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

Sylvia Richardson* Laurent Leblond, Isabelle Jaussent †
Imperial College, London, INSERM, France

Peter J. Green‡
University of Bristol, UK.

June 2001

Summary

This paper focuses on Bayesian 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 normals 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 encountered in epidemiological studies: logistic regression between unknown covariates and outcome, with normal or lognormal 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 ML method of Roeder, Carroll and Lindsay (1996). Finally, the methodology is illustrated on a data set on coronary heart disease and blood level of cholesterol.

Some key words: Bayesian modelling, Epidemiological studies, Errors in variables, Finite mixture distributions, Heterogeneity, Logistic regression, MCMC algorithms, Misspecification.

---

*Imperial College, Department of Epidemiology and Public health, London
Corresponding author: S. Richardson, Email: `sylvia.richardson@ic.ac.uk`
†INSERM U170, 16 avenue Paul Vaillant Couturier, Villejuif, France.
‡Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK.

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 using a mixture of normals, in order to reduce 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-variable 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, 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, Ruppert and Stefanski, 1995 and the references therein).

Bayesian analysis of measurement error problems has been developed from the seminal work of Clayton (1991). 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* $p(X|\pi)$ 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, 1993).

At a second stage, functional forms for the distributions involved in the submodels have to 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 lognormal error models corresponding to additive or multiplicative errors, have been used in epidemiology to model the association between (dichotomous) outcomes, covariates and their surrogates.

On the other hand, in the common situation of observational studies, there are many reasons to suspect heterogeneity in the distribution of the unknown X among the studied population. Exposure variables often have a skew distribution with a high peak corresponding to moderate exposure for the majority of the population and long tail or a second lower peak corresponding to larger exposure for a small fraction of the population. These have traditionally been fitted by lognormal 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 in order 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, that can be misleading if the validation group is small or the measurement error large. Since regression results are known to be sensitive to choice of a particular parametric shape for $p(X|\pi)$, a natural question is how to model $p(X|\pi)$ flexibly, avoiding such choices but allowing the data to speak for themselves.

In the measurement error context, this question has been actively pursued and has led to different lines of development. Carroll, Gail and Lubin (1993) have proposed a pseudolikelihood method, using validation data. Roeder, Carroll and Lindsay (1996) combine a parametric disease model and nonparametric mixtures for $p(X)$ and carry out the estimation via nonparametric maximum likelihood (NPML) algorithms. A similar line is also followed by Schafer (2001), who emphasises the general applicability of NPML to different measurement error contexts. Müller and Roeder (1997) have developed Dirichlet process priors for the joint model of X and U.

Mixture models with variable number 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 recently been developed in a Bayesian context by

Richardson and Green (1997) using novel 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 normals 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 ML method of Roeder, Carroll and Lindsay (1996).

Using mixtures of normals to increase robustness to model misspecification for $p(X|\pi)$ is in line with recent developments. Mixtures with a fixed number of components have been used by Carroll, Roeder and Wasserman (1999) for linear error-in-variable models. Mixtures of normals have also been implemented by Carroll, Maca and Ruppert (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 impact of misspecifying the parametric model for the unknown covariates and the improvement brought by using mixture models, and showing a limited comparison of our approach with the NPML method of Roeder, Carroll and Lindsay (1996). An application to a data set on coronary heart disease and blood level of cholesterol 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 (1993), 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. 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 (1993), the general structure follows from the formulation of local submodels between the components using conditional independence assumptions and information on the design used to evaluate the measurement process.

Submodels:

- $p(Y_i|X_i, C_i, \beta)$ regression model
- $p(U_i|X_i, \lambda)$ (classical) measurement model
- $p(X_i|\pi)$ 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:

$$p(\beta)p(\lambda)p(\pi) \prod_i p(X_i|\pi) \prod_i p(U_i|X_i, \lambda) \prod_i p(Y_i|X_i, C_i, \beta) \tag{1}$$

where $p(\beta), p(\lambda)$ and $p(\pi)$ are the prior distributions for the parameters of the 3 submodels.

In expression (1), the contribution of the validation group is not formally separated out, because in a Bayesian formulation, the only distinction is that X_i is observed and not latent for this group of indices. 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 contributing to the estimation of β and π.

2.2 Mixture model

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

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

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, the model given in (2) is formulated equivalently as:

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

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

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

$$p(k, \theta, w, z, X) = p(k)p(\theta|k)p(w|k)p(z|w, k)p(X|\theta, z). \tag{3}$$

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).

Let us stress that treating k, the number of components, as being unknown, and integrating over its posterior distribution when estimating regresion parameters of interest enhances the adaptivity of the mixture to heterogeneity in the underlying distribution of the X's. This is different to other approachess where a fixed value for k is usually estimated using a BIC approximation, as in Carroll, Roeder and Wasserman (1999).

2.3 Joint model and parametric specifications

Combining (1) and (3) leads to the following expression for the joint distribution of the measurement error problem:

$$p(k)p(\theta|k)p(w|k)p(\beta)p(\lambda)\prod_i p(z_i|w,k)p(X_i|\theta_{z_i},z_i)\prod_i p(U_i|X_i,\lambda)\prod_i p(Y_i|X_i,C_i,\beta). \qquad (4)$$

Note that we are fully benefitting here from the flexibility 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 the following, we shall concentrate on the widely entertained logistic regression with normal or lognormal 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:

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

The dichotomous observations will be referred to as cases ($Y = 1$), and non-cases ($Y = 0$). Note that this regression model corresponds to a prospective formulation, *i.e.* where the likelihood considered is that of disease given exposure. For the measurement model, we investigate gaussian errors on an additive or multiplicative scale, either

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

or

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

For the mixture model, we consider a mixture of gaussian distributions which can approximate any continous 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 will 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 measurement error precision, τ, 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 this former paper, the setting of the normal priors for the component means μ_j as well as the hierarchical gamma priors for the inverse 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: $\widehat{X_i}$ from the observed U_i. Finally, we define "initial values" for the X_i by perturbing the $\widehat{X_i}$ by a Gaussian $(0, \hat{\tau}^{-1})$ 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 $\widehat{X_i}$ to define this range is too restrictive. Once the notional range R is 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 MCMC 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, Markov Chain Monte Carlo (MCMC) methods play now a major role in these computations; see, for example, Besag *et. al.*(1995), Gilks *et. al.*(1996), 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) update of β, α, τ

(b) update 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) update of w, z, θ *conditional on* k

(d) update of k and consequently of the relevant mixture parameters

For (a) the moves are conventional. We used a random walk Metropolis step where the acceptance probabilities are calculated using the expression (4) of the joint distribution. Gibbs sampling was not straightforward due to the logistic form of the regression model.

For (b), this involves updating one at a time all the unknown covariates X_i. Since this is done *conditional on* z_i, this is again straightforward using a random walk Metropolis, 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 one and preserving some moment conditions are proposed, then accepted following a Metropolis ratio. Note that in contrast to 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 need 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 initialisation 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 3 simulated data sets. The design of the three simulated data sets will also be used in the simulation study reported in Section 5. For the first two, we simulate X from a mixture, while for the third, we follow the simulation set-up of Roeder, Carroll and Lindsay (1996) where X has a lognormal 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 $\tau = 0.67$, $\text{logit}P(Y|X) = \beta_0 + \beta_1 X$ with $\beta_0 = 0.5, \beta_1 = 0.6$. The distribution of X is simulated from a well separated symmetric Bimodal mixture: $0.5N(-2.0, 1.0) + 0.5N(2.0, 1.0)$ (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 $\tau = 1.8$, $\text{logit}P(Y|X) = \beta_0 + \beta_1 X$ with $\beta_0 = -0.8, \beta_1 = 0.9$. The distribution of X is simulated from an asymmetric normal mixture : $0.5N(0.19, (0.08)^2) + 0.2N(1.05, (0.2)^2) + 0.3N(2.0, (0.48)^2)$

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 one half of the variance of X (which is equal to 3 and 1.1 for "Bimod" and "Skew" respectively). The value of β_0 ensures that the data sets contain approximatively half cases and half non-cases. Note that there is more information in “Bimod” since the relative size of the validation group to the main study is larger than for “Skew”. 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 from one used by Roeder, Carroll and Lindsay (1996) and is referred to as “RCL”. There is nonetheless a major difference with Roeder, Carroll and Lindsay's work, in that we treat the study as prospective rather than as a retrospective case-control set-up. Implementation of case-control studies in a Bayesian framework are not straightforward, see Seaman and Richardson (2000).

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

Irrespectively of the model used to generate the $X's$, we implement estimation of the joint model equation (4). In particular, for the RCL example, the normal mixture will be used to approximate a lognormal distribution.

4.2 Performance

The MCMC algorithm was run for 50 000 sweeps. As will be seen from the graphical assessment given by the plots in Figure 1, in “Bimod”, convergence to stable cumulative averages for the regression parameter β_1 (posterior mean = 0.63, posterior s.d. = 0.1) or the precison τ (plotted every 20 sweeps) is obtained after 30 000 iterations, giving values which are close to the simulated true ones. Similarly, the cumulative occupancy fractions for k, the number of components in the mixture, have stabilised also after 30 000 sweeps and gives a clear indication that there is posterior support for $k = 2$ in the Bimodal mixture.

Convergence is somewhat slower in the “Skew” example, but nevertheless parameter estimates stabilise after 40 000 sweeps (Figure 1). For this particular data set, we see support for 2 or 3 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 s.d. = 0.2) than in the Bimod case.

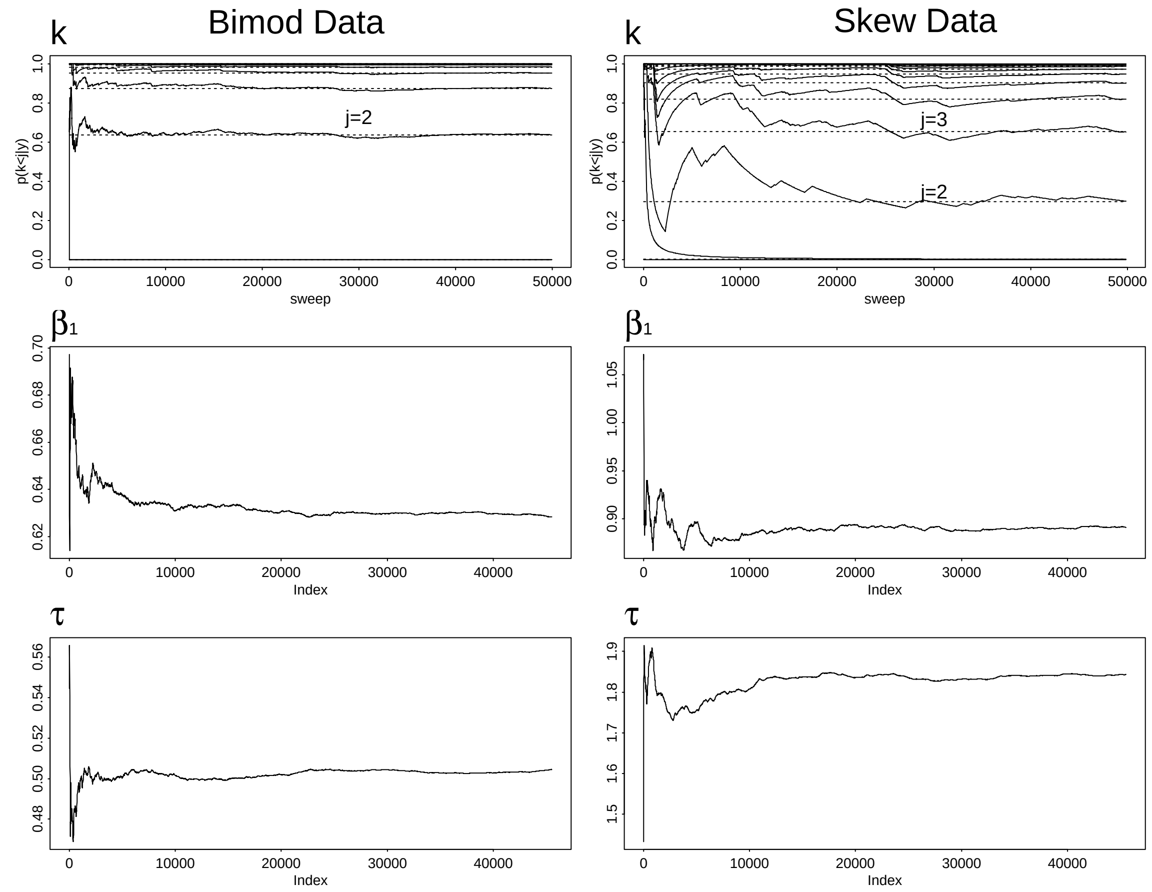

Figure 1: Graphical assessment of the convergence of the MCMC algorithm for "Bimod" and "Skew": cumulative occupancy fractions for k (top panels), cumulative mean of β_1 and τ

In Figures 2 and 3 are displayed the simulated distribution of X in the whole study (top left panel), that observed in the validation group (top right panel) and its posterior density estimation given by the Bayesian implementation using the full model described in equation (4) (bottom right panel). Indeed, from the output of the chain, one obtains an estimate of the predictive densities of the mixture conditional on k, by averaging across the MCMC run conditional 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 line), by averaging across values of k. One sees a good recovery of the true underlying density in both cases, in comparison to the blurred shape of the distribution of U (bottom left panel), even though there was only a small fraction of the sample in the validation group.

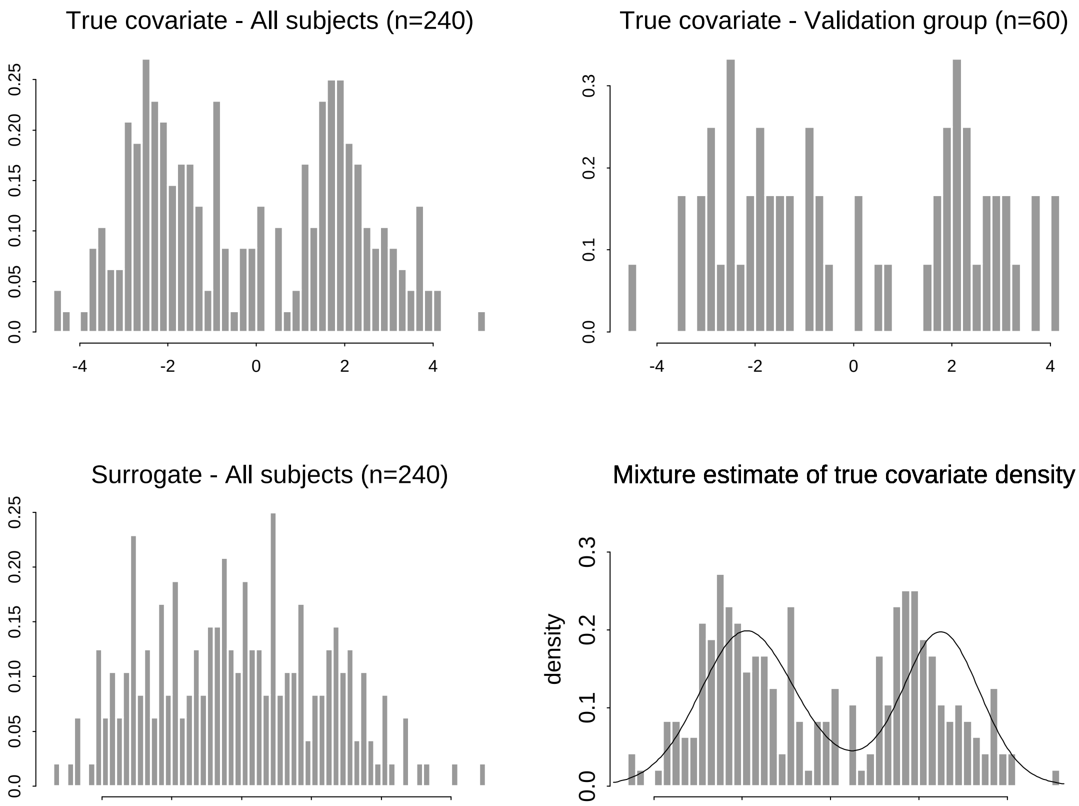

Figure 2: Simulated data set (all subjects and validation group) and comparison with the posterior density estimate given by the mixture model in the "Bimod"case

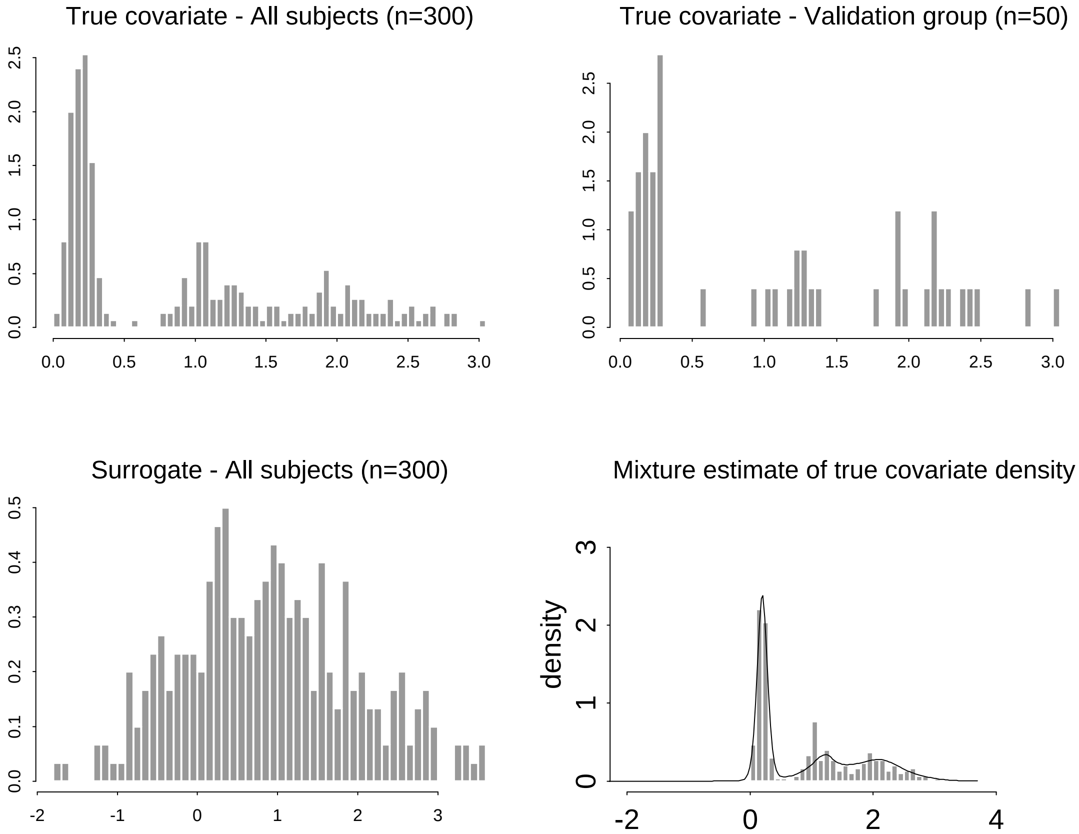

Figure 3: Simulated data set (all subjects and validation group) and comparison with the posterior density estimate given by the mixture model in the “Skew”case

It is also interesting to investigate the performance of the model with respect to the latent variables by comparing the true X_i to the distribution of the X_i simulated through the algorithm. This is done in Figure 4, which displays on the left-hand side, true X_i (dots) versus their posterior mean (crosses) ordered by values of the surrogate U, while on the right hand-side an interval of 95% posterior variability is superimposed. For both data sets, one can see a classical shrinkage effect of the posterior means. This is more noticeable for the skew case. Indeed, due to 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 true X_i are included in the the posterior variability interval.

The ability of the mixture model to estimate the regression coefficients as well as to characterise the underlying density of the true $X's$ is also apparent in the RCL example (Figure 5). There, the mixture uses between 4 and 7 components to approximate the lognormal shape. Again, estimates stabilise quickly after 30 000 sweeps. On the bottom right panel of Figure 5, one sees that the variability (on the log scale) of the posterior estimates matches the true values well.

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 $R = 1/4, 1/2$ 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 line in the Tables corresponds to 50 repetitions of the same simulation set-up. In order to study simultaneously the influence of misspecifying the prior for X and the potential improvement

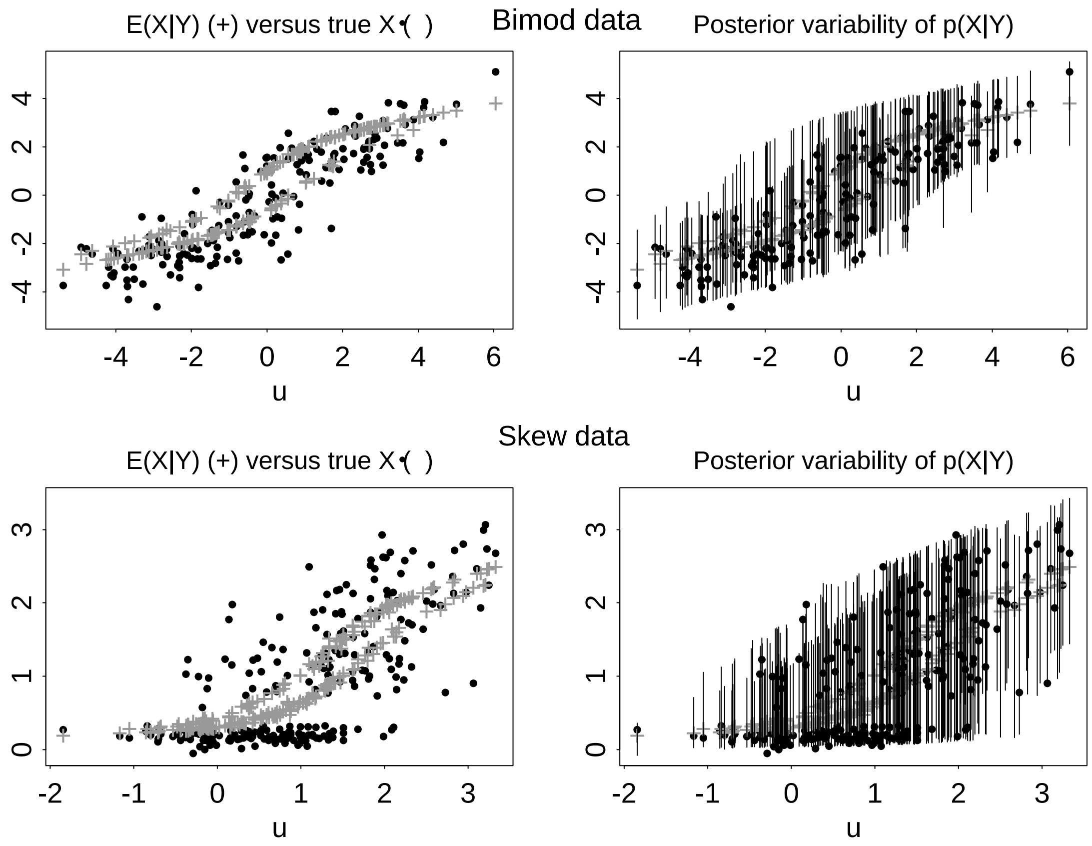

Figure 4: Comparison of the true values of X and their posterior estimates for the “Bimod” and “Skew”

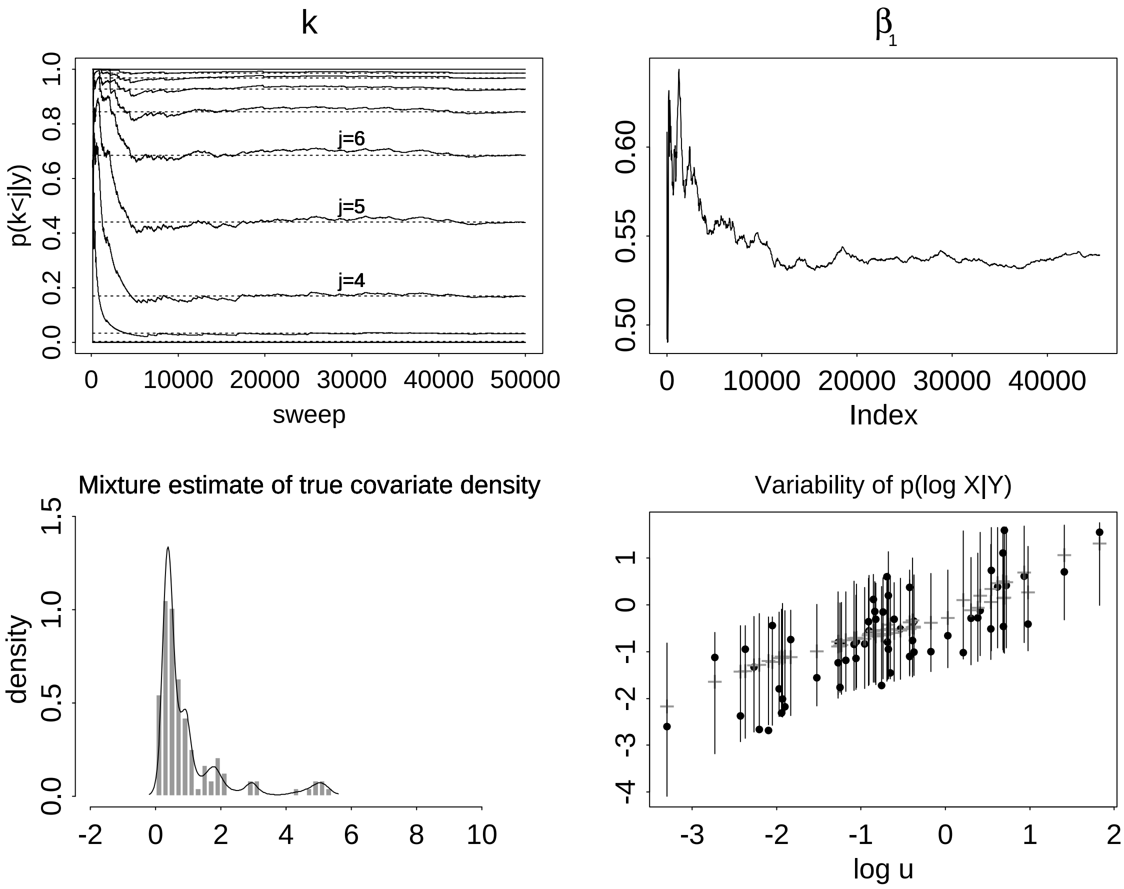

Figure 5: Performance of the mixture model when X is simulated from a lognormal distribution. The bottom right panel displays posterior variability of X on the log scale

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 also report a "benchmark estimate" which corresponds to the situation of no measurement error, *i.e.* where the analysis is carried out using the true simulated values of X_i. This allows to calculate the ratio of Mean Square Error (MSE ratio) between the estimates of β_1 and τ obtained under a measurement error model and the benchmark ones. This MSE ratio quantifies the loss of precision due to using surrogates 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 MSE ratio for β_1 varying from 1.8 to 2.9 in the worst case ($\tau = 0.34, R = 1$). The deterioration of the performance when the exposure prior is misspecified is clearly shown, with MSE ratio for β_1 roughly 1.5 times larger. Similar remarks can be made for τ. Note that the performances of the two middle cases of measurement error ($\tau = 1.34, R = 1/4$ and $\tau = 0.67, R = 1/2$) are fairly similar, while there is a marked deterioration between $\tau = 0.67, R = 1/2$ and $\tau = 0.34, R = 1$, which is a particularly bad scenario.

As expected the results for the Skew case (Table 2) are not as good as those for "Bimod". The estimation of β_1 in the case of large measurement error ($\tau = 0.9, R = 1$) is problematic with an MSE ratio 8 times larger than for the benchmark estimates. But in the other two cases, the performance clearly improves with MSE ratio varying between 3 and 2. Again, misspecifying the prior as a standard Gaussian leads to clear deterioration of the performance.

5.2 Comparison with NPML

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 set-up as in Roeder *et. al.*, concerning the prior distribution used to simulate X and the measurement error model. The simulated distribution of X is lognormal and the

Table 1: Sensitivity to misspecification : Comparison of mixture and Gaussian priors in the "Bimod" case

	$\tau = 1.34$ m.e. with validation group		$\tau = 0.67$ m.e. with validation group		$\tau = 0.34$ m.e. with validation group	
	MIX	NORMAL	MIX	NORMAL	MIX	NORMAL
$\overline{\beta_1}$	0.63	0.66	0.64	0.68	0.63	0.68
$sd(\beta_1)$	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
$\overline{\tau}$	1.37	1.37	0.67	0.68	0.35	0.35
$sd(\tau)$	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
	benchmark estimates+ (no measurement error)		benchmark estimates+ (no measurement error)		benchmark estimates+ (no measurement error)	
$\overline{\beta_1}$	0.61		0.62		0.60	
$sd(\beta_1)$	0.08		0.08		0.08	
$MSE(\beta_1)$	0.006		0.008		0.005	
$\overline{\tau}$	1.34		0.68		0.34	
$MSE(\tau)$	0.02		0.006		0.001	

+ 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

	$\tau = 4.0$ m.e. with validation group		$\tau = 1.8$ m.e. with validation group		$\tau = 0.9$ m.e. with validation group	
	MIX	NORMAL	MIX	NORMAL	MIX	NORMAL
$\overline{\beta_1}$	0.97	1.13	1.06	1.12	1.13	1.13
$sd(\beta_1)$	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
$\overline{\tau}$	3.37	2.24	1.66	1.44	0.85	0.82
$sd(\tau)$	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
	benchmark estimates+ (using whole data set)		benchmark estimates+ (using whole data set)		benchmark estimates+ (using whole data set)	
$\overline{\beta_1}$	0.92		0.94		0.94	
$sd(\beta_1)$	0.16		0.16		0.16	
$MSE(\beta_1)$	0.021		0.030		0.022	
$\overline{\tau}$	3.91		1.81		0.90	
$MSE(\tau)$	0.120		0.023		0.006	

+ estimates calculated using the whole data set, true value of $\beta_1 = 0.9$

Table 3: Comparison of mixture and NPML performance in a case where there is no validation group and the measurement error is known

		M.E. without validation group			
		$n = 80$		$n = 240$	
		$\tau = 16$	$\tau = 4$	$\tau = 16$	$\tau = 4$
MIX	$\overline{\beta_1}$	0.57	0.69	0.49	0.56
	$sd(\beta_1)$	0.28	0.36	0.15	0.17
	MSE ratio (β_1)	1.17	1.80	1.45	2.02
NPML*	MSE ratio (β_1)	1.28	1.82	1.83	2.16
Benchmark estimates +	$\overline{\beta_1}$	0.53	0.58	0.47	0.54
	$sd(\beta_1)$	0.26	0.27	0.14	0.15
	$MSE(\beta_1)$	0.05	0.12	0.02	0.02

* estimates reported in Roeder *et. al.*(1996), Table 1

\+ estimates calculated using the whole data set, true value of $\beta_1 = 0.5$

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

	Known Measurement Error			Unknown Measurement Error		
	$n_0 = 180, n_1 = 60$ m.e. with validation group			$n_0 = 180, n_1 = 60$ m.e. with validation group		
	$\tau = 16$	$\tau = 4$	$\tau = 1.78$	$\tau = 16$	$\tau = 4$	$\tau = 1.78$
$\overline{\beta_1}$	0.54	0.53	0.58	0.56	0.55	0.58
$sd(\beta_1)$	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
	benchmark estimates+			benchmark estimates+		
$\overline{\beta_1}$	0.52	0.50	0.55	0.52	0.51	0.53
$sd(\beta_1)$	0.15	0.14	0.15	0.14	0.14	0.14
$MSE(\beta_1)$	0.02	0.03	0.02	0.02	0.02	0.03

\+ estimates calculated using the whole data set, true value of $\beta_1 = 0.5$

error structure is multiplicative. But our model having been developed in the framework of a prospective study and not a case-control study, a strict comparison is not possible. In order to keep the proportion of cases fairly comparable with Roeder *et. al.*, we have chosen in our simulations a high baseline 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 the Tables 3 and 4.

In the first table (Table 3), the assumption is that the measurement variance is known and that there is no validation group. The model and graph for the Bayesian analysis are thus modified accordingly. Two study sizes are investigated, $n = 80$ and $n = 240$. In the table are also indicated the MSE ratio given by Roeder *et. al.*(1996) in the corresponding cases. We find 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 variance 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 1/4, 1/2 and 3/4 between the measurement error variance and the variance of X. Unfortunately, Roeder *et. al.* do not report the MSE ratio with respect to the no measurement error case for the NPML method, so we cannot compare those between mixture and NPML as in Table 3.

In the known measurement case, we see slightly better performance of the mixture model than in the "Bimod" case with MSE ratio varying between 1.1 and 1.5 for the ratios $R = 1/4$ to 1/2. In the unknown measurement error case, the performance is expectedly 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 mixture approximation of a lognormal prior, the mixture model gives 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 (CHD) as a function of blood cholesterol level, that was discussed and analysed using NPML by Roeder *et. al.*(1996), the data having been kindly communicated to us by R. Carroll. Subjects are considered as cases of CHD if they had a previous heart attack, history of angina pectoris or an abnormal exercise electrocardiogram. They are all non-smokers. The covariate of interest is low-density lipoprotein (LDL) cholesterol, which is one of the components of total cholesterol (TC). Direct measurement of this variable is more costly and technically demanding than that of 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, that is CHD status, LDL and TC levels, are available. In what follows, we will refer to CHD 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.*, we have analysed this data as if coming from a cohort study, *i.e.* using the prospective likelihood. Conditions for the equivalence of Bayesian analysis using prospective and retrospective likelihoods are discussed in Seaman, Clayton and Richardson (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 2 components, $p(k = 2|y) = 0.53$. Recovery of the true value of β_1 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 posterior s.d. equal to 0.34, while that 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 s.d. 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 subject of the validation set, the results would also be quite poor, with $\beta_1 = 1.35$ and a posterior s.d. 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 Figure 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.

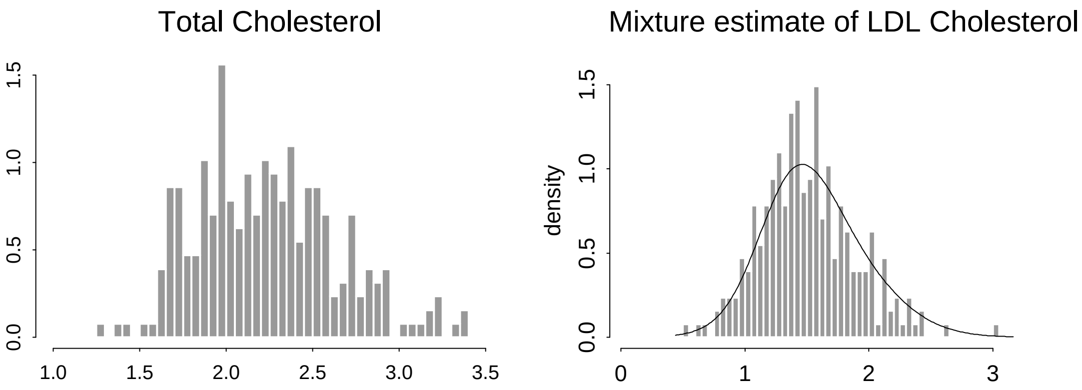

Figure 6: CHD and cholesterol study: TC density (left panel) and mixture density estimate of the underlying LDL density (right panel)

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). 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 variable number of components that we have developed for the unknown covariate prior was a natural candidate to overcome potential misspecification. We have shown that it is feasible to use it in measurement error problems, and that it has a good performance, avoiding attenuation of the regression coefficients of interest. 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 generalised linear mixed models has also been used as an alternative approach, see Aitkin (1999).

There could be some concern that we are overparametrising the model by using a variable number of components. The alternative, as suggested in Carroll, Roeder and Wasserman (1999) would be to fit increasing number of components and to judge the improvement sequentially. This is of course reasonable, but we feel that our global approach avoids the delicate statistical problem of choosing k, and that the performance does not seem to deteriorate with respect to another method, NPML, in the cases where we were able to 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 in order to avoid biases due to misspecification of the distribution of unobserved covariates.

Acknowledgements

We wish to thank Ray Carroll for kindly providing the data set on CHD and cholesterol. We acknowledge the financial support of INSERM, the ESF programme on Highly Structured Stochastic Systems and an EPSRC visiting fellowship (SR).

References

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

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

Caroll, R. J., Gail, M. H. and Lubin, J. H. (1993) Case-control studies with errors in covariates. *Journal of the American Statistical Association* , **88**, 421, 185-199.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). Measurement error in nonlinear models. London, Chapman & Hall, 1st edition, *Monographs on Statistics and Applied Probability.*

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

Carroll, R. J., Maca, J. D., Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika* , **86**, 3, 541-554.

Clayton, D. G (1991).

Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In : Dwyer, J.H., Feinleib, M., Lippert, P., et al *eds. Statistical Models for longitudinal studies of health*, New York : Oxford University Press, , 301–331.

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

Green, P. J. (2001) A primer on MCMC. In : *Complex Stochastic Systems*, eds Barndorff-Nielsen O.E., Cox D.R. and Kluppelberg C. Chapman and Hall/CRC.

Mallick, B. K. and Gelfand, A. E. (1996) Semiparametric errors in variables models : a Bayesian approach. *Journal of Statistical Planning and Inference* , **52**, 307-321.

Müller, P. and Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika* , **84**, 3, 523-537.

Richardson, S. and Gilks, W. R. (1993). A Bayesian approach to measurement error problems in Epidemiology using conditionnal independence models. *American Journal of Epidemiology* , **138**, 6, 430-442.

Richardson, S. and Gilks, W. R. (1993) Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine*, **12**, 1703–1722.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion) *Journal of the Royal Statistical Society*, B, **59**, 731–792.

Richardson, S., Leblond, L. (1997). Some comments on misspecification of priors in Bayesian modelling of measurement error problems. *Statistics in Medicine* , **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. *Journal of the American Statistical Association* , **91**, 722- 732.

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

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

Seaman, S. R. and Richardson, S. (2000). Bayesian analysis of case-control studies with categorical covariates. to appear in *Biometrika.*

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. INSERM.

Watier, L., Richardson, S. and Green, P. J. (1999). Using gaussian mixtures with unknown number of components for mixed model estimation. *Proceedings of the 14th International Workshop on Statistical Modelling.* H Friedl, A Berghold; G. Kauermann eds.