## CHAPTER 9

# THE LINEAR MODEL

**9.1** The Linear Model and its classical analysis are dealt with very fully in Volume 2. This chapter presents some corresponding Bayesian theory. Emphasis is on analysis of the normal linear model under various formulations of prior distributions.

### The normal linear model

**9.2** As in A**28.1**, we write the linear model in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{9.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations, $\mathbf{X}$ is an $n \times p$ matrix of known coefficients, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters and $\boldsymbol{\epsilon}$ an $n \times 1$ vector of random errors. The elements of $\boldsymbol{\epsilon}$ are assumed to have zero mean, to be uncorrelated and to have common variance $\sigma^2$, which is an additional parameter.

If we further assume that the elements of $\boldsymbol{\epsilon}$ are jointly normally distributed then the model is described as the *normal linear model.* The model says simply that the conditional distribution of $\mathbf{y}$ given parameters $(\boldsymbol{\beta}, \sigma^2)$ is the multivariate normal distribution $N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where as usual $\mathbf{I}$ denotes the $(n \times n)$ identity matrix. Therefore the likelihood becomes

$$f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\{-(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/(2\sigma^2)\}. \tag{9.2}$$

**9.3** We can write the quadratic form $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ in the exponent of (9.2) in various ways. Just expanding,

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \mathbf{y}'\mathbf{y} \tag{9.3}$$

If the $p \times p$ matrix $\mathbf{X}'\mathbf{X}$ is non-singular we can go on to complete the square to give

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + S \tag{9.4}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is the classical Maximum Likelihood or Least Squares estimator of $\boldsymbol{\beta}$, and $S = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ is the residual sum of squares.

### The normal-inverse-gamma distribution

**9.4** From (9.2) and (9.4) the natural conjugate family of prior distributions has the form

$$f(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-(d+p+2)/2} \exp[-\{(\boldsymbol{\beta} - \mathbf{m})'\mathbf{V}^{-1}(\boldsymbol{\beta} - \mathbf{m}) + a\}/(2\sigma^2)], \tag{9.5}$$

with hyperparameters $a$, $d$, $\mathbf{m}$ and $\mathbf{V}$. Possible values for the hyperparameters include all those values for which (9.5) is a proper distribution. The reason for using the power $(d + p + 2)/2$ for $\sigma^2$ will soon become apparent.

**9.5** We can first find the marginal prior distribution of $\sigma^2$ by integrating (9.5) with respect to $\boldsymbol{\beta}$.

$$\begin{aligned} f(\sigma^2) &= \int f(\boldsymbol{\beta}, \sigma^2)\,\mathrm{d}\boldsymbol{\beta} \\ &\propto (\sigma^2)^{-(d+p+2)/2} \exp\{-a/(2\sigma^2)\} |\sigma^2 \mathbf{V}|^{1/2} \\ &\propto (\sigma^2)^{-(d+2)/2} \exp\{-a/(2\sigma^2)\}. \end{aligned} \tag{9.6}$$

In doing this integration, the symmetric $p \times p$ matrix $\mathbf{V}^{-1}$ must be positive definite. Otherwise the integral diverges and (9.5) is not a proper distribution. A proper distribution must therefore have $\mathbf{V}$ positive definite.

The marginal distribution (9.6) is an *inverse-gamma* distribution, which is seen to be an appropriate name if we transform from $\sigma^2$ to $\phi = \sigma^{-2}$. Then (9.6) becomes

$$f(\phi) \propto \phi^{(d-2)/2} \exp(-a\phi/2). \tag{9.7}$$

This is a gamma distribution with hyperparameters $a/2$ and $d/2$. Alternatively, $a\phi$ has the chi-square distribution with $d$ degrees of freedom. It is proper if both hyperparameters are positive. Therefore for (9.5) to be proper we require $a > 0$, $d > 0$ as well as positive definite $\mathbf{V}$.

We will therefore say that (9.6) is the inverse gamma distribution with hyperparameters $a$ and $d$, and denote this by $IG(a, d)$. Its normalizing constant can be deduced from that of the gamma distribution (9.7), i.e.

$$f(\sigma^2) = \frac{(a/2)^{d/2}}{\Gamma(d/2)} (\sigma^2)^{-(d+2)/2} \exp\{-a/(2\sigma^2)\}. \tag{9.8}$$

**9.6** Summaries of this distribution may be found directly or via summaries of $\phi$. It is always unimodal with mode at $a/(d+2)$. Its mean is $E(\sigma^2) = a/(d-2)$ provided $d > 2$. If $2 \geqslant d > 0$ then the distribution is proper but the mean does not exist (and is effectively infinite). The mean is greater than the mode, reflecting the fact that the distribution is positively skewed. If $d > 4$, its variance is $\operatorname{var}(\sigma^2) = 2a^2/\{(d-2)^2(d-4)\}$.

**9.7** In integrating out $\boldsymbol{\beta}$ in **9.5** we implicitly used the fact that the conditional distribution of $\boldsymbol{\beta}$ given $\sigma^2$ is $N(\mathbf{m}, \sigma^2\mathbf{V})$. In particular

$$E(\boldsymbol{\beta} \mid \sigma^2) = \mathbf{m}, \qquad \operatorname{var}(\boldsymbol{\beta} \mid \sigma^2) = \sigma^2\mathbf{V}.$$

Therefore

$$E(\boldsymbol{\beta}) = E(E(\boldsymbol{\beta} \mid \sigma^2)) = \mathbf{m}, \tag{9.9}$$

$$\operatorname{var}(\boldsymbol{\beta}) = E(\operatorname{var}(\boldsymbol{\beta} \mid \sigma^2)) + \operatorname{var}(E(\boldsymbol{\beta} \mid \sigma^2)) = E(\sigma^2)\mathbf{V} = \{a/(d-2)\}\mathbf{V} \tag{9.10}$$

provided $d > 2$. Otherwise the variance of $\boldsymbol{\beta}$ is infinite.

**9.8** From the two integrations in **9.5** we derive the normalizing constant of (9.5). The joint distribution becomes

$$f(\boldsymbol{\beta},\sigma^2)=\frac{(a/2)^{d/2}}{(2\pi)^{p/2}|\mathbf{V}|^{1/2}\Gamma(d/2)}(\sigma^2)^{-(d+p+2)/2}\exp[-\{(\boldsymbol{\beta}-\mathbf{m})'\mathbf{V}^{-1}(\boldsymbol{\beta}-\mathbf{m})+a\}/(2\sigma^2)]. \quad (9.11)$$

We will call this the *normal-inverse-gamma* distribution with hyperparameters $a$, $d$, $\mathbf{m}$ and $\mathbf{V}$, and denote it by $NIG(a,d,\mathbf{m},\mathbf{V})$.

**9.9** To find the marginal distribution of $\boldsymbol{\beta}$ we integrate (9.11) with respect to $\sigma^2$. Notice that the conditional distribution of $\sigma^2$ given $\boldsymbol{\beta}$ is $IG((\boldsymbol{\beta}-\mathbf{m})'\mathbf{V}^{-1}(\boldsymbol{\beta}-\mathbf{m})+a,d+p)$. Therefore integrating using (9.8) we have

$$\begin{aligned}f(\boldsymbol{\beta})&=\frac{a^{d/2}\Gamma((d+p)/2)}{|\mathbf{V}|^{1/2}\pi^{p/2}\Gamma(d/2)}\{a+(\boldsymbol{\beta}-\mathbf{m})'\mathbf{V}^{-1}(\boldsymbol{\beta}-\mathbf{m})\}^{-(d+p)/2} \qquad (9.12)\\&\propto\{1+(\boldsymbol{\beta}-\mathbf{m})'(a\mathbf{V})^{-1}(\boldsymbol{\beta}-\mathbf{m})\}^{-(d+p)/2}.\end{aligned}$$

This is a generalization of the Student $t$ distribution in **A16.10**. We will call it the (multivariate) $t$ distribution with degrees of freedom $d$ and hyperparameters $\mathbf{m}$ and $a\mathbf{V}$, and denote it by $t_d(\mathbf{m},a\mathbf{V})$.

The distribution is symmetric around $\mathbf{m}$, with mean and variance given by (9.9) and (9.10).

### Conjugate analysis

**9.10** Now suppose that the $NIG(a,d,\mathbf{m},\mathbf{V})$ distribution (9.5) or (9.11) is adopted as the prior distribution for $(\boldsymbol{\beta},\sigma^2)$. Combining with the likelihood (9.2) gives the posterior distribution

$$f(\boldsymbol{\beta},\sigma^2\,|\,\mathbf{y})\propto(\sigma^2)^{-(d+n+p+2)/2}\exp\{-Q/(2\sigma^2)\}, \qquad (9.13)$$

where

$$\begin{aligned}Q&=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+(\boldsymbol{\beta}-\mathbf{m})'\mathbf{V}^{-1}(\boldsymbol{\beta}-\mathbf{m})+a \qquad (9.14)\\&=\boldsymbol{\beta}'(\mathbf{V}^{-1}+\mathbf{X}'\mathbf{X})\boldsymbol{\beta}-\boldsymbol{\beta}'(\mathbf{V}^{-1}\mathbf{m}+\mathbf{X}'\mathbf{y})-(\mathbf{m}'\mathbf{V}^{-1}+\mathbf{y}'\mathbf{X})\boldsymbol{\beta}+(\mathbf{m}'\mathbf{V}^{-1}\mathbf{m}+\mathbf{y}'\mathbf{y}+a)\\&=(\boldsymbol{\beta}-\mathbf{m}^\star)'(\mathbf{V}^\star)^{-1}(\boldsymbol{\beta}-\mathbf{m}^\star)+\mathbf{a}^\star, \qquad (9.15)\end{aligned}$$

and where

$$\mathbf{V}^\star=(\mathbf{V}^{-1}+\mathbf{X}'\mathbf{X})^{-1}, \qquad (9.16)$$

$$\mathbf{m}^\star=(\mathbf{V}^{-1}+\mathbf{X}'\mathbf{X})^{-1}(\mathbf{V}^{-1}\mathbf{m}+\mathbf{X}'\mathbf{y}), \qquad (9.17)$$

$$a^\star=a+\mathbf{m}'\mathbf{V}^{-1}\mathbf{m}+\mathbf{y}'\mathbf{y}-(\mathbf{m}^\star)'(\mathbf{V}^\star)^{-1}\mathbf{m}^\star. \qquad (9.18)$$

Therefore letting $d^\star=d+n$ the posterior distribution of $(\boldsymbol{\beta},\sigma^2)$ is $NIG(a^\star,d^\star,\mathbf{m}^\star,\mathbf{V}^\star)$. Summaries of this distribution are therefore immediately given by the results of **9.5** to **9.9**, simply changing $a$ to $a^\star$, $d$ to $d^\star$, $\mathbf{m}$ to $\mathbf{m}^\star$ and $\mathbf{V}$ to $\mathbf{V}^\star$. We now consider these in a little more detail.

**9.11** $E(\boldsymbol{\beta} \mid \mathbf{y}) = \mathbf{m}^\star$ is a posterior estimate of $\boldsymbol{\beta}$. In fact, since the posterior distribution of $\boldsymbol{\beta}$ is symmetric, this is also the posterior mode. If $\mathbf{X}'\mathbf{X}$ is non-singular, we can write

$$\mathbf{m}^\star = (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{V}^{-1}\mathbf{m} + \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}) = (\mathbf{I} - \mathbf{A})\mathbf{m} + \mathbf{A}\hat{\boldsymbol{\beta}}, \tag{9.19}$$

where $\mathbf{A} = (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}$. (9.19) expresses the posterior estimate $\mathbf{m}^\star$ as a matrix-weighted average of the prior mean $\mathbf{m}$ and the classical estimate $\hat{\boldsymbol{\beta}}$, with weights $\mathbf{I}-\mathbf{A}$ and $\mathbf{A}$. If prior information is strong, the elements of $\mathbf{V}$ will be small, reflecting small prior variances for the elements of $\boldsymbol{\beta}$. Then $\mathbf{V}^{-1}$ will be large and $\mathbf{A}$ small, so that the posterior mean gives most weight to the prior mean. Conversely, if prior information is weak or the data substantial then most weight will be given to $\hat{\boldsymbol{\beta}}$.

If $\mathbf{X}'\mathbf{X}$ is singular there is no unique solution $\hat{\boldsymbol{\beta}}$ to the classical Least Squares equations (see A**19.13**). The posterior distribution is nevertheless proper with mean (9.19) as long as the prior distribution is proper ($\mathbf{V}$ is positive definite).

**9.12** Posterior uncertainty about $\boldsymbol{\beta}$ is described in part by $\mathbf{V}^\star$. Thus $\operatorname{var}(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}) = \sigma^2\mathbf{V}^\star$ and $\operatorname{var}(\boldsymbol{\beta} \mid \mathbf{y}) = E(\sigma^2 \mid \mathbf{y})\mathbf{V}^\star$. $\mathbf{V}^\star$ also represents a combination of prior information and data. Now $(\mathbf{V}^\star)^{-1} = \mathbf{V}^{-1} + \mathbf{X}'\mathbf{X}$ is in some sense greater than $\mathbf{V}^{-1}$, so $\mathbf{V}^\star$ is 'smaller than' $\mathbf{V}$. The extra information from the data has reduced uncertainty about $\boldsymbol{\beta}$. We can make this vague argument precise as follows. Let $\phi = \mathbf{a}'\boldsymbol{\beta}$ be any linear combination of the elements of $\boldsymbol{\beta}$, $\mathbf{a} \neq \mathbf{0}$. Then

$$\begin{aligned}\sigma^{-2}\operatorname{var}(\phi \mid \sigma^2, \mathbf{y}) &= \mathbf{a}'\mathbf{V}^\star\mathbf{a} = \mathbf{a}'(\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{a} \\ &= \mathbf{a}'\{\mathbf{V} - \mathbf{V}\mathbf{X}'(\mathbf{I} + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1}\mathbf{X}\mathbf{V}\}\mathbf{a} \\ &= \mathbf{a}'\mathbf{V}\mathbf{a} - (\mathbf{X}\mathbf{V}\mathbf{a})'(\mathbf{I} + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1}(\mathbf{X}\mathbf{V}\mathbf{a}).\end{aligned} \tag{9.20}$$

Now $\mathbf{I} + \mathbf{X}\mathbf{V}\mathbf{X}'$ is positive definite, so $(\mathbf{I} + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1}$ is also positive definite. Therefore $\operatorname{var}(\phi \mid \sigma^2, \mathbf{y}) \leqslant \sigma^2\mathbf{a}'\mathbf{V}\mathbf{a} = \operatorname{var}(\phi \mid \sigma^2)$. $\operatorname{var}(\phi \mid \sigma^2, \mathbf{y})$ will be strictly less than $\operatorname{var}(\phi \mid \sigma^2)$ for all $\sigma^2$, showing a real reduction in uncertainty about $\phi$, if $\mathbf{X}\mathbf{V}\mathbf{a} \neq \mathbf{0}$. If $\mathbf{X}$ has full column rank, $\mathbf{X}\mathbf{V}\mathbf{a} = \mathbf{0}$ implies $\mathbf{a} = \mathbf{0}$, in which case the data $\mathbf{y}$ give a reduced variance for every linear function of $\boldsymbol{\beta}$. In particular the variance (given $\sigma^2$) of each element of $\boldsymbol{\beta}$ is reduced.

However, if the rank of $\mathbf{X}$ does not equal its number of columns, $p$, there will be some $\mathbf{a} \neq \mathbf{0}$ for which $\operatorname{var}(\phi \mid \sigma^2, \mathbf{y}) = \operatorname{var}(\phi \mid \sigma^2)$. The data provide no direct information about these functions of $\boldsymbol{\beta}$. This is the case of singular $\mathbf{X}'\mathbf{X}$, when classical Least Squares or Maximum Likelihood fail to produce a unique estimator $\hat{\boldsymbol{\beta}}$. With a proper prior distribution $\boldsymbol{\beta}$ has a proper posterior distribution with a unique posterior mean, but the absence of information in the data about certain aspects of $\boldsymbol{\beta}$ shows through in linear combinations $\mathbf{a}'\boldsymbol{\beta}$ for which $\mathbf{a}'\mathbf{V}^\star\mathbf{a} = \mathbf{a}'\mathbf{V}\mathbf{a}$. These are nonidentifiable in the sense of **3.15**.

*Example 9.1*

The simplest case of a linear model is when $p = 1$ and $\mathbf{X} = \mathbf{1}$, an $n \times 1$ vector of ones. $\boldsymbol{\beta}$ is a scalar, which we will denote by $\mu$, and (9.1) reduces to $y_i = \mu + \epsilon_i$. Therefore the $y_i$s are independent and identically distributed as $N(\mu, \sigma^2)$. Then $\mathbf{X}'\mathbf{X} = \mathbf{1}'\mathbf{1} = n$ and

$\hat{\boldsymbol{\beta}} = \hat{\mu} = n^{-1}\mathbf{1}'\mathbf{y} = n^{-1}\sum y_i = \bar{y}$ is the classical estimator of $\mu$. In the prior distribution, $\mathbf{m}$ and $\mathbf{V}$ reduce to the scalars $m$ and $v$. The posterior mean is

$$m^\star = (v^{-1} + n)^{-1}(v^{-1}m + n\bar{y}) = (1 - a)m + a\bar{y},$$

where $a = (v^{-1} + n)^{-1}n$ is the weight given to the data estimate $\bar{y}$. This weight is large if $n$ is large (strong data) or $v$ is large (weak prior). The posterior variance is $v^\star E(\sigma^2 \,|\, \mathbf{y})$ where $v^\star = (v^{-1} + n)^{-1}$.

*Example 9.2*

The simple regression model $y_i = \alpha + \beta x_i + \epsilon_i$ is a linear model with $p = 2$, $\boldsymbol{\beta} = (\alpha, \beta)'$ and $\mathbf{X} = (\mathbf{1}, \mathbf{x})$. Then

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, \qquad \mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}.$$

and $\hat{\boldsymbol{\beta}}$ is the usual least squares estimator for the simple regression model. The way in which this is modified by the prior information to give the posterior mean $\mathbf{m}^\star$ is no longer as simple as in Example 9.1. If the weight matrix $\mathbf{A}$ in (9.19) is diagonal, then the posterior mean of $\alpha$ will be a weighted average of its prior mean and the classical $\hat{\alpha}$, and the posterior mean of $\beta$ will similarly be a weighted average of its prior mean and $\hat{\beta}$. But this will not generally be the case, and the posterior mean of either $\alpha$ or $\beta$ will depend on both components of the prior mean $\mathbf{m}$ and both components of the classical estimate $\hat{\boldsymbol{\beta}}$.

**9.13** Information about $\boldsymbol{\beta}$ is also obtained indirectly through $E(\sigma^2 \,|\, \mathbf{y})$ in the formula $\text{var}(\boldsymbol{\beta} \,|\, \mathbf{y}) = E(\sigma^2 \,|\, \mathbf{y})\mathbf{V}^\star$. If the data $\mathbf{y}$ suggest that $\sigma^2$ is smaller than its prior estimate so that $E(\sigma^2 \,|\, \mathbf{y}) < E(\sigma^2)$, then $\text{var}(\mathbf{a}'\boldsymbol{\beta} \,|\, \mathbf{y}) = E(\sigma^2 \,|\, \mathbf{y})\mathbf{a}'\mathbf{V}^\star\mathbf{a} \leqslant E(\sigma^2 \,|\, \mathbf{y})\mathbf{a}'\mathbf{V}\mathbf{a} < E(\sigma^2)\mathbf{a}'\mathbf{V}\mathbf{a} = \text{var}(\mathbf{a}'\boldsymbol{\beta})$. Then the posterior variance of every linear function of $\boldsymbol{\beta}$ is less than its prior variance. If, on the other hand, $E(\sigma^2 \,|\, \mathbf{y}) > E(\sigma^2)$ then posterior uncertainty about some elements or functions of $\boldsymbol{\beta}$ may increase despite the reduction in $\mathbf{V}^\star$. Since $E(E(\sigma^2 \,|\, \mathbf{y})) = E(\sigma^2)$, the data are not expected *a priori* either to increase or decrease the expectation of $\sigma^2$, and so $\mathbf{V}^\star$ is the main determinant of posterior uncertainty about $\boldsymbol{\beta}$.

If, therefore, we wish to design an experiment with the primary purpose of estimating $\phi = \mathbf{a}'\boldsymbol{\beta}$, we would do so by choosing $\mathbf{X}$ to maximize the term $(\mathbf{XVa})'(\mathbf{I} + \mathbf{XVX}')^{-1}(\mathbf{XVa})$ in (9.20).

### Optimal design

**9.14** Various other criteria for design of experiments have been proposed. In practice we may have the loose objective of 'obtaining best possible information about $\boldsymbol{\beta}$', and it may be difficult to be so specific as minimizing the posterior variance of a single linear function $\phi = \mathbf{a}'\boldsymbol{\beta}$. As an overall measure of quality of information about $\boldsymbol{\beta}$, the determinant of its posterior variance matrix, or 'generalized variance', is one possibility. We may therefore consider minimizing $|\mathbf{V}^\star|$, or equivalently maximizing $|\mathbf{V}^\star|^{-1} = |\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X}|$. This approach has the drawback that $|\mathbf{V}^\star|$ can be made arbitrarily small, or even zero, by obtaining good information about any element or linear combination of $\boldsymbol{\beta}$, while posterior variances of other elements or combinations might still be large.

An alternative generalized criterion is to choose $\mathbf{X}$ to minimize the trace of $|\mathbf{V}^\star|$, which is the sum of the variances of the elements of $\boldsymbol{\beta}$. This criterion can only go to zero if *all* the variances go to zero, corresponding to perfect information about the whole of $\boldsymbol{\beta}$. On the other hand, this criterion is not invariant under linear reparametrization. If we let $\boldsymbol{\phi} = \mathbf{B}\boldsymbol{\beta}$, then the posterior variance matrix of $\boldsymbol{\phi}$ is $\mathbf{B}\mathbf{V}^{\star}\mathbf{B}'$. If $\mathbf{B}$ is non-singular, $|\mathbf{B}\mathbf{V}^{\star}\mathbf{B}'| = |\mathbf{B}|^2|\mathbf{V}^{\star}|$ and the criterion of minimizing $|\mathbf{V}^{\star}|$ is invariant in the sense that it does not matter whether we parametrize by $\boldsymbol{\beta}$ or $\boldsymbol{\phi}$. But minimizing $tr(\mathbf{B}\mathbf{V}^{\star}\mathbf{B}') = tr(\mathbf{V}^{\star}\mathbf{L})$, where $\mathbf{L} = \mathbf{B}'\mathbf{B}$, is not the same as minimizing $tr\mathbf{V}^{\star}$. Implicitly, the trace criterion sets $\mathbf{L} = \mathbf{I}$ or may be generalized by using any other specific $\mathbf{L}$. (9.15) is the case $\mathbf{L} = \mathbf{a}\mathbf{a}'$. Designs based on the trace and determinant criteria provide Bayesian alternatives to classical $A$-optimal and $D$-optimal designs, respectively. For references see **3.47**.

**9.15** We now examine the posterior mean $E(\sigma^2 \,|\, \mathbf{y})$ of $\sigma^2$. Note that if $\mathbf{X}'\mathbf{X}$ is non-singular we can rewrite (9.18) after a little algebra as

$$a^{\star} = a + (n-p)\hat{\sigma}^2 + (\mathbf{m} - \hat{\boldsymbol{\beta}})'\{\mathbf{V} + (\mathbf{X}'\mathbf{X})^{-1}\}^{-1}(\mathbf{m} - \hat{\boldsymbol{\beta}}), \tag{9.21}$$

where $\hat{\sigma}^2$ is the classical unbiased estimator of $\sigma^2$ as derived in A**19.9**. That is $(n-p)\hat{\sigma}^2 = \mathbf{y}'\{\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\mathbf{y}' = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = S$, the residual sum of squares in classical theory. To interpret the last term in (9.21) we note the following.

$$\begin{aligned} E(\hat{\boldsymbol{\beta}} \,|\, \sigma^2) &= E\{E(\hat{\boldsymbol{\beta}} \,|\, \boldsymbol{\beta}, \sigma^2) \,|\, \sigma^2\} = E(\boldsymbol{\beta} \,|\, \sigma^2) = \mathbf{m}, \\ \operatorname{var}(\hat{\boldsymbol{\beta}} \,|\, \sigma^2) &= E\{\operatorname{var}(\hat{\boldsymbol{\beta}} \,|\, \boldsymbol{\beta}, \sigma^2) \,|\, \sigma^2\} + \operatorname{var}\{E(\hat{\boldsymbol{\beta}} \,|\, \boldsymbol{\beta}, \sigma^2) \,|\, \sigma^2\} \\ &= E\{\sigma^2(\mathbf{X}'\mathbf{X})^{-1} \,|\, \sigma^2\} + \operatorname{var}(\boldsymbol{\beta} \,|\, \sigma^2) = \sigma^2\{(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{V}\}. \end{aligned}$$

$$\begin{aligned} \therefore E[(\mathbf{m} - \hat{\boldsymbol{\beta}})'&\{\mathbf{V} + (\mathbf{X}'\mathbf{X})^{-1}\}^{-1}(\mathbf{m} - \hat{\boldsymbol{\beta}}) \,|\, \sigma^2] \\ &= E[tr\{\mathbf{V} + (\mathbf{X}'\mathbf{X})^{-1}\}^{-1}(\mathbf{m} - \hat{\boldsymbol{\beta}})(\mathbf{m} - \hat{\boldsymbol{\beta}})' \,|\, \sigma^2] \\ &= tr[\{\mathbf{V} + (\mathbf{X}'\mathbf{X})^{-1}\}^{-1}E\{(\mathbf{m} - \hat{\boldsymbol{\beta}})(\mathbf{m} - \hat{\boldsymbol{\beta}})' \,|\, \sigma^2\}] \\ &= tr[\{\mathbf{V} + (\mathbf{X}'\mathbf{X})^{-1}\}^{-1}\sigma^2\{\mathbf{V} + (\mathbf{X}'\mathbf{X})^{-1}\}] \\ &= tr(\sigma^2\mathbf{I}) = p\sigma^2. \end{aligned} \tag{9.22}$$

Therefore

$$E(\sigma^2 \,|\, \mathbf{y}) = \frac{d-2}{d+n-2}E(\sigma^2) + \frac{n-p}{d+n-2}\hat{\sigma}^2 + \frac{p}{d+n-2}t \tag{9.23}$$

in which $t = p^{-1}(\mathbf{m} - \hat{\boldsymbol{\beta}})'\{\mathbf{V} + (\mathbf{X}'\mathbf{X})^{-1}\}^{-1}(\mathbf{m} - \boldsymbol{\beta})$, which from (9.22) has expectation $\sigma^2$. So the posterior mean of $\sigma^2$ is a weighted average of three estimates. The first is the prior mean, the second is the standard classical estimate, and the third is the estimate derived above. The third estimate arises from comparing the prior and classical estimators, $\mathbf{m}$ and $\hat{\boldsymbol{\beta}}$, for $\boldsymbol{\beta}$. Since the prior variance of $\boldsymbol{\beta}$ given $\sigma^2$ is $\sigma^2\mathbf{V}$, a large discrepancy between the two estimates of $\boldsymbol{\beta}$ suggests that the prior estimate may have been poor, which in turn suggests that $\sigma^2$ is large.

The relative weights given to the three estimates in (9.23), i.e. $d-2$, $n-p$ and $p$, reflect the strengths of these information sources. The strength of the prior information of $\sigma^2$ itself is shown by $d-2$, since for a fixed $E(\sigma^2)$ increasing $d$ reduces $\operatorname{var}(\sigma^2)$. The strength of the classical estimate $\hat{\sigma}^2$ is denoted by the classical degrees of freedom, $n-p$. Finally, the third source of information about $\sigma^2$ comes from comparisons between $\mathbf{m}$ and $\hat{\boldsymbol{\beta}}$, and the weight $p$ given to this is the number of dimensions in which these comparisons are made. As the number of observations increases, the weight for the classical estimate increases and $E(\sigma^2 \mid \mathbf{y}) \to \hat{\sigma}^2$.

Notice that essentially the same results apply if we look at the posterior mode rather than $E(\sigma^2 \mid \mathbf{y})$. The weight given to the prior mode is $d+2$ instead of $d-2$, but this is the only change. The posterior mode also tends to $\hat{\sigma}^2$ as $n \to \infty$.

**9.16** $\text{var}(\sigma^2 \mid \mathbf{y}) = (a^\star)^2/\{(d^\star - 2)^2(d^\star - 4)\} = E(\sigma^2 \mid \mathbf{y})^2/(d^\star - 4)$. So the relative variance $\text{var}(\sigma^2 \mid \mathbf{y})/E(\sigma^2 \mid \mathbf{y})^2$, which is the square of the posterior coefficient of variation of $\sigma^2$, is simply $(d^\star - 4)^{-1}$. Since $d^\star > d$, the information from the data decreases the coefficient of variation, representing a reduction in uncertainty about $\sigma^2$ relative to its mean.

### Weak prior information

**9.17** We can represent weak prior information about $(\boldsymbol{\beta}, \sigma^2)$ within the conjugate family by letting prior variances tend to infinity. Letting the prior variances of elements of $\boldsymbol{\beta}$ tend to infinity results in $\mathbf{V}^{-1} \to \mathbf{0}$. Setting $\mathbf{V}^{-1} = \mathbf{0}$ in (9.5) produces

$$f(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-(d+p+2)/2} \exp\{-a/(2\sigma^2)\}. \tag{9.24}$$

In this expression, $\boldsymbol{\beta}$ has an improper uniform distribution and $\sigma^2$ has the $IG(a, d+p)$ distribution. The conventional improper prior distribution $f(\sigma^2) \propto \sigma^{-2}$, which is often recommended for positive parameters, is now obtained by setting $a = 0$, $d = -p$. Then

$$f(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}. \tag{9.25}$$

This is not the only way in which weak prior information can be formulated. In particular, if we begin by looking at $\sigma^2$, its marginal prior distribution is $IG(a, d)$, which we can equate to the $f(\sigma^2) \propto \sigma^{-2}$ form by letting $a = 0$, $d = 0$. Then

$$f(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-(p+2)/2} \exp\{-(\boldsymbol{\beta} - \mathbf{m})'\mathbf{V}^{-1}(\boldsymbol{\beta} - \mathbf{m})/(2\sigma^2)\}. \tag{9.26}$$

If we now let $\mathbf{V}^{-1} \to \mathbf{0}$ we obtain the alternative form

$$f(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-(p+2)}. \tag{9.27}$$

This can be shown to be the Jeffreys prior distribution for the normal Linear Model.

These two representations both set $\mathbf{V}^{-1} = \mathbf{0}$ and $a = 0$, but give alternative values of $-p$ and 0 to $d$. It is also reasonable to argue for the uniform prior $f(\sigma^2) \propto 1$ instead of $f(\sigma^2) \propto \sigma^{-2}$ giving values $d = -(p+2)$ or $d = -2$. We shall take the view, as in **4.35**, that if the different formulations lead to essentially the same posterior inference it clearly does not matter which we use. If different weak prior distributions lead to important differences in posterior inference it is necessary to think about prior information instead of adopting any standard formula.

**9.18** We now consider the effect of weak prior information on the hyperparameters of the posterior $NIG(a^\star, d^\star, \mathbf{m}^\star, \mathbf{V}^\star)$ distribution. Letting $\mathbf{V}^{-1} \to \mathbf{0}$ results in $\mathbf{V}^\star = (\mathbf{X}'\mathbf{X})^{-1}$, $\mathbf{m}^\star = \hat{\boldsymbol{\beta}}$, $a^\star = a + (n-p)\hat{\sigma}^2$. Now it is necessary that $\mathbf{X}'\mathbf{X}$ be non-singular, otherwise the posterior distribution becomes improper. When prior information about $\boldsymbol{\beta}$ is very weak, the data must provide information about all elements and linear functions of $\boldsymbol{\beta}$.

The posterior distribution of $\boldsymbol{\beta}$ given $\sigma^2$ is now $N(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, which is directly analogous to classical inference in which the estimator $\hat{\boldsymbol{\beta}}$ is normally distributed with mean $\boldsymbol{\beta}$ and variance $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The Bayesian estimate is the same as the classical estimate $\hat{\boldsymbol{\beta}}$, and its accuracy is described in the same terms, via a normal distribution with variance $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. However, the prior distribution (9.24) clearly provides information about $\sigma^2$, and this is reflected in the posterior distribution of $\sigma^2$.

$$E(\sigma^2 \mid \mathbf{y}) = a^\star/(d^\star - 2) = \frac{d+p-2}{d+n-2}E(\sigma^2) + \frac{n-p}{d+n-2}\hat{\sigma}^2 \tag{9.28}$$

is a weighted average now of only two estimates. The third term comparing estimates of $\boldsymbol{\beta}$ disappears because there is no strength in the prior information about $\boldsymbol{\beta}$ and hence no comparison to make. The weight on the prior mean $E(\sigma^2)$ increases to $d+p-2$ instead. Strictly, (9.24) is improper as a joint distribution for $(\boldsymbol{\beta}, \sigma^2)$, but if we regard it as a proper $IG(a, d+p)$ marginal distribution for $\sigma^2$ times an improper distribution for $\boldsymbol{\beta}$ then we have the value $E(\sigma^2) = a/(d+p-2)$ used in (9.28).

**9.19** Now letting $a \to 0$ gives $a^\star = (n-p)\hat{\sigma}^2$ and $E(\sigma^2 \mid \mathbf{y}) = \{(n-p)/(d+n-2)\}\hat{\sigma}^2$, which takes different values depending on which value we use for $d$ in representing weak prior information. Setting $d = 2-p$ would lead to agreement between $E(\sigma^2 \mid \mathbf{y})$ and the classical estimate $\hat{\sigma}^2$, but $f(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-4}$ is not generally advocated for any other reasons. Since the posterior distribution of $\sigma^2$ is skew, $E(\sigma^2 \mid \mathbf{y})$ is not the only estimate which might be considered. The mode $\{(n-p)/(d+n+2)\}\hat{\sigma}^2$ or the median, which will lie between mean and mode, might also be used. Values $d = -p$ or $-(p+2)$ will therefore produce posterior estimates which are close to the classical $\hat{\sigma}^2$. The alternatives $d = 0$ or $-2$ will also produce similar results if $n$ is much larger than $p$, but otherwise the different prior formulations (9.25) and (9.27) will not yield essentially the same posterior inferences.

## Interval estimation

**9.20** The discussion so far has concentrated on point estimates such as posterior means, together with posterior variances to indicate the strength of posterior information. Another useful form of inference is an interval estimate. Bayesian interval estimation in the form of highest density regions is considered in **2.50**, and we now develop highest posterior density regions for $\boldsymbol{\beta}$ and $\sigma^2$. In general, a highest posterior density interval for $\theta$ is a region $C$ such that the posterior density for $\theta$ is higher at all points in $C$ than at any point outside $C$. $C$ is therefore bounded by a contour of the posterior density function. If the posterior probability that $\theta \in C$ is $p$, then $C$ is made the smallest region amongst all those that contain $\theta$ with the same probability $p$. (Alternatively, $p$ is higher than the probability that $\theta$ lies in any other region of the same size as $C$.)

**9.21** Turning first to interval estimation for the entire $\boldsymbol{\beta}$ vector, the posterior density for $\boldsymbol{\beta}$ is the multivariate $t$ distribution (9.12) but substituting $a^\star$ for $a$, $d^\star$ for $d$, $\mathbf{m}^\star$ for $\mathbf{m}$ and $\mathbf{V}^\star$ for $\mathbf{V}$. Contours of this density are values of $\boldsymbol{\beta}$ such that $f(\boldsymbol{\beta} \mid \mathbf{y})$ is constant, and are therefore the ellipsoids for which $(\boldsymbol{\beta}-\mathbf{m}^\star)'(\mathbf{V}^\star)^{-1}(\boldsymbol{\beta}-\mathbf{m}^\star)$ is constant. A highest posterior density region for $\boldsymbol{\beta}$ therefore takes the form

$$C=\{\boldsymbol{\beta}:(\boldsymbol{\beta}-\mathbf{m}^\star)'(\mathbf{V}^\star)^{-1}(\boldsymbol{\beta}-\mathbf{m}^\star)\leqslant c\}$$

Then

$$P(\boldsymbol{\beta}\in C\,|\,\mathbf{y})=P((\boldsymbol{\beta}-\mathbf{m}^\star)'(\mathbf{V}^\star)^{-1}(\boldsymbol{\beta}-\mathbf{m}^\star)\leqslant c\,|\,\mathbf{y}) \tag{9.29}$$

and we wish to evaluate this probability and find $c$ such that it equals some appropriate value, such as 0.9 or 0.99.

Now since the posterior distribution of $\boldsymbol{\beta}$ given $\sigma^2$ is $N(\mathbf{m}^\star,\sigma^2\mathbf{V}^\star)$, the posterior distribution of $\phi=\sigma^{-2}(\boldsymbol{\beta}-\mathbf{m}^\star)'(\mathbf{V}^\star)^{-1}(\boldsymbol{\beta}-\mathbf{m}^\star)$ given $\sigma^2$ is $\chi^2_p$, independent of $\sigma^2$ (see **A15.11** and **A15.21**). And since the posterior distribution of $\sigma^2$ is $IG(a^\star,d^\star)$, the relationship between $IG$ and $\chi^2$ distributions developed in **9.5** shows that $a^\star\sigma^{-2}$ is distributed as $\chi^2_{d^\star}$. Therefore $(p^{-1}\phi)/((d^\star)^{-1}a^\star\sigma^{-2})=\{d^\star/(pa^\star)\}(\boldsymbol{\beta}-\mathbf{m}^\star)'(\mathbf{V}^\star)^{-1}(\boldsymbol{\beta}-\mathbf{m}^\star)$ has the $F$ distribution with degrees of freedom $p$ and $d^\star=d+n$. We can now solve (9.29), with $P(\boldsymbol{\beta}\in C\,|\,\mathbf{y})=P(F_{p,d+n}\leqslant(d^\star c)/(pa^\star))$.

We can therefore determine a highest posterior density region $C$ with given probability $1-\alpha$ of containing $\boldsymbol{\beta}$. Letting $F$ be the upper $100\alpha\%$ point of the $F_{p,d+n}$ distribution,

$$C=\{\boldsymbol{\beta}:(\boldsymbol{\beta}-\mathbf{m}^\star)'(\mathbf{V}^\star)^{-1}(\boldsymbol{\beta}-\mathbf{m}^\star)\leqslant pa^\star F/d^\star\}. \tag{9.30}$$

The form of this region is the interior of an ellipsoid centred at $\mathbf{m}^\star$ and with shape matrix proportional to $\mathbf{V}^\star$. Its principal axes are therefore given by the principal components of $\mathbf{V}^\star$ (see **2.15**).

**9.22** Now consider an arbitrary linear transformation $\boldsymbol{\Phi}=\mathbf{A}\boldsymbol{\beta}$, where $\mathbf{A}$ is an $r\times p$ matrix of rank $r$. Then the conditional posterior distribution of $\boldsymbol{\Phi}$ given $\sigma^2$ is $N(\mathbf{Am}^\star,\sigma^2\mathbf{AV}^\star\mathbf{A}')$. The marginal posterior distribution of $\boldsymbol{\Phi}$ will therefore be $t_{d^\star}(\mathbf{Am}^\star,a^\star\mathbf{AV}^\star\mathbf{A}')$. We can immediately deduce the highest posterior density region by analogy with (9.30):

$$C=\{\gamma:(\gamma-\mathbf{Am}^\star)'(\mathbf{AV}^\star\mathbf{A}')^{-1}(\gamma-\mathbf{Am}^\star)\leqslant ra^\star F/d^\star\}, \tag{9.31}$$

where now $F$ is the upper $100\alpha\%$ point of $F_{r,d+n}$. In particular, if $m_i^\star$ is the $i$th element of $\mathbf{m}^\star$ and $v_i^\star$ is the $i$th diagonal element of $\mathbf{V}^\star$, we have the following highest density region for an individual $\beta_i$.

$$C=\{\beta_i:(\beta_i-m_i^\star)^2/v_i^\star\leqslant a^\star F/d^\star\}. \tag{9.32}$$

Now $F$ is the upper $100\alpha\%$ point of $F_{1,d+n}$, so that $F=t^2$, where $t$ is the upper $50\alpha\%$ point of the Student $t$ distribution $t_{d+n}$. So (9.32) becomes a highest posterior density *interval*

$$\begin{aligned}C&=\{\beta_i:|\beta_i-m_i^\star|\leqslant(a^\star v_i^\star/d^\star)^{1/2}t\}\\&=[m_i^\star-(a^\star v_i^\star/d^\star)^{1/2}t,\ m_i^\star+(a^\star v_i^\star/d^\star)^{1/2}t].\end{aligned} \tag{9.33}$$

This corresponds to the fact that the posterior distribution of $\beta_i$ is $t_{d+n}(m_i^\star,a^\star v_i^\star)$ and hence $(\beta_i-m_i^\star)(a^\star v_i^\star/d^\star)^{-1/2}$ has the standard $t_{d+n}$ distribution.

**9.23** In the case of weak prior information, we have seen that $\mathbf{m}^\star = \hat{\boldsymbol{\beta}}$, $\mathbf{V}^\star = (\mathbf{X}'\mathbf{X})^{-1}$ and $a^\star = (n-p)\hat{\sigma}^2$. Then (9.27) becomes the interval $\hat{\beta}_i \pm \{(n-p)/(d+n)\}^{1/2} a_i^{1/2} \hat{\sigma} t$, where $a_i$ is the $i$th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. If we also choose $d = -p$ to represent weak prior information, corresponding to the prior distribution (9.20), the interval is exactly equal to the standard classical confidence interval (A28.27). The more general ellipsoidal region (9.25) is also then a classical confidence region.

A different value of $d$ will not exactly reproduce classical confidence regions, but if $n$ is much larger than $p$ the difference will be negligible.

*Example 9.3*

Following Example 9.1, the posterior marginal distribution of the population mean $\mu$ is a $t$ distribution, and a highest posterior density interval for $\mu$ is, from (9.33)

$$m^\star \pm (a^\star v^\star / d^\star)^{1/2} t,$$

where $t$ is an appropriate upper percentage point of the $t_{d+n}$ distribution. In the case of weak prior information with $v^{-1} = 0$, $a = 0$, $d = -1$ (since $p = 1$), we have the standard classical confidence interval

$$\bar{y} \pm \hat{\sigma} n^{-1/2} t,$$

where $t$ is a percentage point of the $t_{n-1}$ distribution and $\hat{\sigma}^2 = (n-1)^{-1} \sum (y_i - \bar{y})^2$ is the usual unbiased estimator of $\sigma^2$.

*Example 9.4*

With the simple regression model of Example 9.2, we could use the general result (9.30) to construct an elliptical highest posterior density region for $(\alpha, \beta)$. Another use is to construct a highest posterior density interval using (9.31) for $\gamma = \alpha + \beta x$, which is the value of the regression line at the point $x$. If we write

$$\mathbf{m}^\star = \begin{pmatrix} a \\ b \end{pmatrix}, \qquad \mathbf{V}^\star = \begin{pmatrix} v_a & c \\ c & v_b \end{pmatrix},$$

then (9.31) reduces (by a similar argument to the derivation of (9.33)) to the interval

$$a + bx \pm (a^\star v(x) / d^\star)^{1/2} t, \tag{9.34}$$

where

$$v(x) = v_a + 2cx + v_b x^2 \tag{9.35}$$

and $t$ is an appropriate percentage point of $t_{d+n}$. Plotting the interval (9.34) as a function of $x$ will yield hyperbolic limits for the regression line, as in Figure 9.1, where the posterior mean of the regression line $y = E(\alpha + \beta x \mid \mathbf{y}) = a + bx$ is also shown passing through the middle of the intervals.

The width of the interval is minimized at $x = c/v_b$ and becomes wider on either side of this minimum because of uncertainty about the true slope $\beta$ of the line.

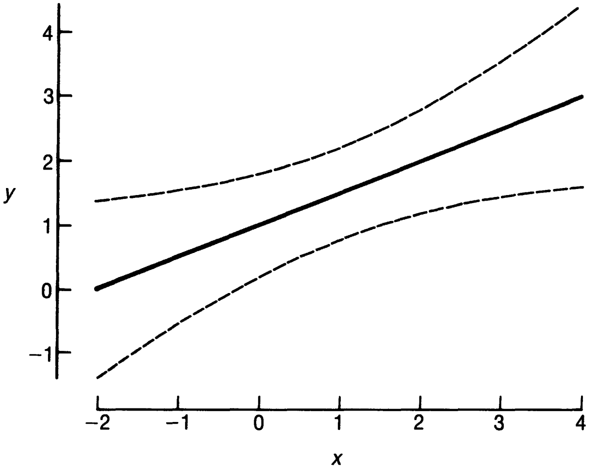

**Fig. 9.1 Typical highest posterior density intervals for a simple regression line**

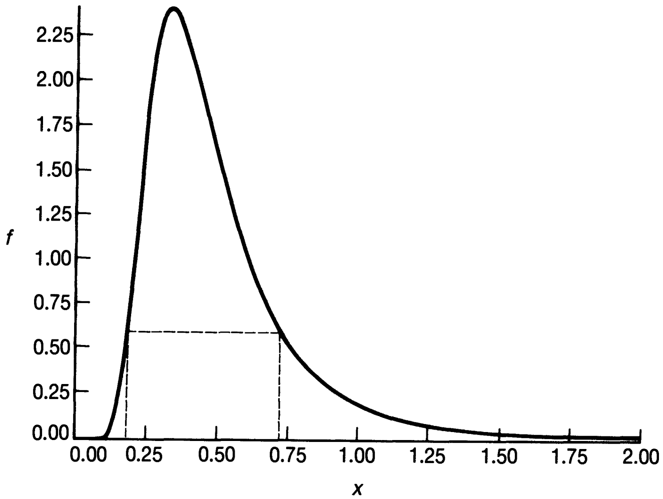

**Fig. 9.2 Highest posterior density interval for $IG(4, 10)$**

**9.24** A highest posterior density interval for $\sigma^2$ is straightforward to construct. The posterior density is the $IG$ density (9.6) with $a$ changed to $a^\star$ and $d$ changed to $d^\star$. The density is unimodal so the highest density region is an interval $[s_1, s_2]$ as in Figure 9.2 such that $f(\sigma^2 = s_1 \mid \mathbf{y}) = f(\sigma^2 = s_2 \mid \mathbf{y})$. However, the density is not symmetric and values of $s_1$ and $s_2$ to obtain an interval with given probability must be obtained numerically. Novick and Jackson (1974) give tables of these highest density regions.

In the case of weak prior information, $a = 0$, $d = -p$ gives $a^\star = (n-p)\hat{\sigma}^2$, $d^\star = n - p$. Then $(n-p)\hat{\sigma}^2\sigma^{-2}$ has a $\chi^2_{n-p}$ posterior distribution which agrees exactly with the classical distribution for $\hat{\sigma}^2$ (given $\sigma^2$). Therefore the highest posterior density interval is also a classical confidence interval. It does not, however, coincide with the usual confidence interval for this problem, because it is clear from Figure 9.1 that the probabilities in the two tails $(0, s_1)$ and $(s_2, \infty)$ are not equal. The shortest confidence interval in Exercise A20.5 is similar to the highest posterior density interval, in being bounded by points of equal density, but with respect to the $\chi^2_{n-p+4}$ density.

### Conditional distributions

**9.25** We have derived the marginal distributions of $\boldsymbol{\beta}$ and $\sigma^2$, and in **9.21** the marginal distribution of an arbitrary linear transform $\boldsymbol{\gamma} = \mathbf{A}\boldsymbol{\beta}$. We have also found conditional distributions for $\boldsymbol{\beta}$ given $\sigma^2$ and for $\sigma^2$ given $\boldsymbol{\beta}$, but now consider conditional distributions given partial specification of $\boldsymbol{\beta}$. First let $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$, and consider distributions conditional on $\boldsymbol{\beta}_2$. Suppose that $(\boldsymbol{\beta}, \sigma^2)$ has the $NIG(a, d, \mathbf{m}, \mathbf{V})$ distribution. Corresponding posterior distributions result if we change $a$ to $a^\star$, $d$ to $d^\star$, $\mathbf{m}$ to $\mathbf{m}^\star$ and $\mathbf{V}$ to $\mathbf{V}^\star$. If $\boldsymbol{\beta}_2$ has $r$ elements write

$$\mathbf{m} = \begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{pmatrix}, \qquad \mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix},$$

where $\mathbf{m}_2$ is $r \times 1$ and $\mathbf{V}_{22}$ is $r \times r$. Now since $\boldsymbol{\beta}$ given $\sigma^2$ is distributed as $N(\mathbf{m}, \sigma^2\mathbf{V})$ we have the following distributions (using general results on multivariate normal distributions, as in A**15.4** and Exercise A15.1)

$$(\boldsymbol{\beta}_2 \mid \sigma^2): \quad N(\mathbf{m}_2, \sigma^2\mathbf{V}_{22}), \tag{9.36}$$

$$(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2, \sigma^2): \quad N(\mathbf{m}_{1.2}, \sigma^2\mathbf{V}_{11.2}), \tag{9.37}$$

where $\mathbf{m}_{1.2} = \mathbf{m}_1 + \mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\boldsymbol{\beta}_2 - \mathbf{m}_2)$ and $\mathbf{V}_{11.2} = \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}$.

From (9.36) and the $IG(a, d)$ marginal distribution of $\sigma^2$ we have the distribution

$$(\boldsymbol{\beta}_2, \sigma^2): \quad NIG(a, d, \mathbf{m}_2, \mathbf{V}_{22})$$

and hence

$$\boldsymbol{\beta}_2: \quad t_d(\mathbf{m}_2, \mathbf{V}_{22}),$$

$$(\sigma^2 \mid \boldsymbol{\beta}_2): \quad IG(a_2, d + r), \tag{9.38}$$

where

$$a_2 = a + (\boldsymbol{\beta}_2 - \mathbf{m}_2)'\mathbf{V}_{22}^{-1}(\boldsymbol{\beta}_2 - \mathbf{m}_2). \tag{9.39}$$

Now (9.37) and (9.38) together give

$$(\boldsymbol{\beta}_1, \sigma^2 \mid \boldsymbol{\beta}_2): \quad NIG(a_2, d + r, \mathbf{m}_{1.2}, \mathbf{V}_{11.2})$$

and finally

$$(\boldsymbol{\beta}_1 \mid \boldsymbol{\beta}_2): \quad t_{d+r}(\mathbf{m}_{1.2}, a_2\mathbf{V}_{11.2}) \tag{9.40}$$

### The general linear hypothesis

**9.26** Consider the hypothesis that $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$, where $\mathbf{A}$ is an $r \times p$ matrix of rank $r$ and $\mathbf{c}$ is any $r \times 1$ vector of constants. Classical testing of this hypothesis is considered in

**A23.25** to **A23.29**. We consider it from a Bayesian point of view in two ways. First suppose that the conjugate prior distribution $NIG(a, d, \mathbf{m}, \mathbf{V})$ is appropriate. This gives zero prior probability to the hypothesis, and so this prior distribution does not formally treat the hypothesis as having a positive probability of being true, and it is not sensible to ask for its posterior probability. Instead we ask more informally whether the posterior distribution suggests that $\mathbf{c}$ is a relatively probable or improbable value for $\gamma = \mathbf{A}\boldsymbol{\beta}$. The posterior distribution of $\gamma$ was found in **9.22** to be $t_{d^\star}(\mathbf{Am}^\star, a^\star\mathbf{AV}^\star\mathbf{A}')$. The ratio of the posterior density at $\gamma = \mathbf{c}$ to its value at the mode $\gamma = \mathbf{Am}^\star$ is

$$\{1 + (\mathbf{c} - \mathbf{Am}^\star)'(a^\star\mathbf{AV}^\star\mathbf{A}')^{-1}(\mathbf{c} - \mathbf{Am}^\star)\}^{-(d^\star+r)/2}$$

and if this is small we should regard $\mathbf{c}$ as a relatively implausible value for $\mathbf{A}\boldsymbol{\beta}$.

To quantify this further, note that the set of all values of $\gamma$ having posterior density higher than at $\gamma = \mathbf{c}$ is a highest posterior density region. Using the argument of **9.22**, the posterior probability that $\gamma$ lies outside this set is

$$\begin{aligned} &P\{f(\gamma \mid \mathbf{y}) \leqslant f(\gamma = \mathbf{c} \mid \mathbf{y}) \mid \mathbf{y}\} \\ &\quad = P\{(\gamma - \mathbf{Am}^\star)'(\mathbf{AV}^\star\mathbf{A}')^{-1}(\gamma - \mathbf{Am}^\star) \geqslant (\mathbf{c} - \mathbf{Am}^\star)'(\mathbf{AV}^\star\mathbf{A}')^{-1}(\mathbf{c} - \mathbf{Am}^\star) \mid \mathbf{y}\} \\ &\quad = P\{F_{r,d+n} \geqslant d^\star(\mathbf{c} - \mathbf{Am}^\star)'(\mathbf{AV}^\star\mathbf{A}')^{-1}(\mathbf{c} - \mathbf{Am}^\star)/(ra^\star)\}, \qquad (9.41) \end{aligned}$$

which can be calculated from tables of the F distribution with $r$ and $d + n$ degrees of freedom. If this is small we can regard the hypothesis $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ as implausible in the sense that $\mathbf{A}\boldsymbol{\beta}$ has a small posterior probability of being so far from its posterior mean $\mathbf{Am}^\star$. We can think of this as an informal Bayesian test of the general linear hypothesis. It is similar to the classical idea of rejecting a hypothesis if the hypothesized value does not lie in a confidence interval with sufficiently large confidence coefficient.

**9.27** In the case of weak prior information we let $\mathbf{V}^{-1} = \mathbf{0}$ and $a = 0$, obtaining $\mathbf{V}^\star = (\mathbf{X}'\mathbf{X})^{-1}$, $\mathbf{m}^\star = \hat{\boldsymbol{\beta}}$ and $a^\star = (n - p)\hat{\sigma}^2$ as discussed in **9.17**. Then (9.41) becomes

$$P\{f(\gamma \mid \mathbf{y}) \leqslant f(\gamma = \mathbf{c} \mid \mathbf{y}) \mid \mathbf{y}\} = P\{F_{r,d+n} \geqslant (d + n)F/(n + p)\}. \qquad (9.42)$$

where

$$F = (\mathbf{c} - \mathbf{A}\hat{\boldsymbol{\beta}})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\mathbf{c} - \mathbf{A}\hat{\boldsymbol{\beta}})/(r\hat{\sigma}^2)$$

is the standard classical test statistic for the general linear hypothesis (developed in a less explicit form in **A23.28**). Several possible values of $d$ are discussed in **9.17**, all claiming to represent weak prior information in some sense. The case $d = -p$, corresponding to the improper prior distribution $f(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$, makes (9.42) the classical observed significance probability, so that the informal Bayesian test then agrees exactly with the usual classical significance test.

**9.28** An alternative Bayesian approach is to give a non-zero prior probability to the hypothesis, and apply the theory of **7.41** and **7.42** for comparing nested models. We would then wish to calculate the Bayes factor $B = f(\mathbf{y} \mid \gamma = \mathbf{c})/f(\mathbf{y})$. We will consider this approach within a general treatment of Bayes factors for comparing linear models.

### Bayes factors for linear models

**9.29** Suppose that in addition to the original model (9.1) we have an alternative model $\mathbf{y} = \mathbf{X}_A\boldsymbol{\beta}_A + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon}$ distributed as $N(\mathbf{0}, \sigma^2\mathbf{I})$ as before. The prior distribution for $(\boldsymbol{\beta}_A, \sigma^2)$ is $NIG(a, d, \mathbf{m}_A, \mathbf{V}_A)$. The two models make the same assumptions about the error term $\boldsymbol{\epsilon}$, including the same $IG(a, d)$ prior distribution for $\sigma^2$. They differ in the matrices $\mathbf{X}$ and $\mathbf{X}_A$ of coefficients, and so try to explain or predict the response variable $\mathbf{y}$ using different regressor variables. Accordingly, they have different parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_A$.

The Bayes factor in favour of the alternative model is the ratio $B = f_A(\mathbf{y})/f(\mathbf{y})$ of the resulting marginal densities for $\mathbf{y}$ under the two models. The denominator is obtained as follows. From (9.2) and (9.11).

$$\begin{aligned} f(\mathbf{y}) &= \int\int f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\beta}, \sigma^2)\, \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\sigma^2 \\ &= k \int\int (\sigma^2)^{-(d+n+p+2)/2} \exp\{-Q/(2\sigma^2)\}\, \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\sigma^2, \end{aligned} \tag{9.43}$$

where

$$k = \frac{(a/2)^{d/2}}{(2\pi)^{(n+p)/2}|\mathbf{V}|^{1/2}\Gamma(d/2)}$$

and $Q$ is given by (9.14). Now the equivalent expression (9.15) allows us to do the integration with respect to $\boldsymbol{\beta}$ in (9.43), to yield

$$\begin{aligned} f(\mathbf{y}) &= k|\mathbf{V}^\star|^{1/2}(2\pi)^{p/2} \int (\sigma^2)^{-(d^\star+2)/2} \exp\{-a^\star/(2\sigma^2)\}\, \mathrm{d}\sigma^2 \\ &= k|\mathbf{V}^\star|^{1/2}(2\pi)^{p/2}(a^\star/2)^{-d^\star/2}\Gamma(d^\star/2) \\ &= \frac{|\mathbf{V}^\star|^{1/2}a^{d/2}\Gamma(d^\star/2)}{|\mathbf{V}|^{1/2}\pi^{n/2}\Gamma(d/2)}(a^\star)^{-d^\star/2}. \end{aligned} \tag{9.44}$$

Notice that $\mathbf{y}$ only appears in (9.44) through $a^\star$. The rest of the expression is the normalizing constant for $f(\mathbf{y})$.

**9.30** The analogous expression for $f_A(\mathbf{y})$ adds subscript $A$ to $\mathbf{V}$, $\mathbf{V}^\star$ and $\mathbf{a}^\star$, so that the Bayes factor is

$$B = \frac{|\mathbf{V}|^{1/2}|\mathbf{V}_A^\star|^{1/2}}{|\mathbf{V}_A|^{1/2}|\mathbf{V}^\star|^{1/2}} \cdot \left(\frac{a^\star}{a_A^\star}\right)^{d^\star/2}. \tag{9.45}$$

The four determinants do not depend on the observed data $\mathbf{y}$, and are concerned with the relative strength of prior information and data information about the parameter vectors, as measured by $\mathbf{V}$, $\mathbf{V}_A$, $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}_A'\mathbf{X}_A$. The term involving $\mathbf{y}$ is an increasing function of $a^\star/a_A^\star$, and so favours the alternative model if it leads to a smaller $a_A^\star$ than the original model's $a^\star$. Since $d^\star = d + n$ is the same in both models, the Bayes factor tends to favour the model producing the lower posterior estimate of $\sigma^2$. This is intuitively reasonable since $\sigma^2$ determines the magnitude of the errors $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ or $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}_A\boldsymbol{\beta}_A$ and so measures the lack of fit of the model to the data. An estimate such as $E(\sigma^2 \mid \mathbf{y}) = a^\star/(d^\star - 2)$ of $\sigma^2$ estimates this lack of fit.

**9.31** We have two expressions for $a^\star$ in (9.18) and (9.21). Several others can be derived. For instance, (9.15) reduces to $Q = a^\star$ if $\boldsymbol{\beta} = \mathbf{m}^\star$, and making this substitution in (9.14) gives

$$a^\star = a + (\mathbf{y} - \mathbf{X}\mathbf{m}^\star)'(\mathbf{y} - \mathbf{X}\mathbf{m}^\star) + (\mathbf{m}^\star - \mathbf{m})'\mathbf{V}^{-1}(\mathbf{m}^\star - \mathbf{m}). \tag{9.46}$$

This is similar to (9.21) but here we have a Bayesian residual sum of squares $(\mathbf{y} - \mathbf{X}\mathbf{m}^\star)'(\mathbf{y} - \mathbf{X}\mathbf{m}^\star)$ in place of the classical $(n-p)\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$. The classical residual vector $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is replaced by Bayesian residuals which are differences between the observations $\mathbf{y}$ and the posterior model fit $\mathbf{X}E(\boldsymbol{\beta} \mid \mathbf{y}) = \mathbf{X}\mathbf{m}^\star$. Similarly, the last term in (9.46) is a comparison between prior and posterior means, instead of the comparison between the prior mean and the classical estimate $\hat{\boldsymbol{\beta}}$ which appears in (9.21).

We obtain another expression for $a^\star$ by deriving $f(\mathbf{y})$ differently. The model (9.1) expresses $\mathbf{y}$ as a sum of $\mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is distributed as $N(\mathbf{0}, \sigma^2\mathbf{I})$ given $\sigma^2$, and $\boldsymbol{\beta}$ is distributed as $N(\mathbf{m}, \sigma^2\mathbf{V})$, also given $\sigma^2$. Therefore the distribution of $\mathbf{y}$ given $\sigma^2$ is $N(\mathbf{X}\mathbf{m}, \sigma^2(\mathbf{I} + \mathbf{X}\mathbf{V}\mathbf{X}'))$. The prior distribution of $\sigma^2$ being $IG(a, d)$, it follows that the joint distribution of $(\mathbf{y}, \sigma^2)$ is $NIG(a, d, \mathbf{X}\mathbf{m}, \mathbf{I} + \mathbf{X}\mathbf{V}\mathbf{X}')$. Therefore the marginal distribution of $\mathbf{y}$ is $t_d(\mathbf{X}\mathbf{m}, a(\mathbf{I} + \mathbf{X}\mathbf{V}\mathbf{X}'))$. By comparing (9.12) in this case with (9.44),

$$a^\star = a + (\mathbf{y} - \mathbf{X}\mathbf{m})'(\mathbf{I} + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\mathbf{m}). \tag{9.47}$$

(We also obtain the matrix identity $|\mathbf{I} + \mathbf{X}\mathbf{V}\mathbf{X}'| = |\mathbf{V}||\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X}|$).

Finally, using (9.17),

$$\mathbf{y} - \mathbf{X}\mathbf{m} = \mathbf{y} - \mathbf{X}\mathbf{m}^\star + \mathbf{X}(\mathbf{m}^\star - \mathbf{m}) = \mathbf{y} - \mathbf{X}\mathbf{m}^\star + \mathbf{X}(\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{m}),$$

so that

$$\mathbf{y} - \mathbf{X}\mathbf{m} = \{\mathbf{I} - \mathbf{X}(\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{m}^\star) = (\mathbf{I} + \mathbf{X}\mathbf{V}\mathbf{X}')(\mathbf{y} - \mathbf{X}\mathbf{m}^\star).$$

Therefore

$$a^\star = a + (\mathbf{y} - \mathbf{X}\mathbf{m}^\star)(\mathbf{I} + \mathbf{X}\mathbf{V}\mathbf{X}')(\mathbf{y} - \mathbf{X}\mathbf{m}^\star). \tag{9.48}$$

**9.32** The various expressions for $a^\star$ show various ways of looking at lack of fit of the model to the data. (9.21) looks at it in terms of the classical residuals and discrepancy between prior and classical estimates of $\boldsymbol{\beta}$. (9.46) expresses lack of fit in terms of the Bayesian residuals and discrepancy between prior and posterior estimates of $\boldsymbol{\beta}$. (9.47) uses a combined measure based on discrepancy between $\mathbf{y}$ and its prior estimate $\mathbf{X}\mathbf{m}$, and (9.48) shows that this can also be turned into a single measure involving the Bayesian residuals (but inflated by the use of $\mathbf{I} + \mathbf{X}\mathbf{V}\mathbf{X}'$ instead of $\mathbf{I}$).

**9.33** If the alternative model is a special case of the original linear model then the models are nested. This arises when the columns of $\mathbf{X}_A$ are all linear combinations of the columns of $\mathbf{X}$, so that $\mathbf{X}_A = \mathbf{X}\mathbf{B}$, where $\mathbf{B}$ is a $p \times p_A$ matrix with $p_A < p$. Then the alternative model says $\mathbf{y} = \mathbf{X}\mathbf{B}\boldsymbol{\beta}_A + \boldsymbol{\epsilon}$, a special case of the original model in which $\boldsymbol{\beta} = \mathbf{B}\boldsymbol{\beta}_A$. Since $p_A < p$ we can define an $r \times p$ matrix $\mathbf{A}$, where $r = p - p_A$, such that $\mathbf{A}\mathbf{B} = \mathbf{0}$. Then the alternative model can be seen as asserting that $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$, a form of the general linear hypothesis. In fact the more general hypothesis $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ corresponds to an alternative model of the form $\mathbf{y} = \mathbf{X}_A\boldsymbol{\beta}_A + \mathbf{X}\mathbf{d} + \boldsymbol{\epsilon}$, when the constant vector $\mathbf{d}$ is any solution of the equation $\mathbf{A}\mathbf{d} = \mathbf{c}$. This is the same alternative model if we simply redefine the response variable to be $\mathbf{y}_A = \mathbf{y} - \mathbf{X}\mathbf{d}$ instead of $\mathbf{y}$. We can go on to show by analogy to (9.39) that the general linear hypothesis results in an alternative model for which

$$a_A^\star = a^\star + (\mathbf{c} - \mathbf{A}\mathbf{m}^\star)'(\mathbf{A}\mathbf{V}^\star\mathbf{A}')^{-1}(\mathbf{c} - \mathbf{A}\mathbf{m}^\star),$$

and hence that the Bayes factor (9.45) depends on $\mathbf{y}$ through the same criterion as the informal Bayesian hypothesis test (9.41).

However, an extra complication is introduced by the fact that $\boldsymbol{\beta}$ and $\sigma^2$ are not independent in the natural conjugate prior distribution. If $(\boldsymbol{\beta}, \sigma^2)$ have the $NIG(a, d, \mathbf{m}, \mathbf{V})$ prior distribution we could follow the development of **9.25** to derive the conditional prior distribution of $(\boldsymbol{\beta}, \sigma^2)$ given the hypothesis $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$. This is a normal-inverse-gamma distribution, but its first two hyperparameters are not $a$ and $d$. The conditioning information $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ provides information about $\sigma^2$ and thereby changes its distribution. This is different from our assumption in **9.29** that the prior marginal distribution of $\sigma^2$ under both models is the same. Consequently we obtain a different result from (9.45) if we begin with a $NIG(a, d, \mathbf{m}, \mathbf{V})$ prior distribution and derive the Bayes factor $B = f(\mathbf{y} \mid \mathbf{A}\boldsymbol{\beta} = \mathbf{c})/f(\mathbf{y})$ as suggested in **7.42** for comparing nested models. This can be seen as an undesirable feature of the normal-inverse-gamma distribution, and the Bayes factor (9.45) seems preferable even in the case of nested models.

### Bayes factors with weak prior information

**9.34** Now consider weak prior information specified by $\mathbf{V}^{-1} = \mathbf{0}$ and $a = 0$, so that $\mathbf{V}^\star = (\mathbf{X}'\mathbf{X})^{-1}$ and $a^\star = (n-p)\hat{\sigma}^2$. For the alternative model we also let $\mathbf{V}_A^{-1} = \mathbf{0}$, yielding $\mathbf{V}_A^\star = (\mathbf{X}_A'\mathbf{X}_A)^{-1}$, and $a_A^\star = (n - p_A)\hat{\sigma}_A^2$ is the residual sum of squares under the alternative model. The Bayes factor (9.45) is indeterminate because of the terms $|\mathbf{V}|^{1/2}$ and $|\mathbf{V}_A|^{1/2}$, both of which are infinite. This is the difficulty noted in **7.43** of using Bayes factors with improper prior distributions. As in **7.54**, we will let their ratio be an undetermined constant $c$ and write

$$B = c\frac{|\mathbf{X}'\mathbf{X}|^{1/2}}{|\mathbf{X}_A'\mathbf{X}_A|^{1/2}}\left(\frac{(n-p)\hat{\sigma}^2}{(n-p_A)\hat{\sigma}_A^2}\right)^{d^\star/2}. \tag{9.49}$$

This Bayes factor depends on the data through the ratio of residual sums of squares $\{(n-p)\hat{\sigma}^2\}/\{(n-p_A)\hat{\sigma}_A^2\}$, which is simply the classical likelihood ratio test statistic (A23.99).

In the case of nested models we can write (9.49) as

$$B = c\frac{|\mathbf{X}'\mathbf{X}|^{1/2}}{|\mathbf{X}_A'\mathbf{X}_A|^{1/2}}\left(1 + \frac{r}{n-p}F\right)^{-d^\star/2}, \tag{9.50}$$

where $F$ is the classical test statistic

$$F = \frac{(n-p_A)\hat{\sigma}^2 - (n-p)\hat{\sigma}^2}{r\hat{\sigma}^2} \tag{9.51}$$

as derived in **A23.28**. A large value of $F$ will lead to a small Bayes factor in favour of the alternative model, and so cause us to favour the original model. This is analogous to the classical procedure in which a large $F$ causes the hypothesis, represented by the alternative model, to be rejected.

**9.35** In order to use the Bayes factor (9.50) in the case of weak prior information we must specify the constant $c$. The approach of **7.56** was advocated by Spiegelhalter and Smith (1982) as a way of specifying $c$ when comparing nested models.

The method requires You first to specify a minimal experiment such that proper posterior distributions are obtained under both the original and alternative models. In general, this will mean an experiment with $n = p + 1$ observations. This is certainly the case for the most commonly used weak prior distribution in which $d = -p$, since then the degrees of freedom $d^* = d + n$ of the posterior inverse-gamma distribution will not be positive for any smaller $n$. In general, $p + 1$ observations allow us to estimate the $p+1$ components of the parameter vector $(\boldsymbol{\beta}, \sigma^2)$. Let $\mathbf{E}$ be the $\mathbf{X}$ matrix for this minimal experiment, under the original model, and let $\mathbf{E}_A$ be the corresponding $\mathbf{X}_A$ matrix for the alternative model.

We now choose the data arising from that hypothetical minimal experiment to maximize $B$. Since the data $\mathbf{y}$ only affect $B$ through the test statistic $F$, choose $\mathbf{y}$ to obtain $F = 0$. Then equate the resulting value of $B$ to one, and solve for $c$:

$$c = |\mathbf{E}_A'\mathbf{E}_A|^{1/2}/|\mathbf{E}'\mathbf{E}|^{1/2}. \tag{9.52}$$

We then insert this value of $c$ into (9.50) to obtain the Spiegelhalter and Smith Bayes factor for the actual data.

*Example 9.5*

If $x_1, x_2, \ldots, x_n$ are identically and independently distributed as $N(\mu, \sigma^2)$ we have the simple linear model of Example 9.1 $\mathbf{X}'\mathbf{X} = n$, and for a minimal experiment we require two observations, so $\mathbf{E}'\mathbf{E} = 2$. A hypothesis that $\mu = c$ corresponds to the alternative model $y_i - c = \epsilon_i$, where there is no $\boldsymbol{\beta}_A$ vector or $\mathbf{X}_A$ matrix. The effect is to set $|\mathbf{X}_A'\mathbf{X}_A| = 1$ (or equivalently we can treat it as undefined and merge it with $c$). Therefore, from (9.52) we have $c = 1/\sqrt{2}$ and a Bayes factor of

$$B = (n/2)^{1/2}\{1 + F/(n-1)\}^{-d^*/2},$$

where in this case $F = n(\bar{y} - c)^2/\hat{\sigma}^2$ with $\hat{\sigma}^2 = (n-1)^{-1}\sum(y_i - \bar{y})^2$.

*Example 9.6*

If the observations follow the simple repression model $y_i = \alpha + \beta x_i + \epsilon_i$ of Example 9.2 then $|\mathbf{X}'\mathbf{X}| = \sum(x_i - \bar{x})^2$. A minimal experiment will have three observations, which we suppose have values $e_i$, $e_2$, $e_3$, of the regressor variable. If the alternative model is that $\beta = 0$, it reduces to independent $N(\alpha, \sigma^2)$ observations and $|\mathbf{X}_A'\mathbf{X}_A| = n$. (9.37) now gives $c = (n/\sum_{i=1}^{3}(e_i - \bar{e})^2)$, but this is not uniquely determined because there is not a unique minimal experiment.

**9.36** In general there may be many possible experiments of minimal size. This is particularly true for regression models, where $\mathbf{X}$ depends on the values of the regressor variables. Spiegelhalter and Smith (1982) deal with regression problems by supposing that not just the observations $\mathbf{y}$ but also the design $\mathbf{X}$ of an experiment of minimal size $n = p+1$

should be chosen to maximize $B$, before then obtaining $c$ by letting $B = 1$. This seems to run counter to their argument that we set $B = 1$ because a minimal experiment should not be able to provide more than negligible support for the alternative model. Choosing $\mathbf{X}$ to maximize $B$ makes the experiment maximally informative. It is less defensible to set $B = 1$ when the data $\mathbf{y}$ are chosen to give maximal support to the alternative model in a maximally informative experiment, even if that experiment has a minimal number of observations. It might even be argued, conversely, that a truly minimal experiment should have $\mathbf{X}$ chosen to *minimize* $B$. The unfortunate consequence of that proposal would be to let $|E'E| \to 0$. (Although it must be strictly positive to obtain a proper posterior distribution, we can make it arbitrarily small.) Therefore $c \to \infty$ if we follow this idea. As the following example shows, there is typically no upper bound on $|\mathbf{X}'\mathbf{X}|$ either, so that Spiegelhalter and Smith's approach leads to the opposite extreme of $c = 0$. In general, their method appears unsuitable for problems where alternative minimal experiments give different possible values for $c$.

*Example 9.7*

Following Example 9.6, we can choose $e_1$, $e_2$ and $e_3$ to obtain an arbitrarily small value of $\sum(e_i - \bar{e})^2$, and have $c \to \infty$, if the minimal experiment is to be minimally informative. For it to be maximally informative, let $\sum(e_i - \bar{e})^2 \to \infty$ and hence $c \to 0$. Spiegelhalter and Smith (1982) actually proposed maximizing $\sum(e_i - \bar{e})^2$ subject to the constraint that $|e_i| \leqslant 1$ for $i = 1, 2, 3$. This gives a solution $\sum(e_i - \bar{e})^2 = 24/9$ and $c = (27/24)^{1/2}$, but the constraint seems to be completely arbitrary.

**Fractional Bayes factors for linear models**

**9.37** An alternative approach presented in **7.62** is to use the fractional Bayes factor based on using a proportion $b$ of the data as a training sample. We specify the prior distribution as $f(\boldsymbol{\beta}, \sigma^2) \propto g(\boldsymbol{\beta}, \sigma^2)$, where $g$ is an improper prior distribution to represent weak prior information. We shall use $g(\boldsymbol{\beta}, \sigma^2) = \sigma^{-2}$ in both models, but the results are easily adapted to other powers of $\sigma^2$ corresponding to other values of the prior degrees of freedom parameter $d$. Under this approach we take as our definition of $f(\mathbf{y})$ the expression

$$f(\mathbf{y}) = \frac{\int\int f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) g(\boldsymbol{\beta}, \sigma^2)\, \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\sigma^2}{\int\int \{f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)\}^b g(\boldsymbol{\beta}, \sigma^2)\, \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\sigma^2}. \tag{9.53}$$

With $g(\boldsymbol{\beta}, \sigma^2) = \sigma^{-2}$, the denominator is

$$I_b = \int\int (2\pi\sigma^2)^{-nb/2} \exp\{-b(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/(2\sigma^2)\}\sigma^{-2}\, \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\sigma^2.$$

Using (9.4) we have

$$I_b = (2\pi)^{-nb/2} \int (\sigma^2)^{-(nb/2)-1} \exp\{-b(n-p)\hat{\sigma}^2/(2\sigma^2)\} J_b\, \mathrm{d}\sigma^2, \tag{9.54}$$

where $(n-p)\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ is the residual sum of squares as before, and

$$\begin{aligned} J_b &= \int \exp\{-b(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/(2\sigma^2)\}\, \mathrm{d}\boldsymbol{\beta} \\ &= (2\pi)^{p/2} |\sigma^2 b^{-1} (\mathbf{X}'\mathbf{X})^{-1}|^{1/2} = (2\pi)^{p/2} (\sigma^2)^{p/2} b^{-p/2} |\mathbf{X}'\mathbf{X}|^{-1/2}. \end{aligned}$$

Substituting into (9.54) gives

$$I_b = (2\pi)^{(p-nb)/2}b^{-p/2}|\mathbf{X}'\mathbf{X}|^{-1/2}\{b(n-p)\hat{\sigma}^2/2\}^{(p-nb)/2}\Gamma((nb-p)/2)$$
$$= \pi^{(p-nb)/2}b^{-nb/2}|\mathbf{X}'\mathbf{X}|^{-1/2}\{(n-p)\hat{\sigma}^2\}^{(p-nb)/2}\Gamma((nb-p)/2).$$

Then (9.53) is

$$f(\mathbf{y}) = I_1/I_b = \pi^{-n(1-b)/2}b^{nb/2}\{(n-p)\hat{\sigma}^2\}^{-n(1-b)/2}\Gamma((n-p)/2)/\Gamma((nb-p)/2). \tag{9.55}$$

Under the alternative model we find the same expression for $f_A(\mathbf{y})$ except that $p$ changes to $p_A$ and $\hat{\sigma}^2$ to $\hat{\sigma}_A^2$. Therefore the fractional Bayes factor is

$$B_q = \frac{\Gamma((nb-p)/2)\Gamma((n-p_A)/2)}{\Gamma((nb-p_A)/2)\Gamma((n-p)/2)}\left(\frac{(n-p)\hat{\sigma}^2}{(n-p_A)\hat{\sigma}_A^2}\right)^{n(1-b)/2}. \tag{9.56}$$

**9.38** Comparing (9.56) with the full Bayes factor (9.49) we notice that the fractional Bayes factor does not have an indeterminate constant $c$, nor does it depend on the matrices $\mathbf{X}$ and $\mathbf{X}_A$. Instead, the constant term depends only on the number of observations, the numbers of parameters in the two models, and the proportion of the data to be regarded as a training sample. The other difference is in the power of the ratio of residual sums of squares. Pericchi (1984), from a very different approach, obtains a Bayes factor which is different from (9.56) but also does not involve $\mathbf{X}$ or $\mathbf{X}_A$. De Vos (1993) obtains a form of intrinsic Bayes factor with the same property.

## Predictive inference

**9.39** A common requirement for regression models is to predict the values of the response variable in some future observations. Suppose that You wish to predict a vector $\mathbf{y}_0$ of future observations, with values of the regressor variables $\mathbf{X}_0$. That is, in terms of the parameters $\boldsymbol{\beta}$ of the original model, the new observations are represented by the equation

$$\mathbf{y}_0 = \mathbf{X}_0\boldsymbol{\beta} + \boldsymbol{\epsilon}_0,$$

where $\boldsymbol{\epsilon}_0$ is a vector of new residuals, independently and identically distributed as $N(0,\sigma^2)$. The posterior predictive distribution for $\mathbf{y}_0$ can be derived in the same way as the prior predictive distribution of $\mathbf{y}$ in **9.29** or **9.31**. Following the latter approach, the posterior conditional distribution of $\mathbf{y}_0$ given $\sigma^2$ is $N(\mathbf{X}_0\mathbf{m}^\star, \sigma^2(\mathbf{I}+\mathbf{X}_0\mathbf{V}^\star\mathbf{X}_0'))$ and the posterior distribution of $\sigma^2$ is $IG(a^\star, d^\star)$. Hence the joint posterior distribution of $(\mathbf{y}_0, \sigma^2)$ is $NIG(a^\star, d^\star, \mathbf{X}_0\mathbf{m}^\star, \mathbf{I}+\mathbf{X}_0\mathbf{V}^\star\mathbf{X}_0')$ and the marginal posterior distribution of $\mathbf{y}_0$ is $t_{d^\star}(\mathbf{X}_0\mathbf{m}^\star, a^\star(\mathbf{I}+\mathbf{X}_0\mathbf{V}^\star\mathbf{X}_0'))$.

*Example 9.8*

For a simple regression model $y_i = \alpha + \beta x_i + \epsilon_i$ consider a single future observation $y_0$ when the regressor variable takes the value $x_0$, so that $\mathbf{X}_0 = (1, x_0)$. Writing

$$\mathbf{m}^\star = \begin{pmatrix} a \\ b \end{pmatrix}, \qquad \mathbf{V}^\star = \begin{pmatrix} v_a & c \\ c & v_b \end{pmatrix}$$

as in Example 9.4, $\mathbf{X}_0\mathbf{m}^\star = a + bx_0$ and $\mathbf{I} + \mathbf{X}_0\mathbf{V}^\star\mathbf{X}_0' = 1 + v_a + 2cx_0 + v_b x_0^2 = 1 + v(x_0)$, where $v(x)$ is as in (9.37). Then a highest posterior density interval for $y_0$ is

$$a + bx_0 \pm \{a^\star(1 + v(x_0))/d^\star\}^{1/2}t,$$

where $t$ is an appropriate percentage point of the $t_{d+n}$ distribution. This differs from the result (9.36) in Example 9.4 only by changing $v(x)$ to $1 + v(x)$. Whereas in Example 9.4 we were making posterior inference about the value of the regression line at a value $x$ (or $x_0$) of the regressor variable, here we are making predictive inference about a new observation at that point, which is the regression line plus a further random error $\epsilon_0$. Hence the increased variance and wider highest posterior density intervals in this case. The distinction is equivalent to that between the classical confidence intervals (A28.29) and (A28.33).

### Ridge regression

**9.40** If we take a normal-inverse-gamma prior distribution with $\mathbf{m} = \mathbf{0}$ and $\mathbf{V} = c^{-1}\mathbf{I}$, then the posterior mean of $\boldsymbol{\beta}$ is

$$\mathbf{m}^\star = (\mathbf{X}'\mathbf{X} + c\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}, \tag{9.57}$$

which is the ridge regression estimator of A**19.12**. Therefore this classical estimator corresponds to a rather special kind of prior information about $\boldsymbol{\beta}$. The elements of $\boldsymbol{\beta}$ are *a priori* independent and identically distributed, with zero mean. The zero prior means cause the posterior estimates (9.57) to be generally shrunk towards zero, relative to the classical least squares estimates $\hat{\boldsymbol{\beta}}$. The larger $c$ is, the smaller is the prior variance, causing the prior to have more influence, and so the shrinkage towards the origin is greater. This is the general behaviour of the ridge regression estimate. It is used in classical statistics when the data are deficient in the sense that $\mathbf{X}'\mathbf{X}$ is nearly singular. This is a situation in which we would expect prior information to be important. Instead of then adopting the simplistic prior $\mathbf{m} = \mathbf{0}$, $\mathbf{V} = c^{-1}\mathbf{I}$ of the ridge regression estimator, it would be appropriate to give careful consideration to Your genuine prior beliefs about $\boldsymbol{\beta}$.

**9.41** The first part of this chapter has comprised a thorough study of the normal linear model with the conjugate normal-inverse-gamma prior. This represents the simplest formulation of proper prior information, and that analysis is therefore also simple enough to provide important insights into how the data might generally modify Your prior beliefs. However, in practice these prior distributions are not realistic. The normal-inverse-gamma family suffers from the usual restrictions of conjugate families for many parameters. In particular, it imposes a specific form of relationship between $\boldsymbol{\beta}$ and $\sigma^2$. The conditional distribution of $\boldsymbol{\beta}$ given $\sigma^2$ is $N(\mathbf{m}, \sigma^2\mathbf{V})$, so the conditional variance must be proportional to $\sigma^2$. This means that if You were to learn that the true value of $\sigma^2$ is small then this would lead You to a small variance for $\boldsymbol{\beta}$, and a correspondingly strong belief that $\boldsymbol{\beta}$ should be close to its prior mean $\mathbf{m}$. But if You learnt that the true value of $\sigma^2$ is very large, then You would instead be very unsure of the value of $\boldsymbol{\beta}$, and would believe it likely to be far from $\mathbf{m}$. Conversely, the prior mean of $\sigma^2$ given $\boldsymbol{\beta}$ is $(d-2)^{-1}\{a + (\boldsymbol{\beta} - \mathbf{m})\mathbf{V}^{-1}(\boldsymbol{\beta} - \mathbf{m})\}$

and if You learnt that $\boldsymbol{\beta}$ is close to $\mathbf{m}$ You would expect $\sigma^2$ to be small, or if You learnt that $\boldsymbol{\beta}$ is far from $\mathbf{m}$ You would then expect $\sigma^2$ to be large.

Such a specific relationship between $\boldsymbol{\beta}$ and $\sigma^2$ will not reflect Your actual prior beliefs in very many practical situations. There is therefore a need to consider other forms of prior distribution. Almost all of the remainder of this chapter consists of an exploration of alternative prior formulations.

### Known variance

**9.42** It is sometimes reasonable to suppose that the residual error variance $\sigma^2$ is known. In that case the difficulty with the conjugate prior distribution disappears. We now regard $\sigma^2$ as fixed in the likelihood (9.2), and denote it by $f(\mathbf{y} \mid \boldsymbol{\beta})$. Only $\boldsymbol{\beta}$ is unknown, and we require a suitable prior distribution $f(\boldsymbol{\beta})$. Using also (9.4), the likelihood simplifies to

$$f(\mathbf{y} \mid \boldsymbol{\beta}) \propto \exp\{-(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/(2\sigma^2)\}, \tag{9.58}$$

and therefore the natural conjugate prior family is the family of normal distributions. Suppose therefore that $\boldsymbol{\beta}$ has the $N(\mathbf{m}, \mathbf{W})$ prior distribution,

$$f(\boldsymbol{\beta}) \propto \exp\{-(\boldsymbol{\beta} - \mathbf{m})'\mathbf{W}^{-1}(\boldsymbol{\beta} - \mathbf{m})/2\}. \tag{9.59}$$

Then $f(\boldsymbol{\beta} \mid \mathbf{y}) \propto f(\mathbf{y} \mid \boldsymbol{\beta})f(\boldsymbol{\beta}) \propto \exp(-Q/2)$, where

$$\begin{aligned} Q &= \sigma^{-2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \mathbf{m})'\mathbf{W}^{-1}(\boldsymbol{\beta} - \mathbf{m}) \\ &= \boldsymbol{\beta}'(\mathbf{W}^{-1} + \sigma^{-2}\mathbf{X}'\mathbf{X})\boldsymbol{\beta} + \boldsymbol{\beta}'(\mathbf{W}^{-1}\mathbf{m} + \sigma^{-2}\mathbf{X}'\mathbf{y}) + (\mathbf{W}^{-1}\mathbf{m} + \sigma^{-2}\mathbf{X}'\mathbf{y})'\boldsymbol{\beta} + R_1 \\ &= (\boldsymbol{\beta} - \mathbf{m}^\star)'(\mathbf{W}^\star)^{-1}(\boldsymbol{\beta} - \mathbf{m}^\star) + R_2, \end{aligned}$$

and where

$$\mathbf{m}^\star = (\mathbf{W}^{-1} + \sigma^{-2}\mathbf{X}'\mathbf{X})^{-1}(\mathbf{W}^{-1}\mathbf{m} + \sigma^{-2}\mathbf{X}'\mathbf{y}), \tag{9.60}$$

$$\mathbf{W}^\star = (\mathbf{W}^{-1} + \sigma^{-2}\mathbf{X}'\mathbf{X})^{-1} \tag{9.61}$$

and $R_1$, $R_2$ are constants. Therefore,

$$f(\boldsymbol{\beta} \mid \mathbf{y}) \propto \exp\{-(\boldsymbol{\beta} - \mathbf{m}^\star)'(\mathbf{W}^\star)^{-1}(\boldsymbol{\beta} - \mathbf{m}^\star)/2\},$$

i.e. the posterior distribution of $\boldsymbol{\beta}$ is $N(\mathbf{m}^\star, \mathbf{W}^\star)$.

**9.43** The analysis is very similar to the case of unknown $\sigma^2$ in several respects. In particular, if we let $\mathbf{W} = \sigma^2\mathbf{V}$, then $\mathbf{m}^\star$ in (9.60) is exactly the same as (9.17) and $\mathbf{W}^\star = \sigma^2\mathbf{V}^\star$, with $\mathbf{V}^\star$ as in (9.16). The explanation of this agreement is that if $\mathbf{W} = \sigma^2\mathbf{V}$ the prior distribution $N(\mathbf{m}, \mathbf{W})$ for $\boldsymbol{\beta}$ is the same as the conditional prior distribution $f(\boldsymbol{\beta} \mid \sigma^2)$ in the case of unknown $\sigma^2$. Then the posterior distribution $N(\mathbf{m}^\star, \mathbf{W}^\star)$ is the same as the conditional posterior distribution $f(\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2)$ in the unknown $\sigma^2$ case, because of Bayes' theorem

$$f(\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2) = f(\boldsymbol{\beta} \mid \sigma^2)f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)/f(\mathbf{y} \mid \sigma^2).$$

Knowing $\sigma^2$ is the same as conditioning on $\sigma^2$, and in particular after observing the data $\mathbf{y}$ our information is $(\mathbf{y}, \sigma^2)$, and so the relevant distribution of $\boldsymbol{\beta}$ is $f(\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2)$.

**9.44** The same analysis can be obtained from a very different perspective by using the Bayes Linear Estimator (see **6.49**). Since only first and second order moments are required in this approach, it is not necessary to assume normality. Suppose then that the model (9.1) simply implies that $E(\mathbf{y} \mid \boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$ and $\operatorname{var}(\mathbf{y} \mid \boldsymbol{\beta}) = \sigma^2\mathbf{I}$. Again assume that $\sigma^2$ is known. Normality is not needed for the prior distribution, either, and we simply have $E(\boldsymbol{\beta}) = \mathbf{m}$, $\operatorname{var}(\boldsymbol{\beta}) = \mathbf{W}$. The Bayes linear estimator (6.29) for a scalar parameter is generalized to the case of estimating a vector parameter in Exercise 6.9. It may then be shown (see Exercise 9.4) that the Bayes linear estimator of $\boldsymbol{\beta}$ based on data $\mathbf{y}$ is $\mathbf{m}^\star$, equation (9.60). Furthermore, $\mathbf{W}^\star$ is the corresponding dispersion matrix (6.33).

If $\mathbf{W} = \sigma^2\mathbf{V}$, $\mathbf{m}^\star$ will still be the Bayes linear estimator in the case of unknown $\sigma^2$, but now the dispersion matrix involves $\sigma^2$. It is not reasonable to use linear functions of $\mathbf{y}$ to estimate $\sigma^2$, and therefore the Bayes linear estimation approach becomes much more complex when $\sigma^2$ is unknown.

### Conditional conjugate analysis

**9.45** Returning to the normal linear model with unknown $\sigma^2$, one way to escape from the difficulies inherent in the conjugate family is to give independent prior distributions to $\boldsymbol{\beta}$ and $\sigma^2$,

$$f(\boldsymbol{\beta}, \sigma^2) = f(\boldsymbol{\beta})f(\sigma^2)$$

Such a prior represents a situation in which learning about $\sigma^2$ would not change Your beliefs about $\boldsymbol{\beta}$, and vice versa. If You assigned a $N(\mathbf{m}, \mathbf{W})$ prior distribution to $\boldsymbol{\beta}$, as in **9.42**, and an $IG(a, d)$ distribution for $\sigma^2$, then

$$f(\boldsymbol{\beta}, \sigma^2) \propto \exp\{-(\boldsymbol{\beta} - \mathbf{m})'\mathbf{W}^{-1}(\boldsymbol{\beta} - \mathbf{m})/2\}(\sigma^2)^{-(d+2)/2} \exp\{-a/(2\sigma^2)\}, \qquad (9.62).$$

The posterior distribution is then proportional to the product of (9.62) and the likelihood (9.2). Notice that for fixed $\sigma^2$ we can proceed exactly as in **9.42**, and note that the posterior conditional distribution $f(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y})$ is $N(\mathbf{m}^\star, \mathbf{W}^\star)$, with $\mathbf{m}^\star$ and $\mathbf{W}^\star$ given by (9.60) and (9.61). We can therefore integrate with respect to $\boldsymbol{\beta}$ to obtain the marginal posterior distribution for $\sigma^2$. However, this $f(\sigma^2 \mid \mathbf{y})$ will not be an inverse-gamma distribution, and it will not be practical to obtain summaries analytically. Nevertheless, summarizing a univariate distribution numerically is a straightforward computational task. Rather than pursue this analysis here, however, we will first generalize to a much larger conditional conjugate family.

**9.46** Consider a prior distribution of the form

$$f(\boldsymbol{\beta}, \sigma^2) \propto p(\sigma^2) \exp[-\{\boldsymbol{\beta} - \mathbf{m}(\sigma^2)\}'\mathbf{W}(\sigma^2)^{-1}\{\boldsymbol{\beta} - \mathbf{m}(\sigma^2)\}/2], \qquad (9.63)$$

when $p(\sigma^2)$, $\mathbf{m}(\sigma^2)$ and $\mathbf{W}(\sigma^2)$ are arbitrary functions of $\sigma^2$, subject only to conditions that $p(\sigma^2)$ is positive for all $\sigma^2$ and $\mathbf{W}(\sigma^2)$ is a positive definite $p \times p$ matrix for all $\sigma^2$. The conditional prior distribution of $\boldsymbol{\beta}$ given $\sigma^2$ is $N(\mathbf{m}(\sigma^2), \mathbf{W}(\sigma^2))$ and the marginal prior distribution of $\sigma^2$ is given by

$$f(\sigma^2) \propto p(\sigma^2)|\mathbf{W}(\sigma^2)|^{1/2}. \qquad (9.64)$$

Further conditions on $p(\sigma^2)$ and $\mathbf{W}(\sigma^2)$ will be required to ensure that this is a proper distribution. The family of distributions (9.63) is a general conditional conjugate family, since we have seen in **9.42** that the normal family of distributions for $\boldsymbol{\beta}$ is conjugate to the likelihood for given $\sigma^2$.

**9.47** Arguing as in **9.42** we obtain the posterior joint density

$$f(\boldsymbol{\beta},\sigma^2 \mid \mathbf{y}) \propto p^{\star}(\sigma^2)\exp[-\{\boldsymbol{\beta}-\mathbf{m}^{\star}(\sigma^2)\}'\mathbf{W}^{\star}(\sigma^2)^{-1}\{\boldsymbol{\beta}-\mathbf{m}^{\star}(\sigma^2)\}/2],$$

where

$$p^{\star}(\sigma^2) = p(\sigma^2)\sigma^{-n}\exp\{-R_2(\sigma^2)/2\}, \tag{9.65}$$

$$R_2(\sigma^2) = \{\hat{\boldsymbol{\beta}}-\mathbf{m}(\sigma^2)\}'\{\mathbf{W}(\sigma^2)+\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\}^{-1}\{\hat{\boldsymbol{\beta}}-\mathbf{m}(\sigma^2)\}+\sigma^{-2}S,$$

$$\mathbf{m}^{\star}(\sigma^2) = \{\mathbf{W}(\sigma^2)^{-1}+\sigma^{-2}\mathbf{X}'\mathbf{X}\}^{-1}\{\mathbf{W}(\sigma^2)^{-1}\mathbf{m}+\sigma^{-2}\mathbf{X}'\mathbf{y}\}. \tag{9.66}$$

$$\mathbf{W}^{\star}(\sigma^2) = \{\mathbf{W}(\sigma^2)^{-1}+\sigma^{-2}\mathbf{X}'\mathbf{X}\}^{-1}, \tag{9.67}$$

and where $S$ is the classical residual sum of squares $(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$. This is clearly a member of the same family, with the 'hyperparameters' $p(\sigma^2)$, $\mathbf{m}(\sigma^2)$ and $\mathbf{W}(\sigma^2)$ replaced by (9.65) to (9.67).

**9.48** To summarize the prior or posterior distribution we exploit the conditional conjugacy property to reduce the dimensionality. That is, we use the normal conditional distribution of $\boldsymbol{\beta}$ given $\sigma^2$ to reduce computations to operations on the one-dimensional marginal distribution (9.64) or its posterior counterpart. For instance,

$$E(\boldsymbol{\beta}) = E(\mathbf{m}(\sigma^2))$$

requires the calculation of $p$ expectations with respect to $f(\sigma^2)$, which can be done with a single one-dimensional numerical integration exercise. var $(\boldsymbol{\beta})$ can also be obtained from the $p^2$ expectations

$$E(\boldsymbol{\beta}\boldsymbol{\beta}') = E(\mathbf{W}(\sigma^2)+\mathbf{m}(\sigma^2)\mathbf{m}(\sigma^2)')$$

and var $(\boldsymbol{\beta}) = E(\boldsymbol{\beta}\boldsymbol{\beta}') - E(\boldsymbol{\beta})E(\boldsymbol{\beta})'$. The mode of the joint density $f(\boldsymbol{\beta},\sigma^2)$ is $(\mathbf{m}(\hat{\sigma}^2),\hat{\sigma}^2)$ where $\hat{\sigma}^2$ maximizes $p(\sigma^2)$.

This dimensionality reduction technique does not, however, give shape summaries of the marginal distribution of $\boldsymbol{\beta}$, or of any single element of $\boldsymbol{\beta}$. This limits the value of being able to compute $E(\boldsymbol{\beta})$ and var $(\boldsymbol{\beta})$ simply.

**9.49** The conditional conjugacy facilitates efficient use of Gibbs sampling. Each iteration can be implemented in just two steps. First, using the current $\sigma^2$ a new $\boldsymbol{\beta}$ is generated from the multivariate normal distribution $N(\mathbf{m}(\sigma^2),\mathbf{V}(\sigma^2))$ for which efficient algorithms exist (see references in Chapter 8). Then for this $\boldsymbol{\beta}$ a new $\sigma^2$ is generated from (9.63) regarded as a function of $\sigma^2$ alone, using a method such as the ratio of uniforms, **8.63**. The values of moments like $E(\boldsymbol{\beta})$, var $(\boldsymbol{\beta})$, $E(\sigma^2)$ or var $(\sigma^2)$, which can all be obtained accurately by one-dimensional quadrature, provide a further check on convergence of the Gibbs sampler.

Although a sample from $f(\boldsymbol{\beta},\sigma^2)$ does not provide accurate shape summaries, it will certainly assist in interpreting moments. It will, for instance, be adequate to identify possible multimodality or marked skewness.

### Generalized error variance

**9.50** A small generalization of the linear model is achieved by allowing the random errors comprising $\boldsymbol{\epsilon}$ to be correlated, but assuming that the correlation structure is known. That is, we replace the variance matrix $\sigma^2\mathbf{I}$ by $\sigma^2\mathbf{D}$, where $\mathbf{D}$ is a known positive definite matrix. If the prior distribution of $(\boldsymbol{\beta},\sigma^2)$ is $NIG(a,d,\mathbf{m},\mathbf{V})$ as before, then the posterior distribution is also normal-inverse-gamma and we denote it as before by $NIG(a^\star,d^\star,\mathbf{m}^\star,\mathbf{V}^\star)$. We still have $d^\star = d+n$ but formulae for the other posterior parameters become

$$\mathbf{V}^\star = (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1},$$

$$\mathbf{m}^\star = (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}(\mathbf{V}^{-1}\mathbf{m} + \mathbf{X}'\mathbf{D}^{-1}\mathbf{y}),$$

$$a^\star = a + \mathbf{m}'\mathbf{V}^{-1}\mathbf{m} + \mathbf{y}'\mathbf{D}^{-1}\mathbf{y} - (\mathbf{m}^\star)'(\mathbf{V}^\star)^{-1}\mathbf{m}^\star.$$

Equation (9.19) can be derived, expressing the posterior mean as a matrix-weighted average of the prior mean and the classical estimator $\hat{\boldsymbol{\beta}}$, but now $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{y}$ (the generalized Least Squares estimator of A**19.19**) and the weight matrix is $\mathbf{A} = (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{D}^{-1}\mathbf{X}')^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}$. It is straightforward to generalize other results to this case.

**9.51** We can now combine these results with those of **9.42**, to consider the case of known but arbitrary error variance. Let the variance matrix of the random errors $\boldsymbol{\epsilon}$ be $\mathbf{C}$ and known. This corresponds to $\mathbf{C} = \sigma^2\mathbf{D}$ above, but $\sigma^2$ is now supposed known as in **9.42**. Therefore the distribution of the data $\mathbf{y}$ given the parameters $\boldsymbol{\beta}$ becomes $N(\mathbf{X}\boldsymbol{\beta},\mathbf{C})$, with likelihood

$$f(\mathbf{y}\mid\boldsymbol{\beta}) \propto \exp\{-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{C}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/2\}.$$

Let the prior distribution of $\boldsymbol{\beta}$ be $N(\mathbf{m},\mathbf{W})$ as in **9.42**. Then simple algebra shows that the posterior distribution of $\boldsymbol{\beta}$ is $N(\mathbf{m}^\star,\mathbf{W}^\star)$, where

$$\mathbf{m}^\star = (\mathbf{W}^{-1} + \mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}(\mathbf{W}^{-1}\mathbf{m} + \mathbf{X}'\mathbf{C}^{-1}\mathbf{y}), \tag{9.68}$$

$$\mathbf{W}^\star = (\mathbf{W}^{-1} + \mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}. \tag{9.69}$$

The posterior mean $\mathbf{m}^\star$ is a matrix-weighted average of the prior mean $\mathbf{m}$ and the generalized least squares $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{y} = (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{y}$.

### Hierarchical linear models

**9.52** The method of hierarchical modelling was introduced in **6.39**. As in **9.51**, we suppose that $\sigma^2$ is known, but allow a general correlation structure. Suppose, then, that the distribution of $\mathbf{y}$ given $\boldsymbol{\beta}$ is $N(\mathbf{X}\boldsymbol{\beta},\mathbf{C})$, where $\mathbf{C}$ is a known $n \times n$ positive definite matrix. A hierarchical prior distribution is now proposed for $\boldsymbol{\beta}$. The prior distribution of $\beta$ is expressed conditional on hyper-parameters $\gamma$ as $N(\mathbf{X}_1\gamma,\mathbf{C}_1)$. The prior distribution of $\gamma$ is finally given as $N(\mathbf{m}_2,\mathbf{C}_2)$.

**9.53** The posterior distribution is

$$f(\boldsymbol{\beta},\boldsymbol{\gamma}\mid\mathbf{y})\propto f(\mathbf{y}\mid\boldsymbol{\beta},\boldsymbol{\gamma})f(\boldsymbol{\beta}\mid\boldsymbol{\gamma})f(\boldsymbol{\gamma})\propto\exp(-Q/2),$$

where

$$Q=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{C}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+(\boldsymbol{\beta}-\mathbf{X}_1\boldsymbol{\gamma})'\mathbf{C}_1^{-1}(\boldsymbol{\beta}-\mathbf{X}_1\boldsymbol{\gamma})+(\boldsymbol{\gamma}-\mathbf{m}_2)'\mathbf{C}_2^{-1}(\boldsymbol{\gamma}-\mathbf{m}_2). \qquad (9.70)$$

Since $Q$ is a quadratic expression in $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, their joint posterior distribution is clearly normal. Let $\boldsymbol{\theta}=(\boldsymbol{\beta}',\boldsymbol{\gamma}')'$ and write $Q=\boldsymbol{\theta}'\mathbf{V}\boldsymbol{\theta}-\boldsymbol{\theta}'\mathbf{z}-\mathbf{z}'\boldsymbol{\theta}+R$, where

$$\mathbf{V}=\begin{pmatrix}\mathbf{C}_1^{-1}+\mathbf{X}'\mathbf{C}^{-1}\mathbf{X} & -\mathbf{C}_1^{-1}\mathbf{X}_1\\ -\mathbf{X}_1'\mathbf{C}_1^{-1} & \mathbf{C}_2^{-1}+\mathbf{X}_1'\mathbf{C}_1^{-1}\mathbf{X}_1\end{pmatrix},\qquad \mathbf{z}=\begin{pmatrix}\mathbf{X}'\mathbf{C}^{-1}\mathbf{y}\\ \mathbf{C}_2^{-1}\mathbf{m}_2\end{pmatrix}, \qquad (9.71)$$

$$R=\mathbf{y}'\mathbf{C}^{-1}\mathbf{y}+\mathbf{m}_2'\mathbf{C}_2^{-1}\mathbf{m}_2.$$

Then the posterior distribution of $\boldsymbol{\theta}$ is $N(\mathbf{V}^{-1}\mathbf{z},\mathbf{V}^{-1})$.

**9.54** This gives the full joint posterior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and from it we can derive marginal posterior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ separately. It is possible to invert $\mathbf{V}$ symbolically in a number of different ways, and so obtain a variety of formulae for the posterior means of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. One simple approach does this indirectly by collapsing the hierarchy in two different ways.

First note that we can express the prior distribution $N(\mathbf{X}_1\boldsymbol{\gamma},\mathbf{C}_1)$ of $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$ by writing $\boldsymbol{\beta}=\mathbf{X}_1\boldsymbol{\gamma}+\boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is distributed as $N(\mathbf{0},\mathbf{C}_1)$ independently of $\boldsymbol{\gamma}$. Then since $\boldsymbol{\gamma}$ has the $N(\mathbf{m}_2,\mathbf{C}_2)$ distribution, we can immediately deduce that the marginal prior distribution of $\boldsymbol{\beta}$ is $N(\mathbf{m},\mathbf{W})$, where

$$\mathbf{m}=\mathbf{X}_1\mathbf{m}_2,\qquad \mathbf{W}=\mathbf{C}_1+\mathbf{X}_1\mathbf{C}_2\mathbf{X}_1'. \qquad (9.72)$$

Now apply the theory of **9.51**, so that the posterior distribution of $\boldsymbol{\beta}$ is $N(\mathbf{m}^\star,\mathbf{W}^\star)$, where $\mathbf{m}^\star$ and $\mathbf{W}^\star$ are given by (9.68) and (9.69) after inserting (9.72). The posterior mean $\mathbf{m}^\star$ is thereby expressed as a matrix-weighted average of the prior mean $\mathbf{X}_1\mathbf{m}_2$ and the generalized least squares $\hat{\boldsymbol{\beta}}$.

The marginal posterior distribution of $\boldsymbol{\gamma}$ may be found similarly by noting that the original model is $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is distributed as $N(\mathbf{0},\mathbf{C})$ and the distribution of $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$ is $N(\mathbf{X}_1\boldsymbol{\gamma},\mathbf{C}_1)$. Therefore the distribution of $\mathbf{y}$ given $\boldsymbol{\gamma}$ is $N(\mathbf{X}\mathbf{X}_1\boldsymbol{\gamma},\mathbf{C}+\mathbf{X}\mathbf{C}_1\mathbf{X}')$. This corresponds to another linear model in which $\mathbf{X}$ is replaced by $\mathbf{X}\mathbf{X}_1$ and $\mathbf{C}$ by $\mathbf{C}+\mathbf{X}\mathbf{C}_1\mathbf{X}'$. We again use the theory of **9.51** to obtain a posterior distribution $N(\mathbf{m}_2^\star,\mathbf{C}_2^\star)$ for $\boldsymbol{\gamma}$, where

$$\begin{aligned}\mathbf{C}_2^\star&=(\mathbf{C}_2^{-1}+\mathbf{X}_1'\mathbf{X}'(\mathbf{C}+\mathbf{X}\mathbf{C}_1\mathbf{X}')^{-1}\mathbf{X}\mathbf{X}_1)^{-1}\\ &=\mathbf{C}_2-\mathbf{C}_2\mathbf{X}_1'\mathbf{X}'(\mathbf{C}+\mathbf{X}\mathbf{C}_1\mathbf{X}'+\mathbf{X}\mathbf{X}_1\mathbf{C}_2\mathbf{X}_1'\mathbf{X}')^{-1}\mathbf{X}\mathbf{X}_1\mathbf{C}_2,\\ \mathbf{m}_2^\star&=\mathbf{C}_2^\star(\mathbf{C}_2^{-1}\mathbf{m}_2+\mathbf{X}_1'\mathbf{X}'(\mathbf{C}+\mathbf{X}\mathbf{C}_1\mathbf{X}')^{-1}\mathbf{y}).\end{aligned} \qquad (9.73)$$

This expresses the posterior mean of $\boldsymbol{\gamma}$ as a matrix-weighted average of its prior mean $\mathbf{m}_2$ and a corresponding generalized least-squares estimator

$$\hat{\boldsymbol{\gamma}}=\{\mathbf{X}_1'\mathbf{X}'(\mathbf{C}+\mathbf{X}\mathbf{C}_1\mathbf{X}')^{-1}\mathbf{X}\mathbf{X}_1\}^{-1}\mathbf{X}_1'\mathbf{X}'(\mathbf{C}+\mathbf{X}\mathbf{C}_1\mathbf{X}')^{-1}\mathbf{y}. \qquad (9.74)$$

**9.55** An alternative derivation by directly inverting $\mathbf{V}$, Exercise 9.7, yields the formulae

$$\mathbf{m}^\star = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{D}^{-1}(\hat{\boldsymbol{\beta}} - \mathbf{X}_1\mathbf{m}_2), \tag{9.75}$$

$$\mathbf{m}_2^\star = \mathbf{m}_2 - \mathbf{C}_2\mathbf{X}_1'\mathbf{D}^{-1}(\mathbf{X}_1\mathbf{m}_2 - \hat{\boldsymbol{\beta}}), \tag{9.76}$$

where

$$\mathbf{D} = \mathbf{C}_1 + \mathbf{X}_1\mathbf{C}_2\mathbf{X}_1' + (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}.$$

(9.75) is a straightforward restatement of the value for $\mathbf{m}^\star$ given by (9.68) and (9.72). (9.76), however, takes a very different form from (9.74), showing more clearly that the inference depends on the data $\mathbf{y}$ only through the sufficient statistic $\hat{\boldsymbol{\beta}}$.

**9.56** Yet another approach is to use the conditional posterior distributions. We note from (9.70) (or from (9.68) with $\mathbf{m} = \mathbf{X}_1\boldsymbol{\gamma}$ and $\mathbf{W} = \mathbf{C}_1$) that the conditional posterior mean of $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$ is

$$\mathbf{m}^\star(\boldsymbol{\gamma}) = (\mathbf{C}_1^{-1} + \mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}(\mathbf{C}_1^{-1}\mathbf{X}_1\boldsymbol{\gamma} + \mathbf{X}'\mathbf{C}^{-1}\mathbf{y}).$$

Therefore

$$\mathbf{m}^\star = E(\mathbf{m}^\star(\boldsymbol{\gamma})) = (\mathbf{C}_1^{-1} + \mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}(\mathbf{C}_1^{-1}\mathbf{X}_1\mathbf{m}_2^\star + \mathbf{X}'\mathbf{C}^{-1}\mathbf{y}). \tag{9.77}$$

Similarly, we obtain an equation for $\mathbf{m}_2^\star$ in terms of $\mathbf{m}^\star$,

$$\mathbf{m}_2^\star = (\mathbf{C}_2^{-1} + \mathbf{X}_1'\mathbf{C}_1^{-1}\mathbf{X}_1)^{-1}(\mathbf{C}_2^{-1}\mathbf{m}_2 + \mathbf{X}_1'\mathbf{C}_1^{-1}\mathbf{m}^\star). \tag{9.78}$$

**Uniform second stage prior distribution**

**9.57** An important special case arises if the prior distribution for $\boldsymbol{\gamma}$, in the final stage of the hierarchy, is an improper uniform distribution representing weak prior information about $\boldsymbol{\gamma}$. This is achieved by letting $\mathbf{C}_2^{-1} \to \mathbf{0}$. (9.73) shows that then $\mathbf{m}_2^\star = \hat{\boldsymbol{\gamma}}$, and from (9.78) $\mathbf{m}_2^\star = (\mathbf{X}_1'\mathbf{C}_1^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{C}_1^{-1}\mathbf{m}^\star$. Inserting the first of these into (9.77) expresses $\mathbf{m}^\star$ as a matrix-weighted average of $\mathbf{X}_1\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$. Inserting the second into (9.77) and solving for $\mathbf{m}^\star$ yields

$$\mathbf{m}^\star = (\mathbf{C}_1^{-1} - \mathbf{C}_1^{-1}\mathbf{X}_1(\mathbf{X}_1'\mathbf{C}_1^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{C}_1^{-1} + \mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{y}. \tag{9.79}$$

The same result is obtained by inverting $\mathbf{W}$ in (9.72) to give $\mathbf{C}_1^{-1} - \mathbf{C}_1^{-1}\mathbf{X}_1(\mathbf{C}_2^{-1} + \mathbf{X}_1'\mathbf{C}_1^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{C}_1^{-1}$, setting $\mathbf{C}_2^{-1} = \mathbf{0}$ and substituting in (9.68).

*Example 9.9*

Independent samples of $m$ measurements are made on each of $p$ subjects. Denote by $y_{ij}$ the $j$th measurement on subject $i$ $(i = 1, 2, \ldots, p,\ j = 1, 2, \ldots, m)$. We propose the model $y_{ij} = \mu_i + \epsilon_{ij}$, where $\mu_i$ is the true mean measurement for subject $i$ and then $\epsilon_{ij}$s are independent measurement errors. We can write this as a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ by letting $\mathbf{y}$ be a single vector of all $n = mp$ observations $y_{ij}$, $\boldsymbol{\beta} = (\mu_1, \mu_2, \ldots, \mu_k)'$. The $n \times p$ matrix $\mathbf{X}$ has elements zero and one arranged as

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_m & \mathbf{0}_m & \ldots & \mathbf{0}_m \\ \mathbf{0}_m & \mathbf{1}_m & \ldots & \mathbf{0}_m \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_m & \mathbf{0}_m & \ldots & \mathbf{1}_m \end{pmatrix},$$

where each $\mathbf{0}_m$ is an $m \times 1$ vector of zeros and each $\mathbf{1}_m$ is an $m \times 1$ vector of ones. Assume that the variance matrix of the errors $\boldsymbol{\epsilon}$ is $\mathbf{C} = \sigma^2 \mathbf{I}_n$, and suppose also that $\sigma^2$ is known. Simple calculations yield $\mathbf{X}'\mathbf{C}^{-1}\mathbf{X} = \sigma^{-2}\mathbf{X}'\mathbf{X} = m\sigma^{-2}\mathbf{I}_k$, $\hat{\boldsymbol{\beta}} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_p)'$, where $\bar{y}_i = m^{-1}\sum_{j=1}^{m} y_{ij}$, or $\hat{\mu}_i = \bar{y}_i$.

For the prior distribution write $\mu_i = \xi + \delta_i$, where $\xi$ is an overall true average observation for all subjects and the $\delta_i$s are independent discrepancies of an individual subject's mean $\mu_i$ from that overall mean. We can write this as $\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\gamma} + \boldsymbol{\delta}$, where $\boldsymbol{\gamma} = (\xi)$ is a scalar hyperparameter and $\mathbf{X}_1 = \mathbf{1}_p$ is a $p \times 1$ vector of ones. Let the variance matrix of $\boldsymbol{\delta}$ be $\mathbf{C}_1 = \tau^2\mathbf{I}_p$ with $\tau^2$ assumed known. Now $\mathbf{X}\mathbf{X}_1 = \mathbf{1}_n$ and

$$\begin{aligned}(\mathbf{C} + \mathbf{X}\mathbf{C}_1\mathbf{X}')^{-1} &= \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{X}(\mathbf{C}_1^{-1} + \mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1} \\ &= \sigma^{-2}\mathbf{I}_n - \sigma^{-4}\mathbf{X}\{(\tau^{-2} + m\sigma^{-2})\mathbf{I}_p\}^{-1}\mathbf{X}' \\ &= \sigma^{-2}\mathbf{I}_n - \sigma^{-4}(\tau^2 - m\sigma^{-2})^{-1}\mathbf{X}\mathbf{X}'.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbf{X}_1'\mathbf{X}'(\mathbf{C} + \mathbf{X}\mathbf{C}_1\mathbf{X}')^{-1} &= \sigma^{-2}\mathbf{1}_n' - \sigma^{-4}(\tau^2 - m\sigma^{-2})^{-1}m\mathbf{1}_n' \\ &= (\sigma^2 + m\tau^2)^{-1}\mathbf{1}_n'.\end{aligned}$$

Hence $\hat{\gamma} = \hat{\xi} = \bar{y} = k^{-1}\sum_{i=1}^{p}\bar{y}_i$.

If we now assume weak prior information about $\xi$, its posterior mean will be $\bar{y}$. Its posterior variance $\mathbf{C}_2^*$ is $\{(\sigma^2 + m\tau^2)^{-1}\mathbf{1}_n'\mathbf{1}_n\}^{-1} = (\sigma^2 + m\tau^2)/n$. Now to find the posterior mean of $\boldsymbol{\beta}$ we apply (9.77) and need first to derive

$$(\mathbf{C}_1^{-1} + \mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1} = (\tau^{-2} + m\sigma^{-2})^{-1}\mathbf{I}_p.$$

Then

$$\mathbf{m}^* = a\bar{y}\mathbf{1} + (1 - a)\hat{\boldsymbol{\beta}},$$

where

$$a = \tau^{-2}(\tau^{-2} + m\sigma^{-2})^{-1} = \sigma^2/(\sigma^2 + m\tau^2).$$

The posterior mean of each $\mu_i$ is $a\bar{y} + (1 - a)\bar{y}_i$, a weighted average of the mean of the observations on subject $i$ and the mean of all observations. This is a *shrinkage* estimator, discussed in **6.42-43**. Indeed, this example is basically Example 6.16 rephrased in linear model terms. The degree of shrinkage is governed by the weight $a$ attached to $\bar{y}$. $a$ will be small if the data provide good information about each subject, either through $\sigma^2$ being small or through having a large number $m$ of observations on each subject. Conversely, $a$ will be large if the information on each subject is not strong and the variability $\tau^2$ between subjects is small.

**9.58** An interesting feature of the hierarchical model with uniform prior distribution for $\gamma$ is that the prior distribution for $\boldsymbol{\beta}$ also becomes improper. It was noted in **9.54** that the marginal prior distribution for $\boldsymbol{\beta}$ is $N(\mathbf{X}_1\mathbf{m}_2, \mathbf{C}_1 + \mathbf{X}_1\mathbf{C}_2\mathbf{X}_1')$. Consider a scalar linear function of $\boldsymbol{\beta}$, $\alpha = \mathbf{b}'\boldsymbol{\beta}$. Its prior distribution is therefore $N(\mathbf{b}'\mathbf{X}_1\mathbf{m}_2, \mathbf{b}'\mathbf{C}_1\mathbf{b} + \mathbf{b}'\mathbf{X}_1\mathbf{C}_2\mathbf{X}_1'\mathbf{b})$. Now as $\mathbf{C}_2^{-1}$ goes to $\mathbf{0}$, $\mathbf{b}'\mathbf{X}_1\mathbf{C}_2\mathbf{X}_1'\mathbf{b}$ will go to infinity unless $\mathbf{b}'\mathbf{X}_1 = \mathbf{0}$, a zero vector. If $\mathbf{b}'\mathbf{X}_1 = \mathbf{0}$, the prior distribution of $\alpha$ is $N(0, \mathbf{b}'\mathbf{C}_1\mathbf{b})$. This results directly from the first stage of the hierarchical prior distribution which asserts that $\alpha$ has this prior distribution independently of $\gamma$. With a uniform prior distribution on $\gamma$, every other linear function $\alpha = \mathbf{b}'\boldsymbol{\beta}$ also has an improper uniform prior distribution.

In this case, therefore, we can interpret the hierarchical model as providing 'structural' information about $\boldsymbol{\beta}$ by giving zero prior mean and finite variance to those linear functions $\mathbf{b}'\boldsymbol{\beta}$ with $\mathbf{b}'\mathbf{X}_1 = \mathbf{0}$, but providing no other information about $\boldsymbol{\beta}$.

*Example 9.10*

In Example 9.9, $\mathbf{X}_1 = \mathbf{1}$. The structural prior information is that every linear contrast $\sum_{i=1}^{p} b_i\mu_i$ with $\sum_{i=1}^{p} b_i = 0$ has zero prior expectation, but no prior information is given about any other functions of the $\mu_i$s. In particular, each $\mu_i$ alone has a uniform prior distribution.

**Hierarchical models with unknown variances**

**9.59** The assumption of known variances in a hierarchical model is mathematically convenient but will rarely be realistic in practice. As in **9.50**, we might let $\mathbf{C} = \sigma^2\mathbf{D}$, where $\mathbf{D}$ is known but $\sigma^2$ is unknown. The case $\mathbf{D} = \mathbf{I}$ corresponds to the usual linear model formulation. Since the first stage of the hierarchical prior distribution is also formulated as a linear model, it is also useful to let $\mathbf{C}_1 = \tau^2\mathbf{D}_1$ with $\mathbf{D}_1$ known but $\tau^2$ unknown. Now all the preceding theory applies in the sense that it gives the conditional posterior distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ given $\sigma^2$ and $\tau^2$, but we need to find also the posterior distributions of $\sigma^2$ and $\tau^2$.

The joint posterior distribution of all the parameters is

$$\begin{aligned} f(\boldsymbol{\beta},\gamma,\sigma^2,\tau^2 \mid \mathbf{y}) &\propto f(\mathbf{y} \mid \boldsymbol{\beta},\gamma,\sigma^2\tau^2)f(\boldsymbol{\beta} \mid \gamma,\sigma^2,\tau^2)f(\gamma \mid \sigma^2,\tau^2)f(\sigma^2,\tau^2) \\ &\propto |\mathbf{C}|^{-1/2}|\mathbf{C}_1|^{-1/2}\exp(-Q/2)f(\sigma^2,\tau^2), \end{aligned} \tag{9.80}$$

where $Q$ is given by (9.70), but now involves $\sigma^2$ and $\tau^2$ through $\mathbf{C}$ and $\mathbf{C}_1$. As before, the conditional posterior distribution of $\boldsymbol{\beta}$ and $\gamma$ given $\sigma^2$ and $\tau^2$ is $N(\mathbf{V}^{-1}\mathbf{z}, \mathbf{V}^{-1})$ with $\mathbf{V}$ and $\mathbf{z}$ as in (9.71). We can therefore integrate out $\boldsymbol{\beta}$ and $\gamma$ to yield the marginal posterior distribution of $\sigma^2$ and $\tau^2$,

$$f(\sigma^2,\tau^2 \mid \mathbf{y}) \propto (\sigma^2)^{-n/2}(\tau^2)^{-p/2}|\mathbf{V}|^{-1/2}\exp(-T/2)f(\sigma^2,\tau^2), \tag{9.81}$$

where

$$T = \mathbf{y}'\mathbf{C}^{-1}\mathbf{y} + \mathbf{m}_2'\mathbf{C}_2^{-1}\mathbf{m}_2 - \mathbf{z}'\mathbf{V}^{-1}\mathbf{z}. \tag{9.82}$$

Both $T$ and $\mathbf{V}$ are typically complex functions of $\sigma^2$ and $\tau^2$. As a result, (9.81) will generally be mathematically intractable. Nevertheless, the underlying linear model structure allows some of the computational methods of Chapter 8 to be applied very efficiently.

**9.60** By integrating out $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ we have reduced dimensionality to the two-dimensional marginal distribution (9.81). By simple quadrature over this distribution we can obtain many summaries of interest. For instance, the posterior means of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are expectations of $\mathbf{m}^{\star}$ and $\mathbf{m}_2^{\star}$ (which are both now functions of $\sigma^2$ and $\tau^2$) with respect to (9.81). Obviously posterior inference about $\sigma^2$ and $\tau^2$ is obtainable directly from (9.81).

Gibbs sampling can be implemented simply and efficiently to provide inferences not obtainable through the dimensionality reduction device. $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be sampled in a single step using their joint multivariate normal distribution given $\sigma^2$ and $\tau^2$. It remains to sample $\sigma^2$ and $\tau^2$ from their conditional distributions, using (9.81) regarded as a function of $\sigma^2$ or $\tau^2$ alone. For a quite general prior distribution $f(\sigma^2, \tau^2)$ this may be done by rejection or 'ratio-of-uniforms' methods, but if $\sigma^2$ and $\tau^2$ are given independent inverse-gamma prior distributions we have a kind of double conditional conjugacy. Then it is easy to find the conditional posterior distributions of $\sigma^2$ and $\tau^2$ given $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, which are also independent inverse-gamma distributions. Gibbs sampling is then even easier to implement.

**9.61** It is tempting now to represent weak prior information about $\sigma^2$ and $\tau^2$ by adopting the improper prior formulation $f(\sigma^2, \tau^2) \propto \sigma^{-2}\tau^{-2}$. Unfortunately this leads to an improper posterior distribution. To see why this is the case, it is instructive to consider why no such problem arises in general if we set $f(\sigma^2) \propto \sigma^{-2}$. The posterior distribution of $\sigma^2$ will be given by

$$f(\sigma^2 \mid \mathbf{y}) \propto f(\mathbf{y} \mid \sigma^2) f(\sigma^2),$$

where $f(\mathbf{y} \mid \sigma^2)$ is the appropriate integrated likelihood function. Now the prior distribution $f(\sigma^2)$ is improper in both tails, by which we mean that $\int_a^b f(\sigma^2) \mathrm{d}\sigma^2$ diverges as either $b \to \infty$ or $a \to 0$. The function $\sigma^{-2}$ tends to infinity too fast as $\sigma^2 \to 0$, and does not tend to zero fast enough as $\sigma^2 \to \infty$, for the prior distribution to be proper. The posterior distribution will only be proper if multiplying by $f(\mathbf{y} \mid \sigma^2)$ remedies these defects. This does indeed happen because the data provide information that $\sigma^2$ is neither zero nor infinite so $f(\mathbf{y} \mid \sigma^2) \to 0$ as $\sigma^2 \to 0$ or $\sigma^2 \to \infty$. In particular, as long as the residual sum of squares $S = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ is positive it is clear that $\sigma^2$ cannot be zero. This manifests itself generally through the appearance of a term like $\exp\{-S/(2\sigma^2)\}$ in $f(\mathbf{y} \mid \sigma^2)$, which tends rapidly to zero as $\sigma^2 \to 0$.

This will happen also in the hierarchical linear model, and no difficulty arises there when we set $f(\sigma^2) \propto \sigma^{-2}$. Consider, however

$$f(\tau^2 \mid \mathbf{y}) \propto f(\mathbf{y} \mid \tau^2) f(\tau^2). \tag{9.83}$$

Now the data $\mathbf{y}$ do not deny the possibility that $\tau^2 = 0$. If $\tau^2 = 0$ then $\boldsymbol{\beta}$ must equal $\mathbf{X}_1\boldsymbol{\gamma}$ for some $\boldsymbol{\gamma}$ and the data cannot refute this entirely. The data may suggest an estimate $\hat{\boldsymbol{\beta}}$ that is very far from $\mathbf{X}_1\boldsymbol{\gamma}$ for any $\boldsymbol{\gamma}$, but this only makes $\tau^2 = 0$ highly improbable rather than actually impossible. As a result, it can be proved that in general $f(\mathbf{y} \mid \tau^2 = 0)$ is positive. When this is multiplied by $f(\tau^2) \propto \tau^{-2}$ in (9.83) the result is an improper posterior distribution.

**9.62** Hierarchical linear models were first introduced by Lindley and Smith (1972). They have been applied and extended in many different ways. See Smith (1973a, 1973b), Fearn (1975), Young (1977), Haitovsky (1987), Lee (1987), Albert (1988a, 1988b), Pericchi and Nazaret (1988), Polasek (1988), Datta and Ghosh (1991), Stroud (1991) and Lange et al. (1992) for various early and more recent examples.

## Hierarchical models and prior beliefs

**9.63** It is interesting to examine the hierarchical linear model in terms of the extent to which it allows a wider range of prior beliefs to be expressed, in the light of the criticism in **9.41** of the natural conjugate family for the linear model. First, note that with known variances the hierarchical model produces a multivariate normal prior distribution for $\boldsymbol{\beta}$, which is a member of the natural conjugate family of **9.42**. In this case the hierarchical modelling does not offer any greater variety of prior distributions. It does, however, provide a framework for thinking about a prior distribution for $\boldsymbol{\beta}$. In Example 9.9, for instance, prior beliefs about the $\mu_i$s would certainly include correlation between them, since if You learn that one subject has a high measurement You will tend to expect a higher measurement of others. It is generally hard to think about correlations, and the hierarchical model simplifies this process by introducing a common mean hyperparameter $\xi$, conditional on which the $\mu_i$s are independent. The primary strength of hierarchical models is in this structuring of possibly complex prior beliefs in terms of simpler constructs.

If we now consider the case of unknown variances it is clear that the hierarchical model with two unknown variances $\sigma^2$ and $\tau^2$, as in **9.59**, is distinct from the simple linear model. The way those variances enter into the conditional distribution of $\boldsymbol{\beta}$ and $\gamma$ given $\sigma^2$ and $\tau^2$ is more rigid than in the most general conditional conjugate family (9.63), but it would be simple to generalize in the same way.

## Heavy-tailed models

**9.64** Another way to broaden the class of prior distributions is to use heavy-tailed priors. The use of heavy-tailed distributions in the context of robustness is discussed in **7.26** to **7.29**. It is certainly convenient, having expressed a prior mean and standard deviation for a parameter $\theta$, to complete the prior specification by assuming a normal prior distribution, but often this gives unrealistically thin tails. With a normal prior distribution, the prior probability that $\theta$ will lie more than, say, two and a half prior standard deviations from its prior mean is very small (0.0124), and in practice we wish to allow somewhat more probability to the event of the prior mean being far from the true value. In other words, prior beliefs are often better represented by a heavier-tailed distribution than the normal.

A useful family of heavy-tailed distributions is the $t$ family. Under the natural conjugate normal-inverse-gamma family, the marginal distribution of $\boldsymbol{\beta}$ is a $t$ distribution, so it seems that the natural conjugate distributions already incorporate hevy tails. However, this is a feature of the relationship between $\boldsymbol{\beta}$ and $\sigma^2$ in the natural conjugate family, which is criticized in **9.41**. The distribution of $\boldsymbol{\beta}$ given $\sigma^2$ is normal, and when in **9.45** we removed the dependence between $\boldsymbol{\beta}$ and $\sigma^2$ we proposed instead a normal marginal distribution for $\boldsymbol{\beta}$, (9.62). The more general conditional conjugate family (9.63) still does not admit a heavy-tailed marginal distribution for $\boldsymbol{\beta}$ except by inducing the same form of correlation between $\boldsymbol{\beta}$ and $\sigma^2$ as is imposed by the natural conjugate family.

**9.65** Let us instead assign a $t_d(\mathbf{m}, \mathbf{W})$ distribution for $\boldsymbol{\beta}$ independently of $\sigma^2$, so that the prior distribution is

$$f(\boldsymbol{\beta}, \sigma^2) \propto \{1 + (\boldsymbol{\beta} - \mathbf{m})'\mathbf{W}^{-1}(\boldsymbol{\beta} - \mathbf{m})\}^{-(d+p)/2} f(\sigma^2). \tag{9.84}$$

This is not a member of the natural conjugate or conditional conjugate families. Multiplying by the likelihood function yields an intractable posterior distribution, in the sense that the conditional distribution of $\boldsymbol{\beta}$ given $\sigma^2$ is not normal and we cannot integrate analytically with respect to $\boldsymbol{\beta}$. Nevertheless, we can achieve a relatively tractable analysis by a simple device that recognizes the derivation of a $t$ distribution as a marginal distribution in a normal-inverse-gamma joint distribution.

Consider a hierarchical model in which we introduce a hyperparameter $\tau^2$ by letting the conditional prior distribution of $\boldsymbol{\beta}$ be $N(\mathbf{m}, \tau^2\mathbf{V})$. Then at the next stage of the hierarchy we give $\tau^2$ an $IG(a, d)$ distribution. Then

$$f(\boldsymbol{\beta}, \sigma^2, \tau^2) \propto (\tau^2)^{-p/2} \exp\{-(\boldsymbol{\beta} - \mathbf{m})'\mathbf{V}^{-1}(\boldsymbol{\beta} - \mathbf{m})/(2\tau^2)\}(\tau^2)^{-(d+2)/2} \exp\{-a/(2\tau^2)\} f(\sigma^2). \tag{9.85}$$

Integrating $\tau^2$ out of (9.85) yields (9.84) with $\mathbf{W} = a\mathbf{V}$. The joint distribution of $\boldsymbol{\beta}$ and $\tau^2$ is of course $NIG(a, d, \mathbf{m}, \mathbf{V})$.

Using the representation (9.85), multiplying by the likelihood (9.2) yields the posterior distribution $f(\boldsymbol{\beta}, \sigma^2, \tau^2 \mid \mathbf{y})$. The conditional posterior distribution of $\boldsymbol{\beta}$ given $\sigma^2$ and $\tau^2$ is $N(\mathbf{m}^\star, \mathbf{W}^\star)$, where $\mathbf{m}^\star$ and $\mathbf{W}^\star$ are given by (9.68) and (9.69) respectively, with $\mathbf{W} = \tau^2\mathbf{V}$ and $\mathbf{C} = \sigma^2\mathbf{I}$. We can then integrate with respect to $\boldsymbol{\beta}$ to obtain

$$f(\sigma^2, \tau^2 \mid \mathbf{y}) \propto |\mathbf{W}^\star|^{-1/2}(\sigma^2)^{-n/2}(\tau^2)^{-(d+p+2)/2} \exp(-T/2) \exp\{-a/(2\tau^2)\} f(\sigma^2), \tag{9.86}$$

where

$$T = \sigma^{-2}\mathbf{y}'\mathbf{y} + \tau^{-2}\mathbf{m}'\mathbf{V}^{-1}\mathbf{m} - (\mathbf{m}^\star)'(\mathbf{W}^\star)^{-1}\mathbf{m}^\star.$$

As with the hierarchical linear model with unknown variances, we have reduced the dimensionality to this two-dimensional posterior distribution for $\sigma^2$ and $\tau^2$. Many inferences about $\boldsymbol{\beta}$ can be made via this representation, such as calculating $E(\boldsymbol{\beta} \mid \mathbf{y})$ as the expectation of $\mathbf{m}^\star$ (a function of $\sigma^2$ and $\tau^2$) with respect to (9.86). Gibbs sampling is also very easy to implement, particularly if $f(\sigma^2)$ is an inverse gamma distribution.

**9.66** This device, of introducing an extra unknown variance, is very generally applicable. The new variance parameter is given an inverse-gamma distribution, and the original parameters have a normal distribution conditional on the unknown variance but a heavy-tailed $t$ distribution unconditionally. However many unknown variances we introduce in this way tractable normal posterior distributions are obtained conditional on all the unknown variances. We can integrate down to the marginal posterior distribution of the unknown variances, which is generally intractable but will be amenable to numerical integration if sufficiently low dimensional. Gibbs sampling is always straightforward, and only requires sampling from normal and inverse-gamma distributions. See references in **7.26**.

*Example 9.11*
The hierarchical linear model with unknown variances, analysed in **9.59**, is an example of this technique. We could also give $\gamma$ a $t$ distribution by letting $\mathbf{C}_2 = \omega^2\mathbf{V}_2$, say, and assuming an inverse-gamma distribution for $\omega^2$.

*Example 9.12*
Instead of an error variance matrix $\sigma^2\mathbf{I}$ in the linear model, where each $\epsilon_i$ has the same variance $\sigma^2$, we could let each $\epsilon_i$ have its own variance $\sigma_i^2$. We now have $n$ unknown error variances, equivalent to assuming a heavy-tailed error distribution. Lindley (1971) presents a hierarchical prior distribution for such a set of error variances.

### Generalizations of the linear model

**9.67** The analysis presented in this chapter can be generalized in a variety of ways. Multivariate linear models, in which each observation of the response variable $y$ is a vector random variable, are introduced in **10.28**, in the simplest case of a multivariate normal sample. Non-normal structures for the error can be considered within a class of generalized linear models, for which Bayesian analysis is given by Albert (1988b), Ibrahim and Laud (1991) and Eaves and Chang (1992).

Dynamic linear models allow the parameter vector $\boldsymbol{\beta}$ to evolve over time. These and other models with variable parameters are considered in **10.42** to **10.47**, and **10.50** to **10.52**. For other variations on the structure and assumptions of the linear model, see for example Reilly and Patino-Leal (1981), Buonaccorsi and Gatsonis (1988), Bagchi et al. (1990), Chib and Tiwari (1991) and Lee (1992). Other specialized forms of linear model are important in econometrics; see Morales (1971), Zellner (1971), Ilmakunnan (1985), Tsurumi (1985), Steel (1991), Steel and Richard (1991), Percy (1992) and Chib (1993).

A good source of theory concerning linear models generally, and covering several generalizations, is Broemeling, (1985).

## EXERCISES

9.1 Consider the simple regression model of Example 9.2 with conjugate prior distribution. Inference is required for $\xi = -\alpha/\beta$, the intercept of the regression line with the $x$-axis. Prove that $E(\xi \mid \mathbf{y})$ does not exist.

In the case of known $\sigma^2$, with a normal prior distribution as in **9.42**, show that $f(\xi \mid \mathbf{y})$ can be obtained explicitly in terms of the standard normal d.f. $\Phi$ by differentiating

$$P(\xi \leqslant t \mid \mathbf{y}) = P(\alpha - \beta t \leqslant 0, \beta \geqslant 0) + P(\alpha - \beta t \geqslant 0, \beta \leqslant 0).$$

9.2 As an alternative to the determinant and trace criteria for experimental design mentioned in **9.14**, one might follow the approach of scoring rules and choose an experiment to minimize the entropy (2.54) of the appropriate posterior distribution.

Prove first that the entropy of the $NIG(a, d, \mathbf{m}, \mathbf{V})$ distribution is

$$[d + p + (d + p + 2)\{\log(a/2) - \psi(d/2)\}]/2,$$

where $\psi(t)$ is the digamma function $\mathrm{d}\log\Gamma(t)/\mathrm{d}t$. Deduce that in the case of a conjugate $NIG(a, d, \mathbf{m}, \mathbf{V})$ distribution the entropy of the full joint posterior distribution is not a useful criterion for experimental design.

Prove also that in the case of known variance presented in (9.42), minimizing the entropy of the posterior density of $\boldsymbol{\beta}$ produces the determinant criterion of maximizing $|\mathbf{W}^{-1} + \sigma^{-2}\mathbf{X}'\mathbf{X}|$.

9.3 Observations $y_{ij}$ $(i = 1, 2,\ j = 1, 2, \ldots, n_i)$ are independently distributed as $N(\mu_i, \sigma^2)$ given $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma^2)$. Write this as a linear model and consider constructing a Bayes factor for this model against the alternative that $\mu_2 = 0$, under weak prior information. Using the approach of **9.35** a minimal experiment must either have $n_1 = 1$, $n_2 = 2$ or $n_1 = 2$, $n_2 = 1$. Show that the Bayes factors in these two cases differ by a factor of $\sqrt{2}$.

9.4 Consider a linear model with known variance as in **9.42**. However, normality is not assumed, and so the model simply states that $E(\mathbf{y} \mid \boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$, $\mathrm{var}(\mathbf{y} \mid \boldsymbol{\beta}) = \sigma^2\mathbf{I}$. Similarly, the only assertions of prior information are $E(\boldsymbol{\beta}) = \mathbf{m}$, $\mathrm{var}(\boldsymbol{\beta}) = \mathbf{W}$. Prove that the Bayes linear estimator (6.34) for $\boldsymbol{\beta}$ as a linear function of $\mathbf{y}$ is (9.60), and that the corresponding dispersion matrix (6.33) is (9.61).

9.5 Express the hierarchical model of **9.52** as a linear model

$$\mathbf{y} = \mathbf{Z}\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix} + \mathbf{e}$$

for appropriate matrix $\mathbf{Z}$ and a non-hierarchical prior distribution for $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and $\sigma^2$. Verify that the generalized posterior mean and variance (9.68) and (9.69) for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are as given in **9.53**.

9.6 With $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ defined as in **9.54**, prove that an alternative expression for $\hat{\boldsymbol{\gamma}}$ is

$$\hat{\boldsymbol{\gamma}} = (\mathbf{X}_1'\mathbf{P}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{P}^{-1}\hat{\boldsymbol{\beta}},$$

where $\mathbf{P} = \mathbf{C}_1 + (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}$.

9.7 Prove the results (9.75) and (9.76) by inverting the matrix $\mathbf{V}$ in (9.71).