and, in general unknown, prior means and precisions these parameters can be estimated using PEB. These covariance components can also be estimated using the ReML algorithm from classical statistics. The relation between PEB and ReML is discussed further in Chapter 22.

We have described two special cases of the PEB algorithm, one for equal variances and one for separable models. Both algorithms are special cases of a general approach described in [Friston et al. 2002a] and in Chapter 24. In these contexts, we have shown that PEB automatically partitions the total degrees of freedom (ie. number of data points) into those to be used to estimate the hyperparameters of the prior distribution and those to be used to estimate hyperparameters of the likelihood distribution. The next Chapter describes how PEB can be used in the context of random effects analysis.

References

[Berger 1985] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, 1985.

[Carlin and Louis 2000] B.P. Carlin and T.A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall, 2000.

[Friston et al. 2002a] K.J. Friston, W.D. Penny, C. Phillips, S.J. Kiebel, G. Hinton, and J. Ashburner. Classical and Bayesian inference in neuroimaging: Theory. *NeuroImage*, 16:465–483, 2002.

[Friston et al. 2002b] K.J. Friston, D.E. Glaser, R.N.A. Henson, S.J. Kiebel, C. Phillips, and J. Ashburner. Classical and Bayesian inference in neuroimaging: Applications. *NeuroImage*, 16:484–512, 2002.

[Friston 2003] K. Friston. Learning and inference in the brain. *Neural Networks*, 16:1325–1352, 2003.

[Gelman 1995] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis.* Chapman and Hall, Boca Raton, 1995.

[Ghahramani 1998] Z. Ghahramani. Learning dynamic bayesian networks. In C.L. Giles and M.Gori, editors, *Adaptive Processing of Temporal Information.* Springer-Verlag, 1998.

[Holmes et al. 1997] A.P. Holmes, J.B. Poline, and K.J. Friston. Characterizing brain images with the general linear model. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, and J.C. Mazziotta, editors, *Human Brain Function*, pages 59–84. Academic Press USA, 1997.

[Jordan and Weiss 2002] M. Jordan and Y. Weiss. Graphical models: Probabilistic inference. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks.* MIT Press, 2002.

[Lee 1997] P. M. Lee. *Bayesian Statistics: An Introduction.* Arnold, 2 edition, 1997.

[Mackay 1992] D.J.C. Mackay. Bayesian Interpolation. *Neural computation*, 4(3):415–447, 1992.

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

[Pearl 1988] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan Kauffman, 1988.

[Penny et al. 2006] W.D. Penny, G. Flandin, and N. Trujillo-Barreto. Bayesian Comparison of Spatially Regularised General Linear Models. *Human Brain Mapping*, 2006. Accepted for publication.

[Winer et al. 1991] B.J. Winer, D.R. Brown, and K.M. Michels. *Statistical principles in experimental design.* McGraw-Hill, 1991.

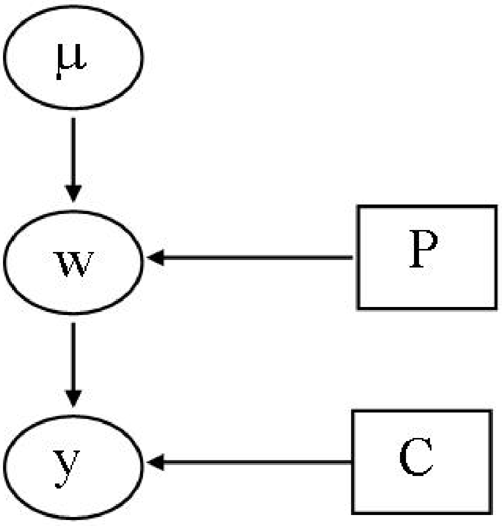

Figure 1: *Two-level hierarchical model. The data y are explained as deriving from an effect w and a zero-mean Gaussian random variation with covariance C. The effects w in turn are random effects deriving from a superordinate effect μ and zero-mean Gaussian random variation with covariance P. The goal of Bayesian inference is to make inferences about μ and w from the posterior distributions $p(\mu|y)$ and $p(w|y)$.*

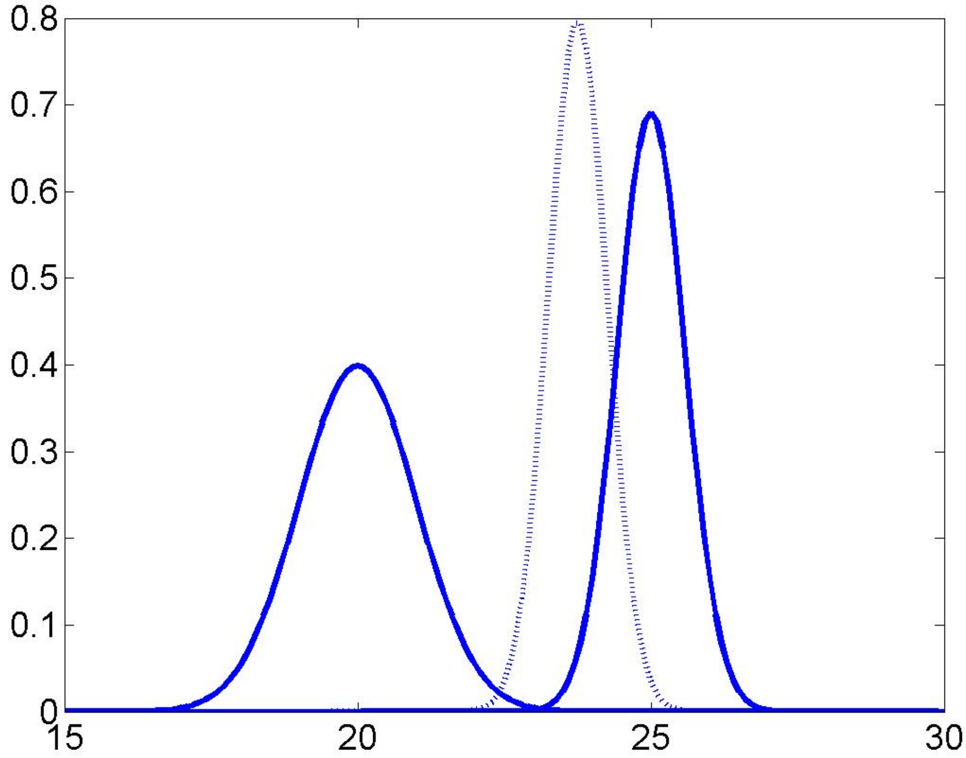

Figure 2: *Bayes rule for univariate Gaussians. The two solid curves show the probability densities for the prior* $p(w) = \mathsf{N}(\mu, \alpha^{-1})$ *with* $\mu = 20$ *and* $\alpha = 1$ *and the likelihood* $p(y|w) = \mathsf{N}(w, \beta^{-1})$ *with* $w = 25$ *and* $\beta = 3$. *The dotted curve shows the posterior distribution,* $p(w|y) = \mathsf{N}(m, \lambda^{-1})$ *with* $m = 23.75$ *and* $\lambda = 4$, *as computed from equation 14. The posterior distribution is closer to the likelihood because the likelihood has higher precision.*

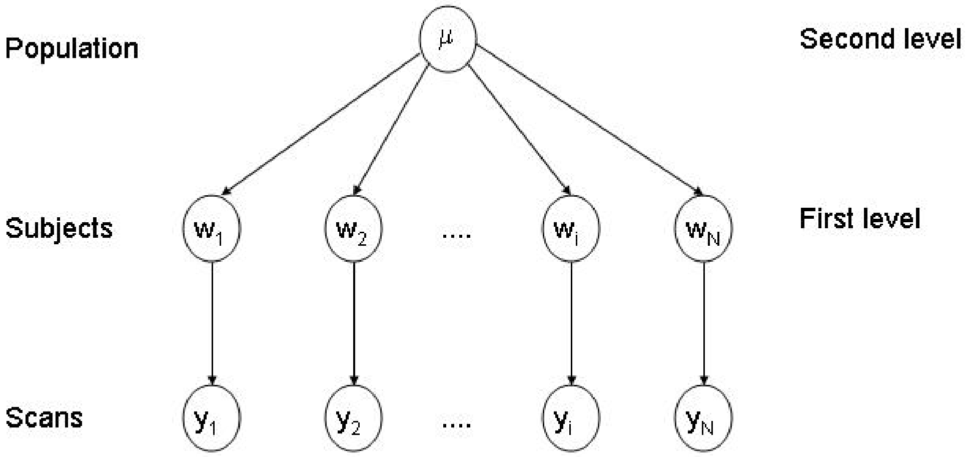

Figure 3: *Generative model for random effects analysis.*

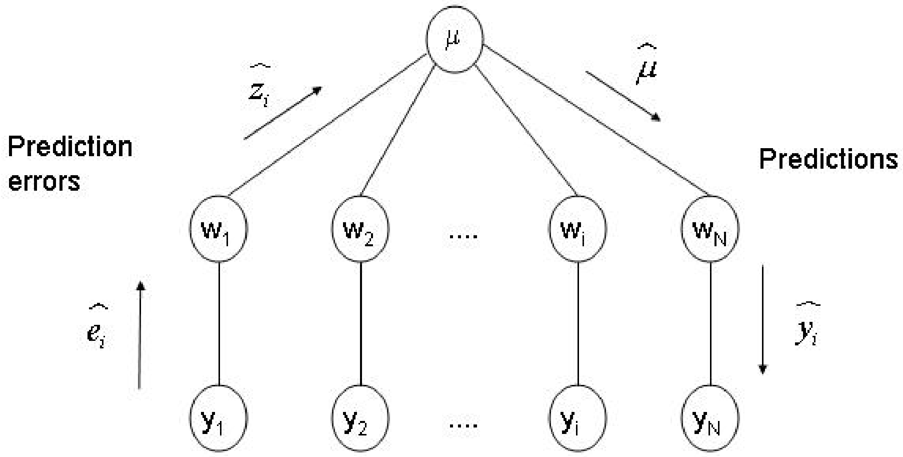

Figure 4: *Part of the PEB algorithm for separable models requires the upwards propagation of prediction errors and downwards propagation of predictions. This passing of messages between nodes in the hierarchy is a special case of the more general belief propagation algorithm referred to in Figure 7*

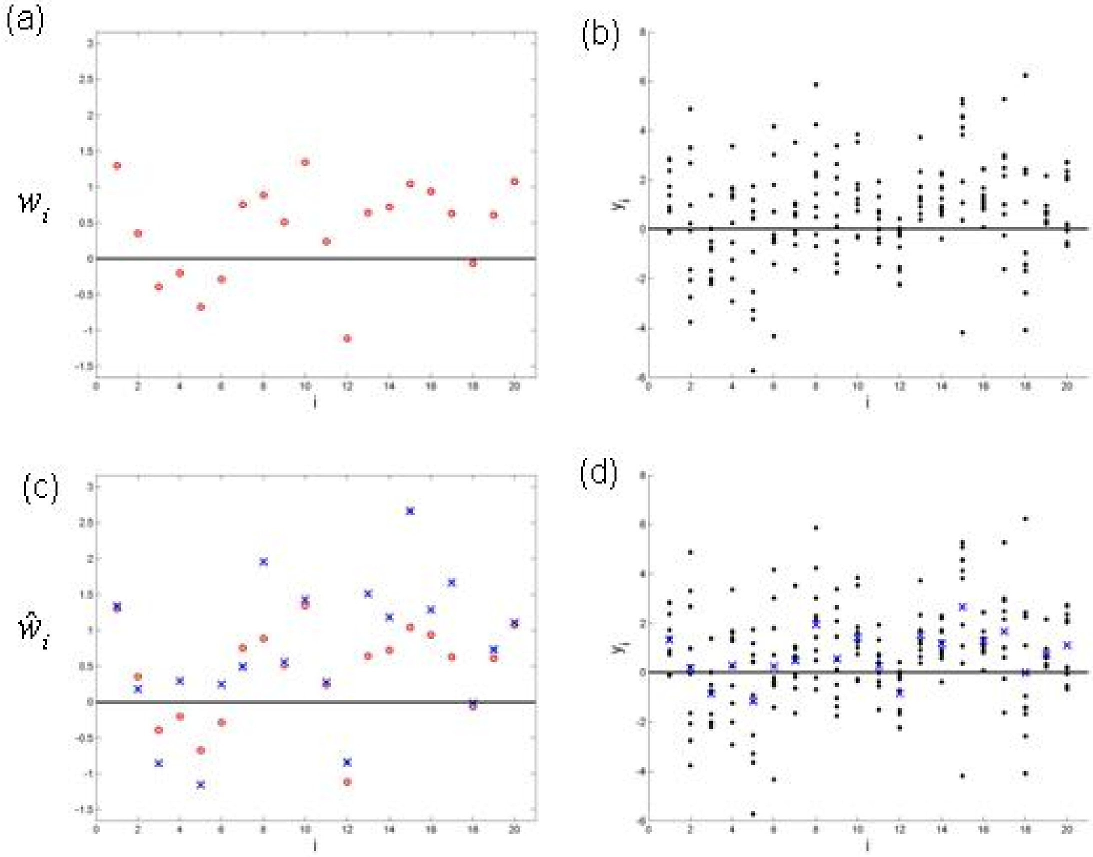

Figure 5: *Data for PEB example. (a) Red circles denote 'true' effect sizes, w_i, for each voxel i, generated from the prior $p(w_i|\alpha) = \mathsf{N}(0, \alpha^{-1})$ with $\alpha = 1$. (b) The black dots denote $n_i = 10$ data points at each voxel generated from the likelihood $p(y_i|w_i) = \mathsf{N}(w_i, \beta_i^{-1})$ with β_i drawn from a uniform distribution between 0.1 and 1. Thus some voxels, eg. voxels 2, 15 and 18, have noisier data than others. Plots (c) and (d) are identical to (a) and (b) but with blue crosses indicating Maximum Likelihood (ML) estimates of the effect size, $\hat{w}_i$. These are simply computed as the mean of the data at each voxel, and are used to initialise PEB - see Figure 6.*

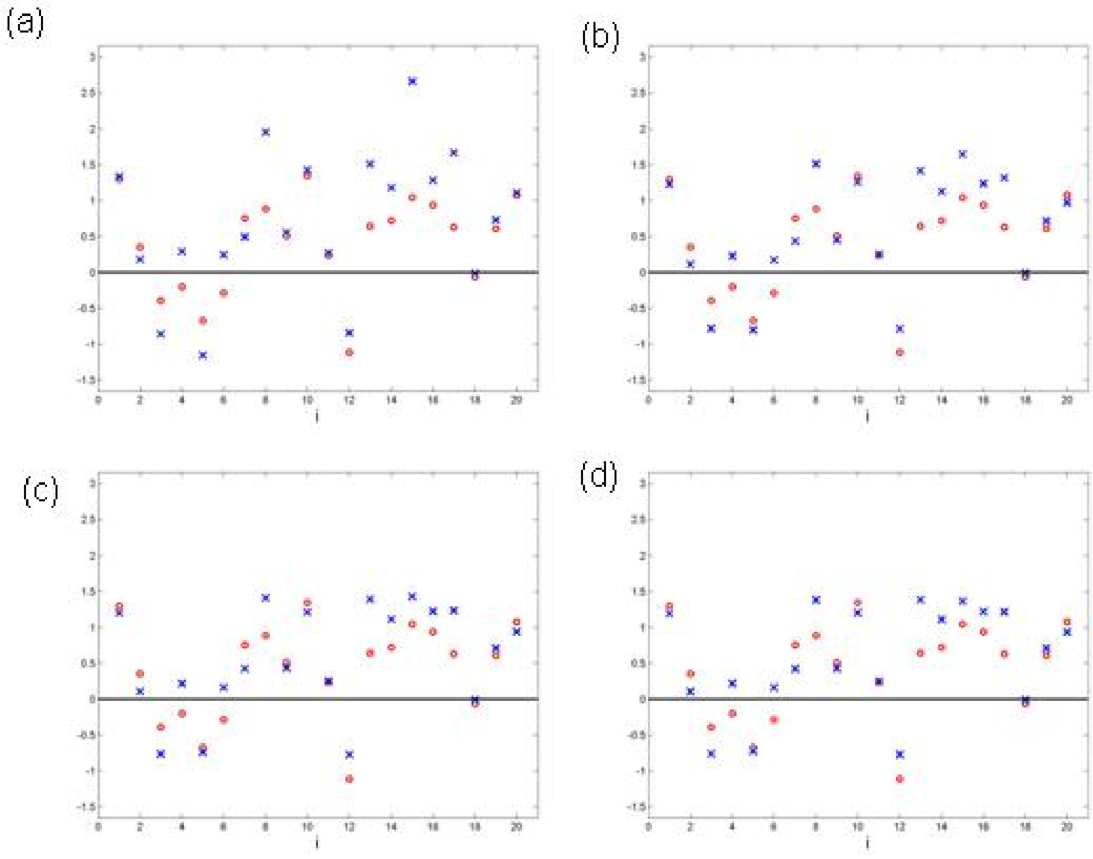

Figure 6: *The plots show the true effect sizes, w_i (red circles) and estimated effect sizes, $\hat{w}_i$, (blue crosses) before PEB iteration number (a) one, (b) three, (c) five and (d) seven. Plot (a) here is the same as plot (c) in the previous figure, as the estimates were initialised using ML.*

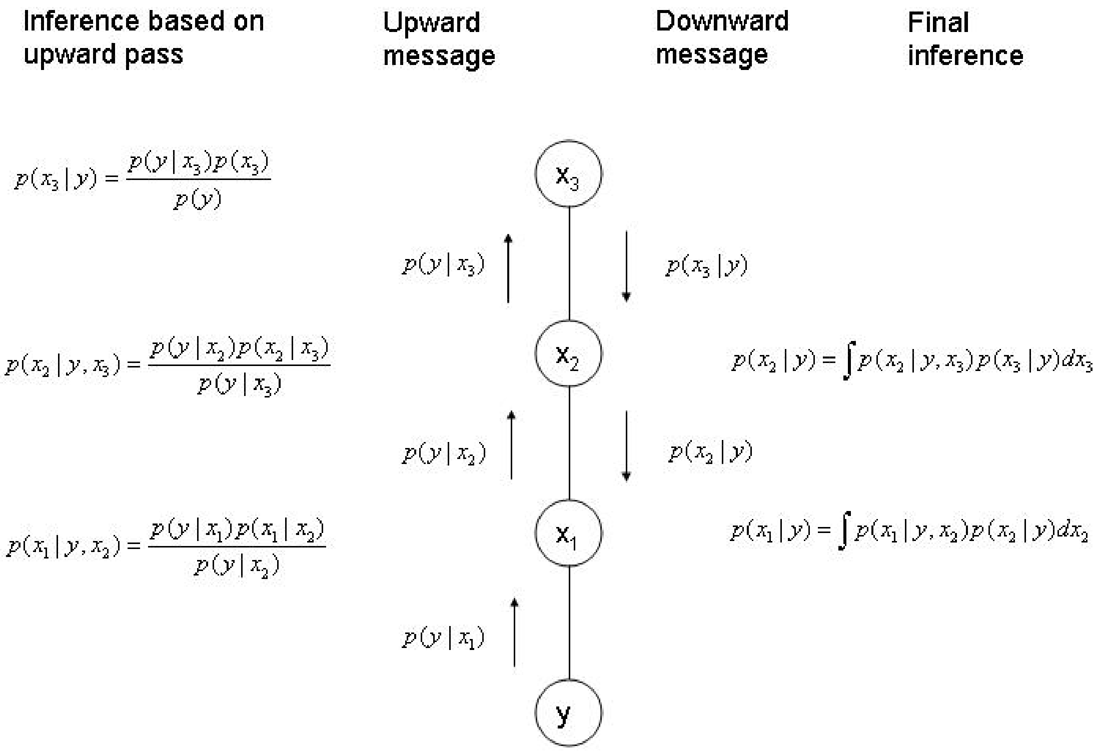
Inference based on upward pass
Upward message
Downward message
Final inference
$p(x_3|y)=\frac{p(y|x_3)p(x_3)}{p(y)}$
x_3
$p(y|x_3)$
$p(x_3|y)$
$p(x_2|y,x_3)=\frac{p(y|x_2)p(x_2|x_3)}{p(y|x_3)}$
x_2
$p(x_2|y)=\int p(x_2|y,x_3)p(x_3|y)dx_3$
$p(y|x_2)$
$p(x_2|y)$
$p(x_1|y,x_2)=\frac{p(y|x_1)p(x_1|x_2)}{p(y|x_2)}$
x_1
$p(x_1|y)=\int p(x_1|y,x_2)p(x_2|y)dx_2$
$p(y|x_1)$
y

Figure 7: *Belief propagation for inference in hierarchical models. This algorithm is used to update the marginal densities ie. to update $p(x_i)$ to $p(x_i|y)$. Inferences based on purely the upward pass are contingent on variables in the layer above whereas inferences based on upwards and downwards passes are not. Completion of the downward pass delivers the marginal density. Application of this algorithm to the two-level Gaussian model will produce the update equations 7 and 11. More generally, this algorithm can be used for Bayesian model averaging, where eg. x_3 embodies assumptions about model structure, and as a model of inference in cortical hierarchies, where eg. completion of the downward pass explains extra-classical receptive field effects [Friston 2003].*