Imperial College of Science Technology and Medicine
Department of Mathematics

EMPIRICAL BAYES INFERENCE AND THE LINEAR MODEL

by

Angela B. Mariotto

Thesis submitted to the University of London for the degree of Doctor of Philosophy (Ph. D.)

November 1988

ABSTRACT

In this thesis we investigate empirical Bayes inference for the normal-theory linear model. Two models are considered: variances changing according to the conjugate prior distribution, inverse-Gamma, and means coming from a normal population. The main interest focuses on the construction of confidence intervals for the means. A combination of numerical work and asymptotic theory is used to investigate the effect of the prior estimation on the confidence levels; a correction for significance levels to allow for the errors in estimating the prior parameters is calculated. Related to this is the estimation of the prior parameters. Estimates are compared by means of their large-sample behaviour and the relation between efficiency and the errors in confidence levels is outlined.

The sensitivity of the empirical Bayes analysis to possible misspecification of the Normal conjugate class of distributions, is investigated by comparing the empirical Bayes estimates from the Edgeworth expansions for the normal distribution.

Also the possibility of discriminating the inverse-Gamma prior distribution for the variances, from other distributions is discussed, and a graphical method developed.

Two illustrations of the theory are discussed: (i) the analysis of several independent 2×2 contingency tables and (ii) a comparative study on the rate of growth of the AIDS epidemic in Europe.

The influence of observations from a training sample on the probability of allocation in discrimination to two normal populations is examined.

ACKNOWLEDGEMENTS

I am deeply grateful to Professor Sir David Cox for his encouragement and guidance throughout this work, for the invaluable insights he has provided on so many subjects and for his ability in clarifying ideas.

I am also thankful to Professor P. J. Brown for the discussions and assistance during my first year.

A warm thank goes to my colleagues from the Ph. D., in particular to David, Silvia and Alan, for the enjoyable discussions (sometimes in the Pub), for helpful suggestions and for their friendship.

I wish also to thank:

my mother, Marilena, for her love and encouragement;

Anna for her friendship;

my 'querido' Len, who has give me so much love and support and has helped me with the English;

CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brasil) and IME-USP (Departamento de Matemática Estatística, Universidade de São Paulo) for their financial support and Deborah who acted on my behalf in securing such support.

CONTENTS

Appendices

Tables and Figures

CHAPTER 1

INTRODUCTION: EMPIRICAL BAYES

Empirical Bayes methods are applied to problems in which some or all the parameters of the model have a prior distribution. The simplest situation is when the observed Y_i's are independently distributed according to

$$Y_i \sim f(y_i \mid \theta_i) \tag{1.1}$$

and the θ_i's themselves are independent unobserved random variables from a common population

$$\theta_i \sim g(\theta_i), \qquad (i=1,\ldots,k), \tag{1.2}$$

with unknown density function $g(.)$.

The essential difference between a pure Bayesian and an empirical Bayes approach lies in the treatment of the prior distribution. In the empirical Bayes approach the prior is unknown and must be estimated, whilst in the pure Bayesian approach the prior is known (to the analyst). The comparison of both methods will be illustrated as we proceed and further discussion of some aspects of this comparison can be found in Deeley and Lindley (1981).

The important feature of the empirical Bayes analysis is that through consideration and estimation of the prior, supplementary data available on the related parameters are employed to improve estimates of θ_i.

To estimate θ_i we apply Bayes rule to assumptions (1.1) and (1.2) and calculate the *posterior* distribution of θ_i

$$f(\theta_i \mid y_i) = f(y_i \mid \theta_i)\, g(\theta_i)/m(y_i) \tag{1.3}$$

where

$$m(y_i) = \int f(y_i \mid \theta_i)\, g(\theta_i)\, d\theta_i, \qquad (i=1,\ldots,k), \tag{1.4}$$

is the *mixture* distribution of y_i (also called the *marginal* distribution of y_i) and can be interpreted as the distribution from which the y_i's will actually occur.

The posterior distribution depends on the unknown prior $g(.)$ and the problem now is to estimate $g(.)$.

A *first approximation* empirical Bayes solution to this problem estimates the prior from the relation (1.4) and then pretends that, given the data y_i, the θ_i are (independently) distributed according to (1.3) with the unknown quantities replaced by their corresponding estimates. This procedure works well when k is large and the estimation of the prior is based on a large amount of data, hence the name *first approximation*. For small or moderate k, however, the errors introduced in the prior estimation may be considerable and will not be reflected in any of the conclusions. It will be our task to investigate the effect of the prior estimation on the models considered. From a sample-theory point of view, a *second order approximation* for the limits of empirical Bayes intervals will be constructed along the arguments of Cox(1975b).

With respect to the prior specification, empirical Bayes methods can be categorized as *parametric empirical Bayes* or as *nonparametric empirical Bayes*. In the former one assumes that θ_i is in some class of distributions indexed by unknown parameters, while in the latter, the only assumption about the prior is the independence of the θ_i's (no distributional form). In nonparametric empirical Bayes two methods are available; one that estimates the prior directly using (1.2) (c.f. Laird(1978)) and one that seeks a representation of the Bayes rule in terms of the mixture distribution (1.2) and uses the data to estimate this marginal distribution; for a reference see Robbins(1955). The methods usually require relatively large data sets.

We will be concerned with the parametric empirical Bayes approach, where the estimation of the prior is reduced to the estimation of the *hyperparameters* (the parameters indexing the prior distribution). Often the prior distribution is chosen to be from a *conjugate* class of distributions, which is categorized by the fact that the posterior distribution will belong to the same class as the prior. The computational simplicity of the conjugate class of distributions makes their use attractive.

Morris(1982) discussed Bayes and empirical Bayes analysis for the class of the natural exponential family with quadratic variance function (on the mean) and conjugate prior distributions; included are the normal with normal mean, Poisson with gamma mean, gamma with inverse-gamma mean and binomial with beta mean.

More realistic models usually involve nuisance parameters. Two situations can be distinguished:

(i) when the parameters of interest (and possibly the nuisance parameters) have a prior distribution;

(ii) when some of the nuisance parameters have a prior distribution but the parameters of interest do not;

Estimation of the parameters of interest is based; in situation (i) on their posterior distribution and in situation (ii) on their marginal distribution with the nuisance parameters eliminated by integration. At some stage, in both (i) and (ii) elimination of the nuisance parameters which do not have a prior distribution has to be considered.

Situations where empirical Bayes methods can be applied appear with moderate frequency in applied statistics. The most obvious situation is when the parameters θ_i arise from some common population. Examples include the two applications of chapter 3 :

(i) the analysis of several 2 ×2 tables, where the θ_i are the effect of smoking, on a logistic scale, in 14 parallel studies;

(ii) the study of the growth of the AIDS epidemic, where the θ_i measure the rate of growth in 18 European countries.

Another related situation is when one is not sure whether to decide on a higher level model with different parameters $(\theta_1,...,\theta_k)$ or on a lower level model with fewer parameters, for example the θ_i's could be related by a regression model. The essential element is that the parameters assigned a single prior distribution, must have similar physical interpretations and in particular be measured in the same units. An important example is when one has to estimate the variance from related treatments and

a choice must be made between the model with constant variance or with different variances, that is, between estimating the treatment variance using the within sample variances or the pooled estimate of these variance estimates. This particular problem is discussed in chapter 2, where it will be shown that the empirical Bayes model represents a compromise between these two.

Problems in classical statistics, such as random effects models and mixture models, can also be considered from an empirical Bayes view point. Although it is assumed that the factor levels θ_i are random, emphasis in the empirical Bayes approach is on the estimation of θ_i, whereas in the random effects model approach, the emphasis is on the estimation of the components of variance.

Empirical Bayes models can also be applied to model the overdispersion relative to some model. The emphasis differs from the problems just discussed in that one is interested in the estimation of the mean of the mixture distribution. The mixture model is a useful mechanism for overdispersion, allowing some 'random heterogeneity' in the mean parameter. Quasi-likelihood methods can be applied to estimate the means of an 'overdispersed' model relative to the exponential family. Firth (1987) investigates asymptotic relative efficiency of the Quasi-likelihood model.

The main interest in this thesis is the construction of parametric empirical Bayes intervals for the means. The approach is general but is set out in terms of its main application the linear model, with errors normally distributed. In chapter 2 the empirical Bayes model, when the variances change from cell to cell according to some distribution, is investigated. The normal linear model, with normal means, is analysed in chapter 3. Focus will be on the estimation of the prior parameters and on the elimination of the errors incurred by the first approximation empirical Bayes solution. The relation between these two issues will also be considered in chapter 2. In chapter 4 robustness of the conjugate class of distributions is investigated.

Although empirical Bayes methodology is much wider than is presented in this thesis, we have tried to present a thorough analysis of the parametric

empirical Bayes methods for the linear model.

A general reference for empirical Bayes methods is Maritz(1970), where emphasis is on non-parametric empirical Bayes. A more recent reference, containing many other references is Berger(1985, Chap.4).

Theoretical issues of the asymptotical behaviour of the empirical Bayes rule, as $k \to \infty$, will not be considered here. For a reference see Morris(1983a).

CHAPTER 2

VARIANCES CHANGING ACCORDING TO AN INVERSE GAMMA DISTRIBUTION

2.1. Formulation

It is common in many schemes of routine measurement to have quite a large number of repeated values on the same individual. For example, blood pressure measurements might be taken on consecutive visits of each of a number of patients.

We shall consider studies where it is reasonable to assume that the measurements are normally distributed, so that we have n_i independent observations from each of k separate populations $N(\mu_i,\sigma_i^2)$, i=1,...,k.

Although measurements may be more or less variable for certain subjects, the measurement process to obtain each observation is the same across subjects. Common methods of analysis require a choice between two assumptions, one of equal and one of unequal variance (for subjects). An empirical Bayes procedure is a compromise between these two extremes, in that it considers different variances coming independently from the same population of variances with unknown density function, $f(\sigma_i^2)$. The empirical Bayes analysis can be performed, at least numerically, for any choice of prior distribution $f(\sigma_i^2)$; however, we adopt here for its computational simplicity, the class of conjugate prior inverse gamma distributions,

$$f(\sigma_i^2) \propto (\sigma_i^2)^{-(\nu/2+1)}\exp[-(\nu\tau/2)/\sigma_i^2], \qquad \tau>0,\ \nu>0, \tag{2.1}$$

with mean $\nu\tau/(\nu-2)$ and variance $2(\nu\tau)^2/[(\nu-2)^2(\nu-4)]$ where the scale parameter τ and dispersion parameter ν are unknown. Note that as $\nu\to\infty$ the prior distribution (2.1) tends to the degenerate density function at τ. The conjugate class of distributions (2.1), whilst convenient, also encompasses a large variety of distributional shapes.

The main focus of interest in this thesis is the construction of confidence

intervals for individual means and contrasts of them, the means being regarded as fixed and unknown parameters.

Denote by $\bar{y}_i$ and S_i the sample mean and the sum of squares about the sample mean of the ith group of observations. Conditionally on the variances σ_i^2, $(\bar{y}_i, S_i)$, are independent sufficient statistics distributed as,

$$\bar{y}_i \sim N(\mu_i, \sigma_i^2/n_i), \tag{2.2}$$

$$S_i \sim \sigma_i^2 \chi^2(f_i), \tag{2.3}$$

i=1,...,k independently, where $f_i=n_i-1$ is the number of degrees of freedom associated with the sum of squares S_i.

Once the nuisance parameters σ_i^2, are eliminated, by integration, from the product of (2.1), (2.2) and (2.3) as

$$f(\bar{y}_i, S_i \mid \nu, \tau) = \int f(\bar{y}_i \mid \sigma_i^2) f(S_i \mid \sigma_i^2) f(\sigma_i^2 \mid \nu, \tau)\, d\sigma_i^2,$$

the conditional distribution of $\bar{y}_i$ given S_i, and the marginal distribution of S_i, become respectively,

$$\bar{y}_i \mid S_i \sim \mu_i + \sqrt{\frac{S_i+\nu\tau}{n_i(f_i+\nu)}}\, T_{f_i+\nu}, \tag{2.4}$$

$$f(S_i; \nu, \tau) = B\left(\frac{f_i}{2}, \frac{\nu}{2}\right)^{-1} \left(1+\frac{S_i}{\nu\tau}\right)^{-(f_i+\nu)/2} \left(\frac{S_i}{\nu\tau}\right)^{f_i/2-1} (\nu\tau)^{-1}, \tag{2.5}$$

(i=1,...,k) independently given ν and τ ($\nu>0$, $\tau>0$), where T is a random variable having the Student-*t* density function with $f_i+\nu$ degrees of freedom and B(. , .) is the Beta function.

The important empirical Bayes points of model (2.1) and the consequential answers given by (2.4) and (2.5) are given below.

(i) If $(\bar{y}_i, S_i)$ have the conditional (on σ_i^2) distributions (2.2) and (2.3) and σ_i^2 is random with density (2.1), then (2.4) and (2.5) are the densities according to

which ($\bar{y}_i$, S_i) will actually occur.

(ii) Given ν and τ, $\bar{y}_i$'s are independently distributed according to a Student-*t* density function with mean μ_i and variance

$$(S_i+\nu\tau)/[n_i(f_i+\nu -2)],$$

which is a weighted average of the sample variance S_i/f_i and the prior mean of the variances, τ ; their corresponding degrees of freedom being the weights. As mentioned before, it represents a compromise between, on the one hand, assuming completely different variances σ_i^2 (estimated by S_i/f_i), and on the other hand, assuming constant variance τ. In particularly, the extreme situations of $\nu\to\infty$ and $\nu=0$ lead respectively, to the normal theory linear model with constant variance τ and to k separate normal models with different variances σ_i^2 estimated by S_i/f_i, (i=1,...,k).

(iii) Expression (2.5) provides the means to estimate ν and τ. The estimate of τ will reflect the common behaviour of the variances between subjects while the estimate of ν will be a measure of the dispersion between the variances, large ν meaning small variance heterogeneity and small ν meaning large variance heterogeneity.

Inferences about the means are based on the conditional distributions of $\bar{y}_i$ given S_i , (2.4). The objective now is to eliminate the dependence on the unknown parameters ν and τ from the answers given by (2.4).

The empirical Bayes approach to this problem is to estimate the prior parameters ν and τ from the marginal distribution of the sample variances (2.5) and to base inferences about μ_i on (2.4) with the prior parameters replaced by their corresponding estimates. This procedure works well for large values of k, when the parameters are accurately estimated. For small k, however, a more careful investigation is needed since the results do not reflect the errors in estimating the prior parameters ν and τ.

In the next section we discuss the estimation of ν and τ comparing maximum likelihood estimates with the method of moments estimates by means of the asymptotic relative efficiency as $k\to\infty$, when $n_i=n$ is fixed. Both a logarithmic and a cube-root transformation of the data are considered, in order to increase the efficiency

of the method of moments estimates.

In section 2.3, empirical Bayes confidence intervals for the means are constructed and a combination of numerical work and asymptotic theory is used to analyse the error incurred for not allowing for the variability in the estimates of ν and τ. More specifically a second order approximation for the coverage probability of the empirical Bayes confidence interval is calculated and a correction for the significance level when the error in estimating ν and τ is not negligible is then constructed. In §2.3.4 the relation between the efficiency of different estimates for the prior parameters and the error incurred in the empirical Bayes confidence interval based on these estimates is investigated for the special case when the parameters are either location or scale. Generalization of the results to a regression model is considered in section 2.5. Section 2.6 discusses a graphical method for examining the agreement of the data with the inverse-gamma prior distribution (2.1).

2.2 Estimation of the prior parameters

2.2.1 *Introduction*

The estimation of ν and τ is based on the marginal distribution of the sample variances (2.5).

It should be mentioned that some information is being ignored, namely the information about ν and τ contained in the distribution of the sample means (2.4). A simultaneous numerical analysis involving the joint distributions of $(\bar{y}_i, S_i)$, $(i=1,\ldots,k)$ could have been performed, with the estimates expressed as integrals, but the information gained is likely to be minimal compared to the difficulty involved.

Maximum likelihood estimates and method of moments estimates with and without data transformation are considered. Particularly important will be the large sample behaviour of these estimates, as $k \to \infty$ with $f_i = f$ $(f=n-1)$ fixed.

2.2.2 *Maximum Likelihood Estimates*

Differentiating the log likelihood function,

$$\sum_{i=1}^{k}\left\{\log\left[\frac{\Gamma[(f_i+\nu)/2]}{\Gamma(\nu/2)}\right]-\frac{1}{2}(f_i+\nu)\log(\nu\tau+S_i)\right\}+\frac{1}{2}k\nu\log(\nu\tau), \tag{2.6}$$

with respect to ν and τ and then equating the derivatives to zero gives

$$\sum_{i=1}^{k}\left\{\psi[(f_i+\hat{\nu})/2]-\psi(\hat{\nu}/2)-\log\left(1+\frac{S_i}{\hat{\nu}\hat{\tau}}\right)+\frac{(S_i-f_i\hat{\tau})}{(S_i+\hat{\nu}\hat{\tau})}\right\}=0, \tag{2.7}$$

and

$$k/\hat{\tau}-\sum_{i=1}^{k}\left(\frac{f_i+\hat{\nu}}{S_i+\hat{\nu}\hat{\tau}}\right)=0, \tag{2.8}$$

respectively, where $\psi(.)$ is the digamma function defined by

$$\psi(\nu)=d\log\Gamma(\nu)/d\nu.$$

Maximum likelihood estimates of ν and τ are the iterative solutions of equations (2.7) and (2.8). Some comments about the solution of these equations are given below.

(i) Equation (2.8) estimates τ^{-1} as the mean of the empirical Bayes estimates of $(\sigma_i^2)^{-1}$.

(ii) An asymptotic approximation of $\psi(x)$ for large x is

$$\psi(x)=\ln(x)-1/(2x)-1/(12x^2),$$

(Abramovitz and Stegun, 1965, p.259) with accuracy to three decimal places for $x\geq 10$. For $x<10$ the approximation can be improved using the recurrence formula

$$\psi(x)=\psi(x+j)-\left\{\frac{1}{x+j-1}+\frac{1}{x+j-2}+.....+\frac{1}{x}\right\}.$$

It can be shown that for large ν

$$\psi\left(\frac{f+\nu}{2}\right)-\psi\left(\frac{\nu}{2}\right)=\frac{f}{\nu}-\frac{f(f-2)}{2\nu^2}+O(\nu^{-3}).$$

(iii) A sufficient condition for $\nu<\infty$ is that

$$\sum_{i=1}^{k}\left\{f_i-\frac{f_i^2}{2}-\frac{S_i^2}{2\tau^2}+\frac{f_iS_i}{\tau}\right\}<0,$$

where the expression on the left is the coefficient of the first order term in the Taylor series expansion of (2.7) for large ν. It is also possible to show that for equal sample variances $S_i/f_i=V$, V is a solution of (2.8) and $\nu=\infty$ a solution of (2.7). This is an intuitively comforting result, since there is then no evidence of variance heterogeneity.

(iv) Equations (2.7) and (2.8) were solved iteratively for ω $(=\nu^{-1})$ and τ, using the Nag10 subroutine , convergence to the solution being very rapid. As a starting point of the algorithm one could use the moments estimates (2.12) and (2.13), discussed in the next section.

The asymptotic covariance matrix of the maximum likelihood estimate of (ν,τ) is given by $I_{\nu\tau}^{-1}$, where

$$I_{\nu\tau}=-E\left\{\frac{\partial^2\log\prod f(S_i|\nu,\tau)}{\partial\nu\,\partial\tau}\right\}=\begin{bmatrix} i_{\nu\nu} & i_{\nu\tau} \\ i_{\nu\tau} & i_{\tau\tau}\end{bmatrix}.$$

is the *Fisher total information matrix* for ν and τ.

Since the asymptotic relative efficiency is invariant with respect to reparameterizations and the calculations are easier for λ $(=\nu\tau)$ and ν,we consider the Fisher information matrix for λ $(=\nu\tau)$ and ν, $I_{\nu\lambda}$, whose elements are

$$i_{\nu\nu}=\sum_{i=1}^{h}\frac{1}{4}\left\{\psi'[(f_i+\nu)]-\psi'[\nu/2]\right\},$$

$$i_{\nu\lambda}=-\sum_{i=1}^{k}\frac{f_i}{2\nu\tau(f_i+\nu)}\quad\text{and}$$

$$i_{\lambda\lambda}=\sum_{i=1}^{k}\frac{\nu f_i}{2(\nu\tau)^2(f_i+\nu+2)},$$

where $\psi'(\nu)=d\psi(\nu)/d\nu$ is the trigamma function. The parameters cannot usefully be orthogonalized, since the expressions of the orthogonal parameters would involve f_i.

The information matrix of the parameters of interest (ν,τ) can then be calculated from the above, using

$$\mathbf{I}_{\nu\tau}=\mathbf{J}^T \mathbf{I}_{\nu\lambda} \mathbf{J} \tag{2.9}$$

where **J**, the jacobian transformation matrix of (ν, λ) to (ν, τ) is,

$$\mathbf{J}=\begin{bmatrix} 1 & 0 \\ \tau & \nu \end{bmatrix}.$$

2.2.3 *Method of Moments Estimates*

In order to have more intuitively accessible estimates ν and τ, consider first the case of a balanced experiment in which f_i =f. The method of moments estimates are calculated by equating the first and second moments of the distribution of S_i, calculated in appendix A (expression (A.1)), to their respective sample moments, giving,

$$\tilde{\tau}=\frac{(\tilde{\nu}-2)\bar{S}}{\tilde{\nu}\,f}, \tag{2.10}$$

$$\tilde{\nu}=\frac{2(f+2)}{f}\left\{\sum_{i=1}^{k}\frac{(S_i-\bar{S})^2}{k\bar{S}^2}-\frac{2}{f}\right\}^{-1}+4, \tag{2.11}$$

where $\bar{S}=\Sigma S_i/k$.

These estimates have a reasonable interpretation even when the σ_i^2 are not inverse gamma distributed; $\tilde{\tau}$ is proportional to the pooled estimate of the common variance, and $\tilde{\nu}$ is an estimate of the variance heterogeneity, a function of the coefficient of variation estimate.

It is possible to shown that if

$$\Sigma(S_i-\bar{S})^2/k\bar{S}^2-2/f<0,$$

then $\tilde{\nu}<2-f$. Also if the left hand side of the condi tion is close to zero $\tilde{\nu}\to-\infty$ and $\tilde{\nu}\to\infty$ are possible solutions. Since a negative estimate of $\tilde{\nu}$ is meaninglessand the condition in this case seems to indicate that the variability between the sample variances is small we should estimate $\tilde{\nu}=\infty$.

The scale parameter τ is positive , thus a solution of (2.10) is not sensible for $\tilde{\nu}<2$. An *ad hoc* procedure is to replace $(\tilde{\nu}-2)$ by $\tilde{\nu}$ in (2.10).

It remains to investigate the situation in which the f_i are different. It can be shown that in this case the estimation equations are

$$\tilde{\tau}=\frac{(\tilde{\nu}-2)}{\tilde{\nu}}\frac{\Sigma S_i}{\Sigma f_i}, \tag{2.12}$$

$$\tilde{\nu}=2\left\{2\frac{\Sigma S_i^2}{\Sigma f_i(f_i+2)}-\frac{(\Sigma S_i)^2}{(\Sigma f_i)^2}\right\}/\left\{\frac{\Sigma S_i^2}{\Sigma f_i(f_i+2)}-\frac{(\Sigma S_i)^2}{(\Sigma f_i)^2}\right\}. \tag{2.13}$$

An application of this problem, considered by Hui & Berger(1983) is to the estimation of regression coefficients (typically slopes) in longitudinal studies. A mixed method was considered; the method of moments to estimate τ and the maximum likelihood method to estimate ν. They also suggested replacing $(\tilde{\nu}-2)$ by $\tilde{\nu}$ in the moments estimate equation of $\tilde{\tau}$. The equations are easier to solve than the maximun likelihood equations.

We proceed now to calculate the asymptotic covariance matrix for the moments estimates when the sample sizes are constant ($f_i=f$).

Writing equations (2.10) and (2.11) as $\mathbf{Z}_k(\theta)=0$, where $\theta'=(\nu,\tau)$, from the first order term of the Taylor series expansion of $\mathbf{Z}_k(\theta)$ at the true parameter value we calculate the asymptotic covariance matrix,

$$\mathrm{cov}(\theta)=\Lambda_k^{-1}\mathrm{cov}\{\mathbf{Z}_k(\theta)\}(\Lambda_k^{-1})^T \tag{2.14}$$

where $\Lambda_k=dE[\mathbf{Z}_k(\theta)]/d\theta$ is evaluated at the true parameter value.

The asymptotic covariance matrix, $\mathrm{cov}\{\mathbf{Z}_k(\theta)\}$, is the covariance matrix of the first two sample moments of S_i which are calculated in appendix A (A.1),

$$\mathrm{Cov}\{Z_k(\nu,\tau)\} = \frac{1}{k}\begin{bmatrix} \kappa_2 & \kappa_3 \\ \kappa_3 & \kappa_4+2\kappa_2^2 \end{bmatrix}. \qquad (2.15)$$

The matrix of derivatives of equations (2.10) and (2.11) has the elements

$$\partial Z_k^1/\partial\nu = -2f\tau/(\nu-2)^2,$$

$$\partial Z_k^1/\partial\tau = f\nu/(\nu-2),$$

$$\partial Z_k^2/\partial\nu = \frac{2f\tau^2\nu\{\nu(\nu-2)(\nu-4) - (f+\nu-2)(\nu^2+2\nu-[illegible])\}}{(\nu-2)^3(\nu-4)^2}$$

and

$$\partial Z_k^2/\partial\tau = \frac{4f\nu^2\tau\,(f+\nu-2)}{(\nu-2)^2(\nu-4)}.$$

Unfortunate ly, no elegant and simple formula is available for the asymptotic covariance matrix of the moments estimates $(\tilde{\nu}, \tilde{\tau})$.

2.2.4 *Asymptotic Relative Efficiency*

We now investigate the large sample behaviour of the method of moments estimates relative to the maximum likelihood estimates as $k\to\infty$, with $f_i = f$ fixed. In particular, we are interested in the behaviour of the asymptotic relative efficiency with respect to the amount of variance heterogeneity measured by ν and with respect to the sample size $(f+1)$. The asymptotic efficiency of the method of moments estimates $(\tilde{\nu}, \tilde{\tau})$ relative to the maximum likelihood estimates $(\hat{\nu}, \hat{\tau})$ is given by

$$\mathrm{are}\{(\tilde{\nu}, \tilde{\tau}) : (\hat{\nu}, \hat{\tau})\} = \{|\,\mathrm{acov}(\hat{\nu}, \hat{\tau})\,| / |\,\mathrm{acov}(\tilde{\nu}, \tilde{\tau})\,|\}^{1/2}, \qquad (2.16)$$

where $|\,\mathrm{acov}(\hat{\nu}, \hat{\tau})\,|$ and $|\,\mathrm{acov}(\tilde{\nu}, \tilde{\tau})\,|$ are the determinants of the asymptotic covariance matrices calculated in (2.9) and (2.14), respectively.

The individual asymptotic relative efficiencies of the estimates of ν and τ when the other parameter is unknown, are

$$\text{are}\{\tilde{\nu} : \hat{\nu}\} = \hat{i}^{\nu\nu}/\tilde{i}^{\nu\nu}$$

and (2.17)

$$\text{are}\{\tilde{\tau} : \hat{\tau}\} = \hat{i}^{\tau\tau}/\tilde{i}^{\tau\tau},$$

respectively, where $\hat{i}^{\nu\nu}$ and $\hat{i}^{\tau\tau}$ are the diagonal elements of the inverse Fisher information matrix $I_{\nu\tau}^{-1}$ (2.9), and $\tilde{i}^{\nu\nu}$ and $\tilde{i}^{\tau\tau}$ are the diagonal elements of the asymptotic covariance matrix of the method of moments estimates (2.14).

Expressions (2.16) and (2.17) are evaluated for some values of f, ν and τ=1, and displayed in Table 1. The last row and last column in each table correspond to the asymptotic relative efficiency when $\nu=\infty$ and $f=\infty$, respectively. Although the values were theoretically calculated by considering expansions of the covariance matrices as $\nu\to\infty$ and $f\to\infty$ we omit their expressions.

It is clear from examination of Table 1 that the loss of efficiency incurred in using the method of moments can be considerable, depending heavily on the amount of variance heterogeneity. For moderate and small values of ν ($\nu<12$) the method of moments has efficiency less than 70%. Full efficiency is attained as $\nu\to\infty$, i.e., for the normal-theory linear model.

Table 1. *Asymptotic efficiency of moments estimates relative to the maximum likelihood estimates*

are$\{(\tilde{\nu}, \tilde{\tau}) : (\hat{\nu}, \hat{\tau})\}$

ν=prior degrees of freedom	f= sample degrees of freedom			
	2	*10*	*20*	∞
10	0.51	0.42	0.38	0.31
12	0.69	0.58	0.54	0.45
14	0.78	0.69	0.64	0.55
16	0.84	0.76	0.71	0.61
18	0.88	0.80	0.76	0.66
20	0.90	0.84	0.90	0.81
30	0.96	0.93	0.90	0.81
∞	1.00	1.00	1.00	1.00

are$\{\tilde{\nu} : \hat{\nu}\}$

	2	*10*	*20*	∞
10	0.29	0.21	0.18	0.14
12	0.49	0.38	0.33	0.26
14	0.62	0.50	0.45	0.35
16	0.71	0.59	0.54	0.42
18	0.77	0.66	0.60	0.48
20	0.82	0.72	0.66	0.52
30	0.92	0.86	0.82	0.67
∞	1.00	1.00	1.00	1.00

are$\{\tilde{\tau} : \hat{\tau}\}$

	2	*10*	*20*	∞
10	0.32	0.28	0.27	0.26
12	0.56	0.52	0.51	0.51
14	0.70	0.67	0.67	0.68
16	0.79	0.78	0.77	0.78
18	0.84	0.83	0.83	0.84
20	0.88	0.87	0.87	0.88
30	0.95	0.95	0.95	0.96
∞	1.00	1.00	1.00	1.00

An interesting feature in Table 1 is that except for very large values of ν and f, it is true that

$$\text{are}\{ (\tilde{\nu}, \tilde{\tau}) : (\hat{\nu}, \hat{\tau})\} > \text{are}\{\tilde{\tau} : \hat{\tau}\} > \text{are}\{\tilde{\nu} : \hat{\nu}\}.$$

That is, the joint estimation of ν and τ by the method of moments is more efficient than the estimation of each parameter separately, compared to the respective maximum likelihood estimates. If we write the joint asymptotic relative efficiency as

$$\text{are}\{ (\tilde{\nu}, \tilde{\tau}) : (\hat{\nu}, \hat{\tau})\} = \text{are}\{\tilde{\nu} : \hat{\nu}\}^{1/2}\text{are}\{\tilde{\tau} : \hat{\tau}\}^{1/2}(1-\hat{\rho}^2_{\nu\tau})^{1/2}(1-\tilde{\rho}^2_{\nu\tau})^{-1/2},$$

where $\hat{\rho}_{\nu\tau}$ and $\tilde{\rho}_{\nu\tau}$ are the correlation coefficients of the maximum likelihood estimates and the moments estimates, respectively. It is clear that when $\hat{\rho}_{\nu\tau}^2 << \tilde{\rho}_{\nu\tau}^2$, the value of $\text{are}\{(\tilde{\nu}, \tilde{\tau}) : (\hat{\nu}, \hat{\tau})\}$ is large compared to both $\text{are}\{\tilde{\nu} : \hat{\nu}\}$ and $\text{are}\{\tilde{\tau} : \hat{\tau}\}$.

2.2.5 *Logarithmic Transformation*

Bartlett & Kendall(1946) show that a logarithmic transformation of the sample variances, $\log(S_i)$, is closer to normality than the distribution of S_i. A similar idea is attempted here, where we investigate the gain in efficiency of the method of moments after the logarithmic transformation is applied to the sample variances S_i.

Consider the random variable $W=\log(S_i)$. The cumulant generating function of W can be expressed as

$$K_W(\xi) = \log\{E[\exp(W\xi)]\} = \xi\log(\nu\tau) + \log\left(E\{[S_i/(\nu\tau)]^{\xi}\}\right),$$

where $E\{[S_i/(\nu\tau)]^{\xi}\} = B(\nu/2-\xi, f/2+\xi) / B(\nu/2, f/2)$.

The cumulants of W are then obtained from the derivatives of the cumulant function at $\xi=0$,

$$\frac{\partial^r K_W(\xi)}{\partial \xi^r} = [\log(\nu\tau)]^{\delta^1_r} + (-1)^r\psi^{(r-1)}(\nu/2) + \psi^{(r-1)}(f/2),$$

with δ^1_r is the dirac function at r=1, and $\psi^{(r)}(.)$ is the r^{th} derivative of the digamma

function $\psi(.)$. In particular the first four cumulants are

$$\begin{aligned}
\kappa_1 &= \log(\nu\tau) + \psi(f/2) - \psi(\nu/2), \\
\kappa_2 &= \psi^{(1)}(f/2) + \psi^{(1)}(\nu/2), \\
\kappa_3 &= \psi^{(2)}(f/2) - \psi^{(2)}(\nu/2), \\
\kappa_4 &= \psi^{(3)}(f/2) + \psi^{(3)}(\nu/2).
\end{aligned} \tag{2.18}$$

Again we consider balanced experiments ($f_i = f$) and equating the first and second moments to their respective sample moments we obtain

$$\psi(f/2) - \psi(\tilde{\nu}/2) + \log(\tilde{\nu}\tilde{\tau}) - \bar{S}_{\log} = 0, \tag{2.19}$$

$$\psi'(f/2) + \psi'(\tilde{\nu}/2) - \Sigma[\log(S_i) - \bar{S}_{\log}]^2/k = 0, \tag{2.20}$$

with $\bar{S}_{\log} = \Sigma\log(S_i)/k$, where $\tilde{\nu}$ and $\tilde{\tau}$ are the moments estimates of ν and τ from the distribution of $\log(S_i)$. Note that, although the same notation as used in §(2.2.3) is being used here to indicate moments estimates, the estimates are not the same.

Some comments are due.

(i) To solve the above equations we first find a solution $\tilde{\nu}$ from (2.20) and with this value substituted in (2.19) we calculate $\tilde{\tau}$. These equations are generally easier to solve than the maximum likelihood equations.

(ii) An asymptotic approximation of $\psi'(x)$ for large x is

$$\psi'(x) \cong 1/x + 1/(2x^2),$$

(Abramovitz and Stegun, 1965, p.260), with accuracy of up to three decimal places for $x>10$. For smaller values of x the approximation can be improved by using the above formula for $\psi'(x+j)$, where $(x+j)$ is large enough to give the desired accuracy, together with the recurrence formula

$$\psi'(x) = \psi'(x+j) + \frac{1}{x^2} + \frac{1}{(x+1)^2} + \ldots + \frac{1}{(x+j-1)^2}.$$

(iii) Asymptotic expansions of (2.19) and (2.20) for large values of ν and τ give:

$$\tilde{\tau} = \frac{1}{f}\prod_{i=1}^{k} S_i^{1/k} \exp\left\{\frac{1}{f} + \frac{1}{3f^2} - \frac{1}{\tilde{\nu}} - \frac{1}{3\tilde{\nu}^2}\right\},$$

$$2(1/f + 1/f^2 + 1/\tilde{\nu} + 1/\tilde{\nu}^2) - (1/k)\sum_{i=1}^{k} (\log(S_i) - \bar{S}_{\log})^2 = 0. \qquad (2.21)$$

The second equation is a quadratic form in $1/\nu$ and it is possible to show that $1/\nu > 0$ if

$$\Sigma[\log(S_i) - \bar{S}_{\log}]^2/k > 2/f + 2/f^2,$$

that is, in the case of heterogeneity of the sample variances.

(iv) The estimate $\tilde{\tau}$ given by the first equation in (2.21) is proportional to the geometric mean of the sample variances S_i.

(v) These approximations can be used as initial values for the solution of (2.19) and (2.20).

The development to calculate the asymptotic covariance matrix of $\tilde{\nu}$ and $\tilde{\tau}$ from (2.19) is essentially the same as in section 2.2.3. The asymptotic covariance matrix, $\mathrm{cov}\{\mathbf{Z}_k(\theta)\}$, is as given by expression (2.15) with the corresponding cumulants of the distribution of $\log(S_i)$ given by (2.18). The matrix of derivatives Λ_k is

$$\partial Z_k^1/\partial\nu = -\psi^{(2)}(\nu/2)/2 + \nu^{-1},$$

$$\partial Z_k^1/\partial\tau = 1/\tau,$$

$$\partial Z_k^2/\partial\nu = \psi^{(2)}(\nu/2)/2,$$

$$\partial Z_k^2/\partial\nu = 0.$$

The asymptotic relative efficiency of the method of moments estimates relative to the maximum likelihood estimates when considering a logarithmic transformation of the sample variances is calculated using expressions (2.16) and (2.17), and the values displayed in Table 2. Note that the asymptotic covariance matrix of the maximum likelihood estimates (2.9) is invariant with respect to data transformations.

Table 2. *Asymptotic efficiency of moments estimates relative to the maximum likelihood estimates , when a logarithm transformation is applied to the sample variances*

$\text{are}\{ (\tilde{\nu}, \tilde{\tau}) : (\hat{\nu}, \hat{\tau})\}$

ν=prior degrees of freedom	f= sample degrees of freedom				
	2	4	10	20	∞
2	0.78	0.94	0.92	0.87	0.81
4	0.57	0.87	0.98	0.97	0.90
10	0.39	0.69	0.94	1.00	0.96
16	0.35	0.62	0.90	0.98	0.98
20	0.33	0.60	0.88	0.97	0.98
30	0.31	0.56	0.84	0.95	0.99
∞	0.27	0.49	0.74	0.86	1.00

$\text{are}\{\tilde{\nu} : \hat{\nu}\}$

	2	4	10	20	∞
2	0.67	0.92	0.87	0.82	0.75
4	0.34	0.78	0.97	0.94	0.85
10	0.16	0.48	0.89	0.99	0.94
16	0.13	0.39	0.80	0.96	0.96
20	0.12	0.36	0.77	0.94	0.97
30	0.11	0.32	0.71	0.90	0.98
∞	0.08	0.25	0.57	0.74	1.00

$\text{are}\{\tilde{\tau} : \hat{\tau}\}$

	2	4	10	20	∞
2	0.69	0.93	0.93	0.88	0.83
4	0.50	0.85	0.99	0.98	0.95
10	0.37	0.72	0.96	1.00	0.99
16	0.34	0.68	0.94	0.99	1.00
20	0.33	0.67	0.93	0.99	1.00
30	0.31	0.65	0.92	0.98	1.00
∞	0.28	0.60	0.89	0.97	1.00

From Table 2 we see that the method of moments estimate based on the log transformation is particularly efficient for large sample sizes (large f). This is a generalization of the result of Bartlett & Kendall (1946), that the distribution of $\log(S_i)$ is closer to the normal for large f under the null hypothesis that the variances are the same, i. e., $H_0: \nu=\infty$. However, for very small sample sizes the method of moments, based on the log transformation, has very low efficiency. The marginal distribution for $f \leq 2$ is *J*-shaped and a log transformation produces a distribution with a long negative tail.

2.2.6 *Hilferty Transformation*

Finally we consider the method of moments estimates after a cube-root transformation of the observed variances, also called the Wilson-Hilferty transformation. This transformation is usually used to approximate the gamma distribution to the normal distribution (Johnson & Kotz, vol.1, 1970, p.176).

The moments estimates based on this transformation are the solutions of the following equations,

$$(\nu\tau)^{1/3} \frac{\Gamma(f/2+1/3)\Gamma(\nu/2-1/3)}{\Gamma(f/2)\,\Gamma(\nu/2)} - \bar{S}_H = 0,$$

$$\frac{\Gamma(\nu/2)\,\Gamma(\nu/2-2/3)}{\Gamma(\nu/2-1/3)^2} = \frac{\Sigma S_i^{2/3}}{k\,(\bar{S}_H)^2} \frac{\Gamma(f/2+1/3)^2}{\Gamma(f/2)\Gamma(f/2+2/3)} \qquad (2.22)$$

where $\bar{S}_H = \Sigma\, S_i^{1/3}/k$.

A solution for the these equations exists for $\nu > 4/3$.

Asymptotic expansions of (2.22) give

$$\left\{1-\frac{4}{9\tilde{\nu}}+\frac{124}{81\tilde{\nu}^2}\right\}=\frac{\Sigma S_i^{2/3}}{k\bar{S}_H^2}\left\{1+\frac{1}{9f}-\frac{7}{81f^2}\right\},$$

$$\tilde{\tau}^{1/3}=\bar{S}_H\left\{1-\frac{1}{9f}+\frac{4}{81f^2}\right\}\left\{1-\frac{4}{9\tilde{\nu}}-\frac{4}{27\tilde{\nu}^2}\right\},$$

where the first expression is a second degree equation in $1/\tilde{\nu}$ $(=\tilde{\omega})$ and the second calculates the estimate of τ as proportional to the cube of $\bar{S}_H$. Initial values to the solution of (2.22) can be found by solving the approximate equations for $\tilde{\omega}$ and $\tilde{\tau}$.

The calculations for the asymptotic covariance matrix are made as in sections 2.2.3 and 2.2.5. We omit the expressions of the moments of $S_i^{1/3}$ about the mean, since no simple and elegant form is available. The moments about the origin of $S_i^{1/3}$ can be calculated from the expressions in appendix A. The matrix of derivatives of equations (2.22), denoted by Z_n, has elements

$$\partial Z_n{}^1/\partial\nu = \partial\mu_1'/\partial\nu,$$

$$\partial Z_n{}^1/\partial\tau = (1/3)\mu_1'/\tau,$$

$$\partial Z_n{}^2/\partial\nu = \partial\mu_2'/\partial\nu - 2\mu_1'(\partial\mu_1'/\partial\nu) \qquad \text{and}$$

$$\partial Z_n{}^2/\partial\tau = (2/3)\mu_2'/\tau,$$

where

$$\partial\mu_1'/\partial\nu = (1/3)\mu_1'/\nu + (\mu_1'/2)[\psi(\nu/2-1/3) - \psi(\nu/2)],$$

$$\partial\mu_2'/\partial\nu = (2/3)\mu_2'/\nu + (\mu_2'/2)[\psi(\nu/2-2/3) - \psi(\nu/2)],$$

and μ_1', μ_2' are the first two moments about the origin of the distribution of $S_i^{1/3}$.

Table 3 gives the asymptotic efficiency of the method of moments estimates relative to the maximum likelihood estimates when the sample variances are transformed according to the Wilson-Hilferty transformation.

Table 3. *Asymptotic efficiency of moments estimates relative to the maximum likelihood estimates , when a Hilferty transformation is applied to the sample variances*

are$\{ (\tilde{\nu}, \tilde{\tau}) : (\hat{\nu}, \hat{\tau})\}$

ν=prior degrees of freedom	f= sample degrees of freedom					
	2	4	6	10	20	∞
4	0.80	0.73	0.73	0.61	0.55	0.47
6	0.93	0.94	0.91	0.87	0.80	0.70
8	0.91	0.97	0.95	0.90	0.90	0.79
10	0.88	0.96	0.98	0.98	0.94	0.84
20	0.78	0.90	0.95	0.96	1.00	0.93
30	0.75	0.86	0.92	0.97	1.00	0.94
∞	0.67	0.78	0.83	0.89	0.94	1.00

are$\{\tilde{\nu} : \hat{\nu}\}$

	2	4	6	10	20	∞
4	0.66	0.56	0.49	0.43	0.37	0.31
6	0.89	0.89	0.84	0.76	0.67	0.54
8	0.85	0.95	0.95	0.90	0.81	0.66
10	0.78	0.93	0.97	0.96	0.89	0.73
20	0.62	0.81	0.90	0.97	0.99	0.87
30	0.57	0.76	0.85	0.94	0.99	0.91
∞	0.45	0.60	0.69	0.79	0.88	1.00

are$\{\tilde{\tau} : \hat{\tau}\}$

	2	4	6	10	20	∞
4	0.79	0.71	0.65	0.58	0.53	0.46
6	0.90	0.95	0.94	0.91	0.87	0.81
8	0.87	0.96	0.98	0.98	0.95	0.91
10	0.83	0.95	0.98	0.99	0.98	0.95
20	0.75	0.90	0.95	0.99	1.00	0.99
30	0.72	0.87	0.93	0.98	1.00	1.00
∞	0.66	0.82	0.89	0.94	0.98	1.00

Comparison of Tables 2 and 3 indicates that in general the Wilson-Hilferty transformation is more efficient than the logarithmic transformation with the exception of the two first rows, i.e.small values of ν. In this case the marginal distribution has long tails and the log transformation is more efficient in approximating the distribution of S_i to normality. However, for small sample sizes (small f), the Wilson-Hilferty transformation makes the method of moments very efficient. This is the case when the marginal distribution is J-shaped and the log transformation produces a distribution with long tail.

2.2.7 *Conclusions*

Computationally, the simplest estimate is the method of moments estimate based on the sample variances $S_1,...,S_k$; however, it has rather low efficiency except for large ν when it approaches full efficiency. For situations in which there is overdispersion, $f>\nu$ and small ν it has rather low efficiency.

The method of moments applied to the log transformation of the sample variances have high efficiency for moderate and large values of f.

In general the maximum likelihood estimate is to be preferred since no substantial gain in simplicity is attained by the method of moments based on some transformation of the sample variances, and a loss of efficiency is almost always incurred.

2.3 Confidence intervals for the means

2.3.1 *Introduction*

The (1-2α) empirical Bayes confidence interval for μ_i, based on (2.4) is

$$\{\bar{y}_i \pm t_\alpha \left(\frac{S_i + \hat{\nu}\hat{\tau}}{n_i(\hat{\nu} + f_i)}\right)^{1/2}\}, \tag{2.23}$$

where t_α is the (1−α)-quantile of the Student-*t* density function with $\hat{\nu}+f_i$ degrees of freedom and $(\hat{\nu},\hat{\tau})$ are estimates of (ν,τ).

The coverage probability of (2.23) is not exactly (1-2α) since the variability in the estimates of ν and τ is not considered. Nevertheless, if $(\hat{\nu},\hat{\tau})$ are consistent estimates based on a large amount of data (large k), then the errors of estimation should be negligible and the required coverage property hold up to higher order in *k*. It remains then to investigate the effect of errors in the estimation of (ν,τ) on the confidence interval for small or moderate values of k. To do this, simulation studies are used and a modification of significance level is calculated. It will be shown that only in very special situations do the errors of estimation have a significant effect on the coverage probability of intervals (2.23).

2.3.2 *Simulation Studies*

Two simulation studies are considered; one with the moderately large value of k=20 (300 simulations), and the other with the quite small value of k=4 (400 simulations). The sample sizes are held constant, so that f_i=f.

In both experiments data are generated according to the structure given in equations (2.1), (2.2) and (2.3), using the appropriate subroutines of the Nag10 library. The maximum likelihood estimate of (ν,τ) is found by solving equations (2.7) and (2.8) iteratively, using the generated sample variances $(S_1,...,S_k)$. Confidence intervals (2.23) are then calculated. Table 4 displays the proportion of intervals in both

simulation studies which cover the true mean μ_1 for some values of ν and τ. The numbers in brackets are the respective coverage proportions of the exact confidence interval, that is, the confidence interval with the true parameter values (ν,τ).

This example serves to illustrate the general point that the errors in estimating (ν,τ) have very little effect on the coverage probability, with the exception of the rather extreme situation of very small k and f. Note that for f=4 in the second simulation study (k=4), (ν,τ) are estimated from S_1, S_2, S_3 and S_4 with 4 degrees of freedom each, a rather extreme example, and the change in the coverage probability is still very small.

Table 4. *Proportions of the simulated 95% empirical Bayes confidence interval and of the confidence interval using the true parameter values (in brackets) containing the true mean* μ_1

300 simulations with k=20

ν=prior degrees of freedom	*f*= sample degrees of freedom		
	2	4	6
2	0.95(*0.95*)	0.96(0.97)	0.95(0.94)
10	0.95(0.95)	0.96(0.96)	0.95(0.96)
20	0.92(0.94)	0.94(0.94)	0.96(0.96)

400 simulations with k=4

ν=prior degrees of freedom	*f*= sample degrees of freedom				
	2	4	6	10	20
2	0.88(0.93)	0.94(0.96)	0.94(0.95)	0.94(0.94)	0.95(0.95)
10	0.91(0.95)	0.93(0.95)	0.94(0.95)	0.94(0.95)	0.95(0.95)
20	0.90(0.95)	0.94(0.96)	0.93(0.94)	0.92(0.95)	0.95(0.96)

The simulations were performed in the most favourable situation, that of

estimating (ν,τ) by the maximum likelihood method. Nevertheless, changes in the confidence limits when other methods are used are likely to be minimal. We pursue this investigation in §2.3.4 in the case of location and scale parameters.

2.3.3 *Correction for Significance Level*

The idea here is to develop a second order approximation for the coverage probability of the empirical Bayes confidence interval. Then, from the relation between the expected and the approximate coverage probabilities, a correction for the significance level is calculated which enables the construction of an empirical Bayes confidence interval with coverage probability $(1-2\alpha)$ up to order $o(n^{-1})$.

The approach was developed in Cox(1975b) as an approximate method of constructing prediction intervals. Here the formulation is restricted to the particular problem being studied: i.e., second order approximation of empirical Bayes confidence intervals.

For convenience consider the case of a single parameter. The generalization to the multidimensional case is immediate and does not provide further insight. Suppose ν is known and consider the correction for the errors in estimating τ.

Write,

$$q_\alpha(\tau) = \bar{y}_k - t_\alpha \left(\frac{S_k + \nu\tau}{n(f+\nu)} \right)^{1/2},$$

as the lower limit of the $(1-2\alpha)$ confidence interval for μ_k such that for τ known

$$Pr[\, \mu_k \le q_\alpha(\tau) \mid S_k \,] = \alpha.$$

For τ unknown we form an estimate W from $S_1, \ldots, S_k$ with

$$E(W-\tau) = a(\tau)/k$$

and (2.24)

$$Var(W) = b(\tau)/k.$$

The first approximation empirical Bayes confidence interval (2.23) takes $q_\alpha(w)$ as the lower limit. The lower tail probability is not exactly the required α, the method requires a second order approximation to the probability,

$$Pr[\,\mu_k \le q_\alpha(W) \mid S_k\,] = E\{Pr[\,\mu_k \le q_\alpha(w) \mid S_k,\, W=w\,]\}. \qquad (2.25)$$

Conditionally on S_k, $\bar{y}_k$ is independent of W and (2.25) is just

$$G[\,q_\alpha(w) \mid S_k\,] = E\{Pr[\,\mu_k \le q_\alpha(w) \mid S_k\,]\}.$$

The first term of the Taylor series expansion of (2.25) at $W=\tau$ is

$$Pr\,[\mu_k \le q_\alpha(w) \mid S_k\,] = G\,[\,q_\alpha(\tau) \mid S_k\,] + \frac{a(\tau)}{k}\left\{\frac{d\,G[q_\alpha(w) \mid S_k]}{d\,w}\Big|_{w=\tau}\right\} +$$

$$\frac{b(\tau)}{2k}\left\{\frac{d^2G\,[q_\alpha(w) \mid S_k]}{d\,w^2}\Big|_{w=\tau}\right\}$$

$$= \alpha + c_\alpha(\tau)/k + o(n^{-1}). \qquad (2.26)$$

where $c_\alpha(\tau)$ gives the correction of the significance level.

Thus if the critical value of the statistic is found from the lower α'-quantile of its distribution, where $\alpha'=\alpha-c_\alpha(w)/k$, the empirical Bayes interval will have confidence limit $(1-2\alpha)$ up to order $o(n^{-1})$.

To calculate the correction factor $c_\alpha(w)$ we have to compute up to order $o(n^{-1})$ the asymptotic bias and the covariance matrix of W, conditionally on S_k. If maximum likelihood estimates are used, general formulae are available for the asymptotic properties in Cox & Hinkley (1974, Ch.9) which can be adapted to take account of the conditioning. However, the difficulties associated with the conditioning can be avoided, at the cost of a slight loss of precision, by basing the estimation of τ on $S_1,\ldots,S_{k-1}$.

While the method is, in principle, very general, problems in which the parameters to be estimated are neither location nor scale tend to be more complicated. This is the case for the parameter ν, in which the derivatives of G(.) involve the

derivatives of the critical value t_α with respect to $\hat{\nu}$.

It is illuminating though to consider the case of the errors in estimating τ. To avoid problems with conditioning we shall in fact consider k+1 groups of observations and estimate the correction for the significance level of the empirical Bayes confidence interval of μ_1.

Denote by g^t_d the Student-*t* density function with *d* degrees of freedom and G^t_d its respective cumulative distribution function.

Also take $W=\hat{\tau}$ the maximum likelihood estimate of τ, based on $S_2,...,S_{k+1}$. The cumulative function at $q_\alpha(\hat{\tau})$ is

$$G\,[q_\alpha(\hat{\tau}) \mid S_1] = G^{\,t}_{\nu+f}\left(-t_\alpha \left\{\frac{S_1+\nu\hat{\tau}}{S_1+\nu\tau}\right\}^{1/2}\right),$$

where t_α is the $1-\alpha$ upper quantile of the Student-*t* density function with $\nu+f$ degrees of freedom.

The first and second derivatives of G with respect to $\hat{\tau}$ are

$$\frac{d\ G[q_\alpha(\hat{\tau}) \mid S_1]}{d\ \hat{\tau}}\Big|_{\hat{\tau}=\tau} = \frac{-\nu\, t_\alpha\, g^{\,t}_{\nu+f}(-t_\alpha)}{2(S_1+\nu\tau)}\ ,$$

and (2.27)

$$\frac{d^{\,2}G[q_\alpha(\hat{\tau}) \mid S_1]}{d\,\hat{\tau}^2}\Big|_{\hat{\tau}=\tau} = \frac{\nu\ t_\alpha\, g^{\,t}_{\nu+f}(-t_\alpha)}{4(S_1+\nu\tau)^2}\left\{1+\frac{\nu\,(f+\nu+1)}{f+\nu}\ \frac{t_\alpha^2}{(f+\nu)+t_\alpha^2}\right\}.$$

The unconditional asymptotic bias and variance of the maximum likelihood estimate can be calculated from the expression given in Cox & Hinkley (1974, p.309),

$$E(\hat{\tau}-\tau) = \{2E[U.(\tau)U'.(\tau)] + E[U.''(\tau)]\}/2(i_{\tau\tau})^2,$$

$$\mathrm{Var}(\hat{\tau}) = (i_{\tau\tau})^{-1},$$

where $U.(\tau)$ is the derivative with respect to τ of the log likelihood function (2.6) and $i_{\tau\tau}$ is the total information for τ as defined by (2.9).

After some calculations the expressions for the bias and variance can be

shown to be

$$E(\hat{\tau}-\tau)=\frac{2\tau(f+\nu+2)(f+2)}{k\nu f(f+\nu+4)}\ ,$$

$$\mathrm{Var}(\hat{\tau})=\frac{2\tau^2(f+\nu+2)}{k\nu f}\ . \qquad (2.28)$$

The correction $c_\alpha(\hat{\tau})$ is calculated by substituting (2.27) and (2.28) in (2.26).

The 'corrected' significance level is $\alpha'=\alpha-c_\alpha(\hat{\tau})/k$. It then follows that the empirical Bayes interval for μ_1 with confidence level $(1-2\alpha)$ up to order $o(n^{-1})$ is calculated from (2.23) with the critical value of the statistic, $t_{\alpha'}$, calculated from the lower α'-quantile of the Student-*t* density function with $\nu+f$ degrees of freedom.

To compare results, the approximate (up to $o(n^{-1})$) confidence level of intervals (2.23), $1-2[\alpha+c_\alpha(\tau)/k]$, were calculated for the same situation as in the second simulation study, (with k=4) of Table 4. The values are displayed in Table 5.

Table 5. *Second order approximation of confidence levels for the 95% empirical Bayes interval allowing for the errors in estimating* τ *by maximumlikelihood.*

ν=prior degrees of freedom	*f*= sample degrees of freedom			
	2	4	6	10
2	0.95	0.95	0.95	0.95
4	0.94	0.95	0.95	0.95
6	0.93	0.94	0.95	0.95
10	0.92	0.94	0.95	0.95
20	0.91	0.94	0.94	0.95

Comparison of Tables 4 and 5 shows immediately that the main difference lies in the first column (f=2) for values of ν smaller than 10. This is no surprise since Table 5 corrects the confidence level only for the errors in estimating τ

while in Table 4 both errors are considered. Also, for small values of ν, the individual sample variance S_i/f_i is the dominant term in the empirical Bayes estimate of the standard deviation, thus the errors in estimating τ are less influential than the errors in estimating ν. On the other hand, for large values of ν, the errors in the estimation of τ seem to explain most of the changes in the confidence level and in this case the dominant term in the empirical Bayes standard deviation is the common variance τ.

The preceding discussion provides some justification for ignoring the errors in estimating ν and τ in most situations. Only in the rather extreme case of small values of ν, large overdispersion, and fewer than 5 replicate observations need a correction be considered. In these cases instead of calculating the correction for estimates of ν a conservative solution would be to consider the limit of the correction for τ when $\nu \to \infty$,

$$\lim_{\nu \to \infty} c_\alpha(\tau) = \frac{g^t_{\nu+f}(-t_\alpha)\, t^3_\alpha}{2kf}.$$

It should be mentioned that other approaches which allow for the variability of the estimates of the prior parameters are possible. The pure Bayesian approach to this problem is to assign a second stage prior distribution $\pi(\lambda)$ for the unknown parameter $\lambda=(\nu,\tau)$. The uncertainty about λ is then described in terms of this second stage prior distribution, which can be a known proper prior, but is more often chosen to be an improper prior distribution. Therefore, the dependence of the results on the unknown parameter λ is eliminated by taking expectations with respect to its (second stage) posterior distribution. The method is called hierarchical Bayes analysis and a general reference is Berger(1985, Chap.4), containing many other references.

To illustrate the method and to compare with empirical Bayes analysis, consider a noninformative prior on the scale parameter τ

$$\pi(\tau \mid \nu) = (1/\tau)d\tau.$$

From (2.1) and (2.3) it can be shown that the posterior distribution of σ_i^2, is inverse-gamma,

$$f(\sigma_i^2 \mid S_i, \nu, \tau) \propto \exp\left\{\frac{S_i + \nu\tau}{2\sigma_i^2}\right\}(\sigma_i^2)^{-(f_i+\nu)/2-1}, \qquad (i=1,\ldots,k) \text{ independently,}$$

and the Bayes estimate of σ_i^2 is the posterior mean

$$E[\sigma_i^2 \mid S_i, \nu, \tau) = (S_i+\nu\tau) / (f_i+\nu - 2).$$

Then the Bayes estimate of the σ_i^2 after elimination of τ is given by

$$E\{\sigma_i^2 \mid S, \nu\} = E_{f(\tau \mid S, \nu)}\{ E[\sigma_i^2 \mid S_i, \nu, \tau)\}$$

where $S=(S_1,\ldots,S_k)$ and $f(\tau \mid S, \nu)$ is the (second stage) posterior distribution of τ calculated from

$$f(\tau \mid S, \nu) = \frac{\prod_{i=1}^{k} f(S_i \mid \nu, \tau)\, \pi(\tau \mid \nu)}{\int \prod_{i=1}^{k} f(S_i \mid \nu, \tau)\, \pi(\tau \mid \nu)\, d\tau},$$

and $f(S_i \mid \nu, \tau)$ is the marginal distribution (2.5).

Calculation of $E\{\sigma_1^2 \mid S, \nu\}$ involves numerical integration in

$$\frac{\nu}{(\nu+f_1-2)} \frac{\int\left\{\prod_{i=2}^{k}\left(1+\frac{S_i}{\nu\tau}\right)^{-(f_i+\nu)/2}\right\}\left(1+\frac{S_1}{\nu\tau}\right)^{-(f_1+\nu)/2+1} \tau^{-\Sigma f_i/2+1}\, \pi(\tau \mid \nu)\, d\tau}{\int\prod_{i=1}^{k}\left\{1+\frac{S_i}{\nu\tau}\right\}^{-(f_i+\nu)/2} \tau^{-\Sigma f_i/2}\, \pi(\tau \mid \nu)\, d\tau}.$$

However, for large ν a rough approximation is

$$E\{\sigma_1^2 \mid S, \nu\} = \frac{\nu}{(\nu+f_1-2)} \frac{\Sigma S_i}{\Sigma (f_i-2)},$$

which is equivalent to the pooled estimate of τ, and illustrates the fact that in general, when k is large, there will be essentially no difference between the empirical Bayes and hierarchical Bayes with vague second prior distribution. Some comments about the comparison of the two approaches are:

(i) The hierarchical Bayes analysis automatically incorporates the errors of the prior estimation.

(ii) From a computational viewpoint, empirical Bayes theory requires the solution of likelihood equations such as (2.7) and (2.8), while the hierarchical Bayes approach requires numerical integration, which can be somewhat more difficult, especially if it is of high dimension. The computational problem can be minimized by considering asymptotic approximations to the posterior moments and marginal densities such as in Tierney and Kadane(1986).

(iii) Clearly the empirical Bayes approach appeals to frequentists while the hierarchical Bayes approach appeals to Bayesians, although from a full subjectivist approach the formal prior $d\tau/\tau$ is not acceptable.

2.3.4 *Effect of Different Estimates on the Confidence Level*

In this section we investigate the effect of different estimates on the empirical Bayes confidence interval. Two aspects of the investigation are considered. The first examines the effect on the confidence level as different methods in estimating the prior parameters are used. The second compares the expected length of 'corrected' confidence intervals for these different estimates.

Denote the change in the lower tail probability of empirical Bayes confidence interval (2.23) when τ is estimated by $\hat{\tau}$ by

$$\hat{\alpha}-\alpha=\hat{c}_{\alpha}(\tau)/k, \tag{2.29}$$

where $\hat{c}_{\alpha}(\tau)$ is calculated from (2.26).

Consider $\hat{\tau}$ and $\tilde{\tau}$, two estimates of τ. Using (2.29), the ratio of the changes in the significance level is

$$(\hat{\alpha}-\alpha)/(\tilde{\alpha}-\alpha)=\hat{c}_{\alpha}(\tau)/\tilde{c}_{\alpha}(\tau).$$

Particularly, in the simplest situation, of unbiased estimates,we have that

$$(\hat{\alpha}-\alpha)/(\tilde{\alpha}-\alpha)=\text{are}(\tilde{\tau}:\hat{\tau}),$$

that is, the ratio of the changes is simply the asymptotic relative efficiency between the two estimates. In general, when the effect of the estimation of more than one parameter is being considered or when they are biased, the correction involves the derivatives of the distribution function and the comparison is not so straightforward.

As an illustration we compare the effect in estimating τ on the confidence interval when using the maximum likelihood estimate $\hat{\tau}$ and the method of moments estimate $\tilde{\tau}$. Since the bias of $\hat{\tau}$ is positive and $\tilde{\tau}$ is unbiased, the ratio of the changes in the significance level is

$$(\hat{\alpha}-\alpha)/(\tilde{\alpha}-\alpha)>\text{are}(\tilde{\tau}:\hat{\tau}).$$

This inequality provides the means of calculating the effect of estimating τ by the method of moments. For example, when $\nu=10$ and $f=2$, the change in the significance level when we allow for the errors in estimating τ by $\hat{\tau}$ is $(\hat{\alpha}-\alpha)=0.015$, as calculated in Table 5. From Table 1 the $\text{are}(\tilde{\tau}:\hat{\tau})=0.32$, hence $(\tilde{\alpha}-\alpha)<0.047$ and the confidence level when τ is estimated by $\tilde{\tau}$ is greater than 90%. Therefore, unless efficiency is very low the errors in estimating τ by the moments estimates are small.

We now study the comparison of two estimates after allowing for the errors in estimating the prior parameters. Since the confidence level of the second order interval is $1-2\alpha$ up to order $o(n^{-1})$, independent of the estimate which is being used, comparison is based on the differences of the expected length of those intervals. The formulation is in terms of the estimation of τ but the results can be applied directly to any location or scale parameter.

Suppose W is an estimate of τ satisfying (2.24). Then the lower limit of the second order empirical Bayes interval is

$$q_{\alpha^*}(w) = \bar{y}_1 - t_{\alpha^*}\, h(w),$$

where $h(w)$ is a known function of w and $\alpha^*=\alpha-c_\alpha(w)/k$.

Since

$$E[\, h(w)] = h(\tau) + h'(\tau)a(\tau)/k + o(k^{-1}), \tag{2.30}$$

and from the fact that

$$c_{\alpha}(w)/k \cong \int_{t_{\alpha}}^{t_{\alpha^*}} d\, G^{t}_{\nu+f},$$

the lower critical value of the statistic near the true parameter value τ is

$$t_{\alpha^*} = t_{\alpha} - \frac{c_{\alpha}(\tau)}{k\, g^{t}_{\nu+f}(-t_{\alpha})}.$$

It then follows that

$$E[q_{\alpha^*}(W)] = q_{\alpha}(\tau) + \frac{b(\tau)}{2k\, g^{t}_{\nu+f}(-t_{\alpha})} \left\{ \frac{d^2 G}{d\, w^2} \right\}_{w=\tau} + o\,(k^{-1}).$$

It is interesting to note that the expected length depends only on the asymptotic efficiency of the considered estimates and not on bias. Therefore, as to be expected on general grounds, estimates with larger efficiency produce second order intervals with smaller expected length.

2.3.5 *Conclusions and remarks*

The discussion provides some justification for ignoring the errors in estimating the prior parameters, in some important situations. Exceptions were for small sample sizes and for those estimates of ν and τ having low efficiency.

One question that would be of interest is the comparison of the empirical Bayes intervals (2.23) with the standard Student-*t* confidence under the assumption of constant variance,

$$y_i \pm t_{\alpha,\Sigma fi}\, (\bar{S}/k)^{1/2}, \qquad (2.31)$$

where $\bar{S}$ is the usual 'pooled' estimate of the common variance. However it revealed to be a difficult problem.

As mentioned before, a practical advantage of the empirical Bayes interval is its flexibility since assumptions (2.1) encompasses a wide range of situations

with different variability in the variances.

2.4 Contrast of Means

If we are interested in confidence intervals for contrasts of means, expression (2.4) leads to the result that, conditionally on the sample variances, the random variable $\Sigma c_i\bar{y}_i$ has the form

$$\sum_{i=1}^{k} c_i\bar{y}_i \sim \sum_{i=1}^{k} c_i\mu_i + \sum_{i=1}^{k} c_i \left\{\frac{S_i + \nu\tau}{n_i(\nu+f_i)}\right\}^{1/2} T_{\nu+f_i}. \qquad (2.32)$$

where T are independent random variables each having the Student-*t* density function with $\nu+f_i$ degrees of freedom, and $n_i=f_i+1$ are the sample sizes.

The distribution of (2.32) is a generalization of the Behrens-Fisher distribution for the difference of two means. A suggestion by Cox(1975a) is approximating the distribution of (2.32) by a Student-*t* with scale parameter θ and degrees of freedom ζ, by equating their corresponding second and fourth cumulants. The second and fourth cumulants of (2.32) are,

$$\kappa_2 = \sum_{i=1}^{k} c_i^2 \frac{(S_i+\nu\tau)}{n_i(f_i+\nu)} \frac{(f_i+\nu)}{(f_i+\nu-2)},$$

$$\kappa_4 = \sum_{i=1}^{k} c_i^4 \left\{\frac{S_i+\nu\tau}{n_i(f_i+\nu)}\right\}^2 \frac{(f_i+\nu)^2}{(f_i+\nu-2)^2(f_i+\nu-4)}.$$

Equating κ_2 and κ_4 to the second and fourth moments of a Student-*t* variable with scale parameter θ and degrees of freedom ζ, gives

$$\zeta^2 = \kappa_2(\kappa_2^2+2\kappa_4)/(\kappa_2^2+4\kappa_4),$$

$$\theta = \kappa_2^2/\kappa_4 + 4. \qquad (2.33)$$

Note that the method requires that $(f_i+\nu-4)>0$.

Thus the (1-2α) empirical Bayes confidence interval for $\Sigma c_i\mu_i$ is

$$\{\Sigma c_i\bar{y}_i \pm \theta^* t_{\zeta^*,\alpha}\}, \qquad (2.34)$$

with θ^* and ζ^* the estimated values of the solutions of (2.33).

A systematic study of the effects of the estimation of (ν,τ) on the confidence level of the above interval (2.34) has not been attempted here, but it would be surprising if the results differ considerably from the results in §2.3.3, and this is confirmed by the simulation experiment below.

The data for this are generated as in the simulation study of §2.3.2, with k=4 and a constant number of replicate observations f+1= 5. Maximum likelihood estimates of ν and τ are calculated from the marginal distribution of $S_1,...,S_4$ and substituted into the expressions for κ_2 and κ_4. Estimates of θ and ζ are then calculated from expressions (2.33) as well as the empirical Bayes 95% confidence interval (2.34) for the difference of the first two means $\mu_1-\mu_2$. We also calculate the exact 95% confidence interval based on the Behrens-Fisher distribution, i.e.,

$$\bar{y}_2 - \bar{y}_1 \pm u_\alpha \left\{\frac{S_1+S_2}{f(f+1)}\right\}^{1/2},$$

where u_α is the (1-2α)-quantile of the Behrens-Fisher distribution, tabulated in Fisher & Yates(1957), (with $f_1=f_2=f$).

The percentage of intervals containing $(\mu_1 - \mu_2)$ in 300 simulations is given in Table 6. The second row refers to the coverage properties of interval (2.34) with the true parameter values.

Table 6. *Comparison of the coverage proportion of 95% confidence intervals for the contrast* ($\mu_1 - \mu_2$) *in 300 simulations*

	f=4		
	ν=2	ν=10	ν=20
Emp. Bayes C.I. (2.32)	0.94	0.94	0.91
Emp. Bayes C.I. with true parameter values	0.95	0.95	0.94
Behrens-Fisher	0.95	0.97	0.95

Again the effect of empirical Bayes estimation on the confidence level when the number of replicate is small (f+1=5) is not very large if we consider that in this study the number of groups is extremely small, k=4. The effect seems to be larger when $\nu \to \infty$, which in turn means that we are losing efficiency if we do not consider common variance in the normal-theory set up.

2.5 Empirical Bayes estimates of regression parameters

In many applications the means $E(Y_i)=\mu_i$ often express dependence on explanatory variables $x_{i0},....,x_{ip}$ whose values are known. A widely used form of dependence is the generalized linear model

$$g(\mu_i) = \sum_{r=0}^{p} x_{ir}\beta_r , \qquad i=1,...,k, \tag{2.35}$$

where $\beta = (\beta_0,...,\beta_p)$ are the regression parameters to be estimated from the data $(y_1,...,y_k)$ and g(.) a specified link function.

The generalized linear predictor together with a distribution for the independent variable Y_i from the linear exponential family, characterizes the class of generalized linear model introduced by Nelder and Wedderburn(1972).

Many standard models fall neatly into the generalized linear framework, as for example, normal linear regression, probit analysis for proportions, logistic

regression and the log-linear model for Poisson counts. Restriction to generalized linear models is not important from a theoretical point of view since a great deal of the theory holds for non-linear regression. However, the generalized linear model has the advantage of easy interpretation whilst retaining flexibility and computational simplicity.

The most familiar model satisfying (2.35) is linear regression where the errors are normally distributed with constant variance τ and the link function is the identity function so that

$$\mu_i = \sum_{r=0}^{p} x_{ir}\beta_r, \quad i=1,...,k. \tag{2.36}$$

Here we assume that there are n_i replicates from the response variable Y_i, distributed according to

$$N(\mu_i, \sigma_i^2), \quad i=1,...,k,$$

and that *p* explanatory variables are measured on each individual. Interest focuses on the estimation of the regression parameters β when the σ_i^2's are unknown. The discussion is set up in terms of the linear form of dependence (2.36) since the generalization to the results for (2.35) is immediate and no further insight is provided.

The least squares estimate of β with estimated sample variances as weights, is the solution to the equation

$$\mathbf{X}^T\mathbf{W}_S^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}_S) = \mathbf{0}, \tag{2.37}$$

where $\mathbf{X}=(x_{ir})$, (i=1,...,k; r=0,...,p) is the design matrix and $\mathbf{W}_S=diag\{S_i/[n_i(n_i-1)]\}$ is the estimated weight matrix.

The main motivation for the use of an empirical Bayes approach to estimate the regression parameters is to produce more efficient estimates. It also avoids the problem of a zero weight in equation (2.37) since for small values of n_i, it might happen that one of the sample variances is nearly zero or, with rounded data, exactly zero. Empirical Bayes estimates of the variances, although not solving completely the problem, reduce the chance of having zero weights.

The empirical Bayes weighted least squares estimate of β, with the assumption of the inverse-Gamma distribution (2.1) for the variances, is the solution of

$$\mathbf{X}^T\mathbf{W}_G^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}_G) = \mathbf{0}, \tag{2.38}$$

where $\mathbf{W}_G$=diag{$[S_i+\nu\tau]/[n_i(n_i+\nu-1)]$} is the matrix of the empirical Bayes estimates of the variances.

Calculation of the asymptotic covariance matrices for both estimates $\hat{\beta}_S$ and $\hat{\beta}_G$, as $k\to\infty$, and with fixed n_i, is considered when the variances follow the distribution assumption (2.1).

Write $\mathbf{Z}_S(\hat{\beta}_S)$ for the left-hand side of equation (2.37). The asymptotic covariance matrix of $\hat{\beta}_S$ is calculated from expression (2.14) in §2.2.3.

Conditional on $S=S_1,...,S_k$, we have that

$$d\mathrm{E}[\, Z_S(\hat{\beta}_S) \mid S] \,/d\hat{\beta}_S = -X^T W_S^{-1} X$$

and

$$\mathrm{Cov}[\, Z_S(\hat{\beta}_S) \mid S] = -X^T W_S^{-1} \mathrm{Var}(\bar{Y}|S_1,...,S_k) W_S^{-1} X$$

$$= X^T \,\mathrm{diag}\left\{ \frac{n_i(n_i-1)^2}{(n_i+\nu-3)} \left[\frac{S_i+\nu\tau}{S_i^2} \right] \right\} X.$$

Substituting both these expressions in (2.14) and calculating the expectation with respect to the marginal distribution of the sample variances (2.5), it follows that the asymptotic covariance matrix of $\hat{\beta}_S$, $\mathrm{Cov}(\hat{\beta}_S)$, is

$$\tau\left\{X^T \mathrm{diag}\left[\frac{n_i(n_i-1)}{(n_i-3)}\right] X\right\}^{-1} \left\{X^T \mathrm{diag}\left[\frac{n_i(n_i-1)^2}{(n_i-3)(n_i-5)}\right] X\right\} \left\{X^T \mathrm{diag}\left[\frac{n_i(n_i-1)}{(n_i-3)}\right] X\right\}^{-1}. \tag{2.39}$$

Repeating this process but with

$$d\mathrm{E}[\, Z_G(\hat{\beta}_G) \mid S] \,/d\hat{\beta}_G = -X^T W_G^{-1} X,$$

and

$$\mathrm{Cov}[Z_G(\hat{\beta}_G) \mid S] = X^T \mathrm{diag}\left\{\frac{n_i(n_i+\nu-1)^2}{(n_i+\nu-3)(S_i+\nu\tau)}\right\} X,$$

leads to the asymptotic covariance matrix of the empirical Bayes weighted least squares estimate $\hat{\beta}_G$,

$$= \tau\{X^T \mathrm{diag}(n_i) X\}^{-1} \left\{X^T \mathrm{diag}\left[\frac{n_i(n_i+\nu-1)}{(n_i+\nu-3)}\right] X\right\} \{X^T \mathrm{diag}(n_i) X\}^{-1}. \tag{2.40}$$

Then the asymptotic efficiency of $\hat{\beta}_S$ relative to $\hat{\beta}_G$ is given by

$$\mathrm{are}(\hat{\beta}_S: \hat{\beta}_G) = \{|\mathrm{Cov}(\hat{\beta}_G)| / |\mathrm{Cov}(\hat{\beta}_G)|\}^{1/p}$$

$$= \left\{\prod_{i=1}^{k} \frac{(n_i+\nu-3)(n_i-5)}{(n_i+\nu-1)(n_i-3)}\right\}^{1/p}. \tag{2.41}$$

The usual weighted least squares estimate $\hat{\beta}_S$ is as efficient as $\hat{\beta}_G$ when $n_i \to \infty$, that is, when the empirical Bayes estimates of the variances coincide with the S_i. Otherwise $\hat{\beta}_G$ is more efficient than $\hat{\beta}_S$, especially for large ν when the model tends to the linear normal regression model. As $\nu \to \infty$ the asymptotic relative efficiency is

$$\{\Pi(n_i-5)/(n_i-3)\}^{1/p},$$

which can be quite small for small values of n_i. This is by no means a surprise, since, because we have assumed that the variances come from (2.1), $\hat{\beta}_G$ is expected to be more efficient than $\hat{\beta}_S$.

The maximum likelihood estimate, $\hat{\beta}$, based on the full empirical Bayes assumption (2.1) and (2.3) provide the means of assessing the efficiency of the least squares empirical Bayes estimate, $\hat{\beta}_G$. To calculate the asymptotic covariance matrix of $\hat{\beta}$ we follow Cox and Hinkley(1968).

Write the log likelihood of β based on (2.4) as

$$l = \Sigma\, h_i(\varepsilon_i;\beta)$$

where

$$h_i(\varepsilon_i;\beta) = -[(n_i+\nu)/2]\ \log\{(n_i+\nu-1)+\varepsilon_i^2/\zeta_i\},$$

with $\varepsilon_i=(\bar{y}_i-\mu_i)$ and $\zeta_i=(S_i+\nu\tau)/[n_i(n_i+\nu-1)]$.

We then have that

$$\frac{\partial^2 l}{\partial\beta_r\partial\beta_s} = \sum_{i=1}^{k} h_i''(\varepsilon_i;\beta)\frac{\partial\mu_i}{\partial\beta_r}\frac{\partial\mu_i}{\partial\beta_s} + \sum_{i=1}^{k} h_i'(\varepsilon_i;\beta)\frac{\partial^2\mu_i}{\partial\beta_r\partial\beta_s}.$$

where $h_i'(.)$ and $h_i''(.)$ are the first and second derivatives of $h_i(.)$.

Since

$$E\{h_i'(\varepsilon_i;\beta)\} = 0,$$

$$E\{E[h_i''(\varepsilon_i;\beta)|S_i]\} = n_i(n_i+\nu)/[\tau(n_i+\nu-2)],$$

it follows that the asymptotic covariance matrix of the maximum likelihood estimate $\hat{\beta}$ is

$$Cov(\hat{\beta}) = \tau\left\{X\ diag\left(\frac{n_i(n_i+\nu)}{n_i+\nu+2}\right)X^T\right\}^{-1}.$$

The asymptotic efficiency of $\hat{\beta}_G$, relative to $\hat{\beta}$ is then

$$are(\hat{\beta}_G:\hat{\beta}) = \prod_{i=1}^{k}\left\{\frac{(n_i+\nu-3)(n_i+\nu+2)}{(n_i+\nu-1)(n_i+\nu)}\right\}^{1/p}.$$

The loss of efficiency of the empirical Bayes weighted least squares estimate, $\hat{\beta}_G$, is small for large n_i or large ν (small overdispersion), as expected; full efficiency is attained when $\nu=\infty$. For example with small $n_i=3$, and reasonable amount of overdispersion, $\nu=2$, the efficiency of $\hat{\beta}_G$ relative to $\hat{\beta}$ is $(0.70)^{k/p}$, which can be very small if k is large compared to p.

Palta & Cook(1987) considered Monte Carlo studies to compare estimates of regression coefficients in longitudinal studies with weights σ_i^2 estimated by the unpooled sample variances, pooled sample variance and empirical Bayes estimate. The results showed that the empirical Bayes method had the advantage of good performance in all situations (balanced and unbalanced designs).

2.6 Graphical method to test adequacy of the prior

2.6.1 *Test for overdispersion*

The overdispersion with respect to the normal-theory model is measured by ν, where small values of ν mean large overdispersion, and large values mean variance homogeneity. It is of interest here to investigate the assumption of the inverse gamma prior distribution for the variances σ_i^2.

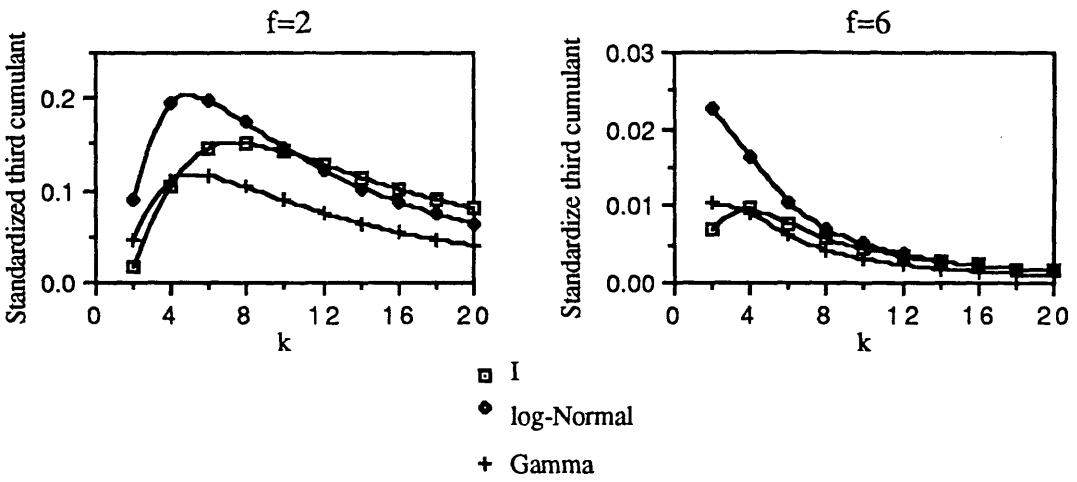
The first step is to check for overdispersion, i.e. as to whether there is any need to consider a more complicated model. This can be done informally by plotting the ordered sample variances S_i against the expected order statistics of the chi-square distribution with f degrees of freedom (Pearson & Hartley, 1970, Table20). More formal procedures are available. Cox(1983) gives a general version of the test for overdispersion. Cox & Solomon (1986) show that the locally most powerful test for $H_0: 1/\nu=0$ is based on the marginal distribution of the dispersion index,

$$I = \frac{f}{2}\left\{\frac{(\Sigma S_i^2)/k}{(\Sigma S_i/k)^2} - 1\right\}, \tag{2.42}$$

which for large k, is asymptotically normal.

A better approximation for the distribution of I, when k is very small is given by the gamma or log-Normal distributions correcting for skewness. The plots in figure 1 compare the standardized third cumulants of I with those of a gamma and a log-normal variable, each with the same coefficient of variation, as k varies from 0 to 20, for f=2 and f=6. It is clear that, for small k ($k<8$), the standardized third cumulant of I is better approximated by the standardized cumulant of the gamma, while as k increases (particularly for small f), it is closer to the standardized cumulant of a log-Normal variable. For larger values of f the normal distribution is a good approximation. The first and second moments of I are calculated in appendix C.

Figure 1. *Comparison of the standardized third cumulants of the distribution of the dispersion index (2.42) with the third cumulants of the gamma and the log-normal with the same coefficient of variation for values of k*



2.6.2 *Probability plot of the inverse-gamma distribution*

If overdispersion is detected it remains to investigate the assumption of inverse gamma prior assumption (2.1). An informal analysis based on a probability plot is suggested here.

Under assumption (2.1) the variable

$$\{S_i/(\nu\tau)\}/\{1+(S_i/\nu\tau)\},$$

follows a Beta distribution with parameters f/2 and ν/2. The cumulative distribution function of S_i is the incomplete Beta function,

$$F(S_i) = Pr(S_i \leq s) = I_x[(f/2),(\nu/2)],$$

where x=[s/(ντ+s)],which is tabulated in Pearson (1934).

Then, assuming continuity of $F(S_i)$, it follows that

$$\Phi^{-1}\{ I_x[(f/2),(\nu/2)]\}$$

is normally distributed. This suggest plotting

$$U_{(i)} = \Phi^{-1}\left(I_{x_{(i)}}\left(\frac{f}{2}, \frac{\nu}{2}\right)\right), \qquad (2.43)$$

against the expected order statistics from the standard normal distribution, where $x_{(i)}=S_{(i)}/(S_{(i)}+\nu\tau)$ and $S_{(i)}$ are the ordered sample sum of squares. So, if the variances vary according to an inverse gamma prior distribution, then a straight line plot is expected, assuming that the model is Normal.

A strong departure from the inverse gamma prior assumption occurs when the variances come from a two point prior distribution, i.e

$$\sigma_i^2 = \begin{cases} \tau_1, \text{ with probability } p \\ \tau_2, \quad " \quad " \quad (1-p). \end{cases} \tag{2.44}$$

To investigate the performance of the graphical method just described we consider two simulation studies where for some value of ν and f (τ=1):

(i) the variances σ_i^2 are simulated in the first one according to the inverse-gamma prior and in the second one according to the two point prior (2.44) (the values of τ_1, τ_2 and p are found by equating the first two moments of both prior distributions given ν and τ);

(ii) given the variances σ_i^2, generated according to (i), S_i are generated as $\sigma_i^2\chi^2(f)$.

For both simulations then the ordered statistics $U_{(i)}$ (2.43) were plotted against the normal order staistics.

In Figure 2, plots (a), (b) and (c) compare the inverse gamma and the two point prior distributions for different values of the overdispersion parameter ν and sample size f. The simulations show that, at least for large overdispersion (small ν) the probability plots are more nearly a straight line when the prior distribution is inverse-gamma.

Clearly as ν increases and the model tends to the normal theory model with constant variance, both plots will get closer to a straight line and it will become more difficult to distinguish between different prior assumptions. Also, for small values of *k*, (the simulation are for k=40) the visual impact of outliers is stronger and

distinction between different shapes more difficult, but this is a handicap of all graphical procedures.

The question of making precise the statement of 'large overdispersion' is now investigated.

Denote the overdispersion parameter by $\omega=1/\nu$. Then the asymptotic variance of $\hat{\omega}_0$, the maximum likelihood estimate of ω under the assumption of no overdispersion, H_0: $\omega=0$, is

$$\mathrm{Var}(\hat{\omega}_0) = 2/\{kf(f+2)\}.$$

If ω_t is the true parameter, then for $\omega_t > 3\times$st.err $(\hat{\omega}_0)$ the chance of detecting overdispersion is large. This means that for k=40, $\nu < k^{1/2}f/6$ is approximately $\nu<f$. This is satisfied by plot (b), where the distinction between the two plots is more evident and the plot from the two point prior distribution reveals two very distinctive variance populations.

A more challenging situation is when the overdispersion has a reasonable but not an overwhelming chance of being detected (Cox, 1983). The variance of S_i/f under the hypothesis of no overdispersion is $2\tau^2/f$. Thus if $\nu \sim k^{1/2}f$ the variance

$$\mathrm{Var}(S_i/f) \cong 2\tau^2\{1+ (f-2)/\nu\}/f$$

is increased by $O(1/(fk^{1/2}))$ and is at the border of detectability.

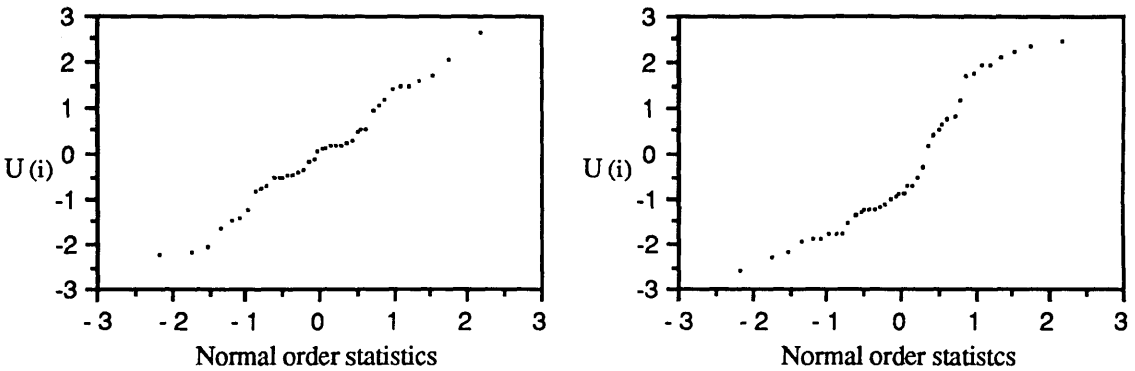
In this situation the probability plot will not be very efficient in discriminating between different prior distributions; it will look like a straight line in most situations.

Figure 2. *Probability plots of the normalized function of the sample variances, $U_{(i)}$ against the expected normal order statistics for σ_i^2 either coming from the inverse-gamma prior or the two point prior*

(a) f=2

Inverse gamma with ν=5.0 and τ=1.0.

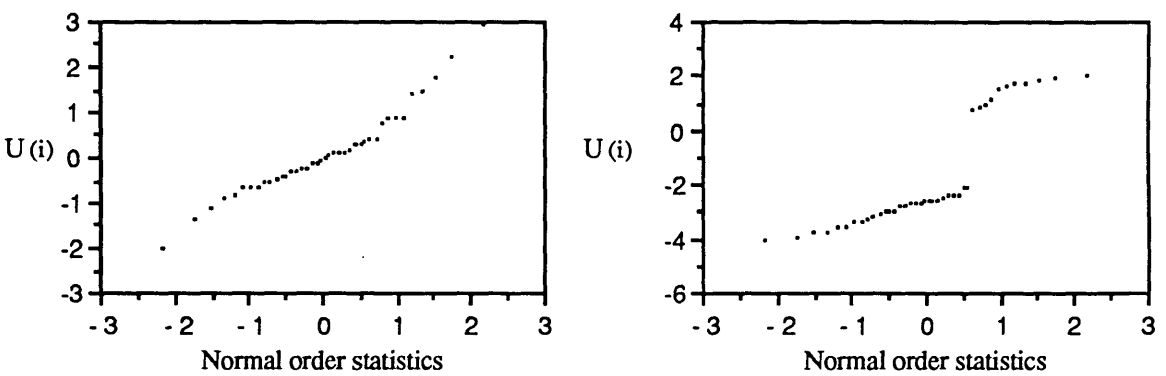
Two-point prior with τ_1=1.0, τ_2=5.2 and p=0.69.



(b) f=10

Inverse gamma with ν=5.0 and τ=1.0.

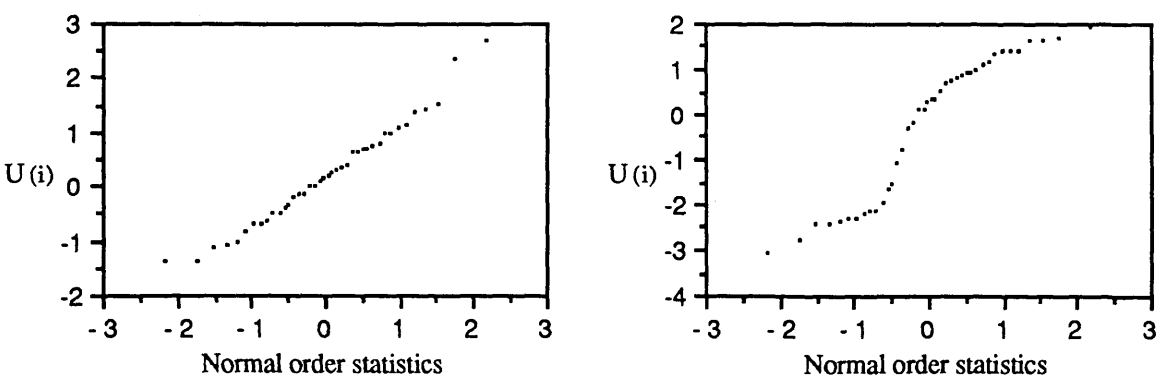
Two-point prior with τ_1=1.0, τ_2=5.2 and p=0.69.



(c) f=4

Inverse gamma with ν=8.0 and τ=1.0.

Two-point prior with τ_1=1.0, τ_2=2.1 and p=0.37



ESTIMATION IN PARALLEL STUDIES

3.1 Introduction

This chapter investigates inferences about the means from k parallel, independent studies. The purpose here is to illustrate empirical Bayes techniques that can be used to summarize evidence in the data (coming from the k studies) about the means, thereby obtaining improved estimates of the mean in each study.

The model considered is the normal model with means normally distributed. An important point is that the variances in each experiment are assumed to be known. The feature of the model is that the means present extra variability along with their respective known variances (usually measuring sampling error).

The model is not new and its derivation is reviewed in §3.2. Empirical Bayes confidence intervals for the individual means are constructed and a correction to allow for the errors in estimating the prior is calculated, along the same lines as in §2.3.3. Generalization to the linear regression model is considered in §3.5.

The theory is discussed and illustrated by considering two applications; the analysis of several 2×2 tables with data from 14 experiments on the effects of smoking and lung cancer (Cox, 1970, p.81) and a comparative study of the rates of growth of the AIDS epidemic in Europe.

3.2 Empirical Bayes estimates

3.2.1 *Formulation*

We shall consider experiments in which it is reasonable to assume that the estimator Z_i, of the i^{th} experiment, has a normal distribution with unknown mean μ_i and known variance V_i, so that

$$Z_i \sim N(\mu_i, V_i) \quad \text{independently } i=1,...,k. \qquad (3.1)$$

If the variances are unknown, they can be replaced by accurate estimates obtained from the data, without greatly affecting the analysis. This usually requires that Z_i is an estimate based on a large amount of data.

The empirical Bayes model assumes that the μ_i themselves are normally distributed with mean m and variance τ :

$$\mu_i \sim N(m, \tau) \qquad \text{independently } i= 1,...k, \quad (3.2)$$

One of the motivations for (3.2) is that, although considering different means μ_i in each experiment, it may be sensible to think of the experiments as chosen from a larger population with normally distributed means. A practical motivation is that it provides estimates with smaller standard errors. The important characteristics of this model are that m will measure the overall mean of the k experiments and τ the extra variability in Z_i, beyond the variability arising from the sample variances V_i.

Interest focuses on inference for the separate means μ_i and for the common mean m. Robustness of the empirical Bayes estimates from assumptions (3.1) and (3.2) will be examined in chapter 4.

Given m and τ, the posterior distributions of the μ_i's are

$$\mu_i \sim N\left(\frac{Z_i/V_i+m/\tau}{1/V_i+1/\tau}, (1/V_i+1/\tau)^{-1}\right), \quad \text{independently, } i=1,...,k. \qquad (3.3)$$

Two features of (3.3) are:

(i) The empirical Bayes estimate of μ_i , the posterior mean, is a compromise between the independent estimate Z_i and the common mean m, with weights inversely proportional to their respective variances, V_i and τ. Hence, each estimate, Z_i, is pulled towards the prior mean m, the extreme values and the values with large variance (compared to τ), experiencing most shift.

(ii) Standard errors are smaller than $V_i^{1/2}$, producing shorter confidence intervals.

The marginal distribution of Z_i , obtained after elimination of the unknown means μ_i from the product of (3.1) and (3.2) by integration, is

$$Z_i \sim N(m, V_i + \tau), \quad \text{independently } i=1,...,k. \quad (3.4)$$

Empirical Bayes intervals for the means μ_i are constructed from (3.3) by replacing the unknown parameters m and τ by their respective estimates, calculated from the marginal distribution of Z_i (3.4). For large values of k this empirical Bayes confidence interval, seems to be adequate, in most cases, whereas for small values of k, it is advisable to investigate the way in which the estimation of m and τ affects the results. This will be done in §3.4 by comparing the coverage probability of the empirical Bayes interval with the expected coverage probability.

3.2.2 *Maximum likelihood estimates of the prior parameters*

The maximum likelihood estimates for m and τ are found from the marginal distribution of Z_i, (3.4) and can be expressed as

$$\hat{m} = \sum_{i=1}^{k} Z_i/(V_i + \hat{\tau}) \Big/ \sum_{i=1}^{k} 1/(V_i + \hat{\tau}) \quad (3.5)$$

and

$$\hat{\tau} = \sum_{i=1}^{k} \left\{ \frac{(Z_i - \hat{m})^2 - V_i}{(V_i + \hat{\tau})^2} \right\} \Big/ \sum_{i=1}^{k} 1/(V_i + \hat{\tau})^2. \quad (3.6)$$

Note that $\hat{m}$ is a weighted average of the individual estimates Z_i with empirical Bayes weights $(V_i+\hat{\tau})$.

Although there is no closed expression for the estimates of m and τ, equations (3.5) and (3.6) provide an easy iterative procedure for their calculation. Simply start with a guess for $\hat{\tau}$ (which can be the unweighted between means variation) and calculate $\hat{m}$, using (3.5). Then calculate a new appproximation for $\hat{\tau}$ and repeat this procedure until the estimates stabilize, which will often be in a few iterations.

If the variances are constant, such that $V_i=V$, then (3.5) and(3.6) become, respectively

$$\hat{m} = \bar{Z},$$

$$\hat{\tau} = \sum_{i=1}^{k} (Z_i - \bar{Z})^2/k - V.$$

In this case, negative values of $\hat{\tau}$ might occur when the 'estimated between means variation' is small compared to the within variation V. Thus, convergence of equation (3.6) to a negative value of $\hat{\tau}$ probably indicates that most of the variation of Z_i is explained by V_i and we should estimate τ by $\hat{\tau}=0$. The analysis reduces then to the analysis of k normal populations with means μ_i and known variances V_i.

The maximum likelihood estimates of m and τ are uncorrelated and their asymptotic variances are, respectively,

$$i^{mm} = \{\sum_{i=1}^{k} 1/(V_i+\tau)\}^{-1},$$

and (3.7)

$$i^{\tau\tau} = \{\frac{1}{2}\sum_{i=1}^{k} 1/(V_i+\tau)^2\}^{-1}.$$

The discussion now continues in the following sections by considering two applications: the analysis of several 2×2 tables and a comparative study of the Aids epidemic in Europe.

3.3 Analysis of several independent 2×2 tables

3.3.1 *Introduction*

Randomized block experiments are a common tool in many contexts to compare treatments. A simple example is of k parallel studies in which individuals are assigned at random, either to a treatment or a control group and a binary response is measured, each separate study consisting of a single 2×2 table. Interest focuses on the estimation of the treatment effect. Usually, simple merging of the k tables is not advisable, since the response may be influenced by factors other than the treatment effect, such as characteristics specific to each study.

The empirical Bayes approach used here, provides a means of combining the evidence of the difference among treatments from all studies. It also produces improved estimates of the treatment effect in each experiment.

3.3.2 *The model and the empirical logistic transform*

Denote the probability of success in study i, for untreated and treated patients, respectively, as

$$\theta_{i0} = e^{\lambda_{i0}}/(1+ e^{\lambda_{i0}}) \quad \text{and} \quad \theta_{i1} = e^{\lambda_{i1}}/(1+ e^{\lambda_{i1}}), \tag{3.8}$$

where

$$\lambda_{ij} = \log[\theta_{ij}/(1-\theta_{ij})], \qquad (j =0,1;\ i =1,...,k)$$

is the logistic transform or the log odds ratio of the probability θ_{ij}. The comparison of the difference among treatments is made in terms of the difference of the log odds ratio

$$\Delta_i = \lambda_{i1}-\lambda_{io} = \log \{[\theta_{i1}(1-\theta_{io})] / [\theta_{io}(1-\theta_{i1})]\}. \tag{3.9}$$

The reason for considering this reparameterization, is that it allows estimation of the treatment effect in retrospective studies, as in the example considered in §3.3.3. For a prospective study, in which the observed proportion of successes is an estimate of the probability of interest, other reparameterizations can be considered, such as the difference in the probabilities of success, $\theta_{i1}-\theta_{io}$.

Three possible models for the analysis of several 2×2 tables, suggested in Cox(1970), are:

(m1) $\lambda_{io}= \alpha_i$ and $\lambda_{i1}= \alpha_i+\Delta_i$,

(m2) $\lambda_{io}= \alpha_i$ and $\lambda_{i1}= \alpha_i+\Delta$ and

(m3) $\lambda_{io}= \alpha$ and $\lambda_{i1}= \alpha+\Delta$

Model (m1) assigns arbitrary probabilities to each of the $2k$ cells, and corresponds to a saturated model. Model (m2) specifies a constant logistic treatment effect but arbitrary probabilities for the k tables. It corresponds to a randomized block experiment where $\alpha_1,...,\alpha_k$ are the nuisance parameters, and is probably the most commonly used in applications. In model (m3) the probabilities are constant for all k studies and the data can be condensed into a single 2×2 table.

Various procedures have been suggested in the literature for the analysis of (m2). To mention two: unconditional maximum likelihood estimation as given by fitting a logistic linear model, and conditional maximum likelihood estimation based on the 'extended' hypergeometric distribution. The method used here is based on a normal approximation of the distribution of the empirical logistic transform,

$$\hat{\lambda}_{ij} = \log\left[(R_{ij}-0.5)/(n_{ij}-R_{ij}-0.5)\right], \qquad (i=1,...,k;\ j=0,1),$$

where R_{i1} is the number of the n_{i1} treated individuals and R_{i0} the number of the n_{i0} untreated individuals who respond positively. For a more detailed discussion of this form of the empirical logistic transform see (Cox, 1970, p.79).

For large n_{i1} and n_{i0}, provided that θ_{i1} and θ_{i0} are not too near 0 or 1, the variable

$$Z_i= \hat{\lambda}_{i1} -\hat{\lambda}_{i0},$$

is normally distributed with mean Δ_i and variance consistently estimated by

$$V_i = \frac{n_{i1}-1}{R_{i1}(n_{i1}-R_{i1})} + \frac{n_{i0}-1}{R_{i0}(n_{i0}-R_{i0})}.$$

If the treatment effect is supposed constant, model (m2), weighted least squares with estimated weights V_i and the techniques from the normal theory model can be used to estimate Δ. The problem in which we are interested, however, concerns the case of different treatment effects.

3.3.3 *Studies on the association between smoking and lung cancer*

We now turn to the data of 14 retrospective studies on the association between smoking and lung cancer Cox(1970, Ch.6). Table 7 gives the values of Z_i and $V_i^{1/2}$ from the separate analysis of each study. The data provide strong evidence against the hypothesis of a constant treatment effect , with the estimates of the treatment effect of studies 6, 7 and 11 standing out.

The question arises as to whether we believe that, for example, the effect of smoking on lung cancer in study 11 is actually 3.81 (on a logistic scale) with standard deviation 0.524, or whether it can be improved upon by using the information from the other studies. The estimation of the overall effect of smoking on lung cancer is another question of interest.

To combine evidence we assume that the treatment effect Δ_i in each study is normally distributed according to (3.2). As mentioned earlier, the motivation for (3.2) is that, although considering different logistic treatment effects, it may be sensible to think of the fourteen studies as chosen from a larger population with normally distributed effects. The results of §3.2, i.e. (3.3) and (3.4) follow immediately.

Maximum likelihood estimates of the common mean m and of the standard deviation $\tau^{1/2}$ obtained from the solutions of (3.5) and (3.6) are

$$\hat{m} = 1.69 \qquad \text{and} \qquad \hat{\tau}^{1/2} = 0.639.$$

The first approximation confidence interval for m is given by

$$\hat{m} \pm k_{\alpha} \left\{ \sum_{i=1}^{k} 1/(V_i + \hat{\tau}) \right\}^{-1/2},$$

where k_{α} is the lower α-quantile of the standard normal. The 95% confidence interval for m is (1.29, 2.09).

The empirical Bayes confidence interval for the mean of each separate study is

$$\left\{ \frac{Z_i/V_i + \hat{m}/\hat{\tau}}{1/V_i + 1/\hat{\tau}} \pm k_{\alpha} (1/V_i + 1/\hat{\tau})^{-1/2} \right\}. \qquad (3.10)$$

Table 7 gives the empirical Bayes estimate of the mean, its standard deviation and the corresponding values of the 95% empirical Bayes interval of the mean (3.10), for each of the 14 studies. An immediate comparison with the results obtained from the standard separate analyses confirms that all point estimates are moved towards the common mean $\hat{\mu}$=1.69 and that the empirical Bayes standard deviations are smaller than the independent standard deviations. For example, in study 11 after combining evidence from the other studies, the treatment effect estimate is 2.96 with standard deviation 0.405, compared to 3.81 and 0.524. Confidence intervals with a treatment effect near the overall effect or with small standard deviation compared to $\hat{\tau}^{1/2}$ are changed only slightly, though are always notionally more efficient, as can be seen in studies 5 and 13.

The results in Table 7 do not allow for the variability of the estimates $\hat{m}$ and $\hat{\tau}$. In the next section a correction for the significance level, to allow for this variability, is calculated along the same lines as in § 2.3.3.

As a final comment, if the V_i's are not accurate estimates of the variances, an *ad hoc* procedure, which in practice works well, is to calculate the critical values of the confidence interval from a Student-*t* distribution with the degrees of freedom associated with the estimated variance V_i.

Table 7: *Confidence intervals for the effect of smoking on lung cancer in each of the 14 studies*

Study	Z_i	$V_i^{1/2}$	95% C.I.	emp.Bayes est. of Δ_i	$(1/\hat{\tau}+1/V_i)^{-1/2}$	95%E.B.C.I.
1	1.83	.653	(0.55, 3.11)	1.76	.456	(0.87, 2.65)
2	1.90	.607	(0.72, 3.09)	1.80	.441	(0.94, 2.67)
3	1.51	.463	(0.61, 2.42)	1.58	.374	(0.83, 2.32)
4	0.64	.327	(0.00, 1.28)	0.86	.292	(0.29, 1.43)
5	1.74	.212	(1.33, 2.16)	1.74	.200	(1.34, 2.13)
6	2.61	.370	(1.88, 3.33)	2.38	.319	(1.75, 3.00)
7	0.25	.546	(-0.82, 1.32)	0.86	.415	(-0.01, 1.73)
8	2.27	.401	(1.49, 3.06)	2.11	.339	(1.44, 2.77)
9	1.79	.625	(0.56, 3.01)	1.74	.447	(0.87, 2.62)
10	1.37	.268	(0.85, 1.90)	1.42	.247	(0.94, 1.91)
11	3.81	.524	(2.78, 4.83)	2.96	.405	(2.17, 3.76)
12	1.18	.275	(0.64, 1.71)	1.26	.251	(0.76, 1.75)
13	1.46	.173	(1.12, 1.80)	1.48	.167	(1.15, 1.80)
14	1.81	.493	(0.85, 2.78)	1.77	.390	(1.00, 2.53)

$\hat{m} = 1.69, \hat{\tau} = 0.639.$

3.4 Correction for the prior estimation.

3.4.1 *Correction for the significance level*

We turn now to the calculations of the correction for the significance level of the empirical Bayes interval (3.10) to allow for the errors in estimating m and τ. With a slight change of notation (we will use again μ_i instead of Δ_i to denote the mean of the i^{th} experiment mean) we proceed as in § 2.3.3.

The calculation, showed in appendix B, of the second order term of the bias of the maximum likelihood estimates $\hat{m}$ and $\hat{\tau}$ revealed that $\hat{m}$ is an unbiased estimate of m and that

$$E(\hat{\tau} - \tau) = -\left\{\sum_{i=1}^{k} 1/(V_i+\tau)\right\}^{-1} = -i^{mm}.$$

Since $\hat{m}$ and $\hat{\tau}$ are uncorrelated (3.7), and the bias of $\hat{m}$, is zero the bivariate extension of the correction factor (2.26) has only three terms different from zero. After some straightforward calculations to obtain the derivatives with respect to $\hat{m}$ and $\hat{\tau}$, of the lower tail probability in interval (3.10), the correction for the significance level α is

$$c_\alpha(m,\tau) = \frac{\phi(-k_\alpha)}{\sum_{i=1}^{k} 1/(1+\varphi_i)} \frac{\varphi_i}{(1+\varphi_i)} \left\{ k_\alpha - \frac{d_i}{1+\varphi_i} \right\} +$$

$$\frac{\phi(-k_\alpha)\,\varphi_i}{(1+\varphi_i)\sum_{i=1}^{k} 1/(1+\varphi_i)^2} \left\{ \frac{k_\alpha d_i^2 \varphi_i}{(1+\varphi_i)^3} - \frac{d_i\,(2+\varphi_i k_\alpha^2)}{(1+\varphi_i)^2} + k_\alpha \left[1 + \frac{\varphi_i\,(k_\alpha^2 - 3)}{4(1+\varphi_i)} \right] \right\}, \quad (3.11)$$

where $\varphi_i = V_i/\tau$ is the ratio between the individual variance and the prior variance, $d_i = (Z_i - m)/\{\varphi_i\tau/(1+\varphi_i)\}^{1/2}$ is the standardized distance between the prior and the sample mean and k_α is the $(1-\alpha)$-quantile of the standard normal distribution. Thus $c_\alpha(m,\tau)$ is the correction of the lower tail probability α when allowing for the variability of the estimates of m and τ. To obtain the coverage probability (up to $o(1/n)$) of $(1-2\alpha)$, the empirical Bayes interval should be calculated from the α'-quantiles of the standard normal, where $\alpha' = \alpha - c_\alpha(\hat{m},\hat{\tau})$ is the corrected significance level.

Table 8 gives the values of the correction $c_\alpha(\hat{m},\hat{\tau})$, when $\alpha=0.025$, for each of the 14 experiments of Table 7. An interesting feature of the results is that the correction for those experiments with a large negative d_i is larger compared to those with a large positive d_i, especially in studies 4 and 7. This is clear from expression (3.11) where all the terms linear in d_i are negative. Interpretation seems to be related to the fact that the method approximates the lower tail probability of the interval, thus for $\hat{m} > Z_i$ (far from the lower tail) the correction for the lower tail probability is large and for $\hat{m} < Z_i$ (near the lower tail) the correction for the lower tail probability is small. The indication is that for extreme d_i the correction for the upper tail probability should be

calculated and the average of the two corrections considered. Table 8 also gives the average of the the corrections for the lower and upper tail probabilities (column 4) and the 95% empirical Bayes interval based on this correction (column 5).

Table 8: *Correction factor (3.10) for 95% empirical Bayes confidence intervals*

Study	d_i	lower tail $c_\alpha(\hat{m},\hat{\tau})$	average $\bar{c}_\alpha(\hat{m},\hat{\tau})$	2^{nd} order E.B.C.I.	Hierarchical Bayes e.B.est. of Δ_i (3.10)	st.error	95% interval
1	0.31	.009	.011	(0.76, 2.76)	1.78	0.512	(0.78,2.78)
2	0.48	.007	.010	(0.84, 2.76)	1.83	0.488	(0.87, 2.79)
3	-0.48	.009	.008	(0.80, 2.36)	1.57	0.401	(0.78, 2.35)
4	-3.60	.021	.010	(0.04, 1.32)	0.81	0.302	(0.22, 1.40)
5	0.25	.001	.002	(1.34, 2.14)	1.74	0.204	(1.34, 2.14)
6	2.89	-.001	.010	(1.69, 3.07)	2.44	0.336	(1.78, 3.10)
7	-3.47	.039	.021	(-0.24, 1.96)	0.73	0.453	(-0.16, 1.62)
8	1.71	.000	.008	(1.39, 2.83)	2.15	0.358	(1.45, 2.85)
9	0.22	.009	.011	(0.76 2.72)	1.76	0.498	(0.78, 2.73)
10	-1.30	.006	.004	(0.92, 1.92)	1.42	0.254	(0.92, 1.91)
11	5.23	.005	.034	–	3.15	0.440	(2.29, 4.01)
12	-2.03	.009	.005	(0.75, 1.81)	1.24	0.258	(0.73, 1.75)
13	-1.38	.003	.002	(1.14, 1.81)	1.47	0.169	(1.14, 1.81)
14	0.31	.006	.008	(0.94, 2.60)	1.78	0.421	(0.96, 2.61)

Comparison of the intervals in Tables 8 and 7 reveals that the correction has some effect; the empirical Bayes intervals, corrected for the estimation of m and τ, are larger compared to the 'first approximation' intervals in Table 7, but smaller than the standard separate 95% confidence intervals. Experiments 7 and 11, with the most extreme values of $|d_i|$ have the largest corrections, with the correction for experiment 11 being larger then 0.025. This is related to the fact that approximation is being calculated for small probabilities, thus, the error in the approximation may be large compared to the small value of the probability. In fact when α=0.010 the average correction also is 0.010. Nevertheless, the method seems to be sensitive to extreme observations and a

robustness analysis, as discussed in chapter 4, is advisable.

Inference about the treatment effect in experiments 5 and 13, in all three analyses, is basically unchanged.

3.4.2 *Hierarchical Bayes*

As mentioned earlier, another approach to the elimination of the dependence of the posterior results on the unknown prior parameters is given by the hierarchical Bayes analysis.

As an illustration, we consider here the hierarchical Bayes linear model with a joint uniform prior distribution for m and τ, $\pi(m, \tau) \equiv 1$.

Conditional on $(Z_1,...,Z_k)$, $(V_1,...,V_k)$ and τ the distribution of m can be shown to be

$$N\{m^*,[\Sigma 1/(V_i+ \tau)]^{-1}\}, \tag{3.12}$$

where $m^*=\Sigma Z_i/(V_i+ \tau)/\Sigma 1/(V_i+ \tau)$.

Note that m* is the maximum likelihood estimate of m when τ, is known. To eliminate the dependence of the posterior mean of μ_i on m, we have to calculate the expectation of the posterior distribution (3.3) with respect to (3.12). It then follows that the posterior distribution of μ_i is normal with mean

$$\frac{Z_i/V_i + m^*/\tau}{1/V_i + 1/\tau} \tag{3.13}$$

and variance

$$(1/V_i + 1/\tau)^{-1} + (1/V_i + 1/\tau)^{-2}\{\Sigma\, 1/(V_i+ \tau)\}^{-1}. \tag{3.14}$$

The covariance of μ_i and μ_j is

$$(1/V_i + 1/\tau)^{-1}(1/V_j + 1/\tau)^{-1}\{\Sigma\, \tau^2/(V_i+ \tau)\}^{-1}.$$

In order to eliminate the dependence on τ we would have to calculate the expectation of the posterior mean and variance with respect to a distribution of τ conditional only on $(Z_1,...,Z_k)$ and $(V_1,...,V_k)$. No analytical expression is available in this case.

Berger(1985, Chapter 4) evaluated this integral when the variances V_i within experiments are the same. Another possibility is to integrate numerically τ from the posterior quantities (3.13) and (3.14) with respect to the likelihood function of τ, calculated from the marginal distribution (3.4)

$$\propto \left\{ \frac{\prod 1/(V_i+\tau)}{\sum 1/(V_i+\tau)} \right\}^{1/2} \exp\left\{ -\frac{1}{2}\left[\sum Z_i^2/(V_i+\tau) - m * \sum 1/(V_i+\tau) \right] \right\}.$$

Table 8 displays the results (columns 6 and 7) of the expectation of the posterior mean (3.13) and the posterior variance (3.14) with respect to the likelihood of τ and the 95% interval (column 8). Comparison with the the second order emprirical Bayes intervals shows that in general both intervals are very similar, the latter being slightly smaller. The largest difference is for experiments 7 and 11 indicating that the approach correcting for the lower tail probability is more robust than the hierarchical Bayes with improper priors.

Lindley and Smith (1972) considered a general hierarchical Bayes model for the linear model with proper conjugate prior distributions.

Morris(1983a) developed approximations to the posterior mean and variance which do take into account the additional errors in estimating the prior (see also Morris(1983b)).

The analysis of the several 2 ×2 tables considered in the last section is not completely satisfactory. It requires large values of n_{i0} and n_{i1} for the normal approximation of the logistic transform. Also it is not valid for probabilities θ_{i0} and θ_{i1} near 0 or 1, which is the case for some medical applications where events can have small probabilities.

A better approach would be to consider an empirical Bayes analysis based on the conditional distribution of the number of success r_{i1} given the marginal total $t_i=r_{i1}+r_{i0}$. This is the 'extended' hypergeometric distribution (Johnson & Kotz, 1969, p.160),

$$f(r_{i1} \mid t_i, \Delta_i) = \frac{\binom{n_{i1}}{r_{i1}} \binom{n_{i0}}{t_i - r_{i1}} \exp(r_{i1}\Delta_i)}{\sum_{s=0}^{\min\{t_i, n_{i1}\}} \binom{n_{i1}}{s} \binom{n_{i0}}{t_i - s} \exp(s\Delta_i)} .$$

which are independent for i=1,...,k.

Because of the intractability of this distribution, the solution would have to be numerical. The use of approximations for the posterior moments and the marginal distributions, as given in Tierney & Kadane (1984), together with the use of the EM algorithm (c.f. Dempster, Laird and Rubin, 1976) might facilitate the solution.

3.5 A more general model

3.5.1 *Formulation*

A natural extension of the i.i.d. assumption (3.2) for the μ_i is to consider that the μ_i arise from a regression model

$$\mu_i = x_i^T\theta + \varepsilon_i$$

where $\theta^T = (\theta_1,...,\theta_p)$ is a vector of unknown regression coefficients and x_i^T a known vector of explanatory variables.

The extension of the results when $\varepsilon_i \sim N(0, \tau)$ is immediate: estimation of the hyperparameters θ and τ is based on the mixture distribution (3.4), with the unknown overall mean replaced by the regression mean $x_i^T\theta$. Thus the maximum likelihood estimates of θ and τ are the iterative solution of equations (3.5) and (3.6). The mean of the posterior distribution (3.3) is then

$$\frac{Z_i/V_i + x_i^T\hat{\theta}/\hat{\tau}}{1/V_i + 1/\hat{\tau}} . \qquad (3.15)$$

As an illustration of this model we consider in the next section, empirical Bayes estimates of the rates of growth of the AIDS epidemic in Europe.

3.5.2 *Rate of growth of the AIDS epidemic in Europe: a comparative analysis*

There is great variation between European countries in the overall incidence of AIDS and detailed interpretation of this variation would be difficult. The rate of increase is, however, more nearly constant, although showing non-trivial variation, and the object of this section is a comparative study of this rate of increase. To improve the estimate of the increase rate for a particular country, by combining the information from all the other countries, empirical Bayes methods are applied to the data.

Countries (Bulgaria, Czechoslovakia, German D.R., Hungary, Iceland, Luxemburg, Malta, Poland, Romania and U.S.S.R.) reporting very few cases have been omitted. For the remainder, quarterly data on new cases in 1986 and 1987 and the total number of cases to March 1986, reported in Table 10, were analysed by fitting a Poisson process with an exponential growth rate:

$$\text{rate of incidence in year } t \;=\; \alpha \exp(\beta t).$$

While the rate of increase may be rather slower than exponential, the resulting estimate of β is a useful summary statistic for fitting the growth in the recent past, although not recommended for extrapolation.

The model specifies that the numbers of AIDS cases, N_j, in $(t_{j-1},t_j]$, $(j=0,...,s)$, with $t_{-1}=-\infty$, follow Poisson distributions with means

$$\int_{-\infty}^{to} \alpha \exp(\beta t)\, dt \;= \alpha \exp(\beta t_o)/\beta = \alpha \varphi_o(\beta)$$

$$\int_{tj-1}^{tj} \alpha \exp(\beta t)\, dt \;=\; \alpha \exp(\beta t_j)(1-e^{-\beta\delta_j})/\beta = \alpha \varphi_j(\beta), \qquad (j=1,...,s),$$

where $\delta_j = t_j - t_{j-1}$.

Maximum likelihood estimates of α and β are the solutions of

$$\alpha - \frac{N_{\cdot}\beta}{\exp(\beta t_s)} = 0,$$
$$\sum_{j=0}^{s} N_j(t_j - t_s) + \sum_{j=1}^{s} \frac{N_j \delta_j \exp(-\beta\delta_j)}{(1-\exp(-\beta\delta_j))} = 0, \qquad (3.16)$$

with $N_{\cdot} = \Sigma N_j$, the total number of cases at t_s.

Note that the maximum likelihood estimate of the increase rate, given by the second equation, is independent of the overall incidence rate α. This is because, conditional on the total number of cases $N_{\cdot}$, $(N_0,...,N_s)$ have a multinomial distribution with probabilities $\varphi_j(\beta)/\Sigma\varphi_j(\beta)$, therefore, not involving α. The maximum likelihood estimate of β, given by (3.16) is equivalent to the maximum likelihood estimate from the multinomial conditional distribution of $(N_0,...,N_s)$ given the total number of cases $N_{\cdot}$.

The elements of the asymptotic covariance matrix of the maximum likelihood estimates are

$$v_\beta = \left\{ \alpha \sum_{j=1}^{s} \exp(\beta t_{j-1})\, \delta_j^2/(1-\exp(-\beta\delta_j)) \right\}^{-1},$$

$$v_\alpha = \alpha\,(t_s - \beta^{-1})^2 \left\{ \sum_{j=1}^{s} \exp(\beta t_{j-1})\delta_j^2/(1-\exp(-\beta\delta_j)) \right\}^{-1} + \alpha\beta\exp(-\beta t_s),$$

$$v_{\alpha\beta} = -(t_s - \beta^{-1}) \left\{ \sum_{j=1}^{s} \exp(\beta t_{j-1})\delta_j^2/(1-\exp(-\beta\delta_j)) \right\}^{-1}.$$

Table 9 (columns (a) and (b)) reports the estimates of β and their estimated standard errors, $v_\beta^{1/2}$. Note that β is in $yr.^{-1}$ and the population doubling time is $0.69/\beta$ yr.. Some systematic features are apparent. To help in interpreting these the proportions of cases that are I.V. drug users (%IVDU), heterosexuals (%HE), homosexual/bisexuals, are also available in Table 9.

(i) Belgium has a very low rate of increase. It has also a high proportion (58%) of heterosexual cases and known special circumstances. There is however some indication that a heterosexual epidemic would grow more slowly than one driven from the other sources.

(ii) The Mediterranean countries have high values of β and also high proportions of IV drug users.

Maximum likelihood estimates for the regression coefficients θ and τ (calculated from equations (3.5) and (3.6)) from the fitting of a normal model with mean

$$\beta = \theta_0 + \theta_1(\%\text{IVDU}) + \theta_2(\%\text{HE}) \tag{3.17}$$

and variance

$$\tau + v_\beta$$

gave

parameter	*estimate* (yr^{-1})	*st. err.*(yr^{-1})
θ_0	67.14	5.47
θ_1	0.62	0.19
θ_2	−0.46	0.23

and estimate of τ (yr^{-2}) = 0.00941.

The fitted rates of increase for each country and corresponding standard errors are displayed in Table 9 (columns (c) and (d)).

The estimated model suggests, but of course does not prove, that the rates of growth of the epidemics amongst heterosexual, homosexual/bisexual and IV drug users are different and in increasing order. A direct check of this hypothesis has not been attempted here, although U.S. data might make this feasible. In fact, however, the hypothesis is in accord with other rather qualitative information on the issue.

Clearly Belgium has a major effect on the inclusion of the proportion of heterosexuals in (3.17). The fitting of the model with Belgium deleted from the data gives

$$\hat{\beta}^* = 64.80\{3.99\} + 0.64\{0.15\}(\%\text{IVDU}),$$

where the numbers in brackets are the standard errors.

Since for almost all countries (excluding Belgium) the largest proportion of cases are either homosexuals/bisexuals or IV drug users, there is the suggestion that the growth rate of the epidemic is determined by the rates of growth amongst IV drug users and homosexual/bisexuals. To check this assumption we consider the fit of a Poisson process with intensity

$$\alpha_0 \exp(\beta_0 t) + \alpha_1 \exp(\beta_1 t),$$

the sum of the two growth rates, subject to the constraint that the ratio of the expected numbers of cases in the two subpopulations at t_S, equals the ratio of the observed proportions. However, there is not enough information in the data for sensible maximum likelihood estimates of the parameters to be achieved. This might be possible if the proportions were available for a further time.

To improve the estimate of the increase rate for a particular country, by combining the information from all the countries, we may calculate the empirical Bayes estimate of β_i given by (3.15) and its standard error;

$$\text{Emp. Bayes est. of } \beta_i = \lambda_i \hat{\beta}_i + (1-\lambda_i)\ (\text{fitted } \hat{\beta}_i^*),$$

$$\text{st. error of E.B. est of } \beta_i = \lambda_i \hat{v}_{\beta i}$$

where $\lambda_i = \hat{\tau}/(\hat{\tau}+\hat{v}_{\beta i})$.

This provides a compromise between the individual estimates (a) and the estimate from the fitting of the 'regression' (c). For countries with small standard error (France, W. Germany, Italy, U.K.) the empirical Bayes estimate is effectively the individual estimate $\hat{\beta}_i$, while for countries with large standard errors the empirical Bayes estimate is near the 'regression' estimate.

An immediate comparison with the results in §3.3.3, Table 7, shows that here the individual estimates experienced less of shift than there. The reason is that here the estimate of the prior variance is large compared to the individual variances V_i . Thus when $V_i << \hat{\tau}$, the empirical Bayes estimate is essentially the

individual estimate Z_i. Also, the estimate of the regression coefficients from the maximum likelihood equations (3.5) and (3.6) are very to close to the ordinary least squares estimates. On the other hand when $V_i >> \hat{\tau}$, the individual estimates Z_i would experience a large shift towards the regression estimate $x_i^T\hat{\theta}$, and maximum likelihood estimation of the coefficient is equivalent to a least squares estimation with weights V_i.

Table 9. *Estimated exponential rates of increase for 18 European countries.*

	est. $\hat{\beta}_i$ (yr^{-1})		fitted β_i (yr^{-1}) from (3.17)		emp. Bayes(yr^{-1})		%ivdu	%he	%hom
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
	est. β	*st. err.*	*est.*	*st. err.*	*est.*	*st.err.*			
Austria	.834	.081	.807	.041	.823	.062	23	2	5
Belgium	.304	.028	.416	.111	.313	.027	2	58	25
Denmark	.585	.048	.661	.046	.601	.043	2	5	84
Finland	.447	.124	.618	041	.553	.076	4	17	71
France	.851	.017	.724	.038	.847	.017	12	5	62
W. Ger.	.731	.021	.714	.042	.731	.021	9	3	75
Greece	.985	.116	.571	.848	.741	.074	1	23	47
Ireland	.803	.161	.846	.045	.834	.083	30	3	27
Israel	.394	.081	.684	.052	.514	.062	2	0	61
Italy	1.033	.030	1.054	.094	1.035	.029	64	4	21
Netherl.	.687	.040	.687	.048	.687	.037	4	2	87
Norway	.681	.098	.677	.041	.679	.069	6	7	79
Portugal	.710	.089	.547	.064	.636	.066	6	35	51
Spain	.997	.039	.999	.076	.997	.036	53	1	25
Sweden	.641	.061	.639	.047	.641	.052	0	7	81
Switz.	.636	.041	.744	.035	.652	.038	19	10	63
U.K.	.717	.024	.665	.047	.714	.024	2	4	85
Yugosl.	1.317	.273	.810	.042	.867	.091	28	8	40

Doubling time $0.693/\beta$ yr.

(Based on report of the WHO Collaborating Centre on AIDS)

Table 10. *Total number of AIDS cases reported in 18 European countries*

Country	Mar. 86	Jun. 86	Sep. 86	Dec. 86	Mar. 87	Jun. 87	Sep. 87	Dec. 87
Austria	34	36	44	54	72	93	120	139
Belgium	160	171	180	207	230	255	277	277
Denmark	80	93	107	131	150	176	202	228
Finland	11	11	14	14	19	19	22	24
France	707	859	1050	1221	1632	1980	2532	3073
W.Ger.	459	538	675	826	999	1133	1400	1669
Greece	14	22	25	35	42	49	78	88
Ireland	8	9	10	12	14	19	19	25
Israel	23	24	31	34	38	39	43	47
Italy	219	300	367	523	664	870	1104	1411
Netherl.	120	146	180	218	260	308	370	420
Norway	21	24	26	35	45	49	64	70
Portugal	24	28	40	46	54	67	81	90
Spain	145	177	201	264	357	508	624	789
Sweden	50	57	76	90	105	129	143	163
Switz.	113	138	170	192	227	266	299	355
U. K.	287	340	512	610	729	870	1067	1227
Yugoslavia	3	3	3	8	10	11	21	26

CHAPTER 4

ROBUSTNESS of EMPIRICAL BAYES ESTIMATES

4.1 Introduction

This section is devoted to studying the sensitivity of the empirical Bayes analysis to possible misspecification of either the prior or the sample distribution. The approach adopted here considers a larger class of distributions with departures from the normal distribution, and then investigates robustness of confidence intervals for the means as the distributions vary in that class. Specifically, empirical Bayes estimates are calculated for the mean and its variance when the distributions of the sample and prior means are not normal and belong to the specified class. The comparison with the empirical Bayes estimates calculated from the normal assumptions gives the correction for departures from normality and indicates the sensitivity to distributional shapes.

4.2 Correction for the empirical Bayes estimate of the mean

With a slight change of notation, the analysis in the previous chapter was based on the normal assumptions,

$$\bar{X}_i \sim N(\mu_i, \sigma_i^2), \tag{4.1}$$

$$\mu_i \sim N(m, \tau^2) \text{ independently } i=1,\ldots,k \tag{4.2}$$

with m and τ^2 estimated from the data and σ_i^2 known.

For convenience the suffix i will be omitted unless otherwise noted.

We suppose now that $\bar{X}$ is the mean of n independent, identically distributed random variables whose distribution has standardized cumulants ρ_r. The Edgeworth expansion of the density of $\bar{X}$ about the normal density is

$$f(\bar{x}|\,\mu) = \sigma^{-1}\phi(z_s)\{1+ \rho_3 h_3(z_s)/(6n^{1/2}) + \rho_4 h_4(z_s)/(24n)$$

$$+\rho_3{}^2 h_6(z_s)/(72n) + O(n^{-3/2})\}, \qquad (4.3)$$

where $\phi(.)$ is the standard normal density and $z_s = (\bar{x} - \mu)/\sigma$ is the standardized version of $\bar{x}$. The hermite polynomials $h_j(z)$ are defined by

$$\phi(z)\,h_j(z) = (-1)^j \frac{d^j\phi(z)}{dz^j},$$

and explicit expressions for the first seven hermite polynomials (see Kendall and Stuart, vol.1, p.167) are

$$h_0(z) = 1,$$

$$h_1(z) = z,$$

$$h_2(z) = z^2 - 1,$$

$$h_3(z) = z^3 - 3z,$$

$$h_4(z) = z^4 - 6z^2 + 3,$$

$$h_5(z) = z^5 - 10z^3 + 15z,$$

$$h_6(z) = z^6 - 15z^4 + 45z^2 - 15.$$

The effect of non normality on the prior distribution will be studied by considering the class of 'perturbed' prior distributions approximated by

$$f(\mu|m) = \tau^{-1}\phi(z_p)\,\{1 + \kappa_3 h_3(z_p)/(6n^{1/2}) + \kappa_4 h_4(z_p)/(24n)$$

$$+ \kappa_3{}^2 h_6(z_p)/(72n) + O(n^{-3/2})\}, \qquad (4.4)$$

where $z_p = (\mu - m)/\tau$ is the standardized version of μ and $\kappa_3/n^{1/2}$, κ_4/n,... are the standardized cumulants of the 'true' prior distribution of μ. The prior distribution does not depend on n; however its cumulants are expressed as powers of n^{-1} as an artefact to simplify the comparison between sample and prior distribution, it being assumed initially that we are interested in situations in which the departures from normality are of roughly equal importance in the two components. The index p in z_p stands for prior and s in z_s stands for sample.

The posterior mean and variance assuming that (4.3) and (4.4) are the

'true' distributions are calculated from the posterior distribution of μ

$$f(\mu|\bar{x}) = f_N(\mu;\mu_N,V_N)\{1 + [\rho_3 h^*_3(z_s)+\kappa_3 h^*_3(z_p)]/(6n^{1/2}) +$$

$$[\rho_4 h^*_4(z_s)+\kappa_4 h^*_4(z_p)]/(24n) + [\rho_3^2 h^*_6(z_s)+\kappa_3^2 h^*_6(z_p)]/(72n) +$$

$$[\rho_3 E(h_3(z_s)|\bar{x})+\kappa_3 E(h_3(z_p)|\bar{x})][\rho_3 h^*_3(z_s)+\kappa_3 h^*_3(z_p)]/(36n)+$$

$$\rho_3\kappa_3[h_3(z_s)h_3(z_p)-E\{h_3(z_s)h_3(z_p)|\bar{x}\}]/(36n) + O(n^{-3/2})\}, \qquad (4.5)$$

where μ_N and V_N are, respectively the mean and variance of the posterior distribution $f_N(\mu;\mu_N,V_N)$ based on the normal assumptions (4.1) and (4.2) and

$$h_j^*(z(\mu)) = h_j(z(\mu)) - E[h_j(z(\mu))|\bar{x}]$$

is a location invariant version of the hermite polynomials where $z(\mu)$ is a linear function on μ. The expectations in (4.5) are taken with respect to the posterior normal density $f_N(\mu;\mu_N,V_N)$.

The calculation of the posterior mean of (4.5) involves the calculation of

$$E[h_j^*(z(\mu))\mu \mid \bar{x}] = E\{h_j[z(\mu)][\mu-\mu_N] \mid \bar{x}\}$$

$$= V_N^{1/2} E\{h_j[z(\mu)]\, h_1[(\mu-\mu_N)/V_N^{1/2}] \mid \bar{x}\}$$

and

$$E\{h_j[z(\mu)] \mid \bar{x}\} = E\{h_j[z(\mu)]\, h_0[(\mu-\mu_N)/V_N^{1/2}] \mid \bar{x}\}.$$

A general formula to calculate the above expectations is given by,

$$E\left\{h_j\left(\frac{\mu-\xi}{\delta}\right) h_r\left(\frac{\mu-\mu_N}{V_N^{1/2}}\right)\Big| \bar{x}\right\} = \frac{j!V_N^{r/2}(\delta^2-V_N)^{(j-r)/2}}{(j-r)!\,\delta^j} h_{j-r}\left(\frac{-(\xi-\mu_N)}{(\delta^2-V_N)^{1/2}}\right), \quad r\le j, \qquad (4.6)$$

(for a proof see appendix D).

In particular for $z(\mu)=z_s$ and $z(\mu)=z_p$ it is possible to show that

$$E\left\{h_j(z_s)\, h_r\left(\frac{\mu-\mu_N}{V_N^{1/2}}\right)\Big|\bar{x}\right\} = \frac{(-1)^j j!\,\tau^r\,\sigma^{j-r}}{(j-r)!\,(\sigma^2+\tau^2)^{j/2}} h_{j-r}(-d),$$

and

$$E\left\{h_j(z_p)\, h_r\left(\frac{\mu-\mu_N}{V_N^{1/2}}\right)\Big|\bar{x}\right\} = \frac{j!\,\sigma^r\,\tau^{j-r}}{(j-r)!\,(\sigma^2+\tau^2)^{j/2}} h_{j-r}(d),$$

where $d = (\bar{x} - m)/(\sigma^2 + \tau^2)^{1/2}$, is the standardized distance between the sample mean and the prior mean.

A better and more illuminating way of writing the results is in terms of the dimensionless ratio $\varphi^2=\tau^2/\sigma^2$ between variances. Thus, after some straightforward calculations we show that the standardized correction for the empirical Bayes estimate μ_N when the 'true' underlying distributions satisfy (4.3) and (4.4) is

$$\frac{E(\mu|\bar{x})-\mu_N}{V_N^{1/2}} = \frac{\varphi(\kappa_3\varphi-\rho_3)}{2n^{1/2}(1+\varphi^2)^{3/2}} h_2(d) + \frac{\varphi(\kappa_4\varphi^2-\rho_4)}{6n(1+\varphi^2)^2} h_3(d) - \frac{\varphi\,(\varphi^4\kappa_3^2 - \rho_3^2)}{2n(1+\varphi^2)^3}\{d^2 - 2d\}$$

$$+ \frac{\varphi^2(\varphi^2 - 1)\rho_3\kappa_3}{2n(1-\varphi^2)^3}\{d^2 - 2d\} + O\,(n^{-3/2}). \qquad (4.7)$$

One important point to note about the form of (4.7) is, as would be expected, that the leading term corrects for skewness whereas kurtosis is corrected by the second order term.

Before commenting on the form of (4.7) with respect to d let us examine its behaviour with respect to φ. The Taylor series expansion of (4.7) for large φ is

$$\varphi^{-1}\left\{\kappa_3 h_2(d)/2n^{1/2} + [2\kappa_4 h_3(d) + \kappa_3^2(d^2 - 2d))]/n\right\} + O(\varphi^{-2}),$$

and for small φ,

$$\varphi\left\{\rho_3 h_2(d)/2n^{1/2} + [2\rho_4 h_3(d) + \rho_3^2(d^2 - 2d)]/n\right\} + O(\varphi^2).$$

Hence, if the prior variance is large compared to the sample variance (large φ) the correction (4.7) depends only on the prior cumulants κ_3 and κ_4, to first order in φ, suggesting that departure of the prior from normality is the more important factor affecting μ_N. In other words, for large φ, μ_N is a robust estimate of the mean with

respect to departures from normality of the sample distribution. If, however, the sample variance is large compared to the prior variance (small φ), the important factor affecting μ_N is departure of the sample distribution with respect to the normal. These results are no surprise since the normal posterior mean μ_N, is a weighted average of the prior and sample mean with the respective variances as weights. The last term of (4.7) is of smaller order in φ (for all cases) and will be ignored in the following discussion.

Robustness arguments clearly depend on the actual observed values and here this is reflected by the form of the correction factor (4.7) with respect to *d*.

(i) When d=0 the only non null term is the first order term correcting for skewness with $h_2(d)$= -1. For skewness of the prior distribution the posterior mean is moved away from its larger tail and for skewness of the sampling distribution it is moved towards its larger tail. But, unless the third cumulants are very large the correction is small.

(ii) The effect of skewness on the posterior mean for moderately large $|d|$ is dominated by the first order term, which moves $E(\mu|\bar{x})$ from μ_N in the direction of the longer tail of the skewed prior distribution and moves $E(\mu|\bar{x})$ from μ_N in the opposite direction away from the longer tail of the skewed sampling distribution (see figure 3). Thus, for example, if $d>0$ ($\bar{x}<m$) the posterior mean will tend towards $\bar{x}$ if $\kappa_3>0$ or towards m if $\rho_3>0$. The latter could be interpreted as a correction for possible outliers in the longer tail of the distribution of $\bar{x}$. The third term, correcting for skewness, has an additive effect when the skewness is negative and a subtractive effect when the skewness is positive.

(iii) The effect of kurtosis in the posterior mean is represented by the second term in (4.7). For moderately large $|d|$, the posterior mean will tend towards the sample mean $\bar{x}$ when the prior distribution has positive kurtosis, $\kappa_4>0$, or will tend towards the prior mean m when the sampling distribution has $\rho_4>0$. These results agree with Dawid(1973) that for large $|d|$ the empirical Bayes estimate from a prior distribution with a flat, i.e. long, tail is closer to $\bar{x}$. This is advisable if we consider that an extreme observation should somehow discredit the prior value m. On the other hand

for a sample distribution with a flat tail, the empirical Bayes estimate is closer to m, suggesting that $\bar{x}$ is an outlier.

Considering both the effect of skewness and kurtosis, the correction is generally larger when the signs of the two first terms in (4.7) are equal. Since we are considering positive kurtosis, the correction will be large when $\bar{x}$ is in the longer tail of the prior distribution (figure 3, (a1)) or when m is in the smaller portion of the sampling distribution (figure 3, (b2)). In the former the posterior mean will tend to $\bar{x}$ and in the latter to the common mean m. This is illustrated by Table 11(a), which shows the values of the correction for positive skewness and kurtosis of the sampling distribution,$\rho_3=1$, $\rho_4=1$ and $\varphi=1$, for some values of d and n; the correction is larger when $m<\bar{x}$. Table 11(b) calculates the correction for $\rho_3=-1$, $\rho_4=1$ and $\varphi=1$ and the values are slightly larger than in the previous table because of the additive affect of the second order term correcting for skewness. The larger correction is when $d<0$, i.e. when m is in the smaller tail of the sampling distribution.

The correction for departures of the prior has the opposite sign to the correction for equivalent departures of the sampling distribution. Thus, for a prior which mimics the sampling distribution the effect will tend to be cancelled and the correction negligible.

Note, however, that (4.7) may give rather bad values as $d\to\infty$, since the hermite polynomials are unbounded. This is due to the fact that Edgeworth expansions, using the three correction terms in (4.3), are not usually good approximations in the tails of the distribution (Barndorff-Nielsen & Cox,1979). In fact, when $(\kappa_4\varphi^2-\rho_4)>0$, $d>0$ and $\rho_3=\kappa_3=0$, $E(\mu|\bar{x})$ may be greater than $\bar{x}$ if

$$d > \{3 + 24n/(\kappa_4\varphi^2-\rho_4)\}^{1/2},$$

that is if $d \sim O(n^{1/2})$. However, this represents quite extreme values of d, as for example, when $n=4$ and $(\kappa_4\varphi^2-\rho_4)=1$ then $d>10$.

In the empirical Bayes approach m and τ^2 are estimated, and can be interpreted, respectively, as the weighted mean of the separate experimental means and

the between sample variation. The error incurred in replacing m and τ by their maximum likelihood estimates is $O(k^{-1/2})$, where k is the number of samples.

While theoretically, expression (4.7) could be used as a correction for μ_N in the case of suspected non normality, it requires a large sample size n and estimates of the cumulants for each separate sample, and so the feasibility is questionable.

Table 11: *Correction (4.7) of the posterior mean* μ_N *when*

(a) $\rho_3=1$, $\rho_4=1$ and $\varphi=1$.

	n=1	n=2	n=4	n=10	n=20	n=30
d						
5.0	-7.89	-4.82	-3.03	-1.71	-1.13	-0.90
4.0	-4.32	-2.71	-1.74	-1.01	-0.68	-0.54
3.0	-1.98	-1.28	-0.85	-0.50	-0.34	-0.28
2.0	-0.61	-0.42	-0.29	-0.18	-0.12	-0.10
1.0	0.02	0.01	0.01	0.00	0.00	0.00
0.0	0.18	0.12	0.09	0.06	0.04	0.03
-1.0	0.10	0.05	0.03	0.01	0.01	0.00
-2.0	0.05	-0.08	-0.12	-0.11	-0.09	-0.08
-4.0	1.02	-0.04	-0.41	-0.47	-0.41	-0.36
-5.0	2.53	0.39	-0.43	-0.66	-0.61	-0.55

(b) $\rho_3=-1$, $\rho_4=1$ and $\varphi=1$.

	n=1	n=2	n=4	n=10	n=20	n=30
d						
5.0	0.60	1.18	1.21	0.98	0.77	0.65
4.0	0.98	1.04	0.91	0.67	0.51	0.43
3.0	0.85	0.72	0.57	0.39	0.29	0.24
2.0	0.45	0.33	0.24	0.16	0.11	0.09
1.0	0.02	0.01	0.01	0.00	0.00	0.00
0.0	-0.18	-0.12	-0.09	-0.06	-0.04	-0.03
-1.0	0.10	0.05	0.03	0.01	0.01	0.00
-2.0	1.11	0.67	0.41	0.23	0.15	0.12
-3.0	3.10	1.84	1.13	0.62	0.40	0.31
-4.0	6.32	3.71	2.24	1.21	0.78	0.61
-5.0	11.01	6.39	3.81	2.02	1.29	1.00

4.3 Correction for posterior variance

To investigate the effect of non normality on confidence intervals we consider the posterior variance of μ,

$$\mathrm{Var}(\mu|\bar{x}) = E[(\mu-\mu_N)^2|\bar{x}] - [E(\mu|\bar{x}) - \mu_N]^2.$$

Since,

$$E[h_j^*[z(\mu)]\ (\mu-\mu_N)^2|\bar{x})\ ) = V_N\, E\{h_j[z(\mu)]\ h_2[(\mu-\mu_N)/V_N^{1/2}]|\bar{x}\},$$

using (4.6) it can be shown that,

$$\mathrm{Var}(\mu|\bar{x}) = V_N\Big\{1+\frac{\varphi(\rho_3\varphi+\kappa_3)}{n^{1/2}(1+\varphi^2)^{3/2}}h_1(d)+\frac{\varphi^2(\rho_4+\kappa_4)}{2n(1+\varphi^2)^2}h_2(d) \tag{4.8}$$

$$-\frac{\varphi^2(\rho_3^2+\kappa_3^2)}{2n(1+\varphi^2)^3}\Big(\frac{3}{2}h_2(d)+\frac{1}{2}\Big)+\frac{\rho_3\kappa_3}{n(1+\varphi^2)^3}\Big\{\frac{\varphi(1+\varphi^4)}{2}h_2(d)+\varphi^3\,(2h_2(d)+1)\Big\}\Big\},$$

where $d=(\bar{x}-m)/(\sigma^2+\tau^2)^{1/2}$.

As in the posterior mean the leading term of the posterior variance corrects for skewness.

When $d=0$, the first order term vanishes. The term correcting for kurtosis is negative and in general $\mathrm{Var}(\mu|\bar{x})$ will be close to V_N and often smaller; in this case V_N overestimates the 'true' variance.

The qualitative behaviour of the leading term $O(n^{-1/2})$ for positive $(\kappa_3\varphi+\rho_3)$ is that $\mathrm{Var}(\mu|\bar{x})<V_N$ when $\bar{x}<m$ and $\mathrm{Var}(\mu|\bar{x})>V_N$ when $\bar{x}>m$. The interpretation for positive skewness of the prior distribution is simply that for $\bar{x}$ in the left portion of the prior distribution (which compared to the Normal is less dispersed) V_N overestimates the 'true' variance and for $\bar{x}$ in the right portion of the prior distribution (which compared to the Normal is more dispersed) V_N underestimates the 'true' variance. The interpretation for positive skewness of the sample distribution, while not so clear, might be that for large d and for m in the left portion of the sample distribution ($\bar{x}>m$) there is conflict between prior and sample information, and the

variance of a robust estimate should be bigger.

When both distributions are symmetrical, that is $\rho_3=\kappa_3=0$, the leading term is $O(n^{-1})$ and involves only the fourth cumulants correcting for kurtosis. The behaviour with respect to$d$ is: for $|d|\leq 1$, V_N overestimates the 'true' variance and for large d, V_N underestimates the 'true' variance. Hence, when there is conflict between sample and prior means V_N may considerably underestimate the true variance, especially for large d and small n, and the empirical Bayes confidence interval based on the normal assumptions may be seriously too small.

The asymptotic expansion of (4.8) for large φ is

$$V_N\{1+\rho_3/\varphi n^{1/2}[h_1(d) + \kappa_3 h_2(d)/2n^{1/2}]\}+O(\varphi^{-2}),$$

and for small φ is

$$V_N\{1+\varphi\kappa_3/n^{1/2}[h_1(d) + \rho_3 h_2(d)/2n^{1/2}]\}+O(\varphi^{2}).$$

It is interesting to note that for large or small φ the term correcting for kurtosis is of smaller order. Also, differenting from the results for the mean, the variance depends on both third cumulants. Thus for large φ ($\tau^2>\sigma^2$) the term $O(n^{-1/2})$ corrects for skewness of the sample distribution while for small φ ($\tau^2<\sigma^2$) it corrects for skewness of the prior distribution.

In contrast to the case of the posterior mean, departures of the same kind for both the sample and the prior distribution give rise to terms with the same sign, increasing the correction for the posterior variance, as illustrated in the example below. Tables 12 (a) and (b) give the values of the ratio between the variances $\mathrm{Var}(\mu|\bar{x})/V_N$ calculated from (4.8), for $\rho_3=\kappa_3=\rho_4=\kappa_4=1$ and $\rho_3=\rho_4=1$, $\kappa_3=\kappa_4=0$ when $\varphi^2=1$, respectively. Because skewness is positive in both cases the correction is larger for positive d when the posterior variance is underestimated by the variance from the normal assumptions. The posterior variance is overestimated for values of d around zero when the ratio is smaller than 1. Only in the very extreme situation, in Table 12(b), of n=1 and d=-5.0, was a negative ratio calculated as a consequence of the poor approximation given by the Edgeworth expansion at the tails of the distributions.

Table 12: *Ratio between posterior variances* $\mathrm{Var}(\mu|\bar{x})/V_N$

(a) $\rho_3=1$, $\rho_4=1$, $\kappa_3=1$, $\kappa_4=1$ and $\varphi=1$.

	n=1	n=2	n=4	n=10	n=20	n=30
d						
5.0	15.10	8.78	5.41	3.17	2.32	2.00
4.0	10.45	6.31	4.07	2.56	1.96	1.74
3.0	6.68	4.28	2.95	2.03	1.65	1.51
2.0	3.79	2.69	2.05	1.58	1.38	1.30
1.0	1.77	1.53	1.37	1.23	1.16	1.13
0.0	0.62	0.81	0.91	0.96	0.98	0.99
-1.0	0.36	0.53	0.66	0.78	0.85	0.87
-2.0	0.96	0.69	0.64	0.69	0.75	0.79
-3.0	2.44	1.28	0.83	0.69	0.70	0.73
-4.0	4.80	2.31	1.24	0.77	0.70	0.70
-5.0	8.03	3.78	1.87	0.94	0.74	0.71

(b) $\rho_3=1$, $\rho_4=1$ and $\varphi=1$.

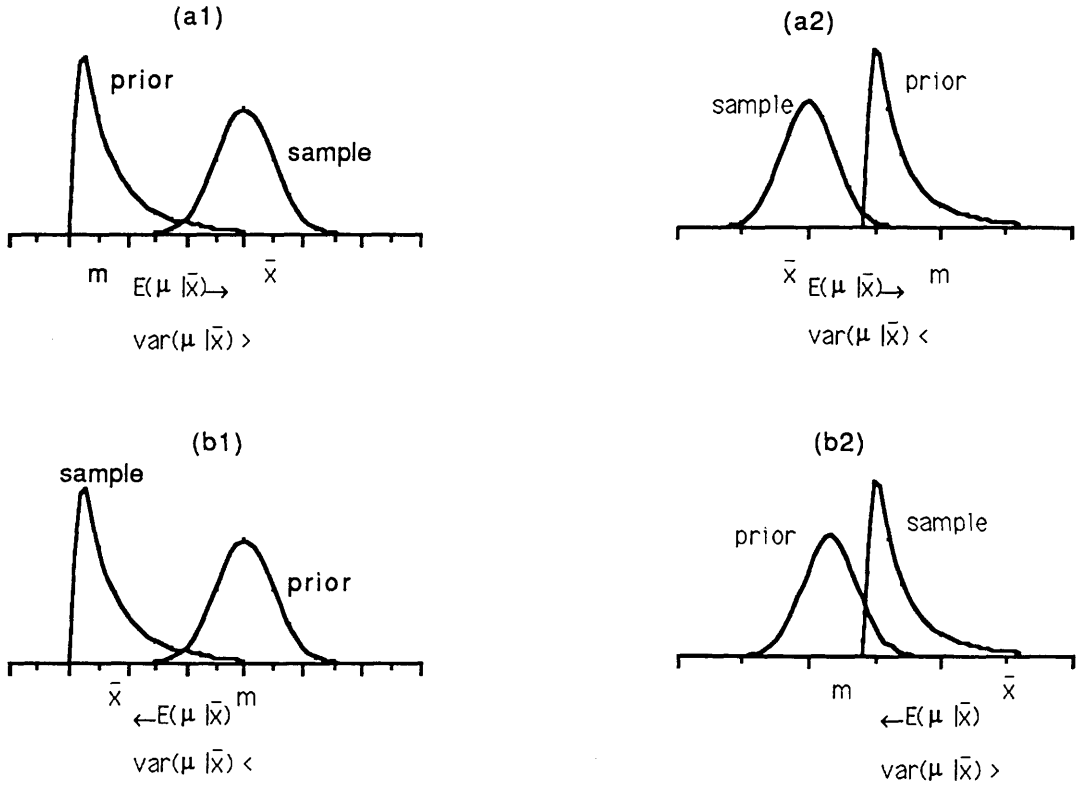
	n=1	n=2	n=4	n=10	n=20	n=30
d						
5.0	3.49	2.61	2.06	1.63	1.43	1.35
4.0	2.85	2.22	1.82	1.49	1.34	1.27
3.0	2.28	1.86	1.59	1.36	1.25	1.20
2.0	1.77	1.53	1.37	1.23	1.16	1.13
1.0	1.32	1.23	1.17	1.11	1.08	1.06
0.0	0.94	0.97	0.98	0.99	1.00	1.00
-1.0	0.62	0.73	0.82	0.89	0.92	0.93
-2.0	0.36	0.53	0.66	0.78	0.85	0.87
-3.0	0.16	0.36	0.52	0.69	0.77	0.81
-4.0	0.02	0.22	0.40	0.60	0.71	0.76
-5.0	-0.05	0.11	0.30	0.51	0.64	0.70

To conclude, we illustrate the results when the prior distribution is non-Normal with $\kappa_3=1$, $\kappa_4=1$ (figure 3, a) and when the sampling distribution is non-Normal with $\rho_3=1$, $\rho_4=1$ (figure 3, b). The figures show the direction of the shift of the posterior mean from μ_N and the size of the posterior variance compared to V_N. The effects on the posterior mean and variance are larger when the first order term correcting for skewness and the correction for kurtosis have the same sign, i.e. when:

(i) in (a1) the sampling distribution is in the larger tail of the prior;

(ii) in (b2) the observed sample is an outlier from its distribution.

Figure 3. *Illustrative example of the effect of skewness and kurtosis in the posterior mean and variance*



4.4 Conclusions and remarks

Two situations can be distinguished in the discussion of robustness; one when d_i is small, i.e. when the likelihood is in the central portion of the prior distribution, and one when d_i is large and the likelihood is in the tail of the prior distribution (subscripts, indicating different populations, are now included for clarity of the discussion).

In the first situation of small distance between $\bar{x}_i$ and m, the results indicate that the use of the normal distribution is reasonably robust to small departures from normality of both the prior and the sample distribution. The posterior mean is close to μ_N and the posterior variance is close to V_N, in some cases being smaller than V_N, as illustrated in Table 12. Thus, in this situation of small d_i the computational

simplicity of the normal distribution makes its use attractive.

However, the situation just discussed is not really of much interest since for m close to $\bar{x}_i$ the empirical Bayes estimate has little effect on the individual estimates $\bar{x}_i$.

In the more interesting situation of large d_i, the empirical Bayes estimate of the mean, from the normal assumptions, can be intuitively unappealing, being close neither to the common mean m nor to the sample mean $\bar{x}_i$ and the variance being substantially underestimated by V_N. The results of the previous section show that priors with flatter tails will tend to be robust in the sense that they are less influential if the likelihood is in the tail; the estimate of the mean will be closer to its sample mean $\bar{x}_i$ and the variance bigger, reflecting the conflict between m and $\bar{x}_i$. On the other hand if the sample distribution has flatter tails the estimate of the mean will be close to the prior mean m, interpreting $\bar{x}_i$ as an outlier and the variance will again be bigger, reflecting the conflict between m and $\bar{x}_i$.

The difficulty now is to determine what is large d_i, and alternative estimates in the case when the normal assumptions are not advisable. Note that d_i is the standardized version of $\bar{x}_i$ with respect to its marginal distribution,

$$\bar{x}_i \sim N(m, \sigma^2+\tau^2).$$

In extreme cases one could be alert to a robustness problem by the surprisingly large value of $|d_i|$. At least as a guide, significance tests based on the marginal distribution could be used to determine what is surprisingly large, although it does not seem possible to determine if the robustness problem is from the model or from the prior. Berger(1985, Ch.4) discusses the role of the marginal distribution in robustness of the prior distribution.

While distributions with flat tails are likely to appear in some applications, implementation is difficult and the results, in most cases, can be obtained only numerically. However, the results indicate that the use of the normal is not advisable for extremely large $|d_i|$. One possible way of overcoming this difficulty is to

consider estimates of the form

$$\hat{\mu}_i = \begin{cases} \mu_N & |d_i| \le d_0 \\ \bar{x}_i & |d_i| > d_0 \end{cases},$$

if τ^2 is larger or of the same order as σ_i^2, or

$$\hat{\mu}_i = \begin{cases} \mu_N & |d_i| \le d_0 \\ \hat{m} & |d_i| > d_0 \end{cases},$$

if σ_i^2 is much larger compared to τ^2. In this last case we are interpreting $\bar{x}_i$ as an outlier from its own distribution.

Perhaps a more relevant discussion about the behaviour of confidence intervals, in the presence of departures from normality, would be in terms of their coverage properties. However, such discussion would involve conditional probabilities and has not been attempted.

CHAPTER 5

INFLUENTIAL OBSERVATIONS IN THE ALLOCATION PROBABILITY TO TWO NORMAL POPULATIONS

5.1 Introduction

In the last chapter the effect of extreme observations on the empirical Bayes estimate of the mean was analysed in a robust framework. In other words the effect of extreme observations was studied comparing the changes in the empirical Bayes estimates with respect to changes in the distributional assumptions. It was shown that, for example, empirical Bayes estimates for the mean of a population with an extreme sample mean from a prior with flat tails (compared to the Normal) would tend to be more robust in the sense that the empirical Bayes estimate would experience less shift towards the prior mean.

Another way of examining the effect of extreme observations is by the use of *Influence measures.*

A Bayesian approach to the problem of influential observations in regression is review ed in Geisser(1984). It compares the posterior distribution of the parameters with and without the set of observations whose influence is to be determined. If the objective is to ascertain the influence of observations in the prediction of a future observation then the predictive distribution is used. As an indicator of the discrepancy between the two distribution functions (with and without the observations whose influence will be measured), the Kullback-Leibler information measure (Kullback & Leibler, 1951) is used. Comparisons with the Cook's statistics (Cook & Weisberg, 1982) for the influence of an observation with regard to the estimation of regression parameters are made.

The same methodology is applied here to Bayesian discriminant

analysis, in which training samples are available from two normal populations, it being required to calculate the probability that a new individual comes from one of the populations. Interest focuses on the influence of individual observations in the training sample on the allocation of future observations. This influence is measured by comparing the posterior probability of allocation with and without the observation whose influence we want to measure.

If one is interested in evaluating the influence of observations in the training sample when the observation we want to classify is not yet observed more relevant approaches are available. For two normal populations, one possible approach is logistic regression and the methods described in Cook & Weisberg(1982) could be applied. Also Copas(1988), in an important paper, discusses influence of outliers in binary regression and develops techniques for robust estimation.

5.2 Formulation

Let the two populations, π_1 and π_2, have p-multivariate normal distributions $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ with mean vectors μ_1, μ_2 and covariance matrices Σ_1, Σ_2. Further, let the training samples be of equal size n and have mean vectors $\bar{x}_1, \bar{x}_2$ and estimated covariances matrices S_1, S_2. We assume that (μ_1, Σ_1) and (μ_2, Σ_2) are independent with known prior distribution $g(\mu, \Sigma)$.

Let z denote the future observation and q_1 and q_2 the *a priori* probabilities of classifying z in populations π_1 and π_2.

On the basis of the full training data the posterior probability that z comes from π_1, given z, is

$$p_1 = \frac{q_1 f\ (z \mid \bar{x}_1, S_1, \pi_1)}{q_1 f\ (z \mid \bar{x}_1, S_1, \pi_1) + q_2 f\ (z \mid \bar{x}_2, S_2, \pi_2)}, \qquad (5.1)$$

where $f(z \mid \bar{x}_j, S_j, \pi_j)$ is the predictive distribution of z, given that z comes from π_j, (j=1,2).

Suppose that the i^{th} observation from the training sample of π_1 is deleted. We denote $\bar{x}_1^i$, S_1^i the new sample mean vector and covariance matrix of π_1, respectively.

The new value of the posterior probability that z comes from π_1, given z, is

$$p_1^i = \frac{q_1 f\ (z \mid \bar{x}_1^i,\ S_1^i,\ \pi_1)}{q_1 f\ (z \mid \bar{x}_1^i,\ S_1^i,\ \pi_1) + q_2 f\ (z \mid \bar{x}_2,\ S_2,\ \pi_2)}. \qquad (5.2)$$

One way of assessing the influence of the deleted observation i in the allocation probability is to measure the difference between (5.1) and (5.2). To do this we use the Kullback-Leibler information measure of distance which is the expected value of the logarithm of the ratio between the posterior probabilities of allocation (5.2) and (5.1),

$$I_{(i)}(z) = p_1^i \ln (p_1^i / p_1) + (1-p_1^i) \ln[(1-p_1^i)/(1-p_1)], \qquad (5.3)$$

with the expectation taken with respect to (5.2).

The influence index (5.3) provides a way of ordering individual observations based on their impact on the allocation probability of z. Note that the index of influence defined above is conditioned on z, so the order of influential observations depends on the particular future observation z.

5.3 Equal and known covariance matrices

To demonstrate the calculations and analyse the dependency of the influence index $I_{(i)}(z)$ on the deleted observation i and the future observation z, we consider first the simplest situation of equal and known covariance matrix Σ. We also assume vague prior distribution for μ_1 and μ_2, so that the predictive distributions of the future observation z are multivariate normal,

$$f(z \mid \bar{x}_j,\ \Sigma,\ \pi_j) = N_p(\bar{x}_j,\ (n+1)\Sigma/n),$$
$$f(z \mid \bar{x}_j^i,\ \Sigma,\ \pi_j) = N_p(\bar{x}_j^i,\ n\Sigma/(n-1)), \qquad (j=1,2). \quad (5.4)$$

Since a simple explicit expression of $I_{(i)}(z)$ is not obtainable when we substitute (5.4) in (5.3), we consider an approximation for large *n*.

Write

$$d_{ij}= \Sigma^{-1/2}(x_{ij} - \bar{x}_j), \tag{5.5}$$

where x_{ij} is the deleted observation from population π_j (j=1,2), and

$$\Sigma^{-1/2} = P^t \, Diag(\lambda_1^{1/2},...,\lambda_p^{1/2}) \, P,$$

with $\lambda_1,...,\lambda_p$ the characteristic roots of Σ^{-1}.

Also let

$$\Delta_j^2(z) = (z - \bar{x}_j)^t \, \Sigma^{-1} \, (z - \bar{x}_j),$$

$$\phi_j(z) = \exp(-(1/2)\Delta_j^2(z)), \qquad (j=1,2). \tag{5.6}$$

The first term of the expansion of (5.3) in powers of $1/n$ in terms of (5.5) and (5.6) is then

$$I^*_{(i)}(z) = \left\{ \frac{q_1 q_2 \phi_1(z) \, \phi_2(z)}{2n^2[\, q_1\phi_1(z) + q_2\phi_2(z)]^2} \right\} \{\Delta_j^2(z)\} \{d_{ji}^t \, d_{ji}\}, \tag{5.7}$$

where the deleted observation $i \in \pi_j$, j=1,2.

The expression above has a nice factorization in terms of the different sources of influence. The first is an overall influence for a given z and has its maximum value when $q_1\phi_1(z) = q_2\phi_2(z)$. The second component gives the population influence, so that if z is near the mean of π_1, that is, $\Delta_1^2(z) \leq \Delta_2^2(z)$, the observations from π_2 will have relatively large values of (5.7). The third component gives the observation influence within a population, so that if $\Delta_1^2(z) = \Delta_2^2(z)$, the most influential observation will be the one most distant from z.

Using an extremely small value of *n* in order to check the adequacy of (5.7) as an approximation to $I_{(i)}(z)$, ten observations were generated from the two bivariate normal distributions,

$$\pi_1 = 0.9 \, N_2\{\binom{0}{0}, I\} + 0.1 \, N_2\{\binom{5}{0}, I\}$$

and

$$\pi_2 = N_2\{\binom{2}{0}, I\}.$$

The data are shown in Figure 4, in the end of this chapter.

The measures $I_{(i)}(z)$ and $I^*_{(i)}(z)$ are calculated for data of figure 4 and z=(0.65, 0.93) and the observations corresponding to the ten largest values of $I_{(i)}(z)$ are then ordered. The results are shown in Table 13.

Table 13. *Comparison of the influence index* $I_{(i)}(z)$ *and its approximation* $I_{(i)}^*(z)$ *for* z*=(0.65, 0.93), with covariance matrix assumed known and equal to* $\hat{\Sigma}$*, the sample pooled covariance matrix*

x_1	x_2	Obs.	Pop.	$\lvert p_1^i - p_1 \rvert \times 10^2$	$I_{(i)}(z)\times 10^3$	$I_{(i)}^*(z)\times 10^3$
1.25	1.30	5	2	5.41	7.24	6.77
2.27	-2.09	3	2	5.52	7.03	7.30
0.81	2.11	9	1	3.78	3.33	2.19
1.45	-1.69	8	2	3.27	2.50	2.58
2.05	-1.06	4	2	2.40	1.35	1.32
2.14	-1.31	10	1	2.16	1.13	1.75
3.70	1.20	7	2	2.01	0.97	0.75
0.23	-0.57	1	2	1.54	0.57	0.50
1.09	-0.15	10	2	1.49	0.54	0.52
-0.30	0.68	7	1	1.45	0.50	0.32

$\bar{x}_1$=(0.71, 0.14), $\bar{x}_2$=(1.91, -0.29), $\hat{\Sigma}=\begin{pmatrix} 3.13 & 0.56 \\ 0.56 & 1.16 \end{pmatrix}$,

$q_1=q_2=0.5$ and $p_1(z\ ) = 0.70$.

The approximation between $I_{(i)}(z)$ and $I^*_{(i)}(z)$ depends on the location of z and the results above are typical for not very extreme z's. For z's very distantly located and small values of n the exact $I_{(i)}(z)$ and the approximation $I^*_{(i)}(z)$ can have quite different values.

Nevertheless, in this example, $I^*_{(i)}(z)$ is on the whole an excellent

approximation to $I_{(i)}(z)$. Note that although observation 8 from π_1 is the most extreme it is not influential for both measures.

Also, the fact that the leading term of $I_{(i)}(z)$ is of order $1/n^2$ explains the perhaps surprisingly small values of $I_{(i)}(z)$. This suggests that in the case of known covariance matrices, for large n and typical z (not extreme) the allocation probability of z is appreciably affected by individual observations only in very extreme cases.

5.4 Unknown covariance matrices

We consider now the more realistic situation of different and unknown covariance matrices. Assuming vague priors for μ_1, μ_2, Σ_1 and Σ_2 the predictive distributions for the future observation z have the form of multivariate Student-t distributions. Because of the difficulties in dealing with these distributions we make the expansions using a 'best' scaled normal approximation to the Student-t. The approximated predictive distributions are then

$$f(z \mid \bar{x}_j, \Sigma, \pi_j) \cong N_p\{ \bar{x}_j, S_j\,[(n-1)(n+1)]/[n(n-3)]\},$$
$$f(z \mid \bar{x}_j^i, \Sigma, \pi_j) \cong N_p\{ \bar{x}_j^i, S_j^i\,[n(n-2)]/[(n-1)(n-4)]\}, \quad (j=1,2). \quad (5.8)$$

Write

$$\Delta_j^*(z) = S_j^{-1/2}\,(z - \bar{x}_j),$$
$$d_{ij}^* = S_j^{-1/2}(x_{ij} - \bar{x}_j)$$

and

$$\phi_j^*(z) = |S_j^{-1/2}|\,\exp(-(1/2)\Delta_j^*(z)^t\Delta_j^*(z)), \qquad (j=1,2). \quad (5.9)$$

where x_{ij} is the deleted observation from population π_j (j=1,2), and

$$S_j^{-1/2} = P^t \mathrm{Diag}(\eta_1^{1/2},\ldots,\eta_p^{1/2})P,$$

with $\eta_1,\ldots,\eta_p$ the characteristic roots of S_j^{-1}.

Then, making the expansions in the same way as before, the approximate influence measure is

$$I^{a}_{(i)}(z) = \left\{ \frac{q_1 q_2\, \phi^{*}_1(z)\, \phi^{*}_2(z)}{2n^2 [\, q_1 \phi^{*}_1(z) + q_2 \phi^{*}_2(z)\,]^2} \right\} \left\{ \Delta^{*}_j(z)^t d^{*}_{ji} + [1 - (d^{*}_{ji})^t d^{*}_{ji}][1 - \Delta^{*}_j(z)^t \Delta^{*}_j(z)] \right\}^2,$$

$i \in \pi_j$, (j=1, 2).

The square of the first term inside the second brackets is equivalent to the last two expressions of $I^{*}_{(i)}(z)$, (5.7), when the covariance matrices are supposed known. The other terms are the effect introduced by the estimation of the unknown covariance matrices.

Using the same data as in Figure 4 we calculate the $I_{(i)}(z)$ and $I^{a}_{(i)}(z)$, supposing now unknown covariance matrices. The results are shown in Table 14 for z=(3.76,1.93).

Table 14. *Comparison of* $I_{(i)}(z)$ *and* $I^{a}_{(i)}(z)$ *when z=(3.76, 1.93), and the covariance matrices are supposed unknown*

x_1	x_2	Obs.	Pop.	$\lvert p_1^i - p_1 \rvert \times 10$	$I_{(i)}(z) \times 10^2$	$I^{a}_{(i)}(z) \times 10^2$
6.44	1.27	8	1	3.97	37.41	15.71
3.70	1.20	7	2	2.43	12.14	11.75
0.81	2.11	9	1	1.46	4.36	6.06
3.59	0.50	6	2	1.16	2.72	5.03
2.14	-1.31	10	1	0.76	1.15	0.36
2.27	-2.09	3	2	0.96	0.97	0.93
0.23	-0.57	1	2	0.46	0.43	0.09
1.09	-0.15	10	2	0.40	0.32	0.24
2.05	-1.06	4	2	0.38	0.30	0.41
-1.17	0.26	3	1	0.28	0.15	0.07

$q_1 = q_2 = 0.5$, $\bar{x}_1 = (0.71, 0.14)$, $\bar{x}_2 = (1.91, -0.29)$, $p_1(z\ ) = 0.48$,

$$S_1 = \begin{bmatrix} 5.10 & 0.77 \\ 0.77 & 1.05 \end{bmatrix} \quad \text{and} \quad S_2 = \begin{bmatrix} 1.15 & 0.35 \\ 0.35 & 1.27 \end{bmatrix}.$$

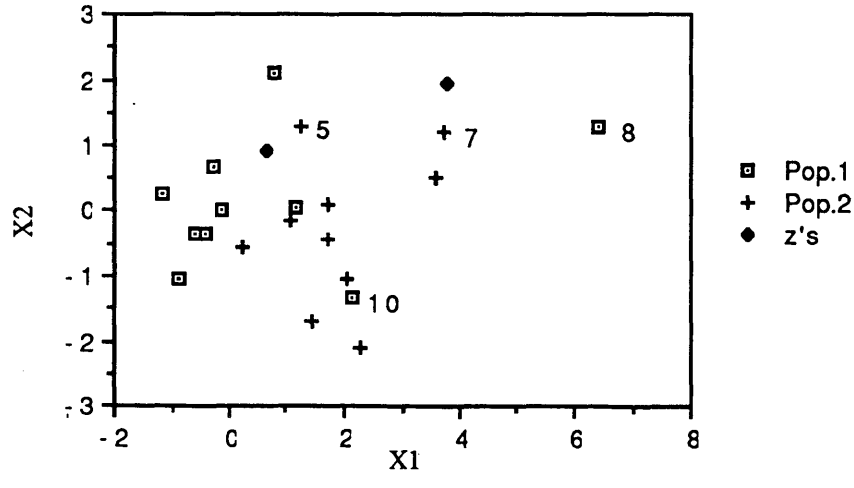
The agreement between $I_{(i)}(z)$ and $I^a_{(i)}(z)$ is not so good as in the previous example, but for such a small value of n the approximation is still reasonable and gives good general guidance. Better agreement of $I_{(i)}(z)$ and $I^a_{(i)}(z)$ should be expected as n increases and in general the effect of individual observations decreases as n increases. As before, the agreement between $I_{(i)}(z)$ and $I^a_{(i)}(z)$ depends on the location of z and on the configuration of the training sample. Note that the values of $I_{(i)}(z)$ are bigger than in Table 13. The same is true for the changes in the allocation probability which changes from 0.48 to 0.08 when observation 8 from π_1 is deleted. This shows that individual observations can be influentional when the covariance matrices are supposed unknown. This is presumably because of the effect, caused by the deletion, on the estimation of the covariance matrices.

For p_1 near p_1^i, $I_{(i)}(z)$ can be approximated by

$$(p_1 - p_1^i)^2 / [2p_1(1-p_1)]$$

which is the standardized distance between the two probabilities of allocation. The approximation is particularly good when p_1 is near 0.5 and can provide some insight about the magnitude of $I_{(i)}(z)$; for example if p_1=0.5 a value $I_{(i)}(z)$ indicates a change of magnitude approximately $(I_{(i)}(z)/2)^{1/2}$ in the probability of allocation on deleting observation *i*.

Figure 4. *Scatter diagram of the generated data*



Appendix A. Properties of the marginal distribution of the sample variances (2.5)

The marginal distribution of S_i (2.5), obtained by integrating out the unknown variances σ_i^2 from the product of (2.1) and (2.3), belongs to the family of Pearson type VI distributions, (Johnson and Kotz, Vol.2, 1970). It is interesting to note that the distribution of $S_i/(\nu\tau)$, can also be obtained as the distribution of the ratio of two independent χ^2-distributions with $f_i/2$ and $\nu/2$ degrees of freedom, being proportional to an *F*-distribution.

The density function of $X=S_i/(\nu\tau)$ is

$$f_X(x) = B^{-1}(a,b)\ \{1+x\}^{-(a+b)}\ x^{b-1}, \qquad x>0,\ a>0,\ b>0,$$

where $a=\nu/2$ and $b=f_i/2=f/2$. The density above is the standard form of a Pearson type VI density function. Moments of X and (1+X) are

$$E[x^r] = \Gamma(a-r)\,\Gamma(b+r)/\{\Gamma(a)\,\Gamma(b)\} \qquad (r<a)$$

$$E[(1+x)^r] = \Gamma(a-r)\,\Gamma(a+b)/\{\Gamma(a+b-r)\,\Gamma(a)\} \quad (r<a).$$

The r^{th} moments of S_i and $S_i+\nu\tau$ are then

$$E[S_i^r] = (\nu\tau)^r \frac{(f+2r-2)(f+2r-4)...f}{(\nu-2)(\nu-4)...(\nu-2r)}, \qquad 0<r<\nu/2,$$

$$E[(1+x)^r] = (\nu\tau)^r \frac{(\nu+f-2)(\nu+f-4)...\nu}{(\nu-2)(\nu-4)...(\nu-2r)}, \qquad 0<r<\nu/2,$$

and the moments of S_i^{-1} and $(S_i+\nu\tau)^{-1}$ are

$$E[S_i^{-r}] = (\nu\tau)^{-r} \frac{(\nu+2r-2)(\nu+2r-4)...\nu}{(f+2r)(f+2r-2)...(f-2)}, \qquad r>0,$$

$$E[(S_i+\nu\tau)^{-r}] = (\nu\tau)^{-r} \frac{(\nu+2r-2)(\nu+2r-4)...\nu}{(\nu+f+2r-2)(\nu+f+2r-4)...(\nu+f)}, \qquad r>0.$$

From the expressions relating cumulants and moments we calculate the first four cumulants of the distribution of S_i,

$$\kappa_1 = \frac{\nu\tau f}{(\nu-2)}$$

$$\kappa_2 = \frac{2(\nu\tau)^2 f(f+\nu-2)}{(\nu-2)^2(\nu-4)}, \tag{A.1}$$

$$\kappa_3 = (\nu\tau)^3\left\{\frac{f(f+2)(f+4)}{(\nu-2)(\nu-4)(\nu-6)} - \frac{3f^2(f+2)}{(\nu-2)^2(\nu-4)} + \frac{2f^2}{(\nu-2)^3}\right\},$$

and

$$\kappa_4 = \frac{(\nu\tau)^4 f}{(\nu-2)}\left\{\frac{(f+2)(f+4)(f+6)}{(\nu-4)(\nu-6)(\nu-8)} - \frac{4f(f+2)(f+4)}{(\nu-2)(\nu-4)(\nu-6)} + \frac{6f^2(f+2)}{(\nu-2)^2(\nu-4)} - \frac{3f^3}{(\nu-2)^3}\right\} - 3\kappa_2^2.$$

The r^{th} cumulant is not defined for $\nu \leq 2r$. It can be shown that as $\nu \to \infty$ the moments tend to the moments of a variable distributed as $\tau\chi^2(f)$ and as $f \to \infty$ to the moments of an inverse gamma distribution, as is to be expected.

The tail of the distribution depends on the term $(S_i/\nu\tau)^{-\nu/2+1}$; for small values of ν the distribution has a very long tail.

The mode of (2.5) is $\nu\tau(f-2)/(\nu+2)$; for $f=2$ the mode is zero and for $f<2$ the density function is *J*-shaped with mode at zero.

Appendix B. Second order bias calculations

Let $l(\theta,Y) = \log f_Y(Y;\theta)$ be the log likelihood function for $\theta=\theta^1,...,\theta^p$ based on the observations Y. The partial derivatives of l with respect to the components of θ may be written as

$$U_r(\theta) = \partial l(\theta,Y)/\partial\theta^r,$$

$$U_{rs}(\theta) = \partial^2 l(\theta,Y)/\partial\theta^r\partial\theta^s; \qquad r,s=1,...,p,$$

and so on. The maximun likelihood estimate of θ satisfies $U_r(\hat{\theta}) = 0$.

For simplicity we use here the index notation described in McCullagh (1987) where summation over any index is implicitly implied when this index is repeated once as a superscript and once as subscript. The range of summation is not explicitly stated but it is clear from the context.

The first order terms of the expansion of $U_r(\hat{\theta})$ are

$$0 = U_r(\hat{\theta})=U_r(\theta) + (\hat{\theta}^s - \theta^s)U_{rs}(\theta) +(1/2)(\hat{\theta}^s - \theta^s)(\hat{\theta}^t - \theta^t)U_{rst}(\theta) + o_p(1/n) \qquad (B.1)$$

Taking expectation we have

$$\begin{aligned} 0 &= E\{(\hat{\theta}^s - \theta^s)U_{rs}(\theta)\} + E\{(1/2)(\hat{\theta}^s - \theta^s)(\hat{\theta}^t - \theta^t)U_{rst}(\theta)\} + o(1/n) \\ &= E\{(\hat{\theta}^s - \theta^s)\}E\{U_{rs}(\theta)\} + \delta_s Cov\{(\hat{\theta}^s - \theta^s), U_{rs}(\theta)\} + \\ &\quad (1/2)E\{(\hat{\theta}^s - \theta^s)(\hat{\theta}^t - \theta^t)\}E\{U_{rst}(\theta)\} + o(1/n), \end{aligned} \qquad (B.2)$$

where $Cov\{(\hat{\theta}^s - \theta^s)(\hat{\theta}^t - \theta^t), U_{rst}(\theta)\}$ is $o(1/n)$.

Assuming the usual regularity conditions for the likelihood function we have from (B.1) that

$$U_r(\theta) = (\hat{\theta}^s - \theta^s)U_{rs}(\theta) + o_p(1)$$

and from the weak law of large numbers that

$$U_{rs}(\theta) \xrightarrow{p} -I_{rs}(\theta),$$

where $I_{rs}(\theta) = -E\{U_{rs}(\theta)\}$, is the Fisher information matrix of θ. We can then write

$$U_r(\theta) = (\hat{\theta}^s - \theta^s)I_{rs}(\theta) + o_p(1). \qquad (B.3)$$

If $I^{rs}(\theta)$ is the inverse matrix of $I_{rs}(\theta)$, inverting expression (B.3) it follows that the maximum likelihood estimate is

$$(\hat{\theta}^s - \theta^s) = U_r(\theta)\, I^{rs}(\theta) + o_p(1). \qquad (B.4)$$

Expression (B.2) can the be written as

$$a(\hat{\theta}^s)\}I_{rs}(\theta) = I^{ts}(\theta)\ \mathrm{Cov}\{U_t(\theta), U_{rs}(\theta)\} + (1/2)\ I^{ts}(\theta)\ E\{U_{rst}(\theta)\} + o(1/n),$$

where $a(\theta^s)$ is the bias of $\hat{\theta}^s$. Since the index s in the above expression indicates the summution we can write the left hand side as $a(\theta^v)\}I_{rv}(\theta)$ and invert to get an expression for the bias as

$$a(\hat{\theta}^v)\}=I^{rv}(\theta)\ I^{ts}(\theta)\ \{\mathrm{Cov}\{U_t(\theta), U_{rs}(\theta)\} + (1/2)\ E[U_{rst}(\theta)]\} + o(1/n). \qquad (B.5)$$

The calculations for the bias of $\hat{\tau}$ are substantially simplified since the information matrix $I_{rs}(\theta)$ is a diagonal matrix. Write $\theta^1=m$ and $\theta^2=\tau$, for the calculations of the bias of $\hat{\theta}^2$ we have v=2, therefore r=2, and then either t=s=1 or t=s=2. Thus

$$a(\hat{\theta}^2) = [I^{22}(\theta)]^2\{\mathrm{Cov}[U_2(\theta), U_{22}(\theta)] + (1/2)\ E[U_{222}(\theta)]\} +$$
$$I^{22}(\theta)I^{11}(\theta)\{\mathrm{Cov}[U_1(\theta), U_{12}(\theta)] + (1/2)\ E[U_{211}(\theta)]\}$$

The first term of the above equation correspond to the bias of $\hat{\theta}^2$ when θ^1 is known and it is not difficult to show that it is zero. The expressions in the secondterm are

$$\mathrm{Cov}[U_1(\theta), U_{12}(\theta) = E\{(\partial l/\partial m)(\partial^2 l/\partial m\partial\tau)\} = -\Sigma\ 1/(V_i+\tau)^2,$$
$$E[U_{211}(\theta)]\} = E\{(\partial^3 l/\partial m^2\partial\tau)\} = \Sigma\ 1/(V_i+\tau)^2.$$

From the covariance matrix of θ, (3.7) we then have that the bias of $\hat{\tau}$ is

$$a(\hat{\tau}) = -\{\Sigma\ 1/(V_i+\tau)\}^{-1}.$$

The maximum likelihood estimate of m is in this case an unbiased estimate.

Apendix C. Moments of the dispersion index I (2.42)

The moments of I (2.42) are calculated using the fact that the statistics $S_i/\Sigma S_j$ (i=1,...,k) are distributed according to a Dirichlet distribution and for equal sample size f, simple calculations give the first two moments of I,

$$\mu_1(I) = \frac{f(k-1)}{kf+2} = 1 - \frac{(f+2)}{fk} + O(k^{-2})$$

and

$$\mu_2(I) = \left(\frac{fk}{2}\right)^2 \left\{ \frac{(f+6)(f+4)(f+2)}{(kf+6)(kf+4)(kf+2)} + \frac{(k-1)(f+2)^2 f}{(kf+6)(kf+4)(kf+2)} - \frac{(f+2)^2}{(kf+2)^2} \right\}$$

$$= \frac{2(f+2)}{fk} + O(k^{-2}).$$

Appendix D. Proof of expression (4.6)

Expression (4.6) can be written as

$$E\left\{h_j\left(\frac{\mu-\xi}{\delta}\right) h_r\left(\frac{\mu-\mu_N}{V_N^{1/2}}\right) \middle| \bar{x}\right\} = \int \phi(z)\, H_j\left(\frac{z-a}{b}\right) H_r(z)\, dz \qquad (D.1)$$

where $z=(\mu-\mu_N)/V_N^{1/2}$, $a=(\xi-\mu_N)/V_N^{1/2}$ and $b=\delta/V_N^{1/2}$.

Because of the orthogonality properties of the Hermite polynomials and from expansion,

$$H_j\left(\frac{z-a}{b}\right) = \sum_{l=0}^{j} c_{j,l}\, H_l(z) \qquad (D.2)$$

it can be shown that

$$E\left\{h_j\left(\frac{\mu-\xi}{\delta}\right) h_r\left(\frac{\mu-\mu_N}{V_N^{1/2}}\right) \middle| \bar{x}\right\} = r!\, c_{j,r}.$$

In order to calculate the coeffients $c_{j,r}$ of expression (D.2) we use the generating function of the Hermite polynomials

$$h(z;\, t) = \exp\{zt - t^2/2\}.$$

The generating function for $x=(z-a)/b$ is then

$$h(x;\, t) = \exp\left\{t\left(\frac{z-a}{b}\right) - t^2/2\right\} = h(z;\, t/b)\, h\left\{-\frac{a}{b}\left(1-\frac{1}{b^2}\right)^{-1/2};\, t\,(1-1/b^2)^{1/2}\right\}. \qquad D.3)$$

The Hermite polynomial of order k given by the k^{th} derivative of (D.3) at t=0, is

$$H_j(x) = \sum_{l=0}^{j} \binom{j}{l} \frac{H_l(z)}{b^l} H_{j-l}\left(-\frac{a}{(b^2-1)^{1/2}}\right) \left(1-\frac{1}{b^2}\right)^{j-l}. \qquad (D.4)$$

Thus substituting the values of a and b in (D.4) we get

$$c_{j,r} = \binom{j}{r} \frac{V_N^{r/2}(\delta^2 - V_N)^{(j-r)/2}}{\delta^j} H_{j-r}\left(-\frac{(\xi-\mu_N)}{(\delta^2-V_N)^{1/2}}\right).$$

References

Abramovitz, M. & Stegun, I.A. (1965), eds. *Handbook of Mathematical Functions*, National Bureau of Standards, U.S. Government Printing Office, Washington, D.C.

Bartlett, M.S. & Kendall, D.G. (1946). The statistical analysis of variance heterogeneity and the logarithm transformation. *J. R. Statist. Soc. Suppl.*, **8**, 128-38.

Barndorff-Nielsen, O. & Cox, D. R. (1979). Edgeworth and sadlle-point approximations with statistical applications. *J. R. Statist. Soc.* B **41**, 279-312.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. Springer-Verlag, New York.

Cook, R.D. & Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, New York.

Copas, J. B.(1988). Binary regression models for contaminated data (with discussion). *J. R. Statist. Soc.* B, **50**, 225-65.

Cox, D. R.(1970). *The Analysis of Binary Data.* Chapman and Hall, London.

Cox, D. R.(1975a). A note on partially Bayes inference and the linear model. *Biometrika*, **62**, 651-54.

Cox, D.R.(1975b). Prediction intervals and empirical Bayes confidence intervals. In *essays in Probability and Statistics*, Ed. J. Gani. London: Academic Press.

Cox, D.R. (1983). Some remarks on overdispersion. *Biometrika*, **70**, 269-74.

Cox, D.R. & Hinkley, D.V. (1968). A note on the efficiency of least-squares estimates. *J. R. Statist. Soc.* B, **30**, 184-89.

Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Cox, D.R. & Solomon, P. (1986). Analysis of variability with large numbers of small samples. *Biometrika*, **73**, 543-54.

Dawid, A. P. (1973). Posterior expectations for large observations. *Biometrika*, **60**, 664- 66.

Deeley, J. J. & Lindley, D. V. (1981). Bayes empirical Bayes. *J. Amer. Statist. Assoc.*, **76**, 833- 41.

Dempster, A. P., Laird, N. M. & Rubin, D. B.(1976). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.*, **39**, 1-38.

Firth, D. (1987). Quasi-Likelihood estimation: efficiency and other aspects. *Biometrika*, **74**, 233-45.

Fisher, R A. & Yates, F. (1957) *Statistical Tables for Biological Agricultural and Medical Research*. Oliver & Boyd, London.

Geisser,S. (1984). On the prediction of observables: a selective update. In *Bayesian Statistics II*, J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.). North Holland, Amsterdam.

Hui, S. & Berger, J. O. (1983). Empirical Bayes estimation of rates in longitudinal studies. *J. Amer. Statist. Assoc.*, **78**, 753-60.

Johnson, N. L. & Kotz, S. (1969). *Discrete Distributions*. Houghton Miflin, Boston.

Johnson, N. L. & Kotz, S. (1970). *Continuous Univariate Distributions*, **1**. John Wiley & Sons, New York.

Kendall, M. G. & Stuart, A. (1977). *The Advanced Theory of Statistics* **1**, 4th edition. Griffin, London.

Kullback, S. & Leibler, R. A. (1951). On the information and sufficiency. *Annals of Mathematical Statistics*, **22**, 2154-58.

Laird, N.M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.*, **73**, 805-11.

Lindley, D.V. & Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. R. Statist. Soc.* B, **34**,1-41.

McCullagh, P. (1987). *Tensor Methods in Statistics*. Chapman and Hall, New York.

Maritz, J.S. (1970). *Empirical Bayes Methods*. Methuen, London.

Morris, C. N. (1982). Natural exponetial families with quadratic variance functions. *Ann. Statist.*, **10**, 65-80.

Morris, C. N.(1983a). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.*, **78**,47-65.

Morris, C. N. (1983b). Parametric Empirical Bayes Confidence Intervals. In *Scientific Inference, Data Analysis, and Robustness* (Box, G. E. P., Leonard, T. & Chien-Fu Wu, Eds.), Academic Press, London.

Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models. *J. R. Statist. Soc.* A, **135**, 370-84.

Palta, M. & Cook, T. (1987) Some considerations in the analysis of rates of change in longitudinal studies. *Statistics in Medicine*, **6**, 599-611.

Robbins, H. (1955). An empirical Bayes approach to statistics. In *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1**, 157-164. University of California Press, Berkeley.

Tierney, L. & Kadane, J. B. (1984). Accurate approximations for posterior moments and marginals densities. *J. Am. Statist. Assoc.*, **81**, 82-6.

Pearson, E. S. & Hartley, H. O. (1970). *Biometrika Tables for Statisticians*, **1**, Cambridge University Press, London.

Pearson K. (Ed.) (1934). *Tables of the Incomplete Beta Function*, 2nd edition, Cambridge University Press, London.