CHAPTER 5

Adjusting for Mismeasured Categorical Variables

The previous chapter described how to carry out analyses which adjust for mismeasurement in continuous explanatory variables, to avoid the sorts of biases described in Chapter 2. Now we turn attention to analyses which adjust for mismeasurement in binary explanatory variables, to avoid the sorts of biases described in Chapter 3. In Sections 5.1 and 5.2 we consider a binary outcome and a binary explanatory variable, with data arising from a case-control study design. In the first instance the extent of misclassification is known exactly, whereas Section 5.2 extends to the more practical case of imperfect prior information about the misclassification. Section 5.3 moves on to consider situations where there is little or no prior information about the misclassification, but several different noisy measurements of the explanatory variable are available. Section 5.4 extends further to situations where additional precisely measured explanatory variables are measured. Some brief concluding remarks appear in Section 5.5, while mathematical details are gathered in Section 5.6.

5.1 A Simple Scenario

Say that the relationship between a binary outcome variable Y and a binary exposure variable V is of interest, but a surrogate binary variable V^* is measured in lieu of V. As a simple initial scenario, say that data are collected via a retrospective case-control study, with n_0 subjects sampled at random from the control ($Y = 0$) population, and n_1 subjects sampled at random from the case ($Y = 1$) population. Thus samples from the conditional distribution of $V^*|Y$ are obtained, for both $Y = 0$ and $Y = 1$. Of course a naive analysis which treats V^* as if it were V will typically yield biased inferences concerning the relationship between V and Y, as described in Chapter 3. However, if enough is known about the mismeasurement process then it can be quite straightforward to make inferences which account for the mismeasurement. Particularly, say that the distribution of $V^*|V, Y$ is known completely. Then, paralleling the analyses presented in Chapter 4, inference could proceed by including each subject's unobserved V variable as an unknown quantity described by the posterior distribution. In fact, an even simpler approach to Bayesian analysis suffices in the present scenario. It is easy to directly simulate draws from the posterior distribution of the unknown parameters given the observed data, without incorporating the V variables explicitly.

To be more specific, let r_0 and r_1 be the prevalence of exposure in the

control and case populations respectively. That is, the chance that a sampled control subject is exposed is $r_0 = Pr(V = 1|Y = 0)$ and the chance that a sampled case subject is exposed is $r_1 = Pr(V = 1|Y = 1)$. As alluded to in Chapter 3, usually the parameter of interest for describing the association between V and Y is the odds-ratio

$$\Psi = \frac{r_1/(1-r_1)}{r_0/(1-r_0)}. \tag{5.1}$$

Of course it is well known that Ψ has a dual interpretation as both the ratio of odds that $Y = 1$ given $V = 1$ to odds that $Y = 1$ given $V = 0$, and the ratio of odds that $V = 1$ given $Y = 1$ to odds that $V = 1$ given $Y = 0$. It is often convenient to think in terms of the log odds-ratio which, for future reference, we denote as $\psi = \log \Psi$.

Under the common assumption of nondifferential misclassification, whereby the conditional distribution of $V^*|V, Y$ does not actually depend on Y, the magnitude of misclassification is characterized in terms of sensitivity and specificity. Whereas in Chapter 3 we used (SN, SP) to denote these quantities, now we use (p, q) for the sake of compactness in mathematical expressions. That is, $p = Pr(V^* = 1|V = 1)$ is the sensitivity and $q = Pr(V^* = 0|V = 0)$ is the specificity. In the present scenario we take (p, q) to be known. Of course in many real scenarios the assumptions of nondifferential misclassification and known sensitivity and specificity will be questionable at best. As noted in Chapter 3, the nondifferential assumption can be particularly tenuous in a retrospective study, particularly when the surrogate V^* is based on subject self-report. Moreover, while investigators may have reasonable estimates or guesses for the values of p and q, scenarios where these values are actually known are rare. Indeed, even if it is reasonable to think p and q are 'known' from a previous study in which both V^* and V were measured, there may be questions about whether (p, q) are transportable across different study populations. Thus one must be circumspect about the extent to which the illustrative analysis in this section is realistic.

In the simple situation at hand, adjusting estimated odds-ratios to account for misclassification has been discussed at length in the literature (see, for instance, Barron 1977, Greenland and Kleinbaum 1983, Morrissey and Spiegelman 1999). Here we present a Bayesian approach to such adjustment. Whereas most of the discussion in the epidemiology literature involves only the adjustment of point estimates, the Bayesian approach has the advantage of automatically adjusting interval estimates simultaneously.

The observed data can be summarized by (S_0, S_1), where S_0 is the number of the n_0 cases who are *apparently* exposed in the sense that $V^* = 1$, while similarly S_1 is the number of the n_1 cases who are apparently exposed. Since an apparent exposure can result from the correct classification of an exposed subject or from the incorrect classification of an unexposed subject, it immediately follows that $S_i \sim Binomial(n_i, \theta_i)$, where

$$\theta_i = Pr(V^* = 1|Y = i)$$

$$
\begin{aligned}
&= Pr(V=1|Y=i)Pr(V^*=1|V=1,Y=i) \; + \\
&\quad Pr(V=0|Y=i)Pr(V^*=1|V=0,Y=i) \\
&= r_i p + (1-r_i)(1-q) \qquad (5.2)
\end{aligned}
$$

is the probability of *apparent* exposure given $Y=i$. Via (5.2) one can regard either (r_0, r_1) or (θ_0, θ_1) as the unknown parameters in the statistical model for the observed data. In going back and forth between these two parameterizations, note that while each r_i can take on any value between zero and one, each θ_i is necessarily constrained to lie in the interval from $l(p,q) = \min\{p, 1-q\}$ to $r(p,q) = \max\{p, 1-q\}$. Note as well that in the second parameterization the odds-ratio is expressed as

$$
\Psi = \frac{(\theta_1 + q - 1)/(p - \theta_1)}{(\theta_0 + q - 1)/(p - \theta_0)}. \qquad (5.3)
$$

In the absence of well-formed prior information an investigator might reasonably take the priors for r_0 and r_1 to be uniform distributions on the unit interval, with an assumption of prior independence. That is, $r_i \sim U(0,1)$ independently for $i = 0, 1$. Clearly this implies that $\theta_i \sim U\{l(p,q), r(p,q)\}$ independently for $i = 0, 1$. Whereas a uniform prior on the unit interval and a binomial likelihood lead to a Beta posterior, a uniform prior on a sub-interval of $(0,1)$ will yield a Beta distribution truncated to the sub-interval as the posterior distribution. Particularly, θ_0 and θ_1 are conditionally independent given the data (S_0, S_1), with $\theta_i|S_0, S_1$ following the $Beta(S_i + 1, n_i - S_i + 1)$ distribution truncated to $\{l(p,q), r(p,q)\}$. Thus it is trivial to simulate from the posterior distribution in the θ parameterization, and thereby obtain a Monte Carlo posterior sample for Ψ via (5.3). We emphasize that this is direct rather than MCMC simulation, so there are no concerns about convergence or mixing associated with the posterior sample.

To test this inferential scheme four synthetic datasets are simulated, each with sample sizes $n_0 = n_1 = 2500$. Datasets (i) and (ii) are based on the same retrospective samples of $V|Y$, with exposure prevalences $(r_0, r_1) = (0.2, 0.25)$. The surrogate V^* values arise from $(p,q) = (0.95, 0.9)$ in (i), and from $(p,q) = (0.85, 0.9)$ in (ii). Datasets (iii) and (iv) are generated with lower underlying prevalences of $(r_0, r_1) = (0.05, 0.0656)$, to mimic the common use of case-control analysis when faced with relatively rare exposures. The misclassification arises from $(p,q) = (0.95, 0.90)$ in (iii) and $(p,q) = (0.85, 0.90)$ in (iv), thereby matching the first two datasets. Note that both choices of prevalences (r_0, r_1) give an odds-ratio of $\Psi = 1.333$, and consequently a log odds-ratio of $\psi = 0.288$. Both the actual exposure totals and the apparent exposure totals for the four datasets appear in Table 5.1

Figure 5.1 displays the posterior distribution of the log odds-ratio ψ for each dataset, as determined in the manner described above. We refer to this as the *adjusted* posterior distribution in the sense that the analysis is adjusted to account for the misclassification. For the sake of comparison, both the *naive* and *ideal* posterior distributions of the log odds-ratio are also presented. The naive posterior results from ignoring the misclassification and treating the

	number actually exposed		number apparently exposed	
dataset	control	case	control	case
(i)	512/2500	641/2500	660/2500	773/2500
(ii)	512/2500	641/2500	637/2500	732/2500
(iii)	133/2500	175/2500	336/2500	394/2500
(iv)	133/2500	175/2500	321/2500	377/2500

Table 5.1 *Summary of the synthetic datasets considered in Section 5.1 Both the number of actual exposures and the number of apparent exposures are given for each sample. The underlying prevalences* (r_0, r_1) *and classification probabilities* (p, q) *are given in the text.*

surrogate exposure status V^* as if it were the actual exposure status V. In terms of the analysis described above, the naive posterior distribution arises from incorrectly assuming that $p = q = 1$. Alternately, the ideal posterior distribution results when V itself can be observed and used as the exposure variable, in which case V^* can be discarded. Of course we only have the luxury of determining the ideal posterior distribution because we simulate the data and thus have measurements of V. Thus the ideal posterior is of interest only for the sake of comparison and assessment of how much inferences are weakened because of misclassification.

For all four datasets the mode of the naive posterior distribution lies closer to zero than that of the ideal posterior distribution. Thus the attenuation that results from ignoring the misclassification is evident. In datasets (i) and (ii) the modes of both the adjusted and ideal posterior distributions are close to the true value of ψ, but the adjusted distribution is slightly wider than the ideal distribution. This reflects the loss of information associated with misclassification. In datasets (iii) and (iv) the adjusted posterior distribution is both wider and shifted to the right in relation to the ideal posterior distribution, with the discrepancy between the adjusted and ideal posteriors being comparable in magnitude to the discrepancy between the naive and ideal posteriors. Thus for datasets (i) and (ii) the adjusted analysis is clearly better than the naive analysis, while the benefit of adjustment is less obvious for datasets (iii) and (iv).

To investigate the benefit associated with adjustment more thoroughly, we turn to some simple comparisons based on asymptotic estimator performance. Let $\hat{\psi}_I$, $\hat{\psi}_A$, and $\hat{\psi}_N$ be estimators of ψ based on the ideal, adjusted, and naive posterior distributions respectively. For the sake of argument say these estimators are modes of the respective posterior distributions, but it is well known that posterior means or medians have the same asymptotic performance, and in particular all these estimators are asymptotically equivalent to the maximum likelihood estimator.

In the case of the ideal analysis, one has a correctly specified model and

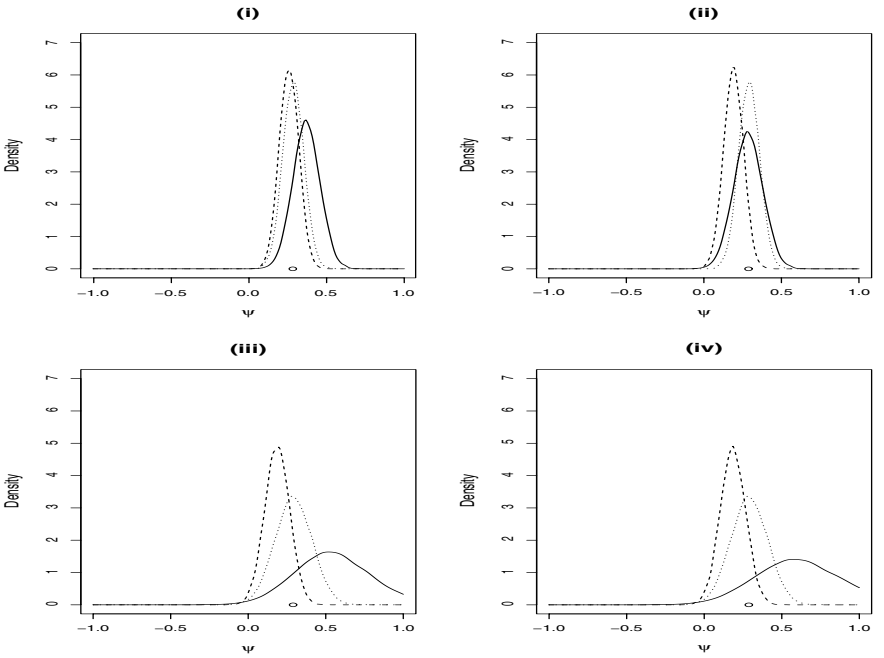

Figure 5.1 *Naive, corrected, and ideal posterior distributions of the log odds-ratio based on the four synthetic datasets given in Table 5.1. For each dataset the dashed curve corresponds to the naive analysis, the solid curve to the adjusted analysis, and the dotted curve to the ideal analysis.*

hence no bias asymptotically. Considering sample proportions of actual exposure we have $Var(\hat{r}_i) = n_i^{-1} r_i(1-r_i)$. Applying the 'delta method' to the logarithm of (5.1) then gives

$$\begin{aligned} MSE(\hat{\psi}_I) &\approx Var(\hat{\psi}_I) \\ &\approx \frac{1}{n_0 r_0 (1-r_0)} + \frac{1}{n_1 r_1 (1-r_1)}. \end{aligned} \tag{5.4}$$

Similarly, the adjusted analysis involves no bias asymptotically, provided correct values of p and q are specified. Considering sample proportions of apparent exposure we have $Var(\hat{\theta}_i) = n_i^{-1}\theta_i(1-\theta_i)$. Viewing $\hat{\psi}_A$ as a function of $(\hat{\theta}_0, \hat{\theta}_1)$ according to (5.3) and then applying the delta method gives

$$\begin{aligned} MSE(\hat{\psi}_A) &\approx Var(\hat{\psi}_A) \\ &\approx \left\{ \frac{p+q-1}{(\theta_0+q-1)(p-\theta_0)} \right\}^2 \frac{\theta_0(1-\theta_0)}{n_0} + \\ &\quad \left\{ \frac{p+q-1}{(\theta_1+q-1)(p-\theta_1)} \right\}^2 \frac{\theta_1(1-\theta_1)}{n_1}. \end{aligned} \tag{5.5}$$

Of course this expression could be re-cast in terms of (r_0, r_1) rather than (θ_0, θ_1), although doing so does not yield any obvious insights in terms of comparison with (5.4).

Finally, the naive analysis does involve a bias, with estimates of θ_i appearing where estimates of r_i 'ought' to appear. Thus

$$\begin{aligned} MSE(\hat{\psi}_N) &= Bias^2(\hat{\psi}_N) + Var(\hat{\psi}_N) \\ &\approx \left[\log\left\{\frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}\right\} - \psi\right]^2 + \\ &\quad \frac{1}{n_0\theta_0(1-\theta_0)} + \frac{1}{n_1\theta_1(1-\theta_1)}. \end{aligned} \tag{5.6}$$

Figure 5.1 plots the approximate root-mean-squared error in estimating the log odds-ratio ψ as a function of total sample size $n = n_0 + n_1$, assuming $n_0 = n_1$. Plots are given for each of the four combinations of (r_0, r_1, p, q) values studied above. The ideal, adjusted, and naive analyses are considered, so the square roots of (5.4), (5.5), and (5.6) are plotted. As one expects, the error arising from the ideal analysis is usually smaller than that of the adjusted and naive analyses, although in scenarios (iii) and (iv) $MSE(\hat{\psi}_N)$ is slightly smaller than $MSE(\hat{\psi}_I)$ at small sample sizes. Of course $\hat{\psi}_N$ involves an asymptotic bias, whereas $\hat{\psi}_A$ and $\hat{\psi}_I$ do not. Thus for any underlying value of (r_0, r_1, p, q) the naive estimator must be the worst of the three estimators at a large enough sample size. What is surprising, however, is that the sample size may have to be extremely large before the adjusted analysis has a lower MSE than the naive analysis. This is certainly the situation in scenarios (iii) and (iv).

In the narrow sense of bias-variance tradeoff then, the reduction in bias achieved by adjusting for misclassification may not fully offset the increased variability incurred by admitting that misclassification is present. However, this should not be viewed as a justification for using a naive rather than adjusted analysis. In particular, as it is based on a correct model, the adjusted analysis will yield credible intervals for the unknown parameters that achieve the nominal coverage probability asymptotically. In general we expect interval estimates generated via the naive analysis to exhibit undercoverage, given that an incorrect model is being used.

5.2 Partial Knowledge of Misclassification Probabilities

Of course the statistical model described in the previous section presumes that the sensitivity and specificity of the exposure assessment are known. This is likely to be unrealistic in many situations. More realistically, an investigator might have good guesses or estimates for these quantities, but not be convinced that these guesses are perfect. In such an instance the most obvious strategy would be to proceed as before, with the best guesses for the sensitivity and specificity substituted in place of the actual but unknown values. That is, one pretends that the best guesses are exactly correct. While this approach is

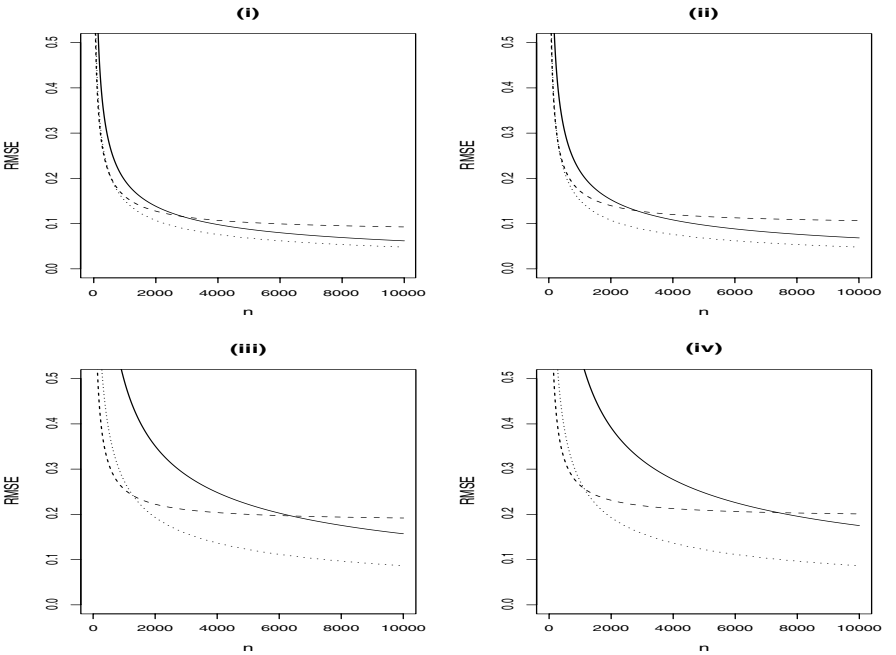

Figure 5.2 *Approximate MSE versus total sample size n for scenarios (i) through (iv). These scenarios underlie datasets (i) through (iv) considered in Table 5.1. In each scenario, the dotted curve corresponds to the ideal posterior, the solid curve to the adjusted posterior, and the dashed curve to the naive posterior.*

formally wrong, one might hope that some robustness obtains. In particular, one might speculate that if the guessed and actual classification probabilities are close to one another, then the corresponding posterior distributions of the log odds-ratio ψ will also be close to one another. Unfortunately, however, Gustafson, Le and Saskin (2001) show that such robustness need not obtain. In particular, they establish the following result.

Let $g(\,\cdot\,; p, q)$ be the function giving apparent exposure prevalences as a function of actual exposure prevalences for given sensitivity p and specificity q. That is, following (5.2),

$$g\left(\begin{pmatrix} r_0 \\ r_1 \end{pmatrix}; p, q\right) \;=\; \begin{pmatrix} r_0 p + (1-r_0)(1-q) \\ r_1 p + (1-r_1)(1-q) \end{pmatrix}.$$

Let (p, q) denote the actual sensitivity and specificity of the exposure assessment, and let (p^*, q^*) be the guessed sensitivity and specificity. Let Ψ be the actual odds-ratio, and let Ψ^* be the large-sample limit of the estimated odds-ratio if the analysis of Section 5.1 is applied with the incorrect values (p^*, q^*).

That is,

$$\Psi^* = \frac{r_1^*/(1-r_1^*)}{r_0^*/(1-r_0^*)},$$

where $r^* = (r_0^*, r_1^*)$ are the incorrectly inferred prevalences of actual exposure $r = (r_0, r_1)$ based on the incorrectly assumed values (p^*, q^*). Specifically,

$$r^* = g^{-1}\left(g(r; p, q); p^*, q^*\right). \tag{5.7}$$

Using (5.7), some algebraic manipulation leads to

$$\frac{\Psi^*}{\Psi} = \frac{\{1 + d/r_1\}/\{1 + c/(1-r_1)\}}{\{1 + d/r_0\}/\{1 + c/(1-r_0)\}}, \tag{5.8}$$

where $c = (p^* - p)/(p + q - 1)$ and $d = (q^* - q)/(p + q - 1)$. Note that the magnitudes of c and d reflect how close the guessed values are to the actual values. Note also that for any values of c and d, no matter how small, the ratio (5.8) can take on values from zero to infinity as the underlying prevalences (r_0, r_1) vary. This is the essence of the lack of robustness alluded to above. Arbitrarily small discrepancies between the guessed and actual classification probabilities can produce an arbitrarily large bias in the estimated odds-ratio, should the underlying prevalences of exposure happen to be 'unfavourable.'

If the guessed sensitivity and specificity are both overestimates, so that c and d are positive, then (5.8) approaches zero or infinity only when one of the prevalences r_0 or r_1 approaches zero or one. This is a minor comfort, as typically studies are not implemented for exposures with prevalences very close to zero or one. The situation is slightly worse, however, if one or both of the guessed values are underestimates. For instance, if d is negative than (5.8) goes to zero as r_1 decreases to $|d|$, and goes to infinity as r_0 decreases to $|d|$. Thus the difference between Ψ^* and Ψ can become arbitrarily large without the underlying exposure prevalences going to zero or one. In all, the form of (5.8) certainly speaks to a lack of inferential robustness when the assumed values of the sensitivity and specificity differ slightly from the actual values.

As an aside, (5.8) also reveals that discrepancies between the assumed and actual sensitivity and specificity can possibly lead one to infer a nonsensical negative odds-ratio. For instance, say $c > 0$, $d < 0$, $r_0 < |d|$, $r_1 > |d|$. It is clear that (5.8) is negative under these conditions. In fact this can be regarded as useful, as the nonsensical inference alerts the investigator to the fact that the assumed values of (p, q) must be incorrect. In reality, however, the situation is not this clear-cut. For a real sample say $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1)$ are the sample proportions of apparent exposure. Then an estimate of Ψ based on estimated prevalences of the form $g^{-1}(\hat{\theta}; p^*, q^*)$ could be negative. However, this would only suggest, and not prove, that (p^*, q^*) are erroneous values. In fact, sampling error in estimating θ by $\hat{\theta}$ can cause $g^{-1}(\hat{\theta}; p, q)$ to yield the same problem even if the correct p and q are used. For the Bayes analysis of Section 5.1 the posterior distribution of θ involves truncation to the interval from $l(p^*, q^*)$ to $r(p^*, q^*)$, which by construction ensures that the posterior distribution on Ψ is restricted to positive values. However, a signal that (p^*, q^*) are wrong may arise if this

truncation is excluding values of θ which would otherwise receive appreciable posterior weight.

More generally, Gustafson, Le and Saskin (2001) define the notion of *asymptotically detectable miscorrection* to describe scenarios where one or both of the apparent prevalences (θ_0, θ_1) do *not* lie between $1 - q^*$ and p^*. Since by definition the apparent prevalences must lie between $1 - q$ and p, if the data support one or both apparent prevalences falling outside $(1 - q^*, p^*)$, then the data equally support the notion that the guessed values (p^*, q^*) differ from the actual values (p, q). It is easy to check that a negative value of (5.8) implies an asymptotically detectable miscorrection, though asymptotically detectable miscorrection does not necessarily imply a negative value of (5.8). Gustafson, Le and Saskin (2001) emphasize that the lack of robustness exhibited in (5.8) is still manifested if scenarios involving asymptotically detectable miscorrection are ruled out. Roughly put, one can obtain arbitrarily wrong answers from arbitrarily good guesses for p and q, without this error being detectable from the observed data.

Having suggested that treating good guesses for (p, q) as if they are exact can be problematic, we now consider the alternate strategy of treating (p, q) as unknown parameters. In particular, prior distributions for (p, q) could be centred at the guessed values (p^*, q^*), with the variation in these distributions taken to reflect the investigator's beliefs about the precision of these guesses. Thus we have two possible strategies for analysis which adjusts for the misclassification. For the sake of clarity we distinguish between them as follows. Pretending the guessed values are exact is referred to as *adjusting with false certainty* (AWFC), while assigning priors centred at the guessed values is referred to as *adjusting with uncertainty* (AWU).

The AWU approach to inference is particularly simple if the four unknown parameters (r_0, r_1, p, q) are considered to be independent *a priori*, with prior distributions

$$\begin{aligned} r_0 &\sim Unif(0,1), \\ r_1 &\sim Unif(0,1), \\ p &\sim Beta(\alpha_p, \beta_p), \\ q &\sim Beta(\alpha_q, \beta_q). \end{aligned}$$

In particular we are contemplating situations where the Beta prior distributions for p and q are fairly peaked, as the investigator may have a reasonable idea about how well the exposure assessment performs. On the other hand, one may want to be objective about r_0 and r_1, particularly if the odds-ratio comparing the two prevalences is the parameter of inferential interest. We note that in this framework one could also assign uniform prior distributions to p and q, by setting $\alpha_p = \beta_p = \alpha_q = \beta_q = 1$. However, the model is nonidentifiable, and it seems impossible to generate reasonable inferences with a nonidentifiable model and flat priors for all the parameters. A more focussed discussion of identifiability is given in Section 6.3, but for now we simply note that the model could be reparameterized with $(\theta_0, \theta_1, p, q)$ rather than

(r_0, r_1, p, q) as the unknown parameters, where, as before, (r_0, r_1) are the actual exposure prevalences and (θ_0, θ_1) are the apparent exposure prevalences. This bring the nonidentifiability to the fore, as with the second parameterization the distribution of the data given all the parameters in fact depends only on (θ_0, θ_1), and not on (p, q).

On first glance MCMC fitting of the AWU model would appear to be extremely straightforward. If one works with the original (r_0, r_1, p, q) parameterization and augments the parameter space to include the true but unobserved exposure status of each subject, then all posterior full conditional distributions are standard distributions. Thus it is trivial to implement Gibbs sampling. The details are given in Section 5.6.1. In turns out, however, that the Gibbs sampler does not work well for all datasets in the present context. In a different context Gelfand and Sahu (1999) report poor mixing of the Gibbs sampler for posterior distributions arising from nonidentifiable models, and we have found this to be a problem here as well. Thus Section 5.6.2 describes an alternate MCMC algorithm developed for the AWU model by Gustafson, Le and Saskin (2001). This algorithm operates with the $(\theta_0, \theta_1, p, q)$ parameterization where the lack of identifiability is clearer, and it does *not* augment the parameter space with the true but unobserved exposure status of the study subjects. Specifically, this algorithm is tailored to take advantage of the fact that (θ_0, θ_1) are identified parameters while (p, q) are not. In Section 5.6.3 we demonstrate that for some synthetic data sets the alternate algorithm works quite well while the Gibbs sampler works quite poorly.

The merits of the AWU model compared to the AWFC model are investigated by way of an example with synthetic data. Say that in the context at hand the investigator's best guess for the sensitivity of the exposure assessment is $p^* = 0.9$, and furthermore he is quite confident that the sensitivity is at least 0.8. Thus he assigns p the Beta distribution which has its mode at 0.9 and gives probability 0.9 to the interval $(0.8, 1)$. This determines $p \sim Beta(29.3, 4.1)$. Also, say the investigator's best assessment is that false positives do not arise in the exposure assessment, so that $q^* = 1$. However, he is not certain of this fact. Therefore he takes the prior density for q to have its mode (with finite density) at one, with probability 0.9 assigned to the interval $q \in (0.95, 1)$. This leads to $q \sim Beta(44.9, 1)$. In fact, this argument leading to this particular prior specification was employed by Gustafson, Le and Saskin (2001) in a real scenario where the investigators were uncertain about the assessment of loss of heterozygosity in a cancer case-control study. The two prior densities appear in the first panel of Figure 5.3.

Now say that unbeknownst to the investigators the true sensitivity and specificity of the exposure assessment are $p = 0.85$ and $q = 0.95$. With respect to these actual values, the specified prior distributions are perhaps of 'typical' quality. That is, Figure 5.3 shows that in both cases the actual value lies in the flank of the prior distribution, rather than near the mode or in the tail. Also, say the true exposure prevalences are $r_0 = 0.05$ and $r_1 = 0.0798$, so that $\psi = 0.5$ is the true log odds-ratio. Under these true values of (r_0, r_1, p, q) synthetic data sets are simulated for three sample sizes. For $n_0 = n_1 = 250$ we

obtain 15/250 and 33/250 apparent exposures in the control and case samples respectively. Samples that are bigger by a factor of four yield 90/1000 and 113/1000 apparent exposures, while samples that are four times larger again give 366/4000 and 430/4000 apparent exposures.

For each synthetic dataset three posterior distributions for ψ appear in Figure 5.3. The AWU posterior results from the specified prior distributions, while the AWFC posterior is based on the assumption that $(p, q) = (p^*, q^*)$. For the sake of comparison we also present posterior distributions arising if the true values of (p, q) are known. This is referred to as *adjusting with correct certainty* (AWCC). That is, the method of Section 5.1 is applied using the correct values of the sensitivity and specificity. Of course the AWCC posterior distributions are included for the sake of comparison only, as we are interested in real scenarios where the correct values are not known.

A first comment about the posterior distributions is that for identifiable statistical models we expect to see inferences which are twice as precise if the sample size is quadrupled. Indeed, this sort of narrowing with sample size is manifested by the AWFC and AWCC posterior distributions, regardless of the fact that the former posterior distribution is based on an incorrect statistical model, while the latter is based on a correct model. For the nonidentifiable AWU model we do see roughly a factor of two improvement in precision going from $n_j = 250$ to $n_j = 1000$; however the further improvement in going from $n_j = 1000$ to $n_j = 4000$ is much more modest. The large-sample behaviour of posterior distributions arising from nonidentifiable models is considered more closely in Section 6.3.

Figure 5.3 indicates that at all three sample sizes the AWU posterior distribution is better than the AWFC posterior distribution in the sense of being closer to the AWCC posterior distribution. Thus there seems to be a benefit to admitting uncertainty about the sensitivity and specificity via a prior distribution rather than simply pretending that the guessed values of these quantities are correct. More specifically, the AWU posterior distributions tend to be only slightly wider than their AWC counterparts, but importantly this extra width is obtained by extending to the right rather than to the left. In particular, at the largest sample size this extension allows the AWU posterior to cover the true value of $\psi = 0.5$ which is missed by the AWFC posterior.

Of course Figure 5.3 describes inferences arising for particular true values of the parameters (r_0, r_1, p, q). A small simulation study is performed to assess whether there is a general benefit in adjusting with uncertainty compared to adjusting with certainty. We retain the prior distributions described above; that is, $r_0 \sim Unif(0, 1)$, $r_1 \sim Unif(0, 1)$, $p \sim Beta(29.3, 4.1)$, $q \sim Beta(44.9, 1)$. But now we consider average performance when each synthetic dataset is constructed by first simulating values of (p, q, r_0, r_1) from this prior distribution, and then simulating data summaries (S_0, S_1) given (p, q, r_0, r_1). Technically speaking, by examining the average performance of an estimator on repeated synthetic datasets where the true underlying parameters vary from dataset to dataset we are quantifying the Bayes performance of the estimator.

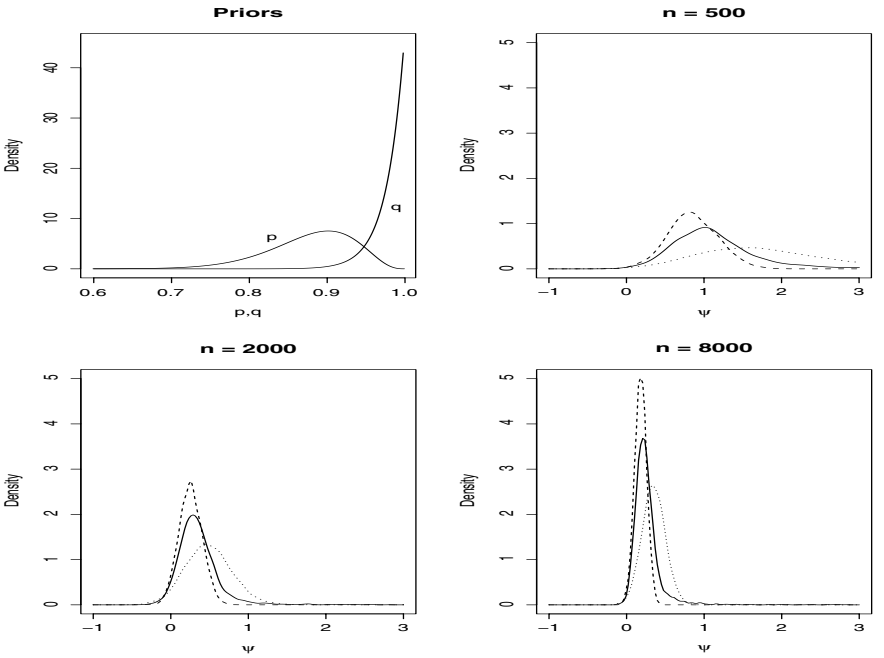

Figure 5.3 *Prior distributions for* (p, q) *and posterior distributions for* ψ *in the synthetic data example of Section 5.2. At each sample size the AWFC posterior density (dashed curve), the AWU posterior density (solid curve), and the AWCC posterior density (dotted curve) are given.*

For each synthetic dataset the posterior mean and equal-tailed 80% credible interval for ψ are computed, for each of the AWFC, AWU, and AWCC posterior distributions. These point and interval estimates are then compared to the true value of ψ, bearing in mind that this true value changes with each repetition in the simulation study. Table 5.2 summarizes the results based on 400 synthetic datasets, in terms of the root-mean-squared error of the point estimate, and the coverage probability and average length of the interval estimate.

Table 5.2 shows that the benefit of adjusting with uncertainty over adjusting with false certainty persists more generally, both in terms of smaller $RMSE$ for the point estimate and closer to nominal coverage for the interval estimate. We should point out that both the AWU and $AWCC$ credible intervals are in fact guaranteed to have actual coverage probabilities equal to their nominal coverage probabilities when the underlying parameter values are repeatedly sampled from the prior distribution. In particular, this fundamental property is entirely unaffected by whether the model is identifiable or not. While the $AWFC$ credible intervals tend, on average, to be only half as

	AWFC	AWU	AWCC
RMSE	0.83	0.74	0.45
Coverage	51%	78%	79%
Avg. Length	0.49	0.95	0.64

Table 5.2 *Bayes performance of the adjusted with false certainty (UWFC), adjusted with uncertainty (AWU), and adjusted with correct certainty (AWCC) analyses. The reported results are based on* 400 *simulated datasets, with the true parameters repeatedly sampled from their prior distribution. The RMSE of the posterior mean and the empirical coverage probability and average length of the* 80% *equal-tailed credible interval are reported. Note that with* 400 *repetitions the reported empirical coverage probability will have a simulation standard error of* 0.02 *if the procedure has an actual coverage probability equal to the nominal* 0.8.

wide as the AWU intervals, the low actual coverage probability indicates that this narrowness is neither justified nor desirable.

5.3 Dual Exposure Assessment

Section 5.1 focussed on treating the sensitivity and specificity of the exposure assessment as known, while Section 5.2 considered assigning highly informative priors to these quantities. It may be possible and desirable to avoid both of these approaches, if two different techniques can be used to assess exposure. Such a strategy is fairly common in epidemiological studies. For instance, Hui and Walter (1980) give an example where the exposure of interest is the presence of tuberculosis, with two different diagnostic tests, the Mantoux test and the Tine test, being applied to study subjects. In a case-control study of sudden infant death syndrome, Drews, Flanders and Kosinski (1993) use patient interviews and medical records to provide two assessments of whether subjects are 'exposed' to various putative binary risk factors. Joseph, Gyorkos and Coupal (1995) consider a study in which both a serology test and stool examination are used to test for a particular parasitic infection. And, in a somewhat more complicated setting, Iversen, Parmigiani and Berry (1999) consider a lab test and a family history 'test' for genetic susceptibility to breast cancer.

Letting V again be the actual but unobserved binary exposure variable, V_1^* and V_2^* are used to denote the two imperfect surrogates for V arising from the two assessment schemes. Initially the utility of having two exposure assessments rather than one may not be so clear. It is not reasonable to view V_1^* and V_2^* as replicates or repeated measurements, as generally the distribution of $V_1^*|V$ may be quite different from that of $V_2^*|V$. For instance, one scheme may be much better than the other in the sense of having considerably higher sensitivity and specificity, although this might not be known to the investigator.

Again we consider the retrospective case-control context. Let $c[0, j, k]$ be the number of the n_0 controls for whom $(V_1^* = j, V_2^* = k)$, and let $c[1, j, k]$ be the

number of the n_1 cases with these (V_1^*, V_2^*) values. Also, let p_i and q_i be the sensitivity and the specificity of the i-th assessment scheme, for $i = 1, 2$. For now we assume that the two surrogate variables V_1^* and V_2^* are conditionally independent given the actual exposure variable V. This can be interpreted as an assumption about the 'separateness' of the two exposure assessments. In fact it is a strong assumption which is not warranted in all situations, a point we return to later in this section. Given the assumption, the likelihood function for the six unknown parameters takes the form

$$L(p_1, p_2, q_1, q_2, r_0, r_1) \quad = \quad \prod_{j=0}^{1} \prod_{k=0}^{1} \theta_{0jk}^{c[0,j,k]} \theta_{1jk}^{c[1,j,k]}, \tag{5.9}$$

where

$$\begin{aligned} \theta_{ijk} \quad = \quad & r_i p_1^j (1-p_1)^{1-j} p_2^k (1-p_2)^{1-k} + \\ & (1-r_i)(1-q_1)^j q_1^{1-j} (1-q_2)^k q_2^{1-k} \end{aligned} \tag{5.10}$$

is the probability of observing $(V_1^* = j, V_2^* = k)$ for an individual sampled from the i-th population, i.e., $i = 0$ for controls and $i = 1$ for cases. The form of (5.10) implies that the likelihood function (5.9) depends on its arguments in a complicated manner. As is often the case, however, Bayes-MCMC fitting of this model is greatly simplified by including each subject's true but unobserved exposure variable V as arguments of the posterior distribution. In particular, if the six unknown parameters are assigned independent Beta prior distributions, then all parameters and unobserved variables can be updated simply via Gibbs sampling. Details are given by Johnson, Gastwirth and Pearson (2001).

Initially it seems remarkable to expect good inferences without either external knowledge about the performance of at least one assessment scheme, or gold-standard measurements for at least some study subjects. However, Hui and Walter (1980) illustrate that subject to some minor caveats the model under consideration is identifiable, and therefore consistent estimators of the parameters can be obtained. That is, as more data accumulate, estimates of all the parameters will converge to the corresponding true values. The intuitive argument for identifiability is as follows. Since the data can be summarized in the form of two 2×2 tables, based on the possible (V_1^*, V_2^*) values for controls and cases respectively, we expect to be able to estimate three parameters from each table, or six in total. Since there are in fact six unknown parameters as the arguments for (5.9), it is plausible that they can all be estimated consistently. Indeed, both Hui and Walter (1980) and Drews, Flanders and Kosinski (1993) discuss maximum likelihood estimation for this model. To be more specific, let

$$\phi \quad = \quad (p_1, p_2, q_1, q_2, r_0, r_1)$$

denote the vector of six unknown parameters, and, with reference to (5.10), let

$$\theta \quad = \quad (\theta_{000}, \theta_{001}, \theta_{010}, \theta_{100}, \theta_{101}, \theta_{110}).$$

Note, in particular, that θ completely determines the conditional distribution of $(V_1^*, V_2^*|V)$, as the omitted probabilities θ_{011} and θ_{111} are both deterministic functions of θ, given that the probabilities associated with the conditional distribution must sum to one for both $V = 0$ and $V = 1$. Of course θ can be viewed as a function of ϕ, so we can write $\theta = g(\phi)$, with the form of $g()$ determined by (5.10). Hui and Walter (1980) demonstrate that $g(\cdot)$ is an invertible function, and give an expression, albeit an extremely complicated one, for $g^{-1}(\cdot)$. Thus in principle one can simply estimate θ by cell proportions in the two 2×2 tables which summarize the data. Then, armed with the estimate $\hat{\theta}$, one can estimate ϕ by $\hat{\phi} = g^{-1}(\hat{\theta})$.

Outwardly, the idea that dual exposure assessment can lead to good estimates without having to quantify knowledge about the sensitivity and specificity of the assessments is quite appealing. There are, however, a number of reasons why this should not be viewed as a 'free lunch.' We have already mentioned that the requisite conditional independence assumption may be dubious in specific applications, and indeed we will look more closely at this later in this section. But there are other concerns above and beyond this issue.

An initial concern about this approach to inference is that strictly speaking the function $g(\cdot)$ is two-to-one, so that two distinct ϕ vectors correspond to the observed cell proportions $\hat{\theta}$. Particularly, it is easy to verify from (5.10) that θ is unchanged if we replace each p_i with $1 - q_i$, each q_i with $1 - p_i$, and each r_i with $1 - r_i$. A simple and reasonable solution to this problem is to restrict the parameter space for ϕ according to $p_i + q_i > 1$ for $i = 1, 2$. Clearly an assessment scheme for which $p_i + q_i < 1$ is worse than guessing, and can be ruled out *a priori* in almost any realistic scenario. On the restricted parameter space $g(\cdot)$ is one-to-one, so that $g^{-1}(\cdot)$ is well defined, and $\hat{\phi}$ can be obtained from $\hat{\theta}$.

A second caveat is that the inversion of $g(\cdot)$ breaks down when $r_0 = r_1$. That is, if $r_0 = r_1$ then ϕ cannot be determined from θ. At face value this may not seem worrisome, as it hard to imagine that r_0 and r_1 would be *exactly* the same in any realistic populations of interest. However, as emphasized by Gustafson (2002b), an implication of inference being 'impossible' when $r_0 = r_1$ is that estimators can perform very poorly if r_0 is close to r_1, even at very large sample sizes. In particular, the Fisher information matrix $I(\phi)$ is singular when $r_0 = r_1$, so that the asymptotic variance of estimators of ϕ diverges to infinity as r_0 and r_1 tend toward one another. Indeed, Gustafson (2002b) shows that $|r_0 - r_1|$ need not be very small before the variance of estimators is very large. This is a particular problem in the case-control context, as typically it is not known *a priori* that $|r_0 - r_1|$ will be large.

A third caveat is that sampling variability and/or model misspecification (violation of the conditional independence assumption in particular) can lead to cell probability estimates of θ which actually fall outside the parameter space. That is, $\hat{\theta}$ can lie outside the image under $g(\cdot)$ of the parameter space for ϕ. In such a dataset one cannot obtain an estimate of ϕ of the form $g^{-1}(\hat{\theta})$. Indeed, this problem motivates the work of Drews, Flanders and Kosinski

(1993) who use the EM algorithm to obtain maximum likelihood estimates of ϕ. In particular, they regard the true but unobserved exposure status of each subject as 'missing' data. This approach to maximum likelihood estimation is similar in flavour to the use of Gibbs sampling on an extended parameter space which included the true exposure variables, as described by Johnson, Gastwirth and Pearson (2001).

Bearing these concerns in mind, if one does not want to incorporate any substantive prior information about the parameters it seems reasonable to assign a uniform prior to ϕ, restricted to the region for which $p_i + q_i > 1$ for $i = 1, 2$. The restriction on the parameter space does not cause particular difficulties for Gibbs sampling, as the full conditional posterior distributions for each p_i and q_i become truncated beta distributions rather than beta distributions. However, while the Gibbs sampler is easy to implement, we find it necessary to monitor its output carefully. The MCMC mixing performance can vary considerably across different data sets. Particularly we have noticed poorer mixing for data sets which do not strongly rule out the troublesome $r_0 = r_1$ scenario mentioned above. We describe the results of fitting the model to data shortly, after describing an extension of the model which relaxes the conditional independence assumption.

5.3.1 Relaxing the Conditional Independence Assumption

The assumption that V_1^* and V_2^* are conditionally independent given V may be quite dubious in some applications (Fryback 1978, Brenner 1996, Vacek 1985). Indeed, it is easy to imagine scenarios whereby a subject who is particularly susceptible to misclassification under one assessment scheme is also particularly susceptible under the other scheme. Moreover, the assumption is not empirically verifiable without any gold-standard measurements, as all the 'degrees of freedom' are used to estimate the six unknown parameters. In light of this, Torrance-Rynard and Walter (1997) use simulation to assess the bias introduced by incorrectly assuming the two exposure assessments are conditionally independent given the true exposure. Their findings are mixed, but scenarios where the bias is considerable are given.

We consider extending the model to allow for conditional dependence between V_1^* and V_2^* given V. It is clear, however, that after relaxing the assumption we will no longer have an identifiable model. That is, the data are still summarized in two 2×2 tables, so no more than six parameters can be estimated consistently. However, allowing an unspecified amount of dependence between V_1^* and V_2^* given V clearly increases the number of unknown parameters beyond six. Nonetheless, we investigate such an extension in the face of prior information which reflects the possibility of *limited* dependence between the two assessments. At the same time we refer the reader ahead to Section 6.3, where some general discussion on the utility of nonidentifiable models is given.

The expanded model postulates that the distribution of (V_1^*, V_2^*) given V depends on six unknown parameters, $(p_1, p_2, q_1, q_2, \delta_0, \delta_1)$, with the form of

		$V_2^*=0$	$V_2^*=1$
$V=0$	$V_1^*=0$	$q_1q_2+\delta_0$	$q_1(1-q_2)-\delta_0$
	$V_1^*=1$	$(1-q_1)q_2-\delta_0$	$(1-q_1)(1-q_2)+\delta_0$
		$V_2^*=0$	$V_2^*=1$
$V=1$	$V_1^*=0$	$(1-p_1)(1-p_2)+\delta_1$	$(1-p_1)p_2-\delta_1$
	$V_1^*=1$	$p_1(1-p_2)-\delta_1$	$p_1p_2+\delta_1$

Table 5.3 *The probabilities describing the conditional distribution of* $V_1^*, V_2^*|V$ *under the model of Section 5.3.1.*

the distribution given in Table 5.3. It is easy to check that this formulation implies $p_i = Pr(V_i^* = 1|V = 1)$ and $q_i = Pr(V_i^* = 0|V = 0)$ for $i = 1, 2$. That is, p_i and q_i retain their interpretation as the sensitivity and specificity of the i-th assessment. As well it is simple to verify that

$$\delta_i \quad = \quad Cov(V_1^*, V_2^*|V = i),$$

so that (δ_0, δ_1) describe the extent to which V_1^* and V_2^* are dependent given V. It is also clear that the observed data in the retrospective study are governed by cell probabilities

$$\begin{aligned}\theta_{ijk} \quad = \quad & r_i \left\{ p_1^j (1-p_1)^{1-j} p_2^k (1-p_2)^{1-k} + (-1)^{|j-k|}\delta_1 \right\} + \\ & (1-r_i) \left\{ (1-q_1)^j q_1^{1-j} (1-q_2)^k q_2^{1-k} + (-1)^{|j-k|}\delta_0 \right\}, \quad (5.11)\end{aligned}$$

which generalizes (5.10) as the probability of observing $(V_1^* = j, V_2^* = k)$ for a subject sampled from the i-th population. The probabilities (5.11) yield a likelihood function for the eight unknown parameters, just as (5.10) leads to the likelihood function (5.9) in the model assuming conditional independence.

In terms of a prior we retain the uniform prior for $\phi = (p_1, p_2, q_1, q_2, r_0, r_1)$ over the appropriately restricted parameter space. Then we consider what might be reasonable as a prior for $(\delta_0, \delta_1|\phi)$. As it is hard to imagine scenarios under which the two surrogates exhibit negative dependence given the true exposure, we use a prior which gives weight only to nonnegative values of δ_0 and δ_1. As well, we must be aware that a legitimate probability distribution for $(V_1^*, V_2^*|V)$ requires

$$\begin{aligned}\delta_0 \quad &\leq \quad \delta_*(q_1, q_2), \\ \delta_1 \quad &\leq \quad \delta_*(p_1, p_2),\end{aligned}$$

where

$$\delta_*(a, b) \quad = \quad \min(a, b) - ab$$

is the maximum possible covariance between two binary random variables with success probabilities a and b. That is, a violation of either inequality

corresponds to having at least one entry in Table 5.3 which lies outside the interval from zero to one.

As a further consideration, dependence is inherently more interpretable when expressed in terms of correlation rather than covariance. Thus we define

$$\begin{aligned}\rho_0 &= \mathrm{Cor}(V_1^*, V_2^* | V = 0) \\ &= \frac{\delta_0}{\{p_1(1-p_1)p_2(1-p_2)\}^{1/2}},\end{aligned}$$

and

$$\begin{aligned}\rho_1 &= \mathrm{Cor}(V_1^*, V_2^* | V = 1) \\ &= \frac{\delta_1}{\{q_1(1-q_1)q_2(1-q_2)\}^{1/2}}.\end{aligned}$$

On the one hand it is easier to think about the magnitude of dependence in terms of (ρ_0, ρ_1), but on the other hand the cell probabilities (5.11) are more easily expressed in terms of (δ_0, δ_1).

Given these considerations, we specify a prior density of the form

$$f(\delta_0, \delta_1 | p_1, p_2, q_1, q_2) = f(\delta_0 | q_1, q_2) f(\delta_1 | p_1, p_2),$$

with

$$f(\delta_0 | q_1, q_2) = \frac{\lambda(q_1, q_2) \exp\{-\lambda(q_1, q_2)\delta_0\} I_{[0, \delta_*(q_1, q_2)]}(\delta_0)}{1 - \exp\{-\lambda(q_1, q_2)\delta_*(q_1, q_2)\}}$$

and

$$f(\delta_1 | p_1, p_2) = \frac{\lambda(p_1, p_2) \exp\{-\lambda(p_1, p_2)\delta_1\} I_{[0, \delta_*(p_1, p_2)]}(\delta_1)}{1 - \exp\{-\lambda(p_1, p_2)\delta_*(p_1, p_2)\}}.$$

In the above we take

$$\lambda(a, b) = \frac{k}{\{a(1-a)b(1-b)\}^{1/2}}, \tag{5.12}$$

where k is a hyperparameter that must be specified. While these expressions are cumbersome, they simply describe truncated exponential priors assigned to the covariance parameters (δ_0, δ_1) given the other parameters. The rate parameters in these exponential distributions, as given by (5.12), are chosen to lend a clear interpretation to the prior in terms of correlation rather than covariance. For instance, consider a value $\delta_0 = \tilde{\delta}_0$ which corresponds to a correlation $\rho_0 = \tilde{\rho}_0$. The ratio of the prior density at $\delta_0 = \tilde{\delta}_0$ to the prior density at $\delta_0 = 0$ can be expressed as $\exp(-k\tilde{\rho}_0)$. Thus k can be chosen to reflect the strength of the belief that the conditional dependence is not likely to be strong. In the examples that follow, for instance, we choose $k = 4 \log 4$ in the priors for both δ_0 and δ_1. Then the prior downweights $\rho_i = 0.25$ by a factor of four relative to $\rho_i = 0$, for $i = 0, 1$.

As alluded to in Section 5.2, there is some literature to suggest that standard MCMC algorithms do not always perform well on posterior distributions arising from nonidentifiable models. This issue seems particularly germane to

exposure		controls		cases	
		$V_2^* = 0$	$V_2^* = 1$	$V_2^* = 0$	$V_2^* = 1$
anemia	$V_1^* = 0$	147	15	125	15
	$V_1^* = 1$	34	20	49	24
UTI	$V_1^* = 0$	190	4	174	14
	$V_1^* = 1$	10	14	14	13
LWG	$V_1^* = 0$	86	4	73	5
	$V_1^* = 1$	3	6	13	11

Table 5.4 *Data from a SIDS case-control study. Three binary exposures are considered: maternal anemia during pregnancy, maternal urinary tract infection (UTI) during pregnancy, and low weight-gain (LWG) during pregnancy. Each exposure is assessed by maternal interview (V_1^*) and by examination of medical records (V_2^*).*

algorithms which update one parameter at a time using the obvious or original parameterization of the problem. Our experience is that MCMC algorithms can perform poorly when applied to the current model, and in Section 5.6.4 we detail an algorithm designed to ameliorate this problem as much as possible. Particularly, this scheme updates (q_1, q_2, δ_0) and (p_1, p_2, δ_1) in blocks, taking into account the structure depicted in Table 5.3.

5.3.2 Example

For the sake of clarity we henceforth refer to the two dual exposure assessment models described in this section as DUAL-IND and DUAL-DEP, with the former assuming conditional independence of the two assessments and the latter placing a prior distribution on the magnitude of dependence. Both models are applied to data given by Drews, Flanders and Kosinski (1993) from a case-control study on sudden infant death syndrome (SIDS). In the first instance the binary exposure variable is maternal anemia during pregnancy, assessed by both an interview with the mother and an examination of the mother's medical records. Neither method can be considered a gold-standard, and it is immediately apparent from the data summarized in Table 5.4 that the two assessments are in disagreement for a substantial fraction of the study subjects.

The upper-left panel of Figure 5.4 gives the posterior distribution of the log odds-ratio $\psi = log\{r_1/(1-r_1)\} - log\{r_0/(1-r_0)\}$ under both the DUAL-IND and DUAL-DEP models. We see that DUAL-DEP yields a wider posterior distribution than DUAL-IND. That is, admitting uncertainty about the dependence leads to more inferential uncertainty *a posteriori*, as one might expect. On the other hand, the posterior mode of ψ is quite similar under the two models. Admitting uncertainty about whether the conditional indepen-

dence assumption is appropriate does not produce a substantial change in the point estimate of ψ.

For the sake of comparison we also try a third, more naive, modelling strategy. In particular we consider using only the data from those subjects with concordant assessments, i.e., $V_1^* = V_2^*$. The concordant value is then treated as if it is the actual exposure V. We label this strategy as DUAL-CON. From Figure 5.4 we see that the widths of the DUAL-CON and DUAL-IND posterior distributions for ψ are very similar, but the DUAL-CON posterior mode lies somewhat closer to zero than the DUAL-IND or DUAL-DEP posterior modes. This attentuation is not surprising, as to some extent the DUAL-CON strategy is ignoring the misclassification rather than adjusting for it.

Given the lack of identifiability in the DUAL-DEP model, there is concern about how well this model can be fit to data via MCMC methods. The upper-right panel of Figure 5.4 gives a traceplot of the MCMC output for ψ under this model. The mixing behaviour seems tolerable, despite the lack of identifiability. Also, the lower panels of Figure 5.4 give posterior scatterplots of ψ versus ρ_0 and ψ versus ρ_1 under the DUAL-DEP model. We note that the posterior correlations are weak in both cases. In part this explains why DUAL-DEP and DUAL-IND point estimates of ψ are similar, as the posterior distribution of ψ given the dependence parameters does not depend strongly on the dependence parameters.

In fact, Drews, Flanders and Kosinksi (1993) considered various binary exposures in this study. Table 5.4 also gives data for the exposures maternal urinary tract infection during pregnancy and low weight gain during pregnancy. Figures 5.5 and 5.6 summarize the posterior analyses for these exposures, in the same format as Figure 5.4. The qualitative findings for these exposures are very similar to those for the maternal anemia exposure. That is, the DUAL-IND and DUAL-DEP posterior distributions for ψ have modes at very similar values, but the former distribution is much narrower than the latter. In comparison, the DUAL-CON posterior mode is closer to zero. The MCMC mixing for the nonidentifiable DUAL-DEP model is tolerable, and the posterior correlation between ψ and ρ_i is weak, for both $i = 0$ and $i = 1$.

As a whole, experience with the three different exposures suggests that the DUAL-IND modelling strategy may have some robustness, in the sense of yielding similar point estimates of ψ to those under the more general DUAL-DEP model. It is also clear, however, that if an investigator is really unsure about the conditional independence assumption then the DUAL-IND posterior distribution is considerably over-confident, as it tends to be much narrower than its DUAL-DEP counterpart. We have also demonstrated that fitting the DUAL-DEP model, which reflects a prior downweighting of larger correlations between V_1^* and V_2^* given V, is a plausible inferential strategy, even though the model is formally nonidentifiable.

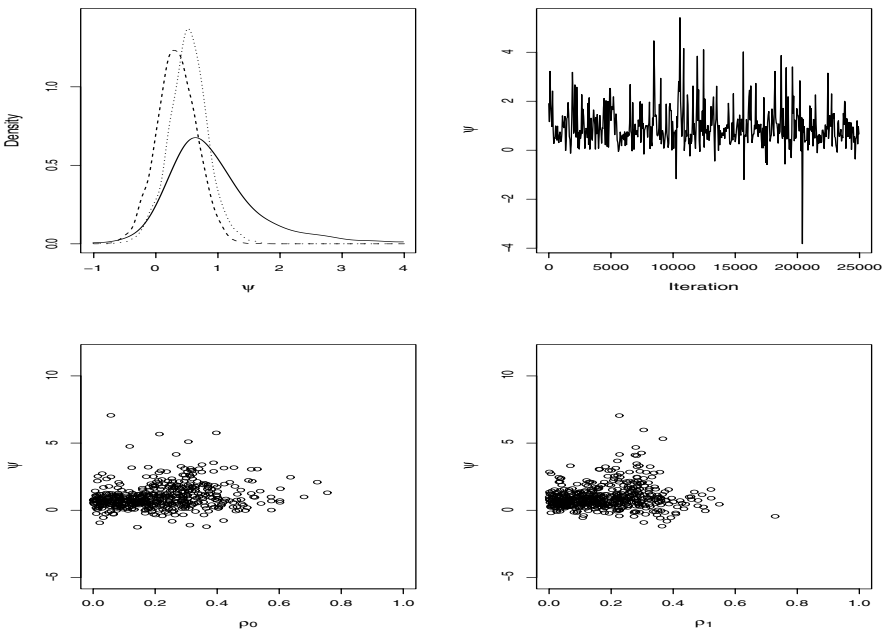

Figure 5.4 *Posterior summaries for the SIDS example, with maternal anemia as the exposure variable. The upper-left panel gives the posterior distribution of the log odds-ratio ψ under the DUAL-IND model (dotted curve), the DUAL-DEP model (solid curve), and the DUAL-CON approach (dashed curve). The upper-right panel gives the MCMC output for ψ under the DUAL-DEP model, while the scatterplots in the lower panels depict the joint posterior distribution of (ρ_0, ψ) and (ρ_1, ψ) under this model.*

5.4 Models with Additional Explanatory Variables

Thus far in this chapter we have examined inferences in the focussed framework of a single binary exposure, with either one or two imperfect assessments of this exposure. Of course many studies involve multiple explanatory variables. In such instances the paradigm of measurement, outcome, and exposure models described in Chapter 4 can be applied equally well to a binary exposure. Consider a random sample of n observations on (V_1^*, V_2^*, Y, Z), where V_1^* and V_2^* are two different surrogates for the actual binary exposure V, Y is the outcome variable, and Z is a vector of precisely measured covariates. A measurement model parameterized by α, an outcome model parameterized by β, and an exposure model parameterized by γ can be combined to yield a posterior distribution

$$f(\alpha, \beta, \gamma, v | v_1^*, v_2^*, y, z) \quad \propto \quad \prod_{i=1}^{n} f(v_{1i}^*, v_{2i}^* | y_i, v_i, z_i, \alpha) \ \times$$

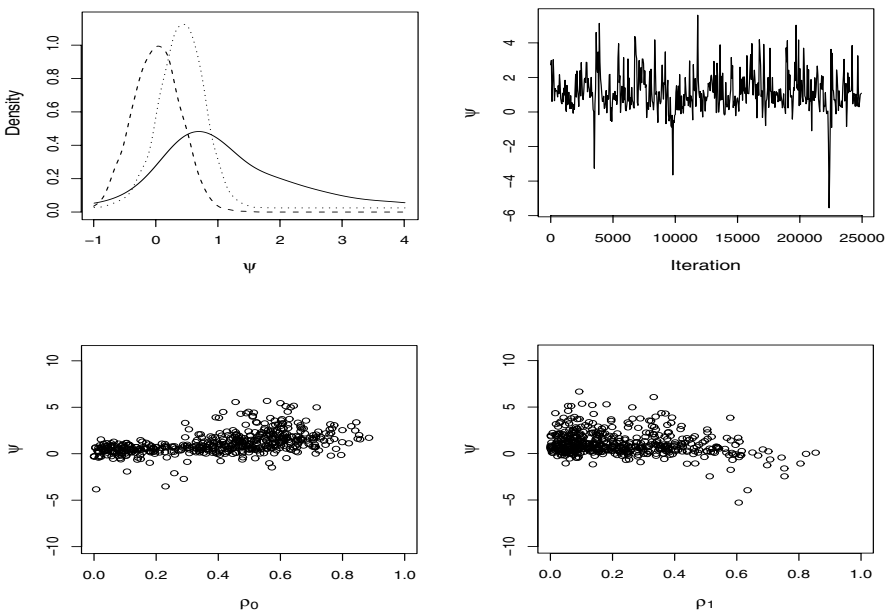

Figure 5.5 *Posterior summaries for the SIDS example, with maternal urinary tract infection as the exposure variable. The format is the same as Figure 5.4.*

$$\prod_{i=1}^{n} f(y_i|v_i, z_i, \beta) \times \prod_{i=1}^{n} f(v_i|z_i, \gamma) \times f(\alpha, \beta, \gamma). \tag{5.13}$$

In the case of a binary exposure V, however, this may not be the most pragmatic way to proceed. In particular, the contribution of the outcome and exposure models may be re-expressed as

$$\begin{aligned} f(y|v,z)f(v|z) &= f(y,v|z) \\ &= f(v|y,z)f(y|z). \end{aligned}$$

Thus an alternative strategy is to use what we will call a *exposure given outcome model*, by specifying a distributional form for $V|Y,Z$ parameterized by λ rather than separate outcome and exposure models as in (5.13). Importantly, in doing so it is not necessary to parameterize the distribution of $Y|Z$, as both

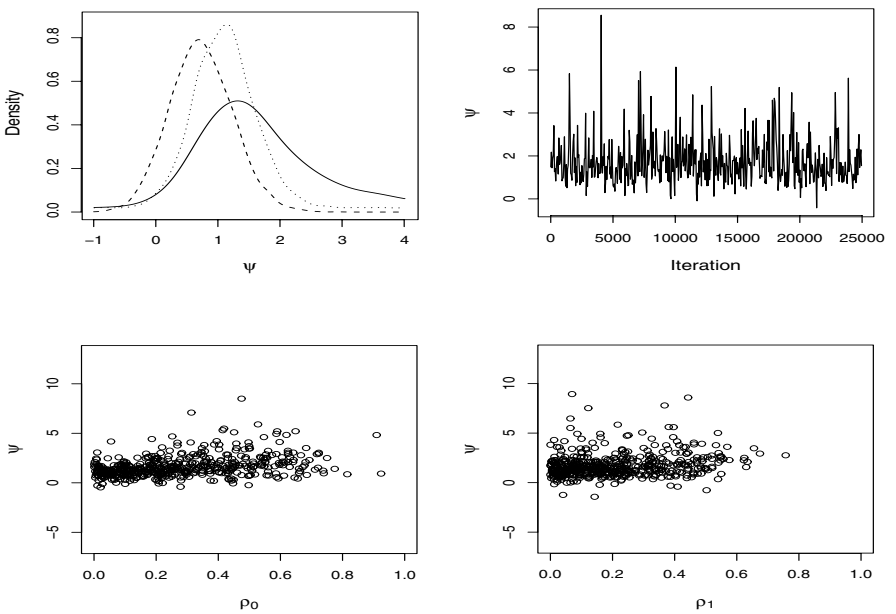

Figure 5.6 *Posterior summaries for the SIDS example, with low pregnancy weight gain as the exposure variable. The format is the same as Figure 5.4.*

Y and Z are observed. Thus we have

$$f(\alpha,\lambda,v|v_1^*,v_2^*,y,z) \quad \propto \quad \prod_{i=1}^{n} f(v_{1i}^*,v_{2i}^*|v_i,y_i,z_i,\alpha) \;\times \prod_{i=1}^{n} f(v_i|y_i,z_i,\lambda) \;\times f(\alpha,\lambda), \tag{5.14}$$

where the $f(y|z)$ term can be omitted, since we need only the posterior density of (α,λ,v) up to a constant of proportionality.

The obvious pragmatic advantage of (5.14) over (5.13) is that one need only estimate two parameter vectors rather than three. Moreover, when Y and V are both binary an exposure given outcome model may be just as good at describing the exposure-outcome relationship as is an outcome model. Particularly, standard logistic regression outcome and exposure models would be

$$\text{logit}Pr(Y=1|V,Z) \quad = \quad \beta_0+\beta_1 V+\beta_2' Z, \tag{5.15}$$

and

$$\text{logit}Pr(V=1|Z) \quad = \quad \gamma_0 + \gamma_1' Z. \tag{5.16}$$

Together these define the joint distribution of (Y, V) given Z, and subsequently the conditional distribution of V given (Y, Z). Indeed, straightforward calculation gives

$$\text{logit}Pr(V=1|Y,Z) \quad = \quad \gamma_0 + \beta_1 Y + \gamma_1' Z + g(Z;\beta), \tag{5.17}$$

where

$$g(z;\beta) \quad = \quad \log\{1+\exp(\beta_0+\beta_2' z)\} - \log\{1+\exp(\beta_0+\beta_1+\beta_2' z)\}.$$

In particular, given Z, β_1 has a well-known dual interpretation. Conditioned on any fixed value of Z it is the log odds-ratio for both Y in terms of X and X in terms of Y. In particular, this suggests that using an exposure given outcome model of the form

$$\text{logit}Pr(V=1|Y,Z) \quad = \quad \lambda_0 + \lambda_1 Y + \lambda_2' Z \tag{5.18}$$

is nearly equivalent to using (5.15) and (5.16), the slight difference being the omission of the nonlinear term $g(Z;\beta)$ in (5.17). Moreover, λ_1 in (5.18) can be interpreted as if it were β_1 in the outcome model (5.15).

We note that in principle one could use an exposure given outcome model instead of separate outcome and exposure models when one or both of V and Y are continuous. This is less attractive, however, as the two approaches will no longer have the similarity and dual interpretation of parameters exhibited in the binary case.

5.4.1 Example

As an example which contrasts the use of an exposure given outcome model with the use of separate outcome and exposure models we consider a dataset presented by Kosinski and Flanders (1999). These data arose from a clinical study of patients with multi-vessel coronary artery disease. The main purpose of the study was to compare two therapies, percutaneous transluminal coronary angioplasty (PTCA) and coronary artery bypass grafting (CABG), with patients randomized to the two therapies. In fact the study did not establish a statistically significant difference between the therapies with respect to the composite primary endpoint. As a secondary issue, Kosinski and Flanders consider whether the presence ($V=1$) of haemodynamically obstructive coronary artery disease (HCAD) at a one-year follow up visit after the initial procedure (PTCA or CABG) is associated with the occurrence of a future cardiac event. A future cardiac event ($Y=1$) is defined as a subsequent procedure, myocardial infarction, or death between 90 days and 2 years after the follow-up visit.

The HCAD status V cannot be measured without error. The first surrogate V_1^* is based on an exercise stress test. The second surrogate V_2^* is based on a single-photon-emission computed tomography (SPECT) thallium test.

Y	Z_1	Z_2	freq.		$V_2^*=0$	$V_2^*=1$
0	0	0	147	$V_1^*=0$	85	46
				$V_1^*=1$	10	6
0	0	1	60	$V_1^*=0$	44	12
				$V_1^*=1$	1	3
0	1	0	66	$V_1^*=0$	36	20
				$V_1^*=1$	3	7
0	1	1	7	$V_1^*=0$	4	1
				$V_1^*=1$	1	1
1	0	0	12	$V_1^*=0$	7	3
				$V_1^*=1$	0	2
1	0	1	8	$V_1^*=0$	5	1
				$V_1^*=1$	0	2
1	1	0	14	$V_1^*=0$	4	6
				$V_1^*=1$	2	2
1	1	1	1	$V_1^*=0$	0	0
				$V_1^*=1$	0	1

Table 5.5 *Data from Kosinski and Flanders (1999). For each combination of Y (non-occurrence/occurrence of cardiac event), Z_1 (weight less/more than 90 kg), and Z_2 (male/female) the observed V_1^* (surrogate for HCAD based on exercise stress test) and V_2^* (surrogate for HCAD based on SPECT thallium test) values are summarized in a 2×2 table. The total sample size is $n = 315$ subjects.*

It is clear from the data, which are summarized in Table 5.5, that the two surrogates are discrepant for a substantial proportion of the study subjects. Drews and Kosinski consider two binary covariates: weight in comparison to a threshold and gender. In particular, weight (Z_1) is coded as zero/one based on a subject's weight being less/more than 90 kg, while gender (Z_2) is coded as zero/one for male/female.

We apply logistic regression models to these data. In the first instance we use outcome model (5.15) and exposure model (5.16), and in the second instance we instead use the exposure given outcome model (5.18). Either way we assume the misclassification is nondifferential and the two surrogates for exposure are conditionally independent given the true but unobserved exposure. Thus, as previously, the misclassification model involves four unknown parameters, namely the sensitivity p_i and specificity q_i of the i-th exposure assessment, $i = 1, 2$. In all, the first approach involves $4+4+3 = 13$ unknown

parameters, while the second involves $4 + 4 = 8$ unknown parameters. The requisite MCMC analysis is easily implemented under either strategy. Particularly, the MCMC updating schemes for logistic regression models described in Section 4.11 can be applied here quite directly.

Selected posterior distributions are displayed in Figure 5.7. In particular, the posterior distribution of p_1, p_2, q_1, q_2 and the log-odds describing the relationship between Y and V given Z are given for each modelling strategy. Recall that the latter is β_1 in the first model and λ_1 in the second model. For all five parameters of interest the posterior distributions based on the two models are virtually the same. This lends credence to opting for the simpler specification of an exposure given outcome model whenever possible.

5.4.2 Beyond Nondifferential and Conditionally Independent Dual Exposure Assessment

With the DUAL-IND model of Section 5.3 we could not relax the assumption of conditionally independent assessments without losing the identifiability of the model. With the availability of other precisely measured covariates, however, the situation may be improved. Kosinski and Flanders (1999) consider a very general approach with the measurement model specified via two logistic regression models,

$$\text{logit}Pr(V_2^* = 1|V_1^*, V, Y, Z) = \alpha_0 + \alpha_1 V_1^* + \alpha_2 V + \alpha_3 Y + \alpha_4 Z \tag{5.19}$$

and

$$\text{logit}Pr(V_1^* = 1|V, Y, Z) = \alpha_5 + \alpha_6 V + \alpha_7 Y + \alpha_8 Z, \tag{5.20}$$

along with a third logistic regression model for the exposure given outcome:

$$\text{logit}Pr(V = 1|Y, Z) = \lambda_0 + \lambda_1 Y + \lambda_2 Z. \tag{5.21}$$

Sub-models of this general model then correspond to commonly used models. For instance, the assumption that the misclassification is nondifferential corresponds to $\alpha_3 = \alpha_4 = \alpha_7 = \alpha_8 = 0$, which makes (V_1^*, V_2^*) and (Y, Z) conditionally independent given V. The other obvious sub-model is that arising when $\alpha_1 = 0$, which corresponds to V_1^* and V_2^* being conditionally independent given (V, Y, Z).

The position taken by Kosinski and Flanders (1999) is that one can actually assess how well supported these various models are in light of the data. In some formal sense this will be the case if the full model is identifiable, so that all the unknown parameters in (5.19), (5.20), and (5.21) can be estimated consistently. To consider whether this might be the case, say that Z is comprised of p covariates, all of which are binary. Moreover, say that the joint distribution of (Y, Z) gives positive probability to all 2^{p+1} possible values. Since the data can be summarized in the form of separate 2×2 tables for (V_1^*, V_2^*) values given each combination of (Y, Z) values, we might expect to be able to estimate as many as $3 \times 2^{p+1}$ unknown parameters consistently. Moreover, a

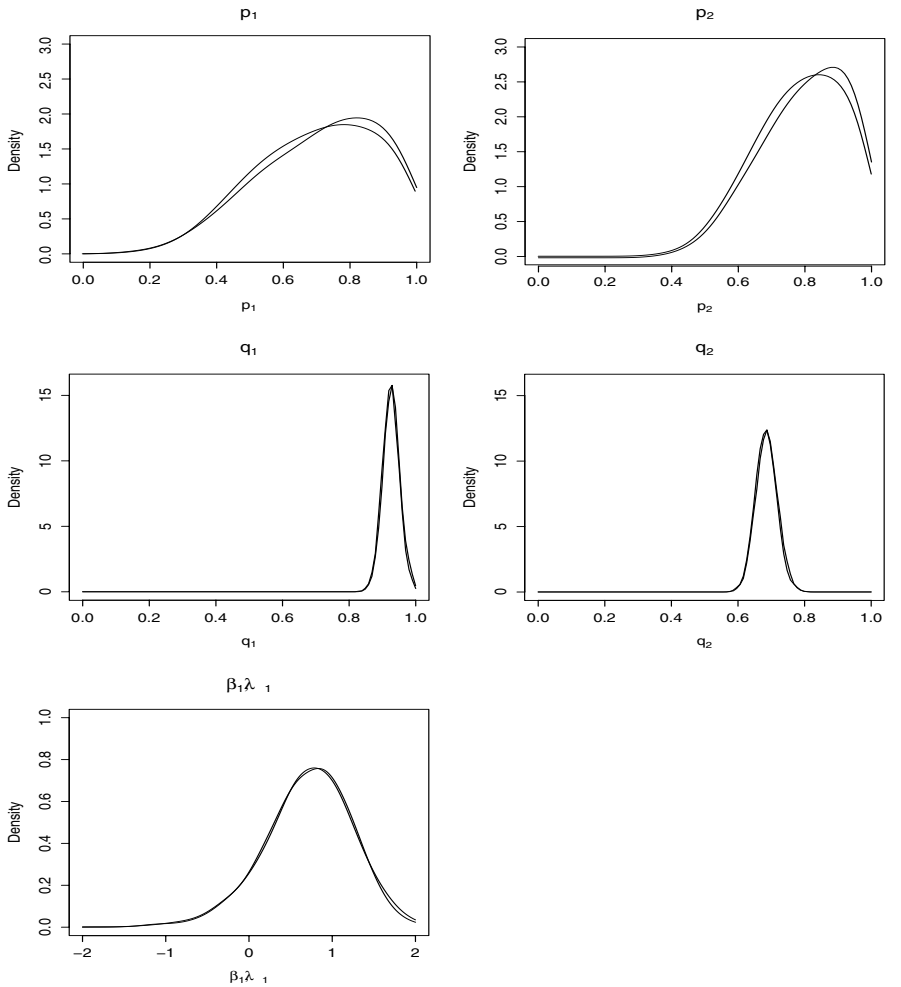

Figure 5.7 *Posterior summaries for the coronary example. Posterior distributions for* p_1, p_2, q_1, q_2 *and the conditional log odds-ratio for the outcome-exposure relationship (*β_1 *in the outcome model or* λ_1 *in the exposure given outcome model) are given, for each of the two models considered.*

quick count reveals a total of $(p+4)+(p+3)+(p+2) = 3p+9$ unknown parameters in the full model. For $p = 1$, the two totals are equal, and for $p > 1$ the upper bound on the number of consistently estimable parameters exceeds the actual number of unknown parameters. While this counting argument does not *prove* that the full model is identifiable provided there is at least one precisely measured covariate, it does suggest this might be the case.

Kosinski and Flanders (1999) report results from maximum likelihood fitting based on (5.19) through (5.21), without assuming the two exposure assessments are nondifferential and conditionally independent. Specifically, they give results for the HCAD dataset considered in Section 5.4.2 as well as for a simulated dataset. Unfortunately, however, our experience with Bayes-MCMC fitting of the same models is quite discouraging. Even assuming nondifferential misclassification ($\alpha_3 = \alpha_4 = \alpha_7 = \alpha_8 = 0$) but allowing dependence (α_1 unconstrained) is extremely problematic. MCMC algorithms tend to mix poorly, and when applied to synthetic datasets very poor results obtain. In particular, the posterior distribution of α_1 tends to concentrate away from zero even if the data are simulated under $\alpha_1 = 0$.

As an alternative strategy we considered using the measurement model from the DUAL-DEP model described in Section 5.3.1, in tandem with the exposure given outcome model (5.21). This involves two parameters (δ_0, δ_1) which describe the dependence between assessments, rather than the single parameter (α_1) in (5.19) and (5.20). Rather than assign an informative prior to (δ_0, δ_1) as before, we now use a flat prior, given that we have enough 'degrees-of-freedom' to estimate all the unknown parameters. Unfortunately this strategy also yields very poor performance, both in terms of poor MCMC mixing and posterior distributions from synthetic datasets which miss the true values of parameters.

The negative experiences from these modelling strategies suggest it is very difficult to fit models which do not either specify the dependence parameters exactly or at least assign them relatively informative prior distributions. Thus the theoretical gain in degrees of freedom afforded by additional precisely measured explanatory variables does not always translate to a practical ability to fit models entailing fewer assumptions.

5.4.3 Dual Exposure Assessment in Three Populations

To shed more light on the difficulty inherent in relaxing the conditional independence assumption, consider the following situation. Say we can sample surrogate exposures (V_1^*, V_2^*) in three distinct populations having exposure prevalences (r_0, r_1, r_2). Thus we can hope to estimate nine parameters from the observed data. If we do assume nondifferential misclassification but do not assume conditional independence of the two assessments then there are nine unknown parameters. In addition to the three unknown prevalences, the six parameters $(p_1, p_2, q_1, q_2, \delta_0, \delta_1)$ govern the distribution of $(V_1^*, V_2^* | V)$, exactly as in the DUAL-DEP model from Section 5.3.1.

In the case of two populations and two conditionally independent exposure assessments we know that the six unknown parameters can be estimated consistently. However, as illustrated by Gustafson (2002) the variance of the estimators can be extremely large if the two population prevalences are not well-separated. The same problem may persist, or in fact be worse, in the present context of three populations and unknown dependence parameters. Indeed, this may explain the difficulties described at the end of the previous subsection.

To investigate, let $S^{(i)}$ be a vector of length four which summarizes the sampled (V_1^*, V_2^*) values in the sample from the i-th population, for $i = 0, 1, 2$. Formally, $S^{(i)}$ is the sum over the sampled subjects of $V_1^* V_2^*$, $V_1^*(1 - V_2^*)$, $(1 - V_1^*)V_2^*$, and $(1 - V_1^*)(1 - V_2^*)$. That is, $S^{(i)}$ simply comprises the cells of the 2×2 table summarizing the data from the i-th sample. Also, let

$$\theta \quad = \quad \begin{pmatrix} p_1 p_2 + \delta_1 \\ p_1(1 - p_2) - \delta_1 \\ (1 - p_1)p_2 - \delta_1 \\ (1 - p_1)(1 - p_2) + \delta_1 \end{pmatrix}$$

be the cell probabilities for a truly exposed subject, while

$$\phi \quad = \quad \begin{pmatrix} (1 - q_1)(1 - q_2) + \delta_0 \\ (1 - q_1)q_2 - \delta_0 \\ q_1(1 - q_2) - \delta_0 \\ q_1 q_2 + \delta_0 \end{pmatrix}$$

are the corresponding cell probabilities for a truly unexposed subject.

Clearly the observed data are distributed as

$$S^{(i)} \quad \sim \quad Multinomial(n_i; r_i\theta + (1 - r_i)\phi).$$

Now let $Z^{(i)}$ denote the (unobservable) counts for only the truly exposed subjects in the i-th sample. Then

$$\begin{pmatrix} Z^{(i)} \\ S^{(i)} - Z^{(i)} \end{pmatrix} \quad \sim \quad Multinomial\left(n_i; \begin{pmatrix} r_i\theta \\ (1 - r_i)\phi \end{pmatrix}\right), \qquad (5.22)$$

which we can regard as the distribution of the 'complete' data. Of course in reality only $S^{(i)}$ is observed, but (5.22) is useful for the purposes of MCMC algorithm construction. Indeed, in Section 5.6.5 we outline the straightforward approach to Gibbs sampling in this scenario.

In light of the apparent identifiability in this problem we consider the use of uniform priors on θ, ϕ, r_0, r_1, and r_2. However, to avoid the trivial non-identifiability that arises by interchanging θ and ϕ as well as r_i and $1 - r_i$, we restrict the parameter space according to $\theta_1 > \phi_1$ and $\theta_4 < \phi_4$. That is, the probability of a subject being classified as exposed by both assessments is greater for a truly exposed subject than for a truly unexposed subject, and similarly the probability of a subject being classified as unexposed by both assessments is greater for a truly unexposed subject than for a truly exposed subject.

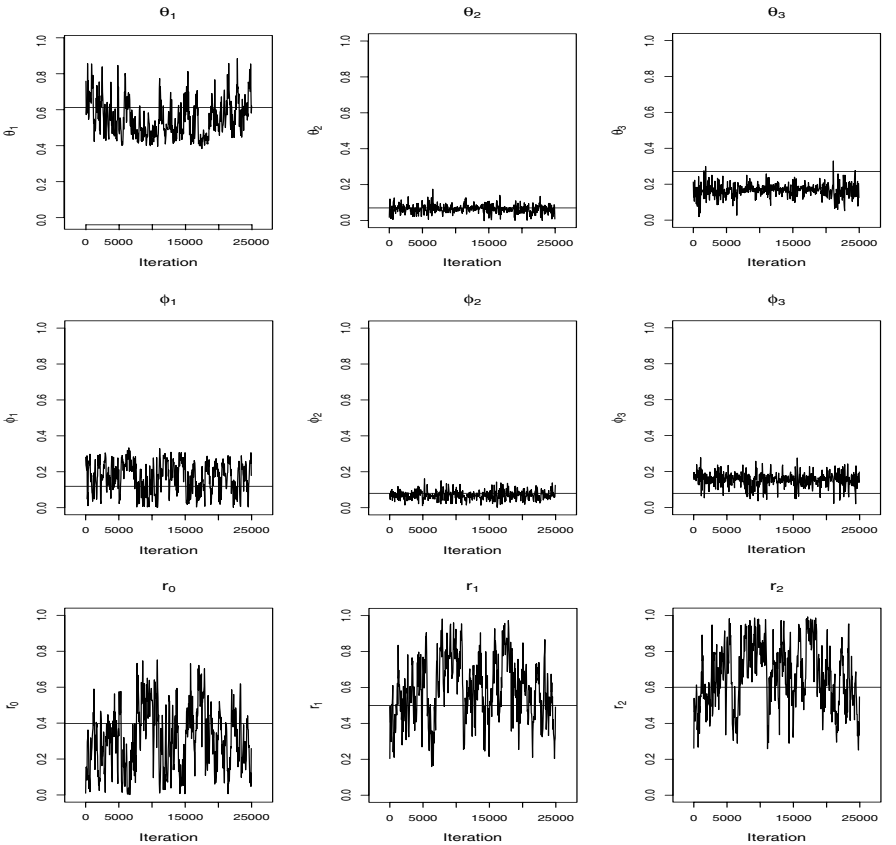

Figure 5.8 *MCMC output for the first synthetic dataset in the three population example of Section 5.4.3. Specifically, 25000 Gibbs sampler iterations are displayed. The superimposed horizontal lines denote true parameter values.*

Three synthetic datasets are generated, each based on $\theta = (.63, .07, .27, .03)$ which corresponds to $p_1 = 0.9$, $p_2 = 0.7$, $\delta_1 = 0$, and $\phi = (.72, .08, .08, .12)$, which corresponds to $q_1 = 0.8$, $q_2 = 0.8$, $\delta_0 = 0.08$. Thus V_1^* and V_2^* are conditionally independent given that $V = 1$, but have a correlation of 0.5 given that $V = 0$.

The first dataset is based on samples of size $n_i = 500$, for $i = 0, 1, 2$, with underlying exposure prevalences of $(r_0, r_1, r_2) = (0.4, 0.5, 0.6)$. MCMC traceplots for the posterior distribution of the parameters given this dataset appear in Figure 5.8. In addition to very poor MCMC mixing, we see the algorithm visiting parameter values very far from the true values. There does not appear to be hope of making reasonable inferences in this scenario.

To see if increased sample sizes lead to reasonable inferences, a larger dataset is simulated under the same parameter values. This time samples of size 2000

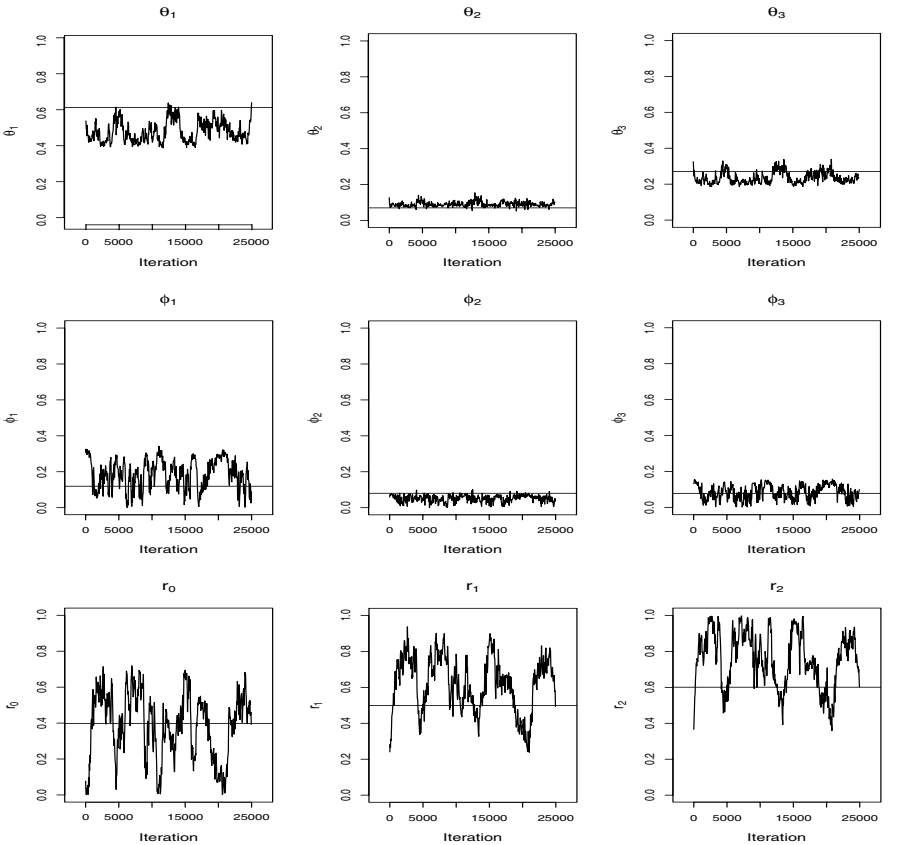

Figure 5.9 *MCMC output for the second synthetic dataset in the three population example of Section 5.4.3. Specifically, 25000 Gibbs sampler iterations are displayed. The superimposed horizontal lines denote true parameter values.*

are drawn from each population, giving a sample size of 6000 in all. The MCMC output based on these data is given in Figure 5.9. Unfortunately we do not see a marked reduction in either statistical error or simulation error. In particular, as best one can tell given the poor mixing, the posterior distribution over the unknown prevalences is very wide.

Given our discussion of the DUAL-IND model in Section 5.3.1, one suspects that the limitations on inference from the second synthetic dataset arise because the underlying exposure prevalences in the three populations are not sufficiently disparate. Thus our third dataset again involves $n_i = 2000$ subjects per population, but now the underlying prevalences are taken to be $(r_0, r_1, r_2) = (0.2, 0.5, 0.8)$. The resultant MCMC output appears in Figure 5.10. While there is some improvement in the MCMC mixing, the posterior distributions over the prevalences are still very wide. Even with a lot of data

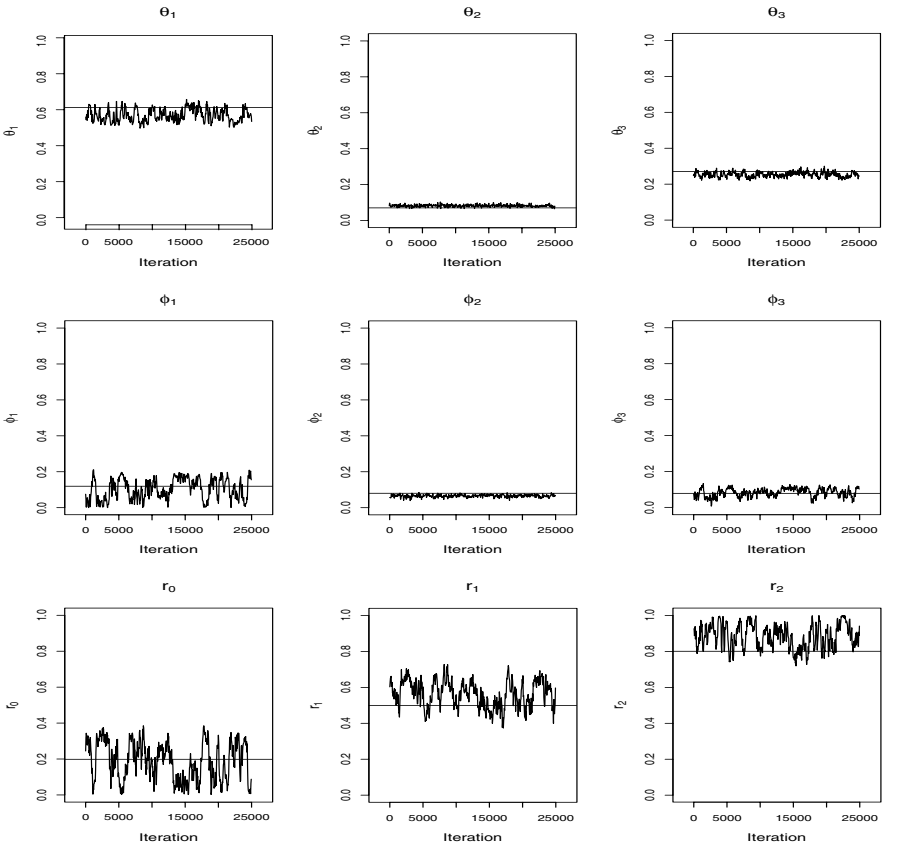

Figure 5.10 *MCMC output for the third synthetic dataset in the three population example of Section 5.4.3. Specifically, 25000 Gibbs sampler iterations are displayed. The superimposed horizontal lines denote true parameter values.*

and *very* disparate prevalences it is difficult to estimate the prevalences with any precision.

We find these results to be quite discouraging with regard to the sort of modelling alluded to at the start of this section. In principle we gain the ability to estimate more parameters when additional precisely measured explanatory variables are available, as each combination of levels of these variables constitutes a population with a distinct exposure prevalence. However, in most practical applications the exposure prevalence will only vary weakly with the covariates and response variable, so one is unlikely to see large differences in prevalences across these populations. And the present findings suggest that good estimation is not possible even with quite substantial variation in prevalence.

5.5 Summary

We have seen that Bayes-MCMC analysis provides a straightforward route to statistical inference which accounts for misclassification of a binary explanatory variable. Roughly speaking, we have considered two routes to obtaining sufficient information for a meaningful analysis to proceed. The first route involves the use of prior information about the misclassification; the second uses dual surrogates for the unobservable variable. We have observed numerous subtleties surrounding the performance of adjusted estimates, and also noted that often something more than 'off-the-shelf' MCMC techniques may be needed for model fitting. Finally, we have seen that simply having a rich enough data structure to potentially estimate the number of parameters at play does not always translate to actual good performance of estimators.

It would be remiss not to mention a number of papers that touch on similar ideas to those exhibited in this chapter. Indeed, a number of authors have considered Bayesian analysis which adjusts for misclassification of categorical variables, though often the focus is not on misclassification in explanatory variables *per se*. Related references include Johnson and Gastwirth (1991), Viana and Ramakrishnan (1992), Viana (1994), Evans, Guttman, Haitovsky, and Swartz (1995), and Tu, Kowalski, and Jia (1999).

5.6 Mathematical Details

5.6.1 Gibbs Sampling for the AWU Model of Section 5.2

The naive Gibbs sampler for fitting the AWU model of Section 5.2 would augment the (r_0, r_1, p, q) parameter space with the actual exposure V for each study subject. We denote the actual and surrogate exposure variables for the i-th subject in the j-th sample as V_{ji} and V^*_{ji} respectively. Clearly the naive Gibbs sampler would involve $n_0 + n_1 + 4$ individual updates at each iteration.

In fact it is more expedient to update sufficient statistics rather than the individual V variables, thereby obtaining an algorithm with an execution time that does not increase with the sample size. If we had complete data, that is observations of both V and V^* for each subject, then sufficient statistics would be (S_0, A_0, B_0) and (S_1, A_1, B_1), where

$$\begin{aligned} S_j &= \sum_{i=1}^{n_j} V^*_{ji} \\ A_j &= \sum_{i=1}^{n_j} V_{ji} V^*_{ji} \\ B_j &= \sum_{i=1}^{n_j} (1 - V_{ji})(1 - V^*_{ji}). \end{aligned}$$

That is, for sample j, S_j is the total number of apparent exposures, A_j is the total number of simultaneous actual and apparent exposures, and B_j is the

total number of simultaneous actual and apparent unexposures. Clearly the distribution of the complete data is given by

$$\begin{pmatrix} A_j \\ B_j \\ S_j - A_j \\ n_j - S_j - B_j \end{pmatrix} \sim Multinomial \left(n_j; \begin{matrix} r_j p \\ (1-r_j)q \\ (1-r_j)(1-q) \\ r_j(1-p) \end{matrix} \right), \tag{5.23}$$

independently for $j = 0, 1$.

From (5.23) and the prior described in Section 5.2 we immediately see the form of the joint posterior density for the unobserved quantities $U = (p, q, r_0, r_1, a_0, a_1, b_0, b_1)$ given the observed quantities $O = (s_0, s_1)$. Specifically,

$$\begin{aligned} f(U|O) \quad \propto \quad & \frac{1}{a_0! b_0! (s_0 - a_0)! (n_0 - s_0 - b_0)!} \times \\ & \frac{1}{a_1! b_1! (s_1 - a_1)! (n_1 - s_1 - b_1)!} \times \\ & r_0^{a_0 + n_0 - s_0 - b_0} (1 - r_0)^{b_0 + s_0 - a_0} \times \\ & r_1^{a_1 + n_1 - s_1 - b_1} (1 - r_1)^{b_1 + s_1 - a_1} \times \\ & p^{\alpha_p - 1 + a_0 + a_1} (1 - p)^{\beta_p - 1 + n_0 + n_1 - s_0 - s_1 - b_0 - b_1} \times \\ & q^{\alpha_q - 1 + b_0 + b_1} (1 - q)^{\beta_q - 1 + s_0 + s_1 - a_0 - a_1}. \end{aligned}$$

Clearly then the full conditional posterior distribution for each of (r_0, r_1, p, q) is a Beta distribution, with parameters that are easily 'read off' from the expression above. For instance, r_0 given the other unobservables and the observables has a $Beta(1 + a_0 + n_0 - s_0 - b_0, b_0 + s_0 - a_0)$ distribution.

Somewhat closer inspection of the expression for $f(U|O)$ reveals that each of (a_0, b_0, a_1, b_1) has a binomial full conditional posterior distribution. For instance, as a function of a_0 alone the expression is proportional to

$$\frac{1}{a_0! (s_0 - a_0)!} r_0^{a_0} (1 - r_0)^{-a_0} p^{a_0} (1 - q)^{-a_0},$$

whence

$$a_0 | a_0^C \quad \sim \quad \text{Binomial} \left(s_0, \frac{r_0 p}{r_0 p + (1 - r_0)(1 - q)} \right),$$

where a_0^C denotes all the elements of (U, O) except a_0. Thus the Gibbs sampler is easily applied to this model. All that is required is sampling from beta and binomial distributions.

5.6.2 A Tailored MCMC Algorithm for the AWU Model of Section 5.2

As alluded to in Section 5.2, Gibbs sampling is not necessarily an effective MCMC algorithm for nonidentifiable models. Here an alternate algorithm which is tailored to the particular model is presented.

Upon reparameterizing from (r_0, r_1, p, q) to $(\theta_0, \theta_1, p, q)$, the prior distribution formulated with respect to the original parameters becomes

$$f(\theta_0, \theta_1, p, q) \propto \frac{b(p; \alpha_p, \beta_p) b(q; \alpha_q, \beta_q)}{(p + q - 1)^2} I_{A(\theta_0, \theta_1)}(p, q), \tag{5.24}$$

where $b(\cdot, \alpha, \beta)$ denote the Beta density function with parameters (α, β), while

$$\begin{aligned} A(\theta_0, \theta_1) &= \left\{(p, q) : 0 < p < \underline{\theta}, 0 < q < 1 - \bar{\theta}\right\} \cup \\ &\quad \left\{(p, q) : \bar{\theta} < p < 1, 1 - \underline{\theta} < q < 1\right\}, \end{aligned} \tag{5.25}$$

with $\underline{\theta} = \min\{\theta_0, \theta_1\}$ and $\bar{\theta} = \max\{\theta_0, \theta_1\}$. Note in particular that the $(p + q - 1)^{-2}$ term in (5.24) arises from the Jacobian associated with the transformation, while the restriction of (p, q) to $A(\theta_0, \theta_1)$ is a re-expression of the restriction that each θ_j by definition must lie between $\min\{p, 1 - q\}$ and $\max\{p, 1 - q\}$. It is easy to verify that the restriction results in (5.24) being bounded, as values of (p, q) summing to one lie outside $A(\theta_0, \theta_1)$.

Thus upon combining the likelihood and prior the joint posterior density from which we wish to sample can be expressed as

$$\begin{aligned} f(\theta_0, \theta_1, p, q | s_0, s_1) \propto\ & \theta_0^{s_0} (1 - \theta_0)^{n_0 - s_0} \theta_1^{s_1} (1 - \theta_1)^{n_1 - s_1} \times \\ & \frac{b(p; \alpha_p, \beta_p) b(q; \alpha_q, \beta_q)}{(p + q - 1)^2} I_{A(\theta_0, \theta_1)}(p, q). \end{aligned} \tag{5.26}$$

There is no obvious way to simulate draws directly from this four-dimensional distribution.

On the other hand, a joint distribution which approximates (5.26) and from which we can sample easily is given by the density

$$\begin{aligned} g(\theta_0, \theta_1, p, q | y_0, y_1) = \ & b(\theta_0; s_0 + 1, n_0 - s_0 + 1) \times \\ & b(\theta_1; s_1 + 1, n_1 - s_1 + 1) \times \\ & \{k(\theta_0, \theta_1)\}^{-1} \times \\ & b(p; \alpha_p, \beta_p) b(q; \alpha_q, \beta_q) I_{A(\theta_0, \theta_1)}(p, q), \end{aligned} \tag{5.27}$$

where

$$k(\theta_0, \theta_1) = \int \int_{A(\theta_0, \theta_1)} b(p; \alpha_p, \beta_p) b(q; \alpha_q, \beta_q) dp\, dq.$$

Thus (5.27) defines a joint distribution for $(\theta_0, \theta_1, p, q)$ under which marginally θ_0 and θ_1 have independent Beta distributions, while conditionally p and q given (θ_0, θ_1) have independent Beta distributions truncated to $A(\theta_0, \theta_1)$. In particular, (5.27) approximates (5.26) in the sense that the (θ_0, θ_1) marginal under (5.26) is the posterior distribution arising from a uniform prior on (θ_0, θ_1), rather than from the intractable prior that results from marginalizing the actual prior on $(\theta_0, \theta_1, p, q)$ given by (5.24).

The simple form of $A(\theta_0, \theta_1)$ as a union of two rectangles in the unit-square implies that one can readily simulate draws from the joint distribution defined by (5.27) by sampling from the marginal $g(\theta_0, \theta_1)$ and then the conditional $g(p, q | \theta_0, \theta_1)$. In particular, the truncation involved in the latter step is readily

handled by evaluating the beta inverse distribution function. Similarly, one can readily compute $k(\theta_0, \theta_1)$ using evaluations of the beta distribution function. This implies that the joint density (5.27) is easily evaluated at any given argument.

Now let $\phi = (\theta_0, \theta_1, p, q)$ denote the entire parameter vector. Our MCMC algorithm for obtaining the next ϕ vector $\phi^{(j+1)}$ given the current vector $\phi^{(j)}$ proceeds by sampling a candidate parameter vector ϕ^* from (5.27) without regard for $\phi^{(j)}$. The Metropolis-Hastings scheme dictates that

$$\phi^{(j+1)} = \begin{cases} \phi^* & \text{with probability} \quad \min\left\{w(\phi^*)/w(\phi^{(j)}), 1\right\}, \\ \phi^{(j)} & \text{with probability} \quad 1 - \min\left\{w(\phi^*)/w(\phi^{(j)}), 1\right\}, \end{cases}$$

where $w(\phi) = f(\phi|s_0, s_1)/g(\phi|s_0, s_1)$ is the ratio of the candidate density to the target density. Note that since the candidate parameter vector is sampled from a distribution not depending on the current parameter vector, the only introduction of serial correlation into the MCMC algorithm is via rejections which lead to the same ϕ vector at successive iterations. Typically we expect that (5.27) will approximate (5.26) well, thereby making the probability of rejection small. By generating candidate (θ_0, θ_1) values from an approximate posterior distribution while generating (p, q) from a modified prior distribution with truncation according to the (θ_0, θ_1) values, we reflect the fact that (θ_0, θ_1) are identifiable, whereas (p, q) are not.

5.6.3 Comparing MCMC Algorithms for the AWU Model of Section 5.2

In the context of the AWU Model of Section 5.2 we compare the performance of the Gibbs sampler and the tailored algorithm for three synthetic datasets. The first dataset consists of 25/250 and 40/250 apparent exposures in the two samples. The second and third datasets involve the same proportions of apparent exposure, but successively four times larger sample sizes. That is, the second dataset comprises 100/1000 and 160/1000 apparent exposures, while the third comprises 400/4000 and 640/4000 apparent exposures. Throughout, $Beta(19, 1)$ priors are assigned to both p and q. This reflects a scenario where the investigator believes the exposure assessment to be very good. Indeed, the prior mode is $(p, q) = (1, 1)$, corresponding to an investigator's best guess that the classification is perfect. However, the prior admits uncertainty about this guess by allowing the possibility that one or both of p and q may be slightly less than one.

Figure 5.11 displays the output of 5000 MCMC iterations for the log odds-ratio ψ, for each dataset and each algorithm. In keeping with our general experience, the performance of the Gibbs sampler worsens as the sample sizes increase. At the smallest of the three sample sizes the performance of the two algorithms seems comparable when assessed 'by eye.' At the largest of sample size, however, the Gibbs sampler is mixing very poorly indeed. On the other hand, there is no obvious change in performance of the tailored algorithm across sample size. In fact the observed rejection rate for the Metropolis-Hastings update does not change appreciably across the three datasets, re-

maining at about 7% in each case. These findings are commensurate with our experience in this model and related models which are also nonidentifiable. In the face of large sample sizes Gibbs sampling can perform very poorly and it may be necessary to consider more specialized algorithms tailored to the nonidentified nature of the posterior distribution at hand.

5.6.4 MCMC Fitting for the DUAL-DEP Model of Section 5.3

The DUAL-DEP model described in Section 5.3 involves eight unknown parameters,

$$\theta \quad = \quad (r_0, r_1, p_1, p_2, q_1, q_2, \delta_0, \delta_1).$$

As is often the case, it is expedient to augment the parameter vector by the true but unobserved values of V for each subject. Then it is trivial to implement Gibbs sampling updates for r_0 and r_1, as these parameters have beta full conditional posterior distributions, akin to the situation for the simpler DUAL-IND model also described in Section 5.3. The full conditional distributions for the other parameters, however, do not have standard forms. In a related model Dendukuri and Joseph (2001) illustrate that updates according to the full conditional distributions can still be implemented. We are concerned, however, that this may not always work well in light of the nonidentifiable model.

As an alternative we consider block updates to (p_1, p_2, δ_1) and (q_1, q_2, δ_0) given the other parameters and the V values in each case. For instance, let $\lambda = (p_1, p_2, \delta_1)$. The way in which λ governs the distribution of $(V_1^*, V_2^* | V = 1)$ is made explicit by reparameterizing from λ to

$$\phi \quad = \quad (p_1 p_2 + \delta_1, p_1(1 - p_2) - \delta_1, (1 - p_1)p_2 - \delta_1).$$

It is easy to verify that the Jacobian of this transformation is simply one, so that the posterior conditional distribution of ϕ given the other parameters and the data has a density of the form

$$f\left(\phi | \phi^C\right) \quad \propto \quad \phi_1^{c_{1,1}} \phi_2^{c_{1,0}} \phi_3^{c_{0,1}} (1 - \phi_1 - \phi_2 - \phi_3)^{c_{0,0}} f_\lambda(\lambda(\phi)),$$

which is the product of a multinomial likelihood function and a complicated prior density. An obvious Metropolis-Hastings update is to generate a candidate ϕ according to

$$\phi^* \quad \sim \quad \text{Dirichlet}(1 + c_{1,1}, 1 + c_{1,0}, 1 + c_{0,1}, 1 + c_{0,0}), \tag{5.28}$$

regardless of the current ϕ value. Note, in particular, that (5.28) would be the posterior conditional distribution if the prior distribution of ϕ were in fact uniform. Following the usual Metropolis-Hastings scheme, the candidate ϕ^* is accepted with probability

$$acc(\phi^*; \phi) \quad = \quad \min\left\{ \frac{f_\lambda(\lambda(\phi^*))}{f_\lambda(\lambda(\phi))}, 1 \right\},$$

which is unlikely to be very small given that only a prior density ratio is involved.

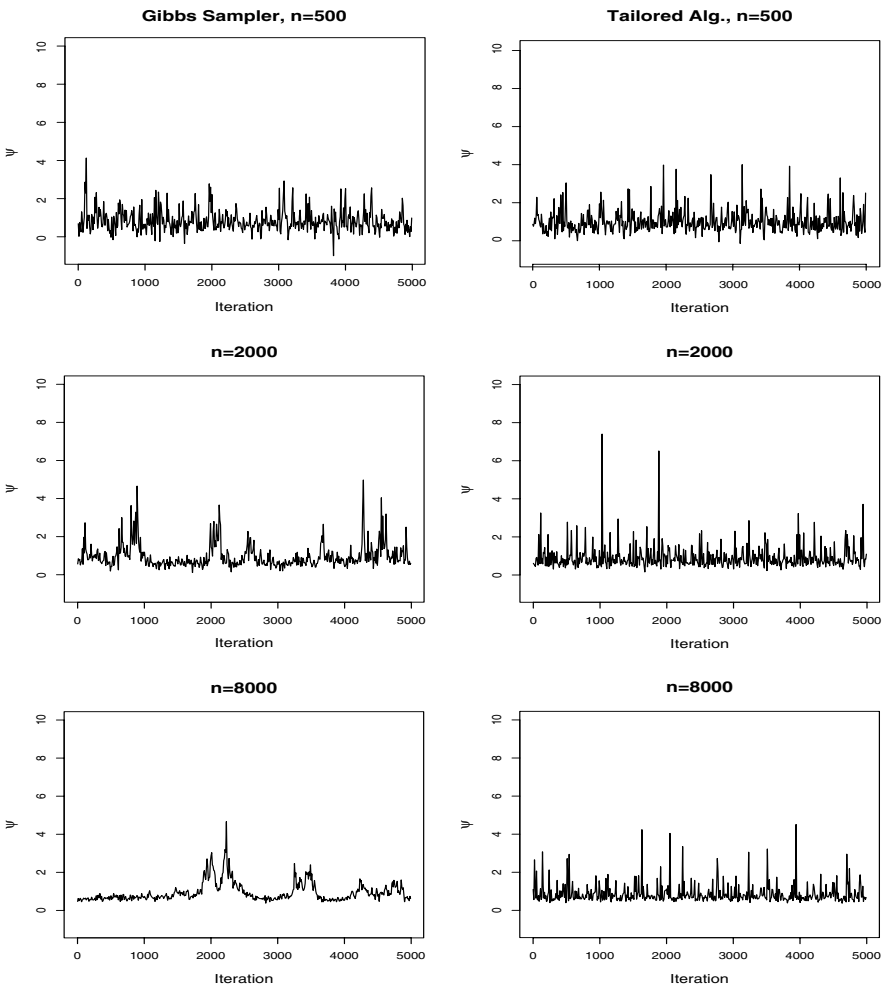

Figure 5.11 *MCMC output for the log odds-ratio ψ under the AWU Model of Section 5.2, using the Gibbs sampler (left panels) and the tailored Metropolis-Hastings algorithm (right panels). The upper panels are based on* 25/250 *and* 40/250 *apparent exposures in the control and case samples respectively. The middle and lower panels are based on the same proportions of apparent exposure but respectively four and sixteen times larger sample sizes. In all cases p and q are assigned Beta(19,1) prior distributions.*

5.6.5 Gibbs Sampling for the Model of Section 5.4.3

Applying uniform priors, or more precisely $\theta \sim Dir(1,1,1,1)$, $\phi \sim Dir(1,1,1,1)$, and $r_i \sim Beta(1,1)$ for $i = 0,1,2$, from the multinomial model (5.22) we have

$$f(z,\theta,\phi,r|s) \quad \propto \quad \prod_{i=0}^{2}\prod_{j=1}^{4}\{r_i\theta_j\}^{z_j^{(i)}}\{(1-r_i)\phi_j\}^{s_j^{(i)}-z_j^{(i)}}.$$

From here we can 'read off' that

$$Z_j^{(i)}|Z_j^{(i)\,C} \quad \sim \quad \text{Binomial}\left(s_j^{(i)}, \frac{r_i\theta_j}{r_i\theta_j+(1-r_i)\phi_j}\right),$$

$$\theta|\theta^C \quad \sim \quad \text{Dirichlet}\left(1+\sum_i z_1^{(i)},\ldots,1+\sum_i z_4^{(i)}\right)$$

truncated to the region $\{\theta_1 > \phi_1, \theta_4 < \phi_4\}$,

$$\phi|\phi^C \quad \sim \quad \text{Dirichlet}\left(1+\sum_i s_1^{(i)}-z_1^{(i)},\ldots,1+\sum_i s_4^{(i)}-z_4^{(i)}\right)$$

truncated to the region $\{\phi_1 < \theta_1, \phi_4 > \theta_4\}$, and finally,

$$r_i|r_i^C \quad \sim \quad \text{Beta}\left(1+\sum_{j=1}^{4} z_j^{(i)}, 1+\sum_{j=1}^{4} s_j^{(i)}z_j^{(i)}\right).$$

Thus the Gibbs sampler can be implemented in a straightforward manner.