J. R. Statist. Soc. A (2002)
165, *Part* 3, *pp.* 549–566

Mixture models in measurement error problems, with reference to epidemiological studies

Sylvia Richardson,

Imperial College School of Medicine, London, UK

Laurent Leblond and Isabelle Jaussent

Institut National de la Santé et de la Recherche Médicale, Villejuif, France

and Peter J. Green

University of Bristol, UK

[Received June 2001. Revised February 2002]

Summary. The paper focuses on a Bayesian treatment of measurement error problems and on the question of the specification of the prior distribution of the unknown covariates. It presents a flexible semiparametric model for this distribution based on a mixture of normal distributions with an unknown number of components. Implementation of this prior model as part of a full Bayesian analysis of measurement error problems is described in classical set-ups that are encountered in epidemiological studies: logistic regression between unknown covariates and outcome, with a normal or log-normal error model and a validation group. The feasibility of this combined model is tested and its performance is demonstrated in a simulation study that includes an assessment of the influence of misspecification of the prior distribution of the unknown covariates and a comparison with the semiparametric maximum likelihood method of Roeder, Carroll and Lindsay. Finally, the methodology is illustrated on a data set on coronary heart disease and cholesterol levels in blood.

Keywords: Bayesian modelling; Epidemiological studies; Errors in variables; Finite mixture distributions; Heterogeneity; Logistic regression; Markov chain Monte Carlo algorithms; Misspecification

1. Introduction

In this paper, we examine the Bayesian formulation of measurement error problems (sometimes known as 'errors-in-variables' problems), and we focus on the question of the specification of the prior distribution of the unknown true covariates. Our aim is to propose a flexible model for this distribution by using a mixture of normal distributions, to reduce the sensitivity to modelling assumptions. Our motivation comes from the field of epidemiology, where measurement error is an important problem and heterogeneity in the distribution of the underlying covariates cannot be ruled out. But the model that we outline applies outside this context, as errors-in-variables problems arise in a variety of domains such as social sciences and econometrics.

Measurement error problems are concerned with inference on regression coefficients for an outcome Y in terms of covariates X, in cases where X is not measured accurately on all subjects,

Address for correspondence: Sylvia Richardson, Department of Epidemiology and Public Health, Imperial College School of Medicine, St Mary's Campus, Norfolk Place, London, W2 1PG, UK.
E-mail: sylvia.richardson@ic.ac.uk

 0964–1998/02/165549

but information on X is available through the recording of an imperfect surrogate U. It is well known that regressing Y on U and ignoring measurement error can be seriously misleading, and many methods have been proposed for countering this (see for example Carroll *et al.* (1995) and the references therein).

Bayesian analysis of measurement error problems has been developed from the seminal work of Clayton (1992). It is based on *structural* and *functional specifications*. In most cases, structural specifications entail the formulation of three submodels: an *outcome model* relating Y to X, a *measurement model* relating U and X and a *prior model* $p(X)$ for the prior distribution of X. These models are then linked by a graphical model using suitable conditional independence assumptions. Information on the measurement error process can be built in a flexible way into the graphical model (Richardson and Gilks, 1993a, b).

At the second stage, functional forms for the distributions involved in the submodels must be chosen. For the outcome and the measurement models, this choice is guided by epidemiological knowledge. Typically, linear or logistic regressions, coupled with normal or log-normal error models corresponding to additive or multiplicative errors, have been used in epidemiology to model the association between (dichotomous) outcomes, covariates and their surrogates.

However, in the common situation of observational studies, there are many reasons to suspect heterogeneity in the distribution of the unknown X among the population studied. Exposure variables often have a skew distribution with a high peak corresponding to moderate exposure for the majority of the population and a long tail or a second lower peak corresponding to larger exposure for a small fraction of the population. These have traditionally been fitted by log-normal distributions, but a mixture model is a plausible alternative (see, for example, the dietary example discussed in Schafer (2001)). Heterogeneity in the distribution of physiological variables can also be expected when these are linked to genetic polymorphism. Heterogeneity may also arise by epidemiological study design, if subgroups with extreme covariate values are oversampled purposely to increase the exposure contrast.

In all these cases, there is little information to model this heterogeneity besides having either some observations of a gold standard in a small validation subgroup or some replicate measures of the surrogate, which can be misleading if the validation group is small or the measurement error large. Since regression results are known to be sensitive to the choice of a particular parametric shape for $p(X)$, a natural question is how to model $p(X)$ flexibly, avoiding such choices and allowing the data to determine the shape of $p(X)$.

In the measurement error context, this question has been actively pursued and has led to different lines of development. Carroll *et al.* (1993) have proposed a pseudolikelihood method, using validation data. Roeder *et al.* (1996) combined a parametric disease model and nonparametric mixtures for $p(X)$ and carried out the estimation via nonparametric maximum likelihood (NPML) algorithms. A similar line was also followed by Schafer (2001) and Aitkin and Rocci (2002), who emphasized the general applicability of NPML to different measurement error contexts. Müller and Roeder (1997) have developed a Bayesian approach using Dirichlet process priors for the joint model of X and U.

Mixture models with variable numbers of components are an alternative natural framework in which to consider a flexible modelling of prior distributions in Bayesian hierarchical models. Such mixture models have been developed in a Bayesian context by Richardson and Green (1997) using novel Markov chain Monte Carlo (MCMC) methods based on the reversible jump algorithm proposed by Green (1995). Our aim is to show how Bayesian mixtures can be used in this context and in particular to discuss the implementation of prior distributions for X based on mixtures of univariate normal distributions with an unknown number of components in typical measurement error situations in epidemiology involving a validation group. We

also include an assessment of the influence of misspecification of the prior distribution of the unknown covariates and a comparison with the semiparametric maximum likelihood method of Roeder *et al.* (1996).

Using mixtures of normal distributions to increase the robustness to model misspecification for $p(X)$ is in line with recent developments. Mixtures with a fixed number of components have been used by Carroll *et al.* (1999) for linear errors-in-variables models. In the closely related context of generalized linear mixed models, mixtures of normal distributions with common standard deviation were used by Magder and Zeger (1996) as prior distributions for the random effects. Mixtures of normal distributions have also been implemented by Carroll *et al.* (1999) in a nonparametric context of regression splines with measurement error.

The model formulation and its estimation via MCMC algorithms are outlined in Sections 2 and 3. In Section 4, we illustrate its implementation and some aspects of its performance. In Section 5 we present simulation results where two aims are pursued: demonstrating the effect of misspecifying the parametric model for the unknown covariates and the improvement that is brought by using mixture models, and showing a limited comparison of our approach with the NPML method of Roeder *et al.* (1996). An application to a data set on coronary heart disease and cholesterol levels in blood is presented in Section 6 and the paper concludes with a brief discussion.

2. Model formulation

Bayesian analysis of measurement error problems has been developed notably by Clayton (1992), Stephens and Dellaportas (1992), Richardson and Gilks (1993a, b), Mallick and Gelfand (1996), Müller and Roeder (1997) and Richardson and Leblond (1997).

2.1. General structure

We shall focus on the case of studies with a validation group, i.e. a subgroup where, by design, the covariate of interest X is accurately recorded by the use of a so-called 'gold standard' method, which is usually costly to implement on a large scale. Note that there are cases of measurement error situations where the measurement error variance is identifiable even in the absence of a validation group or repeated measures. For example, Aitkin and Rocci (2002) discuss linear or generalized linear regression models, normal measurement models and a discrete distribution for X. In most cases though, the identifiability of the measurement error variance will be very weak. Inference on the regression parameters of interest will thus be considerably strengthened by including information on the measurement model parameters through observations in a validation group.

Throughout, let i index the individuals, let Y denote the known outcome, X the true covariate which is unobserved (except in the validation subgroup), U the observed surrogate for X and C the known covariates. For the present work, we assume that X and U are univariate. By the terminology 'main study', we refer to the group of individuals where X has not been recorded, a group which is usually by design larger than the validation group.

As detailed in Richardson and Gilks (1993a), the general structure follows from the formulation of local submodels between the components by using conditional independence assumptions and information on the design used to evaluate the measurement process:

(a) $p(Y_i|X_i, C_i, \beta)$, the regression model;
(b) $p(U_i|X_i, \lambda)$, the (classical) measurement model;
(c) $p(X_i|\pi)$, the prior model for X.

A key assumption which allows the modelling to proceed is that of non-differential measurement error which can be expressed as

$$p(Y_i|U_i, X_i, C_i, \beta) = p(Y_i|X_i, C_i, \beta)$$

or equivalently $Y_i \perp U_i | X_i$ for all i. This leads to the joint distribution

$$p(\pi, \lambda, \beta, X, U, Y) = p(\pi)\, p(\lambda)\, p(\beta) \prod_i p(X_i|\pi) \prod_i p(U_i|X_i, \lambda) \prod_i p(Y_i|X_i, C_i, \beta) \qquad (1)$$

where $p(\beta)$, $p(\lambda)$ and $p(\pi)$ are the prior distributions for the parameters of the three submodels.

In expression (1), the contribution of the validation group is not formally separated out, because our implementation uses directly the product form of expression (1), the only distinction between the contribution of the main study or validation group data being that X_i is observed and not latent. The validation group contributes crucially information on the parameters λ, which is used together with values of U_i to inform about X_i in the main study group; both groups contribute to the estimation of β and π.

If an analytical expression for the posterior distribution of all the parameters based only on the validation group were available, a different implementation could be envisaged. To be specific, let us distinguish data in the main study and the validation group by the superscript index *main* or *val* respectively. Then, the following identity is true (for simplicity, we are omitting the known covariates C from this):

$$\begin{aligned} &p(\pi, \lambda, \beta, X^{\text{val}} | X^{\text{main}}, U^{\text{main}}, Y^{\text{main}}, U^{\text{val}}, Y^{\text{val}}) \\ &\qquad \propto p(\pi, \lambda, \beta | X^{\text{val}}, U^{\text{val}}, Y^{\text{val}})\, p(X^{\text{main}}, U^{\text{main}}, Y^{\text{main}} | \pi, \lambda, \beta). \end{aligned} \qquad (2)$$

The left-hand side is the desired posterior of interest and the first factor on the right-hand side is the posterior based only on the validation group. If this term had an analytical expression, then it could be used as a prior in the implementation, with the second term of the right-hand side of expression (2) as the likelihood. This would have the advantage of having a well-calibrated prior and would increase the efficiency of the sampling. But, as will be apparent from the next section, the mixture prior with an unknown number of components that we propose as the prior for X does not lead to a tractable analytic expression for $p(\pi, \lambda, \beta | X^{\text{val}}, U^{\text{val}}, Y^{\text{val}})$. Thus, our implementation does not make use of decomposition (2) but derives directly from the joint distribution of all the parameters expressed in distribution (1).

2.2. Mixture model

We now expand the structure of the mixture model chosen as a prior for $p(X|\pi)$. Here π denotes generically all the parameters of the mixture model, as detailed below. Following the hierarchical Bayesian formulation of mixtures introduced in Richardson and Green (1997), we suppose that

$$X_i \sim \sum_{j=1}^{k} w_j f(\cdot|\theta_j) \qquad \text{independently for } i = 1, 2, \ldots, n \qquad (3)$$

where $f(\cdot|\theta)$ is a given parametric family of densities, and the parameters $\theta = \{\theta_j\}, w = \{w_j\}$ and k are unknown.

The hierarchical formulation of this mixture model introduces *latent allocation variables* z_i indicating to which mixture component the observation X_i belongs. Hence, model (3) is formulated equivalently as

$$p(z_i = j) = w_j \qquad \text{independently for } i = 1, 2, \ldots, n,$$

$$X_i|z \sim f(\cdot|\theta_{z_i}) \qquad \text{independently for } i = 1, 2, \ldots, n.$$

Assuming natural conditional independence assumptions, the hierarchical mixture model is thus expressed by the joint distribution

$$p(X, k, \theta, w, z) = p(k)\, p(\theta|k)\, p(w|k)\, p(z|w, k)\, p(X|\theta, z). \tag{4}$$

With respect to the generic notation of Section 2, we have $\pi = (k, \theta, w, z)$. For flexibility, we allow the priors for k, θ and w to depend on hyperparameters, drawn from independent hyperpriors (for details, see Richardson and Green (1997)).

We stress that treating k, the number of components, as being unknown and integrating over its posterior distribution when estimating regression parameters of interest enhances the adaptiveness of the mixture to heterogeneity in the underlying distribution of the Xs. This is different from other approaches where a fixed value for k is usually estimated by using a Bayes information criterion approximation, as in Carroll *et al.* (1999).

2.3. Joint model and parametric specifications

Combining distributions (1) and (4) leads to the following expression for the joint distribution of the measurement error problem:

$$p(k)\, p(\theta|k)\, p(w|k)\, p(\beta)\, p(\lambda) \prod_i p(z_i|w, k)\, p(X_i|\theta_{z_i}, z_i) \prod_i p(U_i|X_i, \lambda) \prod_i p(Y_i|X_i, C_i, \beta). \tag{5}$$

We are fully benefiting here from the flexibility that is offered by Bayesian modelling, coupled with conditional independence assumptions, which allows us to insert the hierarchical specification of the mixture model coherently within the measurement error model. The structure outlined above is generic and can accommodate many parametric specifications of the submodels. In what follows, we shall concentrate on the widely entertained logistic regression with normal or log-normal errors, a set-up which will allow us to make comparisons with other approaches.

For the regression model, we consider a binary outcome Y related to the unobserved X by a logistic link involving an intercept β_0 and a slope β_1:

$$Y \sim B(p) \qquad \text{with } \text{logit}(p) = \beta_0 + \beta_1 X.$$

With respect to the generic notation of Section 2, we have $\beta = (\beta_0, \beta_1)$. The dichotomous observations will be referred to as cases ($Y = 1$) and non-cases ($Y = 0$). This regression model corresponds to a prospective formulation, i.e. where the likelihood considered is that of a disease given a risk factor.

For the measurement model, we investigate Gaussian errors on an additive scale:

$$U_i \sim N(\alpha_0 + \alpha_1 X_i, \tau^{-1}). \tag{6}$$

The measurement model thus involves three parameters; in the notation of Section 2, we have $\lambda = (\alpha_0, \alpha_1, \tau)$. Note that we have used a parameterization in terms of the precision τ which is the inverse variance of the measurement error model, a common choice of parameterization in Bayesian analyses. The interpretation of the variance τ^{-1} depends on the context. It can be related to the accuracy of the measuring instrument, or more commonly in epidemiology arises from within-person variability of the underlying biological quantity or risk factor that we are trying to study. For α_0 and α_1 to be identifiable, it is necessary to have additional information

on the measurement model provided, for example in our set-up, by the gold standard in the validation group. In the absence of validation data, only the measurement error variance could be identifiable, and this only when the regression model is not Gaussian. In Section 5, we consider one situation of a measurement error model with no validation data, but in this case the estimation is carried out with known measurement error variance and known values of α_0 and α_1.

For the mixture model, we consider a mixture of Gaussian distributions which can approximate any continuous distribution well; thus $\theta_j = (\mu_j, \sigma_j^2)$ and $f(\cdot|\theta_j) = N(\mu_j, \sigma_j^2)$. We assume that there are n_0 observations of $\{Y_i, U_i\}$ in the main study and n_1 observations of $\{Y_i, X_i, U_i\}$ in the validation group. We let $n = n_0 + n_1$.

2.4. Prior specifications

Our aim is to define hyperprior settings which use the available information derived from the validation group, while remaining only weakly informative. Estimates for the regression parameters, β_0 and β_1, as well as the measurement parameters, α_0 and α_1, can be obtained by standard analyses based only on the validation group. Of course, if the validation group is small, these will be highly variable and we shall only use them for centring the priors. For these parameters, we thus choose weakly informative Gaussian priors centred on these estimates, with a large variance equal to 100. For the precision τ that is involved in the measurement model (6), we use a gamma distribution $\Gamma(0.01, 0.01)$. This quasi-'non-informative' choice is possible here because the validation group provides information on τ.

We also need to specify the hyperparameters of the mixture model. There, we follow Richardson and Green (1997), to which the reader is referred for details. In Richardson and Green (1997), the setting of the normal priors for the component means μ_j as well as the hierarchical gamma priors for the inverse of the variances σ_j^2 depend on the notion of a 'range' R of the mixture, which is usually taken to be the observed range of the data. In our case, the mixture concerns 'hidden variables', so there is no observed range. Again, we make use of the available information to define a notional range. To be precise, from the validation group data, we obtain an estimate of the error variance $\hat{\tau}^{-1}$ and of the regression model between X_i and U_i. Then, this regression is used in the main study group to predict values of X_i: $\hat{X}_i$ from the observed U_i. Finally, we define 'initial values' for the X_i by perturbing the $\hat{X}_i$ by a Gaussian $(0, \hat{\tau}^{-1})$ distribution randomly. The range R of these initial values is used as the notional range for X. Other choices would be suitable so long as they do not restrict too much the range for X, in particular, using directly the $\hat{X}_i$ to define this range is too restrictive. In the case of known measurement error, the same procedure is followed using the known rather then the estimated value of τ.

Once the notional range R has been defined, we then adopt the detailed specifications of Richardson and Green (1997) for the mixture parameters, based on this range.

Finally for k we choose a uniform prior on $1, \ldots, 30$, and for w a Dirichlet distribution $D(\delta, \delta, \ldots, \delta)$ with $\delta = 1$ which corresponds to a uniform prior on the weights. In practice, the mixture will rarely use more than 10 components and k could be defined on a smaller range without any loss of flexibility.

3. Estimation via Markov chain Monte Carlo algorithms

Bayesian methods for estimating model parameters are based on the posterior distributions of the parameters given the data. Given the analytic intractability of these distributions, MCMC methods now play a major role in these computations; see, for example, Besag *et al.* (1995), Gilks

et al. (1996) and Green (2001). Here, apart from standard Gibbs or Metropolis algorithms, we shall also make use of the reversible jump algorithm introduced by Green (1995).

In brief, at each sweep of the algorithm, the following moves are used:

(a) updates of β, α and τ;
(b) updates of X_i using $(z_i, \theta_{z_i}, U_i, Y_i, C_i, \beta, \alpha, \tau)$ for i corresponding to the individuals in the main study;
(c) updates of w, z and θ *conditional on* k;
(d) updates of k and consequently of the relevant mixture parameters.

For step (a) the moves are conventional. We used a random walk Metropolis step where the acceptance probabilities are calculated by using expression (5) for the joint distribution. Gibbs sampling was not straightforward owing to the logistic form of the regression model.

For step (b), this involves updating one at a time all the unknown covariates X_i. Since this is done *conditionally on* z_i, this is again straightforward by using a random walk Metropolis step, the distribution $X_i|z_i$ being $N(\mu_{z_i}, \sigma^2_{z_i})$.

The moves for updating the mixture parameters and changing k, the number of components, by using reversible jump *split–merge* proposals have been described in detail in Richardson and Green (1997). Basically, local changes increasing or decreasing the number of components by 1 and preserving some moment conditions are proposed, then accepted following a Metropolis ratio. In contrast with the usual setting for mixtures, the values of X_i for the main study group will change at each iteration of the algorithm. Thus, teasing out the mixture is challenging in this context and its feasibility needs to be tested in a simulation study. Slower convergence of the mixture parameters than in a case where the X_i are fixed can be expected.

The only initialization of the algorithm that is useful to mention is that of the unobserved X_i in the main study. There, we use the initial values obtained by the procedure described in Section 2.4 for setting the notional range of the mixture.

4. Implementation and diagnostics

4.1. The data sets

In this section, we show details of the performance of the algorithm on three simulated data sets. The design of the three simulated data sets will also be used in the simulation study that is reported in Section 5. For the first two data sets, we simulate X from a mixture, whereas for the third we follow the simulation set-up of Roeder *et al.* (1996) where X has a log-normal distribution.

The first example is referred to as 'bimod'. It corresponds to $n_0 = 180$, $n_1 = 60$, $U \sim N(X, \tau^{-1})$, with $\tau = 0.67$, and $\text{logit}\{P(Y|X)\} = \beta_0 + \beta_1 X$ with $\beta_0 = 0.5$ and $\beta_1 = 0.6$. The distribution of X is simulated from a well-separated symmetric bimodal mixture: $0.5\, N(-2.0, 1.0) + 0.5\, N(2.0, 1.0)$ (with numbers from each component drawn binomially from the stated weights).

The second example is referred to as 'skew'. It corresponds to $n_0 = 250$, $n_1 = 50$, $U \sim N(X, \tau^{-1})$, with $\tau = 1.8$, and $\text{logit}\{P(Y|X)\} = \beta_0 + \beta_1 X$ with $\beta_0 = -0.8$ and $\beta_1 = 0.9$. The distribution of X is simulated from an asymmetric normal mixture: $0.5\, N(0.19, 0.08^2) + 0.2\, N(1.05, 0.2^2) + 0.3\, N(2.0, 0.48^2)$.

These two examples are cases of what is considered a substantial measurement error, i.e. where the measurement error variance τ^{-1} is roughly equal to half of the variance of X (which is equal to 3 and 1.1 for the bimod and skew examples respectively). The value of β_0 ensures that the data sets contain approximatively half cases and half non-cases. There is more information in the bimod example since the relative size of the validation group to the main study is larger than

for the skew example. Moreover two of the components in the skew mixture have substantial overlap and hence inference is more challenging in this case.

The third example is inspired by an example that was used by Roeder *et al.* (1996) and is referred to as 'RCL'. There is none-the-less a major difference from the work of Roeder *et al.* (1996), in that we treat the study as prospective rather than as a retrospective case–control set-up. The implementation of case–control studies in a Bayesian framework is not straightforward; see Seaman and Richardson (2001).

In the RCL example, $n_0 = 180$ and $n_1 = 60$ as in the bimod example, but the measurement model is multiplicative: $\log(U) \sim N\{\log(X), \tau^{-1}\}$, with $\tau = 4$, and $\text{logit}\{P(Y|X)\} = \beta_0 + \beta_1 X$ with $\beta_0 = -0.7$ and $\beta_1 = 0.5$. The distribution of X is defined as $\log N(-0.43, 1.08^2)$.

Irrespectively of the model that is used to generate the Xs, we implement an estimation of the joint model equation (5) below. In particular, whenever there is a validation group, we estimate α_0 and α_1 even though their simulation settings were 0 and 1 respectively. Note that, for the RCL example, the normal mixture will be used to approximate a log-normal distribution.

4.2. Performance

The MCMC algorithm was run for 50000 sweeps. The graphical assessment of convergence includes time plots of the parameters indexed by sweep, cumulative averages or occupancy fraction for continuous or discrete-valued parameters. The stability of these plots as the number of sweeps increases gives a useful graphical check on convergence. As can be seen from the plots in Fig. 1, in the bimod example convergence to stable cumulative averages for the regression parameter β_1 (posterior mean 0.63; posterior standard deviation 0.1) or the precison τ (plotted every 20 sweeps) is obtained after 30000 iterations, giving values which are close to the simulated true values. Similarly, the cumulative occupancy fractions for the number of components in the mixture, $\{p(k < j|y), 2 \leqslant j \leqslant 10\}$, have stabilized also after 30000 sweeps. On this graph, $p(k < 2|y)$ is not represented because the state $\{k = 1\}$ was never visited. We see that $p(k = 2|y) = p(k < 3|y)$ is around 0.6, giving a clear indication that there is posterior support for two components in the bimodal mixture.

Convergence is somewhat slower in the skew example, but nevertheless the parameter estimates stabilize after 40000 sweeps (Fig. 1). For this particular data set, again $\{k = 1\}$ was never visited and we see support for two or three components, with a mode of $p(k|y)$ at $k = 3$. It is interesting to see that the increased uncertainty in the skew example is reflected in a larger posterior standard deviation of the regression parameter β_1 (posterior mean 0.89; posterior standard deviation 0.2) than in the bimod case.

In Figs 2 and 3 are displayed the empirical distribution of the simulated Xs in the whole study with an empirical density estimate provided by the density function in S-PLUS (Figs 2(a) and 3(a)), that observed in the validation group (Figs 2(b) and 3(b)) and the posterior density estimation given by the Bayesian implementation using the full model described in equation (5) (Figs 2(d) and 3(d)). Indeed, from the output of the chain, we obtain an estimate of the predictive densities of the mixture conditional on k, by averaging across the MCMC run conditionally on k, and an 'overall' 'Bayesian predictive density estimate of the distribution of X integrated with respect to k, $E\{f(\cdot|k, w, \theta)|X\}$ (full curve), by averaging across values of k. To summarize, the density curves in Figs 2(a) and 3(a) are based on X^{main} and X^{val}, whereas those in Figs 2(d) and 3(d) are based on X^{val}, U^{main}, U^{val}, Y^{main} and Y^{val}. Both curves are superimposed on the histogram of X^{main} and X^{val}. We see a good recovery of the true underlying density in both the bimod and the skew cases, in comparison with the blurred shape of the distribution of U (Figs 2(c) and 3(c)), even though there was only a small fraction of the sample in the validation group.

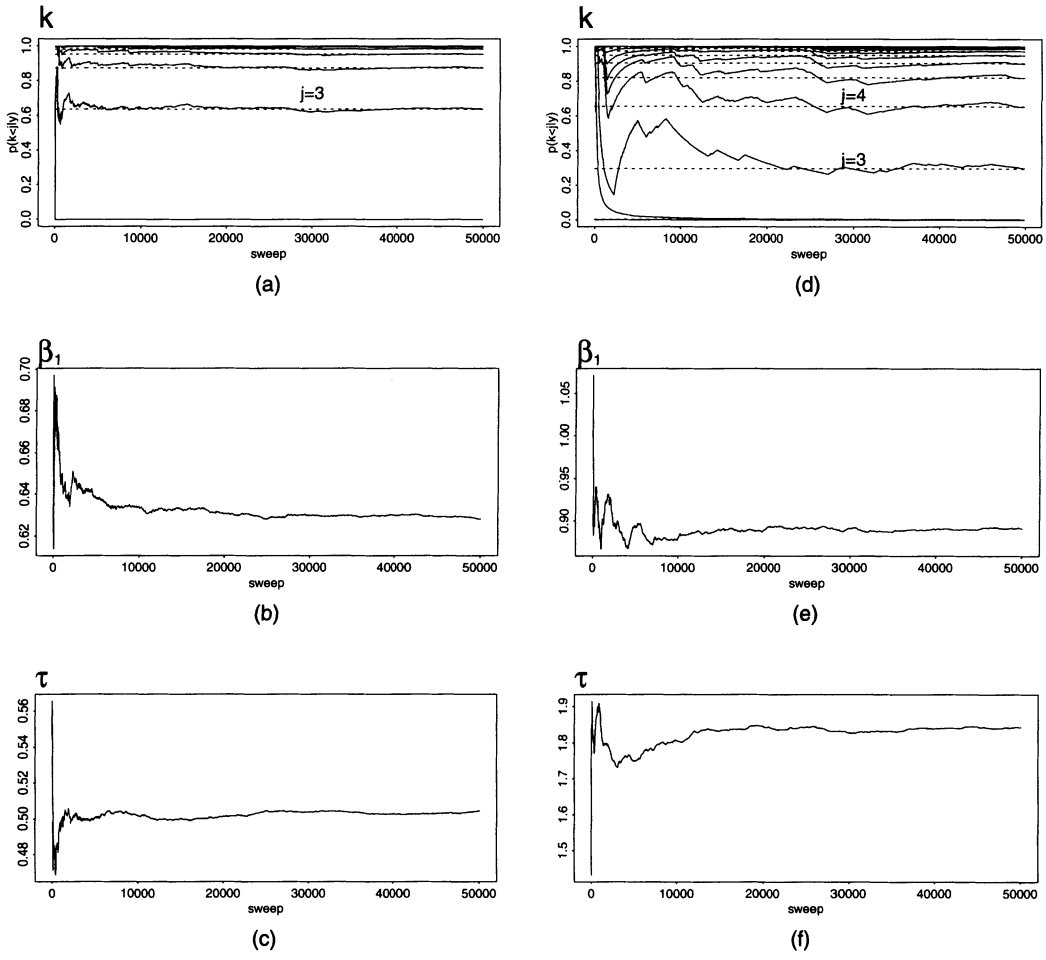

Fig. 1. Graphical assessment of the convergence of the MCMC algorithm for (a)–(c) the bimod example and (d)–(f) the skew example: (a), (d) cumulative occupancy fractions $p(k < j|y)$ for j ranging from 2 to 10; (b), (e) cumulative mean of β_1; (c), (f) cumulative mean of τ

It is also interesting to investigate the performance of the model with respect to the latent variables by comparing the true X_i with the distribution of the X_i simulated through the algorithm. This is done in Fig. 4, which displays in Figs 4(a) and 4(c) the true X_i (dots) *versus* their posterior mean (crosses) ordered by values of the surrogate U, whereas in Figs 4(b) and 4(d) an interval of 95% posterior variability is superimposed. For both data sets, we can see a classical shrinkage effect of the posterior means. This is more noticeable for the skew case. Indeed, because of the large weight of the first component and the substantial measurement error, the observations in the second component are pulled down on average. Nevertheless, nearly all the true X_i are included in the posterior variability interval.

The ability of the mixture model to estimate the regression coefficients as well as to characterize the underlying density of the true Xs is also apparent in the RCL example (Fig. 5). There, the mixture uses between four and seven components to approximate the log-normal shape. Again, the estimates stabilize quickly after 30000 sweeps. In Fig. 5(d) we see that the variability (on the log-scale) of the posterior estimates matches the true values well.

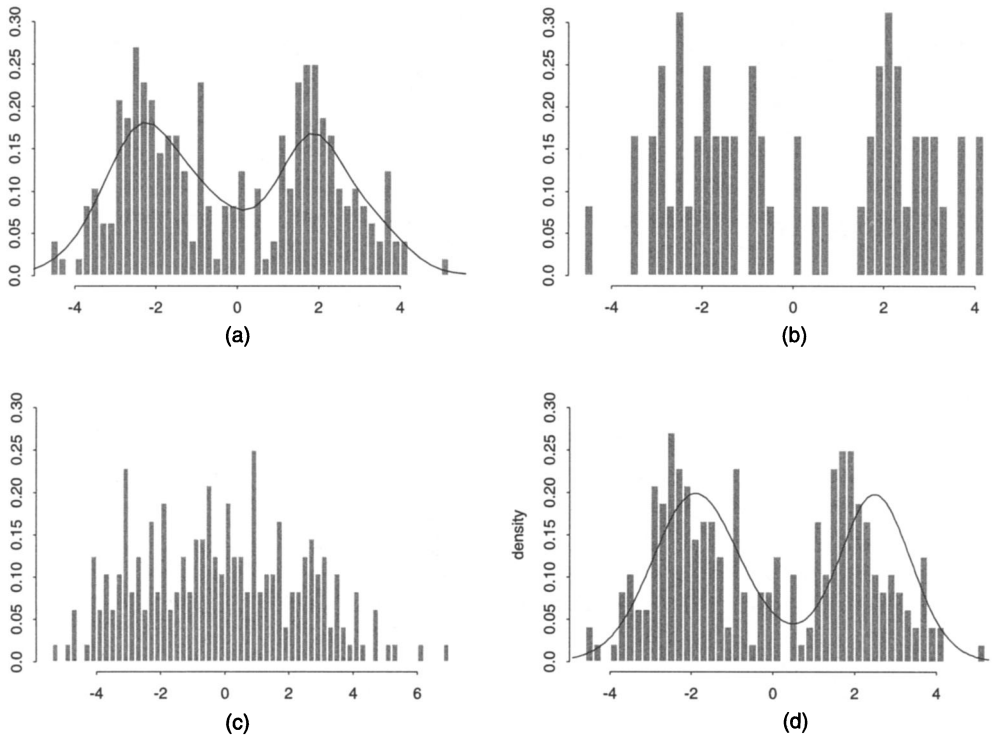

Fig. 2. Bimod case—simulated data set (all subjects and validation group) and a comparison with the posterior density estimate given by the mixture model based only on the validation, surrogate and outcome data: (a) true covariate X—all subjects (n = 240); (b) true covariate X—validation group (n = 60); (c) surrogate U—all subjects (n = 240); (d) mixture density of the true covariate density

5. Simulation study

We now report results from a simulation study based on the three examples described in Section 4.

5.1. Comparison between mixture and standard Gaussian priors

The results corresponding to the bimod and the skew cases are presented in Table 1 and Table 2 respectively. Three different values of the measurement error precision τ were chosen in each case, which correspond to ratios R of $\frac{1}{4}$, $\frac{1}{2}$ and 1 between the measurement error variance and the variance of X. Thus in the worst case ($R = 1$) the size of the noise is as large as the underlying variability which is a very unfavourable situation. Each row in Tables 1 and 2 corresponds to 50 repetitions of the same simulation set-up. To study simultaneously the influence of misspecifying the prior for X and the potential improvement brought by using a flexible mixture prior for X, we report estimates for β_1 and τ obtained on the same 50 simulated data sets using both the mixture model as described in Section 2.3 and a simple Gaussian prior (which corresponds to the mixture with fixed $k = 1$). We do not report posterior distributions for α_0 or α_1; these were concentrated close to 0 and 1 respectively. We also report a 'bench-mark estimate' which corresponds to the situation of no measurement error, i.e. where the analysis is carried out by using the true simulated values of X_i. This allows us to calculate the mean-square error (MSE) ratio between the estimates of β_1 and τ obtained under a measurement error model and the bench-mark estimates. This MSE ratio quantifies the loss of precision due to using surrogates

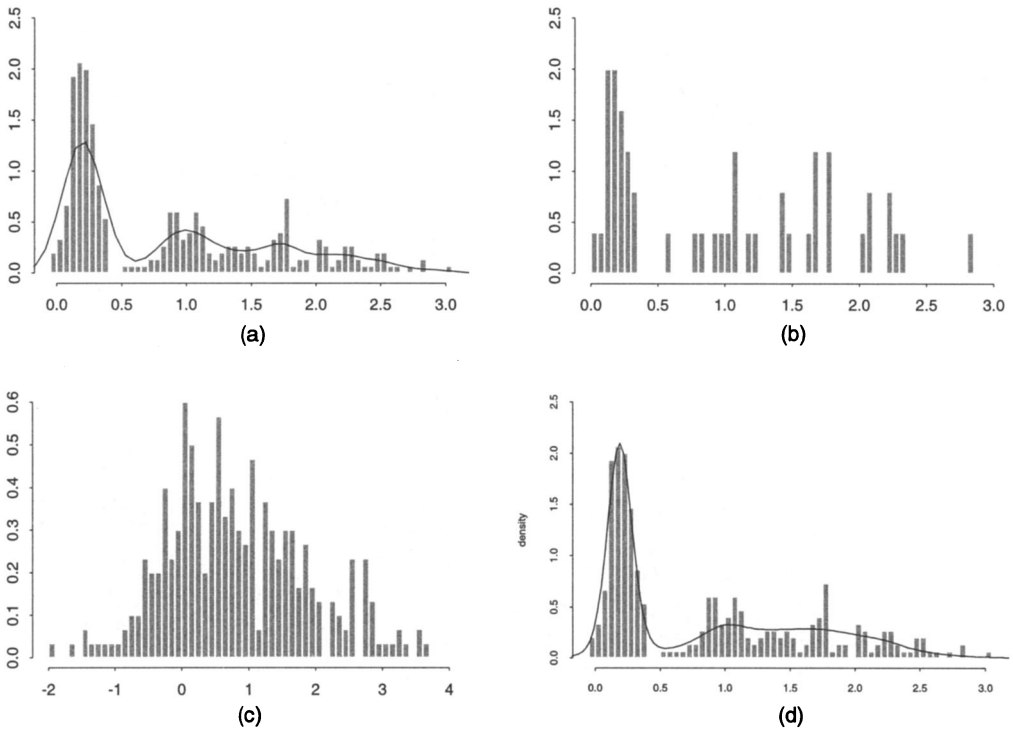

Fig. 3. Skew case—simulated data set (all subjects and validation group) and a comparison with the posterior density estimate given by the mixture model based only on the validation, surrogate and outcome data: (a) true covariate X—all subjects ($n = 300$); (b) true covariate X—validation group ($n = 50$); (c) surrogate U—all subjects ($n = 300$); (d) mixture density of the true covariate density

rather than the true values of the covariate. Of course, many factors influence the MSE ratio, in particular the size of τ and that of the validation group.

In Table 1, we see good performance of the mixture model with the MSE ratio for β_1 varying from 1.8 to 2.9 for the worst case ($\tau = 0.34$ and $R = 1$). The deterioration of the performance when the exposure prior is misspecified is clearly shown, with the MSE ratio for β_1 roughly 1.5 times larger. Similar remarks can be made for τ. The performances of the two middle cases of measurement error ($\tau = 1.34$ and $R = \frac{1}{4}$, and $\tau = 0.67$ and $R = \frac{1}{2}$) are fairly similar, whereas there is a marked deterioration between $\tau = 0.67$ and $R = \frac{1}{2}$, and $\tau = 0.34$ and $R = 1$, which is a particularly bad scenario.

As expected the results for the skew case (Table 2) are not as good as those for the bimod case. The estimation of β_1 in the case of large measurement error ($\tau = 0.9$ and $R = 1$) is problematic with an MSE ratio that is eight times larger than for the bench-mark estimates. But in the other two cases the performance clearly improves with MSE ratios varying between 3 and 2. Again, misspecifying the prior as a standard Gaussian distribution leads to a clear deterioration in the performance.

5.2. Comparison with nonparametric maximum likelihood

Our aim here is to try to compare the performance of the Bayesian mixture model with the NPML approach developed by Roeder *et al.* (1996). Thus, we reproduce the same simulation

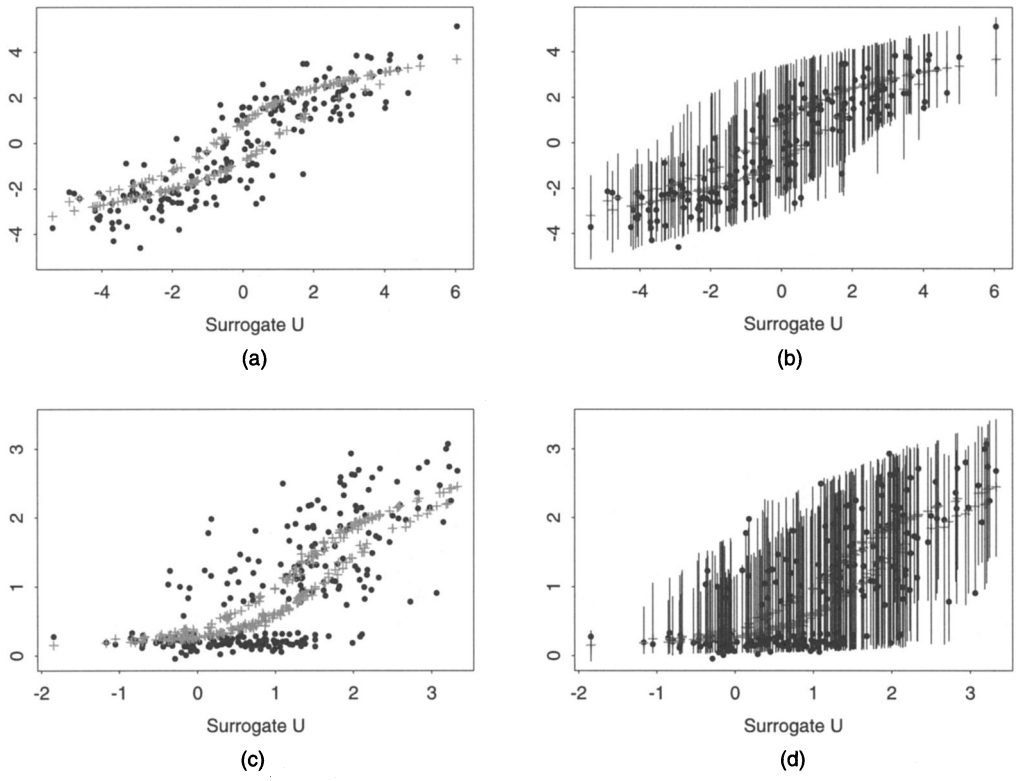

Fig. 4. Comparison of the true values of X and their posterior estimates for (a), (b) bimod and (c), (d) skew cases (the plots do not display values for the validation groups): (a), (c) posterior estimate (+) and true X; (b), (d) superimposed 95% posterior variability

set-up as in Roeder *et al.* (1996) concerning the prior distribution that is used to simulate X and the measurement error model. The simulated distribution of X is log-normal and the error structure is multiplicative. But, as our model was developed in the framework of a prospective study and not a case–control study, a strict comparison is not possible. To keep the proportion of cases fairly comparable with those of Roeder *et al.* (1996), we have chosen in our simulations a high base-line risk $\beta_0 = -0.7$ which, together with the range of values of X, ensures that we obtain approximately equal numbers of cases and non-cases. Thus, our set-up is as detailed for the example RCL described in Section 4.1. Two sets of simulations are presented in Tables 3 and 4.

In Table 3, the assumption is that the measurement variance is known and that there is no validation group. Model (5) for the Bayesian analysis is thus modified accordingly and the values of α_0 and α_1 are assumed known. Two study sizes are investigated, $n = 80$ and $n = 240$. In Table 3 are also indicated the MSE ratios given by Roeder *et al.* (1996) in the corresponding cases. We find a similar MSE ratio between the mixture and the NPML model. This is a useful point of comparison.

Next, we investigate cases with a validation group, distinguishing as in Roeder *et al.* (1996) a case where the measurement error model is known and a case when it is also estimated. The three values of τ chosen (16, 4 and 1.78) correspond on the log-scale approximatively to ratios R of $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{3}{4}$ between the measurement error variance and the variance of X. Unfortunately, Roeder *et al.* (1996) did not report the MSE ratio with respect to the case of no measurement

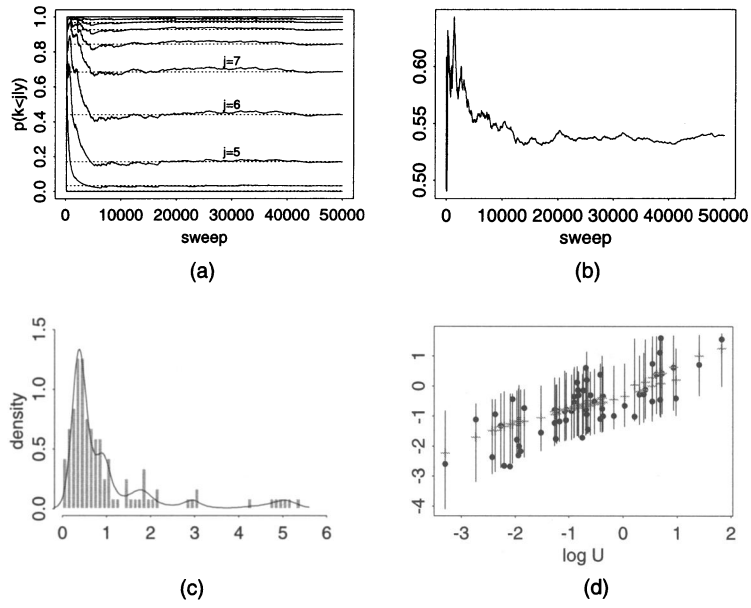

Fig. 5. Performance of the mixture model when X is simulated from a log-normal distribution: (a) k; (b) β_1; (c) mixture estimate of the true covariate density; (d) 95% posterior variability and true X (log-scale)

Table 1. Sensitivity to misspecification: comparison of mixture and Gaussian priors in the bimod case†

Parameter	*Results for the following models with validation group and values of* τ *and R:*					
	$\tau = 1.34\ (R = \frac{1}{4})$		$\tau = 0.67\ (R = \frac{1}{2})$		$\tau = 0.34\ (R = 1)$	
	Mixture	*Gaussian*	*Mixture*	*Gaussian*	*Mixture*	*Gaussian*
$\bar{\beta}_1$	0.63	0.66	0.64	0.68	0.63	0.68
sd(β_1)	0.09	0.10	0.10	0.12	0.12	0.13
MSE ratio (β_1)	1.90	2.60	1.83	3.31	2.94	4.61
$\bar{\tau}$	1.37	1.37	0.67	0.68	0.35	0.35
sd(τ)	0.22	0.23	0.10	0.11	0.05	0.06
MSE ratio (τ)	2.73	4.03	2.83	3.85	2.27	3.28
	Bench-mark estimates‡ (no measurement error)					
$\bar{\beta}_1$	0.61		0.62		0.60	
sd(β_1)	0.08		0.08		0.08	
MSE(β_1)	0.006		0.008		0.005	
$\bar{\tau}$	1.34		0.68		0.34	
MSE(τ)	0.02		0.006		0.001	

†Results are averaged over 50 repetitions.
‡Estimates calculated using the whole data set; true value of $\beta_1 = 0.6$.

Table 2. Sensitivity to misspecification: comparison of mixture and Gaussian priors in the skew case†

Parameter	*Results for the following models with validation group and values of τ and R:*					
	$\tau = 4.0$ $(R = \frac{1}{4})$		$\tau = 1.8$ $(R = \frac{1}{2})$		$\tau = 0.9$ $(R = 1)$	
	Mixture	*Gaussian*	*Mixture*	*Gaussian*	*Mixture*	*Gaussian*
$\bar{\beta}_1$	0.97	1.13	1.06	1.12	1.13	1.13
sd(β_1)	0.20	0.30	0.25	0.31	0.32	0.34
MSE ratio (β_1)	1.93	5.82	3.08	4.40	8.38	9.69
$\bar{\tau}$	3.37	2.24	1.66	1.44	0.85	0.82
sd(τ)	0.52	0.57	0.24	0.27	0.11	0.12
MSE ratio (τ)	5.00	26.42	2.82	7.09	2.30	2.91
	Bench-mark estimates‡					
$\bar{\beta}_1$	0.92		0.94		0.94	
sd(β_1)	0.16		0.16		0.16	
MSE(β_1)	0.021		0.030		0.022	
$\bar{\tau}$	3.91		1.81		0.90	
MSE(τ)	0.120		0.023		0.006	

†Results are averaged over 50 repetitions.
‡Estimates calculated using the whole data set; true value of $\beta_1 = 0.9$.

Table 3. Comparison of mixture and NPML model performance in a case where there is no validation group and the measurement error parameters are known†

Parameter	*Results for the measurement error without validation group and the following values of n, τ and R:*			
	$n = 80$		$n = 240$	
	$\tau = 16$ $(R = \frac{1}{4})$	$\tau = 4$ $(R = \frac{1}{2})$	$\tau = 16$ $(R = \frac{1}{4})$	$\tau = 4$ $(R = \frac{1}{2})$
Mixture model				
$\bar{\beta}_1$	0.57	0.69	0.49	0.56
sd(β_1)	0.28	0.36	0.15	0.17
MSE ratio (β_1)	1.17	1.80	1.45	2.02
NPML model‡				
MSE ratio (β_1)	1.28	1.82	1.83	2.16
Bench-mark estimates§				
$\bar{\beta}_1$	0.53	0.58	0.47	0.54
sd(β_1)	0.26	0.27	0.14	0.15
MSE(β_1)	0.05	0.12	0.02	0.02

†Results are averaged over 50 repetitions.
‡Estimates reported in Roeder *et al.* (1996), Table 1.
§Estimates calculated by using the whole data set; true value of $\beta_1 = 0.5$.

Table 4. Performance of the mixture model in a similar set-up to that described in Roeder *et al.* (1996)†

Parameter	*Results for the known measurement error model and the following values of* τ *and R*‡*:*			*Results for the unknown measurement error model and the following values of* τ *and R*‡*:*		
	$\tau = 16$ $(R = \frac{1}{4})$	$\tau = 4$ $(R = \frac{1}{2})$	$\tau = 1.78$ $(R = \frac{3}{4})$	$\tau = 16$ $(R = \frac{1}{4})$	$\tau = 4$ $(R = \frac{1}{2})$	$\tau = 1.78$ $(R = \frac{3}{4})$
$\bar{\beta}_1$	0.54	0.53	0.58	0.56	0.55	0.58
sd(β_1)	0.16	0.17	0.19	0.16	0.19	0.21
MSE ratio (β_1)	1.15	1.46	2.03	1.60	2.20	2.04
	Bench-mark estimates§					
$\bar{\beta}_1$	0.52	0.50	0.55	0.52	0.51	0.53
sd(β_1)	0.15	0.14	0.15	0.14	0.14	0.14
MSE(β_1)	0.02	0.03	0.02	0.02	0.02	0.03

†Results are averaged over 50 repetitions.
‡Model with validation group, $n_0 = 180$ and $n_1 = 60$.
§Estimates calculated using the whole data set; true value of $\beta_1 = 0.5$.

error for the NPML method, so we cannot compare those between the mixture and NPML as in Table 3.

In the known measurement parameters case, we see a slightly better performance of the mixture model than in the bimod case with MSE ratios varying between 1.1 and 1.5 for ratios R of $\frac{1}{4}$ and $\frac{1}{2}$. In the unknown measurement error model, the performance is as expected a little worse, but the MSE ratios stay close to 2. These ratios are well in keeping with those reported in Table 3. Thus, even in this case of a mixture approximation of a log-normal prior, the mixture model gives a satisfactory performance for estimating the regression coefficient of interest.

6. Application to study of cholesterol and coronary heart disease

In this example, we reanalyse a data set concerning the risk of coronary heart disease as a function of blood cholesterol level, which was discussed and analysed using NPML by Roeder *et al.* (1996); the data were kindly supplied to us by R. Carroll. Subjects are considered as cases of coronary heart disease if they had a previous heart attack, a history of *angina pectoris* or an abnormal exercise electrocardiogram. They are all non-smokers. The covariate of interest is low density lipoprotein cholesterol LDL, which is one of the components of total cholesterol TC. The direct measurement of this variable is more costly and technically demanding than that of the TC level. The question investigated here is thus whether TC could be a useful surrogate for LDL. We refer to Roeder *et al.* (1996) for more details concerning the data set.

The sample analysed consists of 113 cases and 143 controls, for whom full data, i.e. coronary heart disease status and LDL and TC levels, are available. As in Roeder *et al.* (1996), we shall refer to coronary heart disease status as Y, LDL/100 as X and TC/100 as U. From the complete data set, we selected at random 32 cases and 40 controls to use in the validation set. For comparability with Roeder *et al.* (1996), we have analysed these data as if they had come from a cohort study, i.e. by using the prospective likelihood. Conditions for the equivalence of Bayesian analysis using prospective and retrospective likelihoods are discussed in Seaman *et al.* (2000).

The Bayesian analysis was carried out along the same lines as described earlier. The skewness in the LDL distribution was fitted by a mixture with mostly two components, $p(k = 2|y) = 0.53$. Recovery of the estimated value of β_1 based on the complete data set was good: the logistic regression analysis carried out between Y and X on the full data set gave a value of the posterior mean of β_1 equal to 0.66 with the posterior standard deviation equal to 0.34, whereas that from using the surrogate U for most cases and controls and only the true X on the validation group gave values of 0.64 and 0.37 respectively. Note that, as expected, the posterior standard deviation is increased. As a point of comparison, the logistic regression analysis ignoring measurement error, carried out between Y and U on the whole data set, leads to an attenuated value for β_1 of 0.55. If the analysis had been carried out only on the 72 subjects of the validation set, the results would also be quite poor, with $\beta_1 = 1.35$ and a posterior standard deviation of 0.6. Thus, we see the additional information obtained by combining the information from the validation group with that of the surrogate on a larger group. In Fig. 6, the good recovery of the LDL density obtained by the mixture model is shown alongside the density of TC. Note the contrasts in shape between the two densities.

7. Discussion

The model that we are proposing illustrates the flexibility of the Bayesian modelling approach which exploits conditional independence structure. Thus, we have been able to link two complex models: that relating to the measurement error problem and that concerning the underlying structure of the latent unobserved variable, models which were both built separately.

In structural measurement error problems, there is a general concern about the parametric specification of the unknown covariates. Recent work has approched this problem via NPML methods that have been shown to perform well (Roeder *et al.*, 1996; Schafer, 2001; Aitkin and Rocci, 2002). We have addressed this concern in a Bayesian hierarchical framework. This framework offers many other possibilities for building in additional complexity of epidemiological data sets, like informative missingness or subject-specific random effects. In this framework, the mixture model with a variable number of components that we have developed for the unknown covariate prior was a natural candidate to overcome potential misspecifications.

We have shown that it is feasible to use it in measurement error problems with validation data and that it has a good performance, avoiding attenuation of the regression coefficients of interest. Mixture priors could also be envisaged in other measurement error situations where information on the measurement model parameters is built in the design, e.g. in designs with repeated measures of an unbiased instrument in a subgroup of individuals. But, in the absence

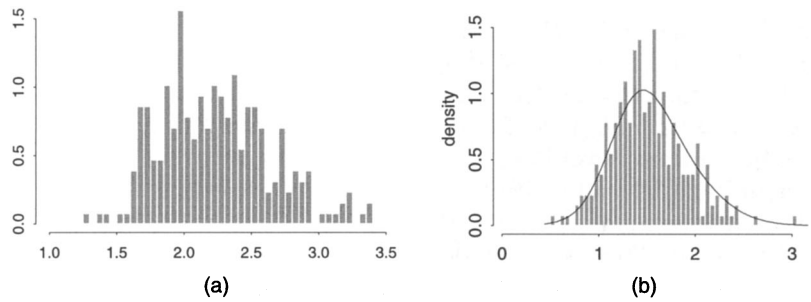

Fig. 6. Coronary heart disease and cholesterol study: (a) TC/100; (b) mixture density estimate of the underlying density of LDL/100

of validation data, it would be inadvisable to use our approach as it would require very strong priors on the mixture parameters to obtain reasonable estimates.

More generally, our mixture approach could be extended to other cases where there is uncertainty about the parametric specification of a latent variable, as in random-effects models. Preliminary work along these lines is reported in Watier *et al.* (1999). In this context, a semiparametric maximum likelihood for generalized linear mixed models has also been used as an alternative approach; see Aitkin (1999).

There could be some concern that we are overparameterizing the model by using a variable number of components. The alternative, as suggested in Carroll *et al.* (1999), would be to fit increasing numbers of components and to judge the improvement sequentially. This is of course reasonable, but our global approach avoids the delicate statistical problem of choosing k, and the performance does not seem to deteriorate with respect to another method, NPML, in the cases where we could make comparisons. Note also the good recovery of the underlying shape of the density of the unknown covariate, based on limited information.

We have thus been able to show how Bayesian measurement error models can make full use of the flexible semiparametric nature of mixture models with an unknown number of components to avoid biases due to a misspecification of the distribution of unobserved covariates.

Acknowledgements

We thank Ray Carroll for kindly providing the data set on coronary heart disease and cholesterol. We acknowledge the financial support of the Institut National de la Santé et de la Recherche Médicale, the European Science Foundation programme on 'Highly structured stochastic systems' and an Engineering and Physical Sciences Research Council Visiting Fellowship (for SR).

References

Aitkin, M. (1999) A general maximum likelihood analysis of variance components in generalised linear models. *Biometrics*, **55**, 117–128.

Aitkin, M. and Rocci, R. (2002) A general maximum likelihood analysis of measurement error in generalised linear models. *Statist. Comput.*, **12**, 163–174.

Besag, J., Green, P. J., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems (with discussion). *Statist. Sci.*, **10**, 3–66.

Carroll, R. J., Gail, M. H. and Lubin, J. H. (1993) Case-control studies with errors in covariates. *J. Am. Statist. Ass.*, **88**, 185–199.

Carroll, R. J., Maca, J. D. and Ruppert, D. (1999) Nonparametric regression in the presence of measurement error. *Biometrika*, **86**, 541–554.

Carroll, R. J., Roeder, K. and Wasserman, L. (1999) Flexible parametric measurement error models. *Biometrics*, **55**, 44–54.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995) *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Clayton, D. G. (1992) Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In *Statistical Models for Longitudinal Studies of Health* (eds J. H. Dwyer, M. Feinleib, P. Lippert and H. Hoffmeister), pp. 301–331. New York: Oxford University Press.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds) (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

——(2001) A primer on MCMC. In *Complex Stochastic Systems* (eds O. E. Barndorff-Nielsen, D. R. Cox and C. Kluppelberg). Boca Raton: Chapman and Hall–CRC.

Magder, L. S. and Zeger, S. L. (1996) A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *J. Am. Statist. Ass.*, **91**, 1141–1151.

Mallick, B. K. and Gelfand, A. E. (1996) Semiparametric errors in variables models: a Bayesian approach. *J. Statist. Planng Inf.*, **52**, 307–321.

Müller, P. and Roeder, K. (1997) A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika*, **84**, 523–537.

Richardson, S. and Gilks, W. R. (1993a) A Bayesian approach to measurement error problems in Epidemiology using conditional independence models. *Am. J. Epidem.*, **138**, 430–442.

———(1993b) Conditional independence models for epidemiological studies with covariate measurement error. *Statist. Med.*, **12**, 1703–1722.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Statist. Soc.* B, **59**, 731–792.

Richardson, S. and Leblond, L. (1997) Some comments on misspecification of priors in Bayesian modelling of measurement error problems. *Statist. Med.*, **16**, 203–213.

Roeder, K., Carroll, R. J. and Lindsay, B. G. (1996) A semiparametric mixture approach to case-control studies with errors in covariables. *J. Am. Statist. Ass.*, **91**, 722–732.

Schafer, D. W. (2001) Semiparametric maximum likelihood for measurement error model regression. *Biometrics*, **57**, 53–61.

Seaman, S. R., Clayton, D. G. and Richardson, S. (2000) Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. *Technical Report*. Institut National de la Santé et de la Recherche Médicale, Villejuif.

Seaman, S. R. and Richardson, S. (2001) Bayesian analysis of case-control studies with categorical covariates. *Biometrika*, **88**, 1073–1088.

Stephens, D. A. and Dellaportas, P. (1992) Bayesian analysis of generalised linear models with covariate measurement error. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press.

Watier, L., Richardson, S. and Green, P. J. (1999) Using gaussian mixtures with unknown number of components for mixed model estimation. In *Proc. 14th Int. Wrkshp Statistical Modelling, Graz* (eds H. Friedl, A. Berghold and G. Kauermann).