Hierarchical Models

W.D. Penny and K.J. Friston

Wellcome Department of Imaging Neuroscience,
University College London.

February 28, 2003

1 Introduction

Hierarchical models are central to many current analyses of functional imaging data including random effects analysis, models using fMRI as priors for EEG source localization and spatiotemporal Bayesian modelling of imaging data [3]. These hierarchical models posit linear relations between 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 17 and in [4]. The algorithm is sufficiently general to accomodate multiple levels in the hierarchy 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.

We provide an introduction to hierarchical models and focus on some relatively simple examples. Each model and PEB algorithm we present is a special case of that described in [4]. Whilst there are a number of tutorials on hierarchical modelling [9],[2] 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 [5].

We restrict our attention to two-level models and show, in section 2, how one computes the posterior distributions over the first- and second-level parameters. These are derived, initially, for completely general design and error covariance matrices. We then consider two special cases (i) models with equal error variances and (ii) separable models. In section 3 of the chapter we show how the parameters and covariance components can be estimated using PEB.

In section 4 we show how a two-level hierarchical model can be used for Random-Effects Analysis. For equal subject error variances at the first level and the same first-level design matrices (ie. balanced designs) we show that the resulting inferences are identical to those made by the Summary-Statistic (SS) approach [11]. If either of the criteria are not met then, strictly, the SS approach is not valid. In section 5, however, we show that a modified SS approach can be used for unbalanced designs and unequal error variances if the covariance structure of the model at the second-level is modified appropriately. The chapter closes 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.

2 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, [3].

The aim of Bayesian inference is to make inferences about w and μ (if we don't already know them) 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

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

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

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

where the covariance term

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

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 a number of special cases.

2.1 Sensor Fusion

The first special case is the univariate model

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

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 posseses 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.

2.2 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 [12] 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 is therefore 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 [7]. The posterior distribution of the second level coefficient is given by equation 11 with

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

2.3 Separable model

We now consider 'separable models' which can be used, for example, for random effects analysis. In these models, the first-level splits into N separate sub-models. For each sub-model, i, there are n_i observations y_i giving information about the parameter w_i via the design vector x_i (this would typically be a boxcar or, for event-related designs, a vector of delta functions). 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 $C = \mathsf{diag}[\beta_1, \beta_1, \beta_2, \beta_2, \beta_2, \beta_3, \beta_3]$. 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 of the first level coefficient 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 of the second level coefficient is, from equation 11, given by

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

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

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

3 Parametric Empirical Bayes

In section 2 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 which in general are unknown. For the equal variance model and the separable model, the hyperparameters α and β_i are generally unknown. In [4] Friston et al. decompose the covariances 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 [3]. The hyperparameters $\lambda = [\{\lambda_j^1\}, \{\lambda_j^2\}]$ are unknown but can be estimated using the PEB algorithm described in [4]. Variants of this algorithm are known as the *evidence framework* [10] or *Maximum Likelihood II (ML-II)* [1]. 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 [2]. 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.

3.1 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

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

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 in section 2 where we were only interested in w-dependent terms) the integral can be written as

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

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'

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

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

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

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

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

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

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

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

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

and then let

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

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.

3.2 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

$$\begin{aligned}
\hat{e}_i &= y_i - \hat{w}_i x_i && (37)\\
\hat{z}_i &= \hat{w}_i - \hat{\mu} \\
\lambda_i &= \beta_i x_i^T x_i \\
\gamma_i &= \frac{\lambda_i}{\lambda_i + \alpha} \\
\gamma &= \sum_i \gamma_i \\
\beta_i &= (n_i - \gamma_i)/\hat{e}_i^T \hat{e}_i \\
\alpha &= \gamma/\hat{z}^T \hat{z} \\
\hat{w}_i &= (\beta_i x_i^T y_i + \alpha \mu)/(\lambda_i + \alpha) \\
d_i &= (\alpha_i^{-1} x_i x_i^T + \beta_i^{-1} I_{n_i})^{-1} \\
\sigma_\mu^2 &= 1/(\sum_i x_i^T d_i x_i) \\
\hat{\mu} &= \sigma_\mu^2 \sum_i x_i^T d_i y_i
\end{aligned}$$

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 we never required more than ten iterations).

The PEB algorithms we have described show how Bayesian inference can take place when the variance components are unknown (in section 2 we assumed the variance components were known). We now turn to an application.

4 Random-Effects Analysis

To make contact with the summary statistic and ML approaches (see [11]) we desribed the statistical model underlying random effects analysis as follows. The model described in this section is identical to the separable model but with $x_i = 1_n$ and $\beta_i = \beta$. Given a data set of contrasts from N subjects with n scans per subject, the population contrast can be modelled by the two level process

$$\begin{aligned} y_{ij} &= w_i + e_{ij} \\ w_i &= w_{pop} + z_i \end{aligned} \tag{38}$$

where y_{ij} (a scalar) is the data from the ith subject and the jth scan at a particular voxel. These data points are accompanied by errors e_{ij} with w_i being the size of the effect for subject i, w_{pop} being the size of the effect in the population and z_i being the between subject error. This may be viewed as a Bayesian model where the first equation acts as a likelihood and the second equation acts as a prior. That is

$$\begin{aligned} p(y_{ij}|w_i) &= \mathsf{N}(w_i, \sigma_w^2) \\ p(w_i) &= \mathsf{N}(w_{pop}, \sigma_b^2) \end{aligned} \tag{39}$$

where σ_b^2 is the between subject variance and σ_w^2 is the within subject variance. We can make contact with the hierarchical formalism by making the following identities. We place the y_{ij} in the column vector y in the order - all from subject 1, all from subject 2 etc (this is described mathematically by the *vec* operator and is implemented in MATLAB (Mathworks, Inc.) by the colon operator). We also let $X = I_N \otimes 1_n$ where $\otimes$ is the Kronecker product and let $w = [w_1, w_2, ..., w_N]^T$. With these values the first level in equation 2 is then the matrix equivalent of equation 38 (ie. it holds for all i, j). For $y = Xw + e$ and eg. $N = 3$,$n = 2$ we then have

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix} \tag{40}$$

We then note that $X^T X = nI_N$, $\hat{\Sigma} = \mathsf{diag}(\mathsf{Var}[w_1], \mathsf{Var}[w_2], ..., \mathsf{Var}[w_N])$ and the ith element of $X^T y$ is equal to $\sum_{j=1}^n y_{ij}$.

If we let $M = 1_N$ then the second level in equation 2 is then the matrix equivalent of the second-level in equation 38 (ie. it holds for all i). Plugging in our values for M and X and letting $\beta = 1/\sigma_w^2$ and $\alpha = 1/\sigma_b^2$ gives

$$\mathsf{Var}[\hat{w}_{pop}] = \frac{1}{N}\frac{\alpha + \beta n}{\alpha\beta n} \tag{41}$$

and

$$\begin{aligned} \hat{w}_{pop} &= \frac{1}{N}\frac{\alpha + \beta n}{\alpha\beta n}\frac{\alpha\beta}{\alpha + \beta n}\sum_{i,j} y_{ij} \\ &= \frac{1}{Nn}\sum_{i,j} y_{ij} \end{aligned} \tag{42}$$

So the estimate of the population mean is simply the average value of y_{ij}. The variance can be re-written as

$$\mathsf{Var}[\hat{w}_{pop}] = \frac{\sigma_b^2}{N} + \frac{\sigma_w^2}{Nn} \tag{43}$$

This result is identical to the maximum-likelihood and summary-statistic results. The equivalence between the Bayesian and ML results derives from the fact that there is no prior at the population level. Hence, $p(Y|\mu) = p(\mu|Y)$ as indicated in section 2.

4.1 Unequal variances

The model described in this section is identical to the separable model but with $x_i = 1_{n_i}$. If the error covariance matrix is non-isotropic ie. $C \neq \sigma_w^2 I$, then the population estimates will change. This can occur, for example, if the design matrices are different for different subjects (so-called 'unbalanced-designs'), or if the data from some of the subjects is particularly ill-fitting. In these cases, we consider the within subject variances $\sigma_w^2(i)$ and the number of scans n_i to be subject-specific.

If we let $M = 1_N$ then the second level in equation 2 is then the matrix equivalent of the second-level in equation 38 (ie. it holds for all i). Plugging in our values for M and X gives

$$\mathsf{Var}[\hat{w}_{pop}] = \left(\sum_{i=1}^{N} \frac{\alpha\beta_i n_i}{\alpha + n_i\beta_i}\right)^{-1} \tag{44}$$

and

$$\hat{w}_{pop} = \left(\sum_{i=1}^{N} \frac{\alpha\beta_i n_i}{\alpha + \beta_i n_i}\right)^{-1} \sum_{i=1}^{N} \frac{\alpha\beta_i}{\alpha + \beta_i n_i} \sum_{j=1}^{n_i} y_{ij} \tag{45}$$

This reduces to the earlier result if $\beta_i = \beta$ and $n_i = n$. Both of these results are different to the summary statistic approach which we note is therefore invalid for unequal variances.

4.2 Parametric Empirical Bayes

To implement the PEB estimation scheme for the unequal variance case we first compute the errors $\hat{e}_{ij} = y_{ij} - X\hat{w}_i$, $\hat{z}_i = \hat{w}_i - M\hat{w}_{pop}$. We then substitute $x_i = 1_{n_i}$ into the update rules derived in section 3 to obtain

$$\sigma_b^2 \equiv \frac{1}{\alpha} = \frac{1}{\gamma}\sum_{i=1}^{N} \hat{z}_i^2 \tag{46}$$

$$\sigma_w^2(i) \equiv \frac{1}{\beta_i} = \frac{1}{n_i - \gamma_i}\sum_{j=1}^{n_i} \hat{e}_{ij}^2 \tag{47}$$

where

$$\gamma = \sum_{i=1}^{N} \gamma_i \tag{48}$$

and

$$\gamma_i = \frac{n_i\beta_i}{\alpha + n_i\beta_i} \tag{49}$$

For balanced designs $\beta_i = \beta$ and $n_i = n$ we get

$$\sigma_b^2 \equiv \frac{1}{\alpha} = \frac{1}{\gamma}\sum_{i=1}^{N} \hat{z}_i^2 \tag{50}$$

$$\sigma_w^2 \equiv \frac{1}{\beta} = \frac{1}{Nn - \gamma}\sum_{i=1}^{N}\sum_{j=1}^{n} \hat{e}_{ij}^2 \tag{51}$$

where

$$\gamma = \frac{n\beta}{\alpha + n\beta}N \tag{52}$$

Effectively, the degrees of freedom in the data set (Nn) are partitioned into those that are used to estimate the between-subject variance, γ, and those that are used to estimate the within-subject variance, $Nn - \gamma$.

The posterior distribution of the first-level coefficients is

$$p(w_i|y_{ij}) \equiv p(\hat{w}_i) = \mathsf{N}(\bar{w}_i, \mathsf{Var}[\hat{w}_i]) \tag{53}$$

where

$$\mathsf{Var}[\hat{w}_i] = \frac{1}{\alpha + n_i\beta_i} \tag{54}$$

$$\hat{w}_i = \frac{\beta_i}{\alpha + n_i\beta_i}\sum_{j=1}^{n_i} y_{ij} + \frac{\alpha}{\alpha + n_i\beta_i}\hat{w}_{pop} \tag{55}$$

Overall, the EB estimation scheme is implemented by first initialising $\hat{w}_i$, $\hat{w}_{pop}$ and α, β_i (for example to values given from the equal error-variance scheme). We then compute the errors $\hat{e}_{ij}$, $\hat{z}_i$ and re-estimate the α and β_i's using the above equations. The coefficients $\hat{w}_i$ and $\hat{w}_{pop}$ are then re-estimated and the last two steps are iterated until convergence. This algorithm is identical to the PEB algorithm for the separable model but with $x_i = 1_{n_i}$.

5 Second-level modelling

The results at the beginning of section 4 and in section 4.1 show that the SS approach is equivalent to PEB for equal first level error variances and balanced designs, but that SS is otherwise invalid. In this section we show that a modified SS approach that uses a non-isotropic covariance at the second level is equivalent to PEB. Firstly, we re-write the first-level equation in 2 as

$$w = X^{+}(y - e) \tag{56}$$

and substitute w into the second level and re-arrange to give

$$X^{+}y = M\mu + z + X^{+}e \tag{57}$$

By letting $c = X^{+}y$ and $r = z + X^{+}e$ we can write the above equation as

$$c = M\mu + r \tag{58}$$

where

$$R \equiv \mathsf{Cov}[r] = P + X^{+}C(X^{+})^{T} \tag{59}$$

The estimation of μ can then proceed based solely on c, R and M

$$\begin{aligned} p(\mu|y) &= \mathsf{N}(\hat{\mu}, \Sigma_{\mu}) \\ \hat{\mu} &= (M^{T}R^{-1}M)^{-1}M^{T}R^{-1}c \\ \Sigma_{\mu} &= (M^{T}R^{-1}M)^{-1} \end{aligned} \tag{60}$$

This implies that if we bring forward OLS parameter estimates from the first-level (ie. $c = X^{+}y$) then we can take into account the, as yet unaccounted for, non-sphericity at the first-level by using an appropriately corrected covariance matrix at the second-level (the matrix R). Jenkinson et al. [8] have proposed a similar strategy but based on Weighted Least Squares (WLS) parameter estimates from the first-level. The problem with this 'plug-in approach', however is that P is unknown. Note that SS estimates of hyperparameters in P contain contributions from both within and between subject error, as shown in [11], so these could not be used directly.

5.1 Separable models

For the case of unequal variances at the first level (described at the beginning of this section), we have $C = \sum_{i=1}^{N} \beta_i^{-1} I^i$, $P = \alpha^{-1} I_N$ and $X = I_N \otimes 1_n$. This gives $X^{+} = n^{-1}(I_N \otimes 1_n^T)$ and results in a diagonal matrix for R with entries

$$R_{ii} = \sigma_b^2 + \frac{1}{n_i}\sigma_{w(i)}^2 \tag{61}$$

The fact that R is diagonal for separable models is no surprise as subjects are drawn independently from the population. Re-assuringly, plugging in the above value of R_{ii} into the expression for Σ_{μ} above gives the same estimate of population variance as before (cf. equations 43 and 44).

Thus, in principle, one could bring forward both OLS estimates, c, *and* first-level variances ($\sigma_{w(i)}^2$) to a second-level analysis. However, as we have already mentioned, the hyperparmeter of P, ie. σ_b^2, is unknown.

An alternative strategy is to estimate the hyperparameters R_{ii} using PEB based solely on a second level model. Ordinarily this would be impossible as there are more hyperparameters and parameters $(N+1)$ than (second-level) data points (N). But by pooling data over voxels, as described in [6], this becomes feasible.

6 Example

We now give an example of random effects analysis on simulated data. The purpose is to compare the PEB and SS algorithms. We generated data from a three-subject, two-level model with population mean $\mu = 2$, subject effect sizes $w = [2.2, 1.8, 0.0]^T$ and within subject variances $\sigma_w^2(1) = 1$, $\sigma_w^2(2) = 1$. For the third subject $\sigma_w^2(3)$ was varied from 1 to 10. The second level design matrix was $M = [1, 1, 1]^T$ and the first-level design matrix was given by $X = \mathsf{blkdiag}(x_1, x_2, x_3)$ with x_i being a boxcar. This model conforms to the notion of a separable model defined in section 2.3.

Figure 3 shows a realisation of the three time series for $\sigma_w^2(3) = 2$. The first two time series contain stimulus-related activity but the third does not. We then applied the PEB algorithm (section 3.2) to obtain estimates of the population mean $\hat{\mu}$ and estimated variances, $\sigma_{\hat{\mu}}^2$. For comparison, we also obtained equivalent estimates using the SS approach. We then computed the accuracy with which the population mean was estimated using the criterion $(\hat{\mu} - \mu)^2$. This was repeated for 1000 different data sets generated using the above parameter values, and for 10 different values of $\sigma_w^2(3)$. The results are shown in figures 4 and 5.

Firstly we note that, as predicted by theory, both PEB and SS give identical results when the first level error variances are equal. When the variance on the 'rogue' time series approaches double that of the others we see different estimates both $\hat{\mu}$ and $\sigma_{\hat{\mu}}^2$. With increasing rogue error variance the SS estimates get worse but the PEB estimates get better (with respect to the true values, as shown in Figure 4, and with respect to the variability of the estimate, as shown in Figure 5). This is because the third time series is more readily recognised by PEB as containing less reliable information about the population mean and is increasingly ignored. This gives better estimates $\hat{\mu}$ and a reduced estimation error, $\sigma_{\hat{\mu}}^2$.

We created the above example to reiterate a key point of this chapter, that SS gives identical results to PEB for equal within subject error variances (homoscedasticity) and unbalanced designs, but not otherwise. In the example, divergent behaviour is observed when the error variances differ by a factor of two. For studies with more subjects (12 being a typical number), however, this divergence requires a much greater disparity in error variances. In fact we initially found it difficult to generate data sets where PEB showed a consistent improvement over SS ! It is therefore our experience that the vanilla SS approach is particularly robust to departures from homoscedasticity. This conclusion is supported by what is known of the robustness of the t-test that is central to the SS approach. Lack of homoscedasticity only causes problems when the sample size (ie. number of subjects) is small. As sample size increases so does the robustness (see eg. [13]).

7 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 and with means that are the sum of the data and prior means, 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.

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 [4] and in Chapter 17. 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 hyperparamaters of the prior distribution and those to be used to estimate hyperparameters of the likelihood distribution.

We have shown, both theoretically and via computer simulation, that a random effects analysis based on PEB and one based on the summary statistic approach are identical given that the first-level error variances are equal. For unequal error variances we have shown, via simulations, how the accuracy of the summary-statistic approach falls off.

Finally, we have noted that the standard summary statistic approach assumes an isotropic error covariance matrix at the second level. If, however, this matrix is changed to reflect both first and second level covariance terms then this ‘modified’ summary statistic approach will give identical results to PEB. This requires that we make estimates of the ‘non-sphericity’ (see Chapter 10) at the second level.

References

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

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

[3] K.J. Friston, D. Glaser, R. Henson, S. Kiebel, C. Phillips, and J. Ashburner. Classical and Bayesian inference in neuroimaging: Applications. *Neuroimage*, 16:484–512, 2002.

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

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

[6] D.E. Glaser and K.J. Friston. Pooling and covariance component estimation in SPM. Manuscript in preparation, 2003.

[7] A. 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.D. Frith, R.J. Dolan, and J.C. Mazziotta, editors, *Human Brain Function*, pages 59–84. Academic Press USA, 1997.

[8] M. Jenkinson, M. Woolrich, D. Leibovici, S. Smith, and C. Beckmann. Group analysis in fMRI using genral multi-level linear modelling. In *HBM: Eighth International Conference on Functional Mapping of the Human Brain, Sendai, Japan*, page 417, 2002.

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

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

[11] W.D. Penny, A. Holmes, and K.J. Friston. Random effects analysis. Technical report, In Human Brain Function II.

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

[13] B.S. Yandell. *Practical data analysis for designed experiments.* Chapman and Hall, 1997.

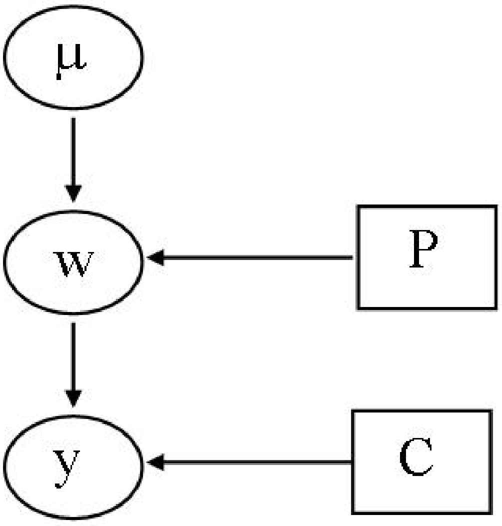

Figure 1: *Two-level hierarchical model. The data y are explained as deriving from an effect w and a zero-mean Gaussian random variation with covariance C. The effects w in turn are random effects deriving from a superordinate effect μ and zero-mean Gaussian random variation with covariance P. The goal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

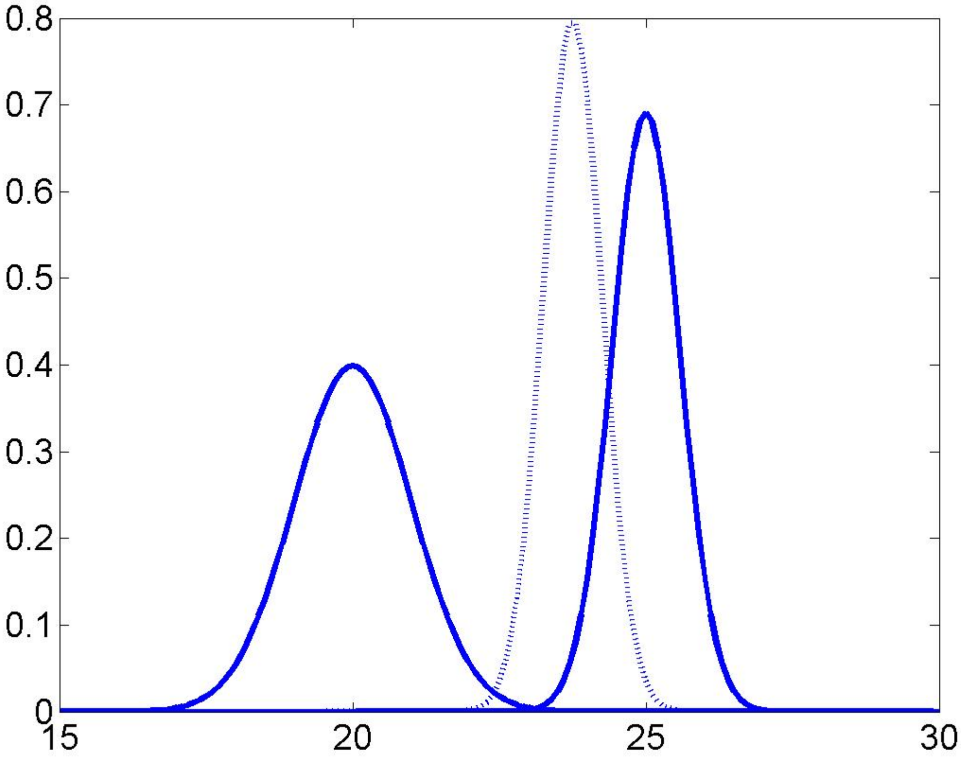

Figure 2: *Bayes rule for univariate Gaussians. The two solid curves show the probability densities for the prior* $p(w) = \mathsf{N}(\mu, \alpha^{-1})$ *with* $\mu = 20$ *and* $\alpha = 1$ *and the likelihood* $p(y|w) = \mathsf{N}(w, \beta^{-1})$ *with* $w = 25$ *and* $\beta = 3$*. The dotted curve 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.*

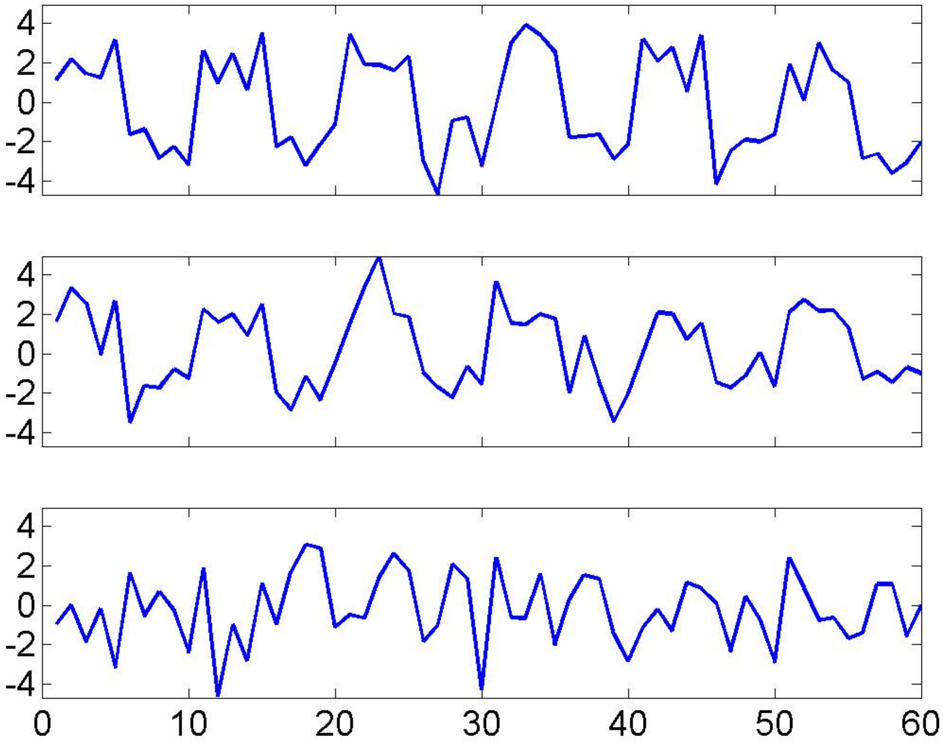

Figure 3: *Simulated data for random effects analysis. Three representative time series produced from the two-level hierarchical model. The first two time-series contain stimulus-related activity but the third does not.*

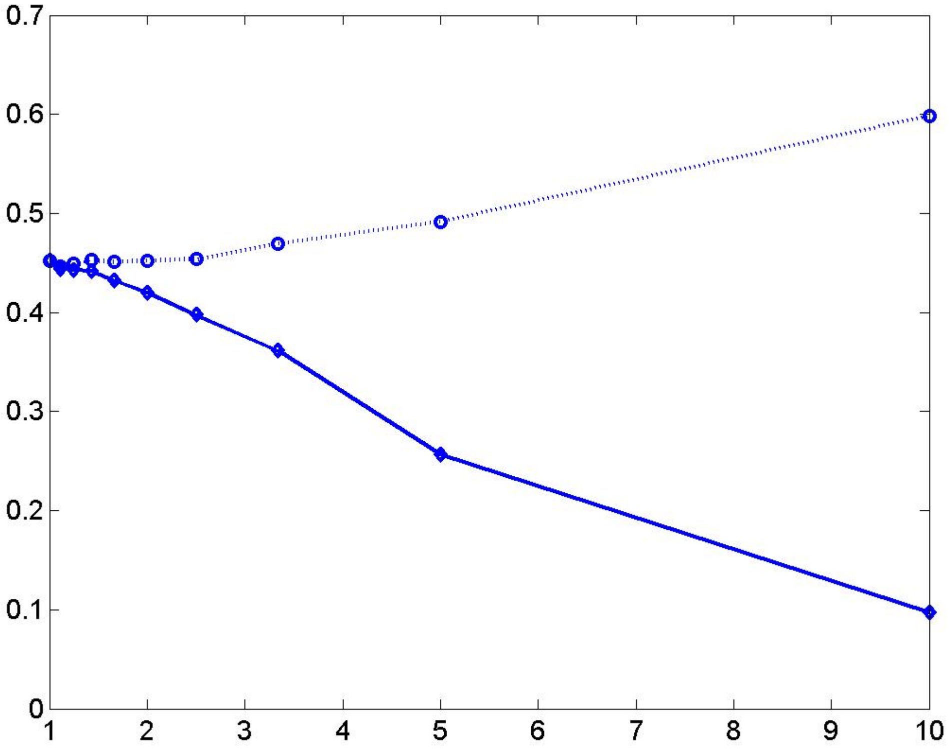

Figure 4: *A plot of the error in estimating the population mean* $E = <(\hat{\mu} - \mu)^2>$ *versus the observation noise level for the third subject,* $\sigma_w^2(3)$*, for the Empirical Bayes approach (solid line) and the Summary-Statistic approach (dotted line).*

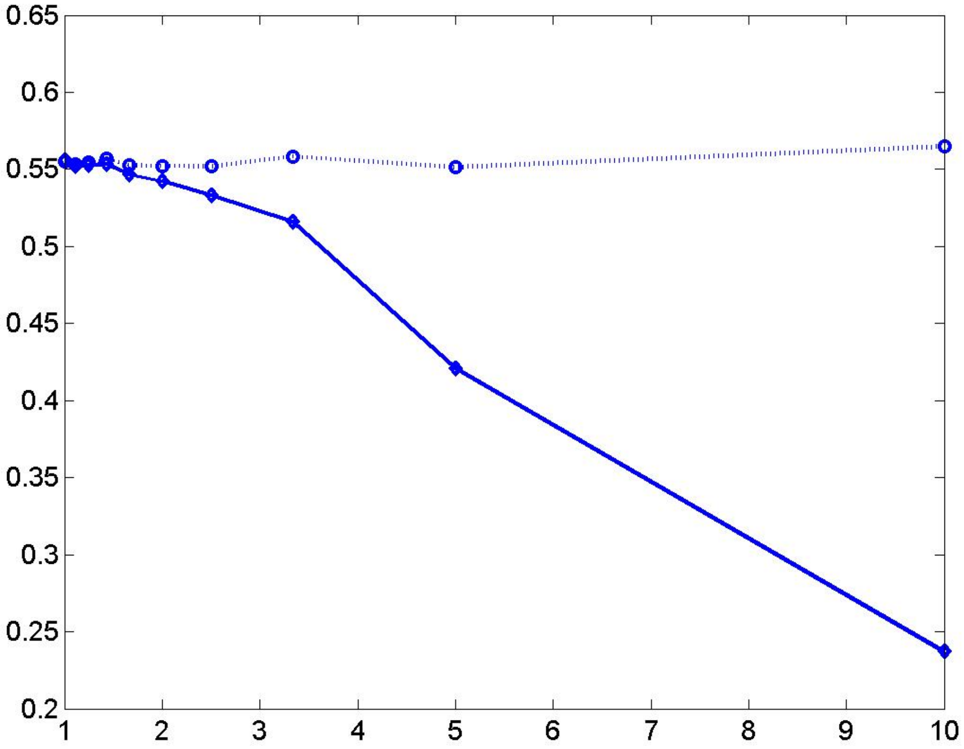

Figure 5: *A plot of the estimated variance of the population mean, σ_μ^2, versus the observation noise level for the third subject, $\sigma_w^2(3)$, for the Empirical Bayes approach (solid line) and the Summary-Statistic approach (dotted line).*