Bayesian Adjustment for Exposure Misclassification in Case-Control Studies

by

Rong Chu

B.Sc.(H), McMaster University, 2005

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

The Faculty of Graduate Studies

(Statistics)

The University Of British Columbia

April, 2007

Abstract

Measurement error occurs frequently in observational studies investigating the relationship between exposure variables and the clinical outcome. Error-prone observations on the explanatory variable may lead to biased estimation and loss of power in detecting the impact of an exposure variable. The mechanism of measurement error, such as whether or in what way the quality of data is affected by the disease status, is seldom completely revealed to the investigators. This increases uncertainty in assessing the consequences of ignoring measurement error associated with observed data, and brings difficulties to adjustment for mismeasurement.

In this study, we consider situations with a correctly specified binary response, and a misclassified binary exposure. We propose a solution to conduct Bayesian adjustment to correct for measurement error subject to varying differentiality, including the nondifferential misclassification, differential misclassification and nearly nondifferential misclassification. Our Bayesian model incorporates the randomness of exposure prevalences and misclassification parameters as prior distributions. The posterior model is constructed upon simulations generated by Gibbs sampler and Metropolis-Hastings algorithm. Internal validation data is utilized to insure the resulting model is identifiable.

Meanwhile, we compare the Bayesian model with maximum likelihood estimation (MLE) and simulation extrapolation (MC-SIMEX) methods, using simulated datasets. The Bayesian and MLE models produce accurate and similar estimates for odds ratio in describing the association between the disease and exposure, when appropriate as-

sumptions regarding the differentiality of misclassification are made. The 90% credible or confidence intervals capture the truth approximately 90% of the time. A Bayesian method corresponding to nearly nondifferential prior belief compromises between the loss of efficiency and loss of accuracy associated with other prior assumptions. At the end, we look at two case-control studies with misclassified exposure variables, and aim to make valid inference about the effect parameter.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my special gratitude to my supervisor, Professor Paul Gustafson, and my co-supervisor Dr. Nhu Le, for their thoughtful guidance and support throughout my study in the Department of Statistics. Their indicatives and insight helped me to complete this thesis.

Sincere thanks to Professors John Petkau, Lang Wu, Harry Joe, Ruben Zamar, Will Welch, Raphael Gottardo and Matias Salibián-Barrera for their excellent teaching and constant support during my Master's study.

I am also grateful for the encouragement from all graduate students and staff members in the department. Special thanks to Kenneth, Zhong, Juxin, Hui, Jean-Franois, Lawrence, Wei Liu and Mike Danilov.

Rong Chu

The University of British Columbia

April 2007

To My Dear Parents

Chapter 1

Introduction

In biomedical research, people are often interested in learning the relationship between a health-related outcome Y and an explanatory variable measuring some kind of exposure status, denoted by T. Sometimes, in practice, the exposure variable or clinical outcome is not precisely recorded. For instance, instead of T, an approximate measurement or a *surrogate*, X is obtained. Carroll, Ruppert, Stefanski and Crainiceanu (2006) found that measurement error in the explanatory variables had triple effects:

- It causes bias in parameter estimation for statistical models.
- It leads to a loss of power, sometimes profound, for detecting interesting relationship among variables.
- It masks the features of the data, making graphical model analysis difficult.

The first problem caused by replacing T with X without accounting for the measurement error in data analysis has most serious impact on subsequent statistical inference. Hence the goal of adjustment for mismeasurement is to achieve roughly unbiased estimates to reveal the relationship between Y and T indirectly, based on the measurements of Y, X and perhaps other correctly recorded covariates $\mathbf{Z}$.

As it is important to have adequate knowledge of the nature and type of measurement errors, some examples from epidemiology are listed below.

The NHANES-I dataset was created in a prospective study consisting of nutrition habits and incidences of breast cancer concerning a cohort of 8596 females (Jones et al.,

1987). The response (Y) represents the occurrence of breast cancer. The covariates of primary interest are the long-term nutrition variables ($\mathbf{T}$). Other explanatory variables $\mathbf{Z}$ include demographic and clinical factors such body mass index (BMI), age, alcohol consumption, age at menarche, etc. The response Y and $\mathbf{Z}$ were assumed to be correctly recorded, whereas $\mathbf{T}$ was not measured due to the difficulty in observing diet of a large cohort over a long period of time. Instead, a surrogate $\mathbf{X}$ recording the nutrition intakes of study subjects in the previous 24 hours was retrieved during the interview. The longitudinal variation of diet results in a major measurement error in this study. The seasonal diet behavior and day-to-day nutrition intake differences make $\mathbf{X}$ an inadequate approximation of $\mathbf{T}$. The nonnegligible mismeasurement was discussed in epidemiological literature (Beaton et al., 1979; Wu et al., 1986). Some measurement error models based on a subset of the cohort were proposed by Carroll et al. (2006).

It is sensible to assume the conditional distribution of $\mathbf{X}$ given $\mathbf{T}$ and Y does not depend on Y in the first example, which is known as the *nondifferential* measurement error. However, in other circumstances, this condition does not hold. In case-control studies, explanatory variables are retrieved after the diagnosis. Two group of study subjects with positive (*cases*) or negative (*controls*) clinical outcomes are first recruited, and consecutively the explanatory variables about their exposure history are measured. This type of sampling scheme may well lead to the so called *differential* measurement error, i.e. the conditional distribution of the surrogate X given the unobservable exposure T also depends on the response Y. When information about covariates is collected through some "self-report" mechanism, subjects with positive clinical outcomes may erroneously "blame" a set of risk factors for their conditions, or "ignore" previous experience with exposure variables to avoid any connection between behaviour and disease. The controls on the other hand may pay less attention or make less efforts to provide precise information about their past actions, as they do not suffer from the disease.

A small portion of subjects for whom both the rough measurements and "gold standard" measurements are acquired (the *validation* sample) is sometimes obtained to monitor the severeness of measurement error. The herpes and cervical cancer study serves as an example to demonstrate the utilization of validation study. A case-control study consisting of 732 subjects of cervical cancer and 1312 community or hospital controls with negative cervical cancer diagnosis was conducted to investigate the impact of herpes simplex virus type 2 (HSV-2, a binary variable) in the development of invasive cervical cancer (Hildesheim et al., 1991). The exposure status was detected by the western blot assay, which produced error-prone measurements. A refined, more accurate procedure was performed on a randomly selected sample whose disease statuses were blinded, in order to assess the misclassification rates. Carroll, Gail and Lubin (1993) observed from the validation sample that the misclassification differ between cases and control (Fisher's exact two-sided test implied a greater sensitivity for the cases, p=0.049), and proposed a pseudo-likelihood model to adjust for the differential measurement error.

1.1 Problem setup

The second example reflects how mismeasurement phenomena arise from biomedical studies with categorical covariates. The measurement error in this situation is often referred as a *misclassification* problem. In this thesis, we restrict ourself to misclassification problems on a binary exposure variable T(=0, 1) in case-control studies with no other covariates at play. Discussions of measurement error on continuous or *polychotomous* (with more than 2 categories) exposure variable can be found in statistical literature (Gustafson, 2004; Carroll et al., 2006).

We assume no measurement error arising for the outcome of interest Y (=0, 1). Complete information on Y, T and the surrogate exposure variable X (=0, 1) is available

Table 1.1: Validation data and main data

	Validation Data				*Main Data*			
	Y=1		Y=0		Y=1		Y=0	
T	X=1	X=0	X=1	X=0	X=1	X=0	X=1	X=0
T=1	a_{11}	a_{12}	a_{01}	a_{02}	b_{11}	b_{12}	b_{01}	b_{02}
T=0	a_{13}	a_{14}	a_{03}	a_{04}	b_{13}	b_{14}	b_{03}	b_{04}
N	$a_{11}+a_{13}$	$a_{12}+a_{14}$	$a_{01}+a_{03}$	$a_{02}+a_{04}$	a_{15}	a_{16}	a_{05}	a_{06}

for a small proportion of data (*validation sample*), whereas the true exposure status for the majority of study subjects (*main study*) is unobservable or cannot be precisely measured. The validation data and incomplete main data are presented in Table 1. While each cell a_{ij} in the validation data is fully specified ($i = 0, 1, j = 1, 2, 3, 4$), only margins $a_{05}, a_{06}, a_{15}, a_{16}$ in the main misclassification table are recorded. Our goal is therefore to recover the main table and ultimately make inference on the relationship between the clinical outcome Y and the actual exposure variable T.

Let us denote the true exposure prevalences amongst cases and controls by r_1 and r_0 respectively, where $r_i = P(T = 1|Y = i), i = 0, 1$. The retrospective odds ratio describing the correlation between the response and explanatory variable is defined as

$$\Phi_r = \frac{\frac{r_1}{1-r_1}}{\frac{r_0}{1-r_0}},$$

which is equal to the prospective odds ratio,

$$\Phi_p = \frac{\frac{P(Y=1|T=1)}{P(Y=0|T=1)}}{\frac{P(Y=1|T=0)}{P(Y=0|T=0)}},$$

via Bayes rule and simple algebraic manipulations. The odds ratio is sometimes adopted to approximate the relative risk $\Psi = P(Y = 1|T = 1)/P(Y = 1|T = 0)$ of having a

disease in two exposure groups, when the disease incidence rate is small (*rare disease*). Sensitivity (SN) and specificity (SP) jointly measure the magnitude of exposure misclassification. In the scenarios subject to *differential* misclassification, the surrogate X given the unobserved true exposure T and the response Y are not mutually independent. The sensitivities and specificities among cases and controls can be formulated as, $SN_i = P(X = 1|T = 1, Y = i), SP_i = P(X = 0|T = 0, Y = i)$, $i = 0, 1$. Prevalences of the *apparent* exposure for diseased and non-diseased individuals are denoted by r_1^* and r_0^*. Another way of expressing the degree of misclassification is to facilitate the positive predictive value (PPV) and negative predictive value (NPV). Their relationships are presented below.

$$\begin{aligned} r_i^* &= P(X = 1|Y = i) \\ &= \Sigma_{j=0}^{1} P(X = 1|T = j, Y = i)P(T = j|Y = i) \\ &= r_i SN_i + (1 - r_i)(1 - SP_i) \end{aligned} \tag{1.1}$$

$$\begin{aligned} PPV_i &= P(T = 1|X = 1, Y = i) \\ &= \frac{P(X = 1|T = 1, Y = i)P(T = 1|Y = i)}{P(X = 1|T = 1, Y = i)P(T = 1|Y = i) + P(X = 1|T = 0, Y = i)P(T = 0|Y = i)} \\ &= \frac{SN_i r_i}{SN_i r_i + (1 - SP_i)(1 - r_i)} \end{aligned} \tag{1.2}$$

$$\begin{aligned} NPV_i &= P(T = 0|X = 0, Y = i) \\ &= \frac{P(X = 0|T = 0, Y = i)P(T = 0|Y = i)}{P(X = 0|T = 1, Y = i)P(T = 1|Y = i) + P(X = 0|T = 0, Y = i)P(T = 0|Y = i)} \\ &= \frac{SP_i(1 - r_i)}{SP_i(1 - r_i) + (1 - SN_i)r_i} \end{aligned} \tag{1.3}$$

It is easy to justify that, in the main study the actual number of subjects of positive exposure status (b_{i1}) amongst those who are apparently exposed in either case or control

group (a_{i5}) follows a Binomial distribution, i.e. $b_{i1} \sim Binomial(a_{i5}, PPV_i)$. Similarly, conditioning on the number of cases or controls with negative apparent exposure status (a_{i6}), the number of truly unexposed subjects (b_{i4}) follows $Binomial(a_{i6}, NPV_i)$, for $i = 0, 1$.

When the nondifferential misclassification condition is fulfilled, meaning the conditional distribution of $X|T, Y$ does not depend on Y, it follows immediately that $SN_0 = SN_1 = SN$, $SP_0 = SP_1 = SP$. However it is worth pointing out that, nondifferential misclassification does not imply equality of cases and controls regarding the apparent exposure prevalence (PPV_i, NPV_i).

The bias caused by misclassification of the explanatory variable in case-control studies can be evaluated by the attenuation factor (AF),

$$AF = \frac{\Phi^*}{\Phi},$$

with a numerator representing limit of the error-prone "apparent" odds ratio,

$$\Phi^* = \frac{\frac{r_1^*}{1 - r_1^*}}{\frac{r_0^*}{1 - r_0^*}}.$$

Under the circumstances of nondifferential misclassification, the impact of bias imputed by measurement error is well understood (Gustafson, 2004). Under a weak condition that the possibility of having an accurately measured surrogate exposure is over 0.5 (i.e. $SN + SP > 1$), AF always moves towards the direction of $\Phi^* = 1$. In other words, there is a tendency to report an artificially weak association between the exposure and response in ignoring measurement error on the exposure. Furthermore, flattening effect on true odds ratios far away from the unity is more manifest. When the exposure prevalence is approaching 0 or 1, measurement error induces serious and sensitive attenuation on

the odds ratio. Nevertheless, Greenland and Gustafson (2006) found that, no general conclusion could be made regarding the direction of estimated association when the non-differential misclassification on a binary exposure is not satisfied, or when the exposure variable is polychotomous.

1.2 Review of currently available methods

There is a large literature on the correctness for mismeasurement in biomedical research. Most of the work concentrates on building measurement error models using frequentist methods (Walter and Irwig, 1987; Bashir and Duffy, 1997; Carroll, Ruppert, Stefanski and Crainiceanu, 2006, for example). When the problem of misclassification on exposure variable is encountered in epidemiologic studies, the matrix method (Barron, 1977) estimates the expectations of cell counts (obtained by cross tabulating Y and T) by utilizing the margins of the main sample in Table 1.1. The odds ratio is estimated subsequently based on the cell counts, and the asymptotic variance of $\widehat{OR}$ is derived by Greenland (1988) via the Delta method. Marshall (1990) reparameterized the misclassification by PPV and NPV and proposed an inverse matrix method to retrieve the true odds ratio. Lyles (2002) proved that Marshall's formula was in fact the maximum likelihood estimate (MLE) subject to differential misclassification. Other approaches including the simulation extrapolation method and latent class logistic regression model are developed to tackle the same problem (Küchenhoff, Mwalili and Lesaffre, 2006; Skrondal and Rabe-Hesketh, 2004). On the other hand, the dramatic improvement of computational capability of electronic computers and the development of indirect simulation techniques such as Markov chain Monte Carlo (MCMC) make it possible to explore misclassification problems from a Bayesian perspective (Gustafson, 2004). In fact, partial knowledge of misclassification probabilities is often accessible to medical researchers before the conduc-

tion of a study. This makes Bayesian analysis an appealing approach, for the inference will be based on the combination of prior knowledge and data thoroughly.

Therefore, in this thesis, we primarily introduce a series of Bayesian methods suitable for different misclassification assumptions. Their performance will be closely compared to those of the maximum likelihood estimates (MLEs) and simulation extrapolation (SIMEX) method, using simulation studies and two real life examples. The thesis is organized as follows. Chapter 2 presents fundamental concepts and algorithms of the Bayesian paradigm, and provides detailed methodology for the proposed Bayesian methods. Chapters 3 and 4 review the MLE and SIMEX methods under differential and nondifferential misclassifications. Chapter 5 discusses the comparative behaviours of the three methods based on two simulation studies. Chapter 6 presents the performances of Bayesian and other methods via two case-control studies with misclassified exposure variables and validation sub-samples. Chapter 7 provides overall conclusion and further remarks for the research.

Chapter 2

Bayesian data analysis

2.1 Bayes theorem

Assume the observable random variables $\mathbf{Y}$ are distributed according to a joint density function $f(\mathbf{y}|\theta)$, where θ is an unknown parameter vector. One's prior belief on the value of θ before observing the data $\mathbf{y}$ is described by a prior distribution $f(\theta)$. A joint probability distribution can be written as a product of the data distribution $f(\mathbf{y}|\theta)$ and the prior distribution $f(\theta)$. Given the observed data, the posterior distribution of the unobserved parameters can be determined using Bayes' rule:

$$f(\theta|\mathbf{y}) = \frac{f(\theta,\mathbf{y})}{f(\mathbf{y})} \tag{2.1}$$

$$= \frac{f(\theta)f(\mathbf{y}|\theta)}{\int f(\theta^*)f(\mathbf{y}|\theta^*)d\theta^*} \tag{2.2}$$

This equation can further be simplified by omitting the fixed factor $f(\mathbf{y})$ for a given $\mathbf{y}$ that does not depend on θ:

$$f(\theta|\mathbf{y}) \propto \mathbf{f}(\theta)\mathbf{f}(\mathbf{y}|\theta)\,.$$

When it is too complicated to derive a normalized form for the above unnormalized posterior distribution, Markov Chain Monte Carlo (MCMC) and other numerical

techniques should be applied.

2.2 Prior distributions

The exposure prevalances r_0, r_1, sensitivities SN_0, SN_1, and specificities SP_0, SP_1 are the parameters of interest in this study, ranging from 0 to 1. We make them cover the whole real line by converting into a logit scale, $logit(x) = \log \frac{x}{1-x}$, for x between 0 and 1. The prior information concerning these parameters can then be modeled using bivariate normal distributions. The actual exposure prevalences (r_i), sensitivities(SN_i) and specificities (SP_i) of X as a surrogate for T are assumed to be uncorrelated in this circumstance.

$$\begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} \equiv \begin{pmatrix} \log \frac{r_0}{1-r_0} \\ \log \frac{r_1}{1-r_1} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_1\sigma_1\sigma_2 \\ \rho_1\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right),$$

$$\begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \equiv \begin{pmatrix} \log \frac{SN_0}{1-SN_0} \\ \log \frac{SN_1}{1-SN_1} \end{pmatrix} \sim N\left(\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \rho_2\tau_1\tau_2 \\ \rho_2\tau_1\tau_2 & \tau_2^2 \end{pmatrix} \right),$$

$$\begin{pmatrix} q_1 \\ q_2 \end{pmatrix} \equiv \begin{pmatrix} \log \frac{SP_0}{1-SP_0} \\ \log \frac{SP_1}{1-SP_1} \end{pmatrix} \sim N\left(\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}, \begin{pmatrix} \delta_1^2 & \rho_3\delta_1\delta_2 \\ \rho_3\delta_1\delta_2 & \delta_2^2 \end{pmatrix} \right).$$

The prior knowledge of differentiality for misclassification is reflected by ν_1, ν_2, γ_1, γ_2 and varying values of ρ_2 and ρ_3. Theory tells us that normality is retained from linear

combination of normal random variables. We then have

$$p_1 - p_2 \sim N(\nu_1 - \nu_2, \tau_1^2 + \tau_2^2 - 2\rho_2\tau_1\tau_2) \tag{2.3}$$

$$q_1 - q_2 \sim N(\gamma_1 - \gamma_2, \delta_1^2 + \delta_2^2 - 2\rho_3\delta_1\delta_2) \tag{2.4}$$

Parameters in the prior distributions are often named *hyperparameters*. In this context, they include μ_i, σ_i, ν_i, τ_i, γ_i, δ_i and ρ_j. Hyperparameters reflect one's prior belief on the target parameters based on previous research, literature review or even personal opinions. We set $\mu_1 = \mu_2$, $\nu_1 = \nu_2$ and $\gamma_1 = \gamma_2$ to be"unbiased", a priori, indicating the prevalences and misclassification probabilities amongst cases and controls are centred around equal means; and $\sigma_1 = \sigma_2$, $\tau_1 = \tau_2$ and $\delta_1 = \delta_2$, asserting that, the variabilities associated with prevalences, sensitivities and specificities in two populations are consistent. As a result, the condition $\rho_2 = \rho_3 = 1$ implies $p_1 - p_2 \sim N(0,0)$, $q_1 - q_2 \sim N(0,0)$, and furthermore $p_1 = p_2$, $q_1 = q_2$. Since the logit transformation is a 1-to-1 function, it follows immediately that $SN_0 = SN_1$ and $SP_0 = SP_1$, which corresponds to nondifferential misclassification. Conversely, a zero-valued correlation coefficient ρ_2 or ρ_3 indicates the independence between p_1, p_2 or q_1, q_2, and hence implies SN_0 and SN_1 or SP_0 and SP_1 are independent. This intuitively reflects the fact that sensitivities or specificities are free to vary by themselves, and can be interpreted as, fully differential misclassification is achieved. Situations in between the two cases are defined as the "nearly nondifferential" misclassification to reflect a certain level of dependence between the misclassification parameters. In addition, by assigning particular values to hyperparameters, for instance μ_i and σ_i, we can postulate a priori that r_i lies between 0.02 and 0.5 say, on logit scale with 95% probability. In fact, this is a "wide" interval containing most feasible prevalence values in epidemiological applications.

Because the marginal and conditional densities of a multivariate normal variable

remain normal, the joint, marginal and conditional prior distributions for the original parameters r_1, r_2, SN_1, SN_2, SP_1, SP_2 can be easily derived applying standard variable transformation techniques.

2.3 Posterior simulation

Let us define the parameter vector θ as

$$\theta \equiv (r_1, r_2, SN_1, SN_2, SP_1, SP_2).$$

The sampling distribution (ignoring constant terms) is

$$\begin{aligned} f(\mathbf{y}|\theta) &= \prod_{i=0}^{1} \Big\{ \Big[SN_i r_i + (1-SP_i)(1-r_i)\Big]^{a_{i5}} \Big[(1-SN_i)r_i + SP_i(1-r_i)\Big]^{a_{i6}} \\ &\quad \Big[SN_i r_i\Big]^{a_{i1}} \Big[(1-SP_i)(1-r_i)\Big]^{a_{i3}} \Big[(1-SN_i)r_i\Big]^{a_{i2}} \Big[SP_i(1-r_i)\Big]^{a_{i4}} \Big\}, \quad (2.5) \end{aligned}$$

and the posterior density $f(\theta|\mathbf{y})$ is proportional to $f(\theta)f(\mathbf{y}|\theta)$.

With this complex unnormalized posterior density function, one cannot (a) simulate independent and identically distributed realizations from the posterior, or (b) simulate dependent but identically distributed realizations from the posterior distribution. Fortunately, under this circumstance, the family of MCMC techniques provides us with an approach to simulate from a Markov chain that converges to the posterior density as its stationary distribution (Gustafson, 2004). An alternative term referring to MCMC methods is the Metropolis Hastings (MH) algorithm. Together with Gibbs sampler, a special case of the MH algorithm, MCMC is commonly used in practice to build Markov chains,through drawing samples from the Bayesian posterior distributions.

2.4 The Metropolis Hastings Algorithm

The Metropolis Hastings (MH) algorithm was originally introduced by Metropolis et al. (1953) and was further generalized by Hastings (1970). The MH algorithm enjoys a simple form. Consider the situation when we try to construct a Markov chain converging to a stationary distribution (target density of interest) $p(\xi|z)$. First, an arbitrary starting point ξ^1 is drawn so that $p(\xi^1|z) > 0$. A jumping distribution (or proposal distribution) $J(\xi^*|\xi^t)$ is then selected to simulate a candidate state ξ^* at time $t+1$ provided the current state ξ^t, for $t = 1, 2, 3, \ldots$. The acceptance probability is defined as

$$\alpha = \min\left\{\frac{p(\xi^*|y)}{p(\xi^t|y)}\frac{J(\xi^t|\xi^*)}{J(\xi^*|\xi^t)}, 1\right\}. \tag{2.6}$$

Next simulate u from the *Uniform*(0,1) distribution. If $u \leq \alpha$, set $\xi^{t+1} = \xi^*$; otherwise, set $\xi^{t+1} = \xi^t$. The transition distribution of the Markov chain is therefore a mixture of the proposal distribution and a point mass at $\xi^{t+1} = \xi^t$. The iteration continues until a sample size n is reached after the "burn-in" period of size m. It is important to select sufficiently large m and n to guarantee accurate estimation of the quantities associated with the target distribution. The choice of the jumping distribution has great impact on the convergence of the Markov chain $\xi^{m+1}, \xi^{m+2}, \ldots, \xi^{m+n}$ to the posterior density, and mixing of the parameter values.

2.5 The Gibbs Sampler

The Gibbs sampler can be treated as a special case of the MH algorithm of acceptance probability being 1 at every jump. It is useful when the target posterior distribution is multidimensional, such as the problem being studied here. When it is difficult to sample from the complicated desired joint distribution, the parameter vector

$\boldsymbol{\theta}$ can be divided into say d subvectors or blocks $(\theta_1, \ldots, \theta_d)$. At every iteration, all d subvectors are updated in sequence, each being sampled from the conditional distribution given all other blocks (*the full conditional distribution*). In other words, at step i, θ_i is sampled from $f(\theta_i|\theta_{-i}^{t-1}, y)$, provided the latest values of other subvectors $\theta_{-i}^{t-1} = (\theta_1^t, \ldots, \theta_{i-1}^t, \theta_{i+1}^{t-1}, \ldots, \theta_d^{t-1})$, With conjugate priors, it is possible to simulate directly from the full conditional target distributions. Otherwise, the MH algorithm can be used to update parameters within a single block.

2.6 The hybrid algorithm

When the Markov chain under differential or nearly nondifferential misclassification reaches some state θ in the parameter space, the cell counts in main data of Table 1.1 (b_{ij}), $i = 0, 1$, $j = 1, 2, 3, 4$ are generated from Binomial distributions described in section 1.1. The data is updated as $\mathbf{y} = \{\mathbf{y}_{obs}, \mathbf{y}_{unobs}\} = \{(a_{i1}, \cdots, a_{i4}), (b_{i1}, \cdots, b_{i4})\}_{i=0}^{1}$. The sampling distribution or the likelihood function becomes

$$\begin{aligned} f(\mathbf{y}_{obs}, \mathbf{y}_{unobs}|\theta) &=. L(\theta|\mathbf{y}_{obs}, \mathbf{y}_{unobs}) \\ &= \prod_{i=0}^{1} \Big\{ (SN_i r_i)^{a_{i1}+b_{i1}} \left((1-SP_i)(1-r_i)\right)^{a_{i3}+b_{i3}} \\ &\qquad \left((1-SN_i) r_i\right)^{a_{i2}+b_{i2}} \left(SP_i (1-r_i)\right)^{a_{i4}+b_{i4}} \Big\} \\ &= \prod_{i=0}^{1} \Big\{ r_i^{a_{i1}+a_{i2}+b_{i1}+b_{i2}} (1-r_i)^{a_{i3}+a_{i4}+b_{i3}+b_{i4}} SN_i^{a_{i1}+b_{i1}} \\ &\qquad (1-SN_i)^{a_{i2}+b_{i2}} SP_i^{a_{i4}+b_{i4}} (1-SP_i)^{a_{i3}+b_{i3}} \Big\} . \end{aligned} \tag{2.7}$$

As it is not plausible to simulate θ from the joint posterior distribution directly, we sample the 6-dimension parameter sequentially from the conditional posterior distribu-

tions,

$$f(r_0, r_1|SN_0, SN_1, SP_0, SP_1, \mathbf{y}) \propto f(r_0, r_1)\prod_{i=0}^{1}\left\{r_i^{a_{i1}+a_{i2}+b_{i1}+b_{i2}}(1-r_i)^{a_{i3}+a_{i4}+b_{i3}+b_{i4}}\right\}, \quad (2.8)$$

$$f(SN_0, SN_1|r_0, r_1, SP_0, SP_1, \mathbf{y}) \propto f(SN_0, SN_1)\prod_{i=0}^{1}\left\{SN_i^{a_{i1}+b_{i1}}(1-SN_i)^{a_{i2}+b_{i2}}\right\}, \quad (2.9)$$

$$f(SP_0, SP_1|r_0, r_1, SN_0, SN_1, \mathbf{y}) \propto f(SP_0, SP_1)\prod_{i=0}^{1}\left\{SP_i^{a_{i4}+b_{i4}}(1-SP_i)^{a_{i3}+b_{i3}}\right\}. \quad (2.10)$$

In above equations, $f(r_0, r_1)$, $f(SN_0, SN_1)$ and $f(SP_0, SP_1)$ are bivariate normal prior densities for the prevalences, sensitivities and specificities over cases and controls.

As the densities are not conditionally conjugate, the MH algorithm is performed at next step. We apply univariate MH jumps embedded in Gibbs sampling to update each component in the paired of parameters, (r_0, r_1), (SN_0, SN_1) and (SP_0, SP_1). It is important to choose a jumping distribution that leads to satisfactory performance of the Markov chain. One intuitive solution is to let the jumping distribution follow a Beta density as appeared in the likelihood function in Equation (2.7). This simplifies calculation of the acceptance rate by cross canceling the ratio of proposed *vs.* current likelihoods and the ratio between two jumping densities. As a result, we are left with merely the ratio between two prior distributions. To be more specific, let us take the acceptance probability in one dimensional MH jump on r_0 in Equation (2.8) for example. The jumping rule is specified as $r_0^* \sim Beta(a_{01}+a_{02}+b_{01}^t+b_{02}^t+1, a_{03}+a_{04}+b_{03}^t+b_{04}^t+1)$, close to the conditional sampling distribution. The ratio of ratios in Equation (2.6)

becomes

$$\begin{aligned}
&\frac{f(r_0^*|r_1^t, SN_0^t, SN_1^t, SP_0^t, SP_1^t, \mathbf{y}^t)}{f(r_0^t|r_1^t, SN_0^t, SN_1^t, SP_0^t, SP_1^t, \mathbf{y}^t)} \\
&\frac{J(r_0^*|r_0^t, r_1^t, SN_0^t, SN_1^t, SP_0^t, SP_1^t)}{J(r_0^t|r_0^*, r_1^t, SN_0^t, SN_1^t, SP_0^t, SP_1^t)} \\
= &\frac{f(r_0^*|r_1^t, SN_0^t, SN_1^t, SP_0^t, SP_1^t)L(r_0^*, r_1^t, SN_0^t, SN_1^t, SP_0^t, SP_1^t)}{f(r_0^t|r_1^t, SN_0^t, SN_1^t, SP_0^t, SP_1^t)L(r_0^t, r_1^t, SN_0^t, SN_1^t, SP_0^t, SP_1^t)} \\
&\frac{L(r_0^*, r_1^t, SN_0^t, SN_1^t, SP_0^t, SP_1^t)}{L(r_0^t, r_1^t, SN_0^t, SN_1^t, SP_0^t, SP_1^t)} \\
= &\frac{f(r_0^*|r_1^t, SN_0^t, SN_1^t, SP_0^t, SP_1^t)}{f(r_0^t|r_1^t, SN_0^t, SN_1^t, SP_0^t, SP_1^t)} \\
= &\frac{f(r_0^*|r_1^t)}{f(r_0^t|r_1^t)}. \qquad (2.11)
\end{aligned}$$

The last step is obtained because of the mutual independence among prevalence, sensitivity and specificity in prior information.

We now summarize the posterior simulation procedure as follows.

1. Acquire initial values of $(r_0^0, r_1^0, SN_0^0, SN_1^0, SP_0^0, SP_1^0)$, each lying between 0 and 1.

2. At the t^{th} iteration,

   - Given parameters $\theta^t = (r_0^t, r_1^t, SN_0^t, SN_1^t, SP_0^t, SP_1^t)$, update the unobserved actual exposure data $\mathbf{y_{unobs}}^t = \{b_{ij}^t\}$ based on *Binomial* distributions, for $i = 0, 1$, $j = 1, 2, 3, 4$.

   - Based on the updated cell counts $\{b_{ij}^t\}$ at the t^{th} iteration, model parameters are generated alternately via Gibbs sampler and MH algorithm.

     (a) Simulate r_0^t conditioning on $(r_1^{t-1}, SN_0^{t-1}, SN_1^{t-1}, SP_0^{t-1}, SP_1^{t-1})$ using MH algorithm. A proposed jumping rule is $r_0^* \sim Beta(a_{01} + a_{02} + b_{01}^t + b_{02}^t + 1, a_{03} + a_{04} + b_{03}^t + b_{04}^t + 1)$, with acceptance rate $\min\left\{\frac{f(r_0^*|r_1^{t-1})}{f(r_0^{t-1}|r_1^{t-1})}, 1\right\}$.

(b) Simulate r_1^t conditioning on $(r_0^t, SN_0^{t-1}, SN_1^{t-1}, SP_0^{t-1}, SP_1^{t-1})$. A proposed jumping rule is $r_1^* \sim Beta(a_{11} + a_{12} + b_{11}^t + b_{12}^t + 1, a_{13} + a_{14} + b_{13}^t + b_{14}^t + 1)$, with acceptance rate $\min\left\{\frac{f(r_1^*|r_0^t)}{f(r_1^{t-1}|r_0^t)}, 1\right\}$.

(c) Simulate SN_0^t conditioning on $(r_0^t, r_1^t, SN_1^{t-1}, SP_0^{t-1}, SP_1^{t-1})$ using MH algorithm. A proposed jumping rule is $SN_0^* \sim Beta(a_{01}+b_{01}^t+1, a_{02}+b_{02}^t+1)$, with acceptance rate $\min\left\{\frac{f(SN_0^*|SN_1^{t-1})}{f(SN_0^{t-1}|SN_1^{t-1})}, 1\right\}$.

(d) Simulate SN_1^t conditioning on $(r_0^t, r_1^t, SN_0^t, SP_0^{t-1}, SP_1^{t-1})$. A proposed jumping rule is $SN_1^* \sim Beta(a_{11} + b_{11}^t + 1, a_{12} + b_{12}^t + 1)$, with acceptance rate $\min\left\{\frac{f(SN_1^*|SN_0^t)}{f(SN_1^{t-1}|SN_0^t)}, 1\right\}$.

(e) Simulate SP_0^t conditioning on $(r_0^t, r_1^t, SN_0^t, SN_1^t, SP_1^{t-1})$ using MH algorithm. A proposed jumping rule is $SP_0^* \sim Beta(a_{04} + b_{04}^t + 1, a_{03} + b_{03}^t + 1)$, with acceptance rate $\min\left\{\frac{f(SP_0^*|SP_1^{t-1})}{f(SP_0^{t-1}|SP_1^{t-1})}, 1\right\}$.

(f) Simulate SP_1^t conditioning on $(r_0^t, r_1^t, SN_0^t, SN_1^t, SP_0^t)$. A proposed jumping rule is $SP_1^* \sim Beta(a_{14} + b_{14}^t + 1, a_{13} + b_{13}^t + 1)$, with acceptance rate $\min\left\{\frac{f(SP_1^*|SP_0^t)}{f(SP_1^{t-1}|SP_0^t)}, 1\right\}$.

- Calculate the log odds ratio ϕ^t at the t^{th} iteration.

3. Repeat step (2) at subsequent iterations, for $t = 1, \ldots, m + n$, to simulate target parameters alternately using the hybrid algorithm.

The procedure is stopped after accomplishing $m+n$ iterations, where m is the number of burn-in iterations and n stands for the number of target iterations. The multivariate Markov chain is generated and converges to the joint posterior distribution at sufficiently large m. The marginal Markov chains for individual parameters including the log odds ratio are constructed at the meantime and statistical inference can be performed thereafter.

Under the circumstances of nondifferential misclassification, the parameter space becomes 4 dimensional with $\theta = (r_0, r_1, SN, SP)$. The prior densities and likelihood function are revised as

$$p \equiv \left(\log \frac{SN}{1-SN}\right) \sim N\left(\nu, \tau^2\right)$$
$$q \equiv \left(\log \frac{SP}{1-SP}\right) \sim N\left(\gamma, \delta^2\right)$$

$$L(\theta|\mathbf{y}_{obs}, \mathbf{y}_{unobs}) = \prod_{i=0}^{1} \left\{ r_i^{a_{i1}+a_{i2}+b_{i1}+b_{i2}} (1-r_i)^{a_{i3}+a_{i4}+b_{i3}+b_{i4}} \right.$$
$$\left. SN^{a_{i1}+b_{i1}} (1-SN)^{a_{i2}+b_{i2}} SP^{a_{i4}+b_{i4}} (1-SP)^{a_{i3}+b_{i3}} \right\}. \quad (2.12)$$

This leads to the combinations of steps (c) and (d), (e) and (f) in the above procedure. Only one MH jump is required to simulate Markov chain for the sensitivity or specificity within the Gibbs sampling structure. Nevertheless, the ideas on posterior simulation can be addressed and utilized as before.

Chapter 3

Maximum Likelihood Estimation

The maximum likelihood estimate of a parameter θ given a sample point $\mathbf{x}$ is a parameter value at which the likelihood $L(\theta|\mathbf{x})$ attains its maximum as a function of θ (Casella and Berger, 2002). Often candidate maximum likelihood estimators (MLEs) can be obtained through solving $\frac{\partial}{\partial\theta} logL(\theta|\mathbf{x}) = \mathbf{0}$, when the likelihood or log likelihood function is differentiable in θ. The left hand side of the equation is sometimes named the *score function* associated with the likelihood. When θ is multidimensional and a closed-form solution does not exist, iterative numerical procedures such as Newton Raphson algorithm are required. Because of some optimality quantities of MLE including the consistency and efficiency, the method of maximum likelihood estimation is currently the most popular method for point estimation.

3.1 MLEs under differential misclassification

A standard way to express the likelihood in terms of $(r_0, r_1, SN_0, SN_1, SP_0, SP_1)$ for problems consisting of a main study (misclassification data) and internal validation data is provided in Equation (2.5) under differential misclassification. Unfortunately the score functions do not form a tractable system and no closed-form solution exists under this parameterization. Lyles (2002) proposed an alternate to parameterize the likelihood in terms of $\theta^* = (r_0^*, r_1^*, PPV_0, PPV_1, NPV_0, NPV_1)$. This reparameterization leads to closed-form MLEs, and is proven to be equivalent to the inverse matrix method of Mar-

shall (1990).

The likelihood function is of the form

$$\begin{aligned} L(\theta^*|\mathbf{y}) &= \prod_{i=0}^{1} \Big\{ r_i^{*a_{i5}} (1-r_i^*)^{a_{i6}} (PPV_i r_i^*)^{a_{i1}} \\ &\quad \big((1-PPV_i) r_i^*\big)^{a_{i3}} \big((1-NPV_i)(1-r_i^*)\big)^{a_{i2}} \\ &\quad \big(NPV_i(1-r_i^*)\big)^{a_{i4}} \Big\}. \end{aligned} \tag{3.1}$$

By setting the score function associated with each parameter zero, we obtain MLEs

$$\begin{aligned} \widehat{PPV}_i &= \frac{a_{i1}}{a_{i1}+a_{i3}}, \\ \widehat{NPV}_i &= \frac{a_{i4}}{a_{i2}+a_{i4}}, \\ \hat{r}_i^* &= \frac{a_{i1}+a_{i3}+a_{i5}}{a_{i1}+a_{i2}+a_{i3}+a_{i4}+a_{i5}+a_{i6}}. \end{aligned} \tag{3.2}$$

If we apply the invariance property, the MLEs of the original parameters and the log odds ratio ϕ can be written as

$$\begin{aligned} \hat{r}_i &= \widehat{PPV}_i \hat{r}_i^* + (1-\widehat{NPV}_i)(1-\hat{r}_i^*), \\ \widehat{SN}_i &= \frac{\widehat{PPV}_i \hat{r}_i^*}{\hat{r}_i}, \\ \widehat{SP}_i &= \frac{\widehat{NPV}_i (1-\hat{r}_i^*)}{1-\hat{r}_i}, \\ \hat{\phi} &= log \frac{\hat{r}_1(1-\hat{r}_0)}{\hat{r}_0(1-\hat{r}_1)}. \end{aligned} \tag{3.3}$$

The Hessian matrix composed of mixed second derivatives is estimated at the MLE $\hat{\theta}^*$ in form of

$$\mathbf{H}(\hat{\theta}^*) = \left[\frac{\partial^2}{\partial {\theta^*}^2} logL\left(\hat{\theta}^*|\mathbf{y}\right) \right].$$

In fact, the zero cross partial derivatives simplifies the expression of Hessian matrix, and

the observed Fisher information is obtained as,

$$\begin{aligned}\mathbf{i}(\hat{\theta}^*) &= -\mathbf{H}(\hat{\theta}^*) \\ &= -\begin{pmatrix} \frac{\partial^2 logL(\theta^*|\mathbf{y})}{\partial r_0^{*2}} & 0 & & \cdots & & 0 \\ 0 & \frac{\partial^2 logL(\theta^*|\mathbf{y})}{\partial r_1^{*2}} & 0 & & \cdots & 0 \\ 0 & 0 & \frac{\partial^2 logL(\theta^*|\mathbf{y})}{\partial PPV_0^2} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \frac{\partial^2 logL(\theta^*|\mathbf{y})}{\partial PPV_1^2} & 0 & 0 \\ 0 & & \cdots & 0 & \frac{\partial^2 logL(\theta^*|\mathbf{y})}{\partial NPV_0^2} & 0 \\ 0 & & \cdots & & 0 & \frac{\partial^2 logL(\theta^*|\mathbf{y})}{\partial NPV_1^2} \end{pmatrix}\Bigg|_{\theta^*=\hat{\theta}^*}\end{aligned} \tag{3.4}$$

The asymptotic covariance matrix of MLEs in the Cramér-Rao lower bound $\mathbf{\Sigma}_{\hat{\theta}^*}$ is approximated by the inverse of the observed Fisher informatin at $\hat{\theta}^*$, which bears the form of a 6x6 diagonal matrix with reciprocals of non-zero second derivatives of the log likelihood on the diagonal. As a result, variances of the "apparent" exposure, positive and negative predictive values are estimated as,

$$\begin{aligned}\widehat{var}\left(\hat{r}_i^*\right) &= \frac{\hat{r}_i^2(1-\hat{r}_i)^2}{a_{i1}+a_{i2}+a_{i3}+a_{i4}+a_{i5}+a_{i6}} \\ \widehat{var}\left(\widehat{PPV}_i\right) &= \frac{\widehat{PPV}_i(1-\widehat{PPV}_i)}{a_{i1}+a_{i3}} \\ \widehat{var}\left(\widehat{NPV}_i\right) &= \frac{\widehat{NPV}_i(1-\widehat{NPV}_i)}{a_{i2}+a_{i4}}.\end{aligned}$$

The asymptotic variance for the log odds ratio is further estimated as (Morrissey and

Spiegelman, 1999; Lyles, 2002),

$$\widehat{var}\left(\hat{\phi}\right) = \sum_{i=0}^{1} \frac{\left(\widehat{PPV}_i + \widehat{NPV}_i - 1\right)^2 \widehat{var}\left(\hat{r}_i^*\right) + \left(\hat{r}_i^*\right)^2 \widehat{var}\left(\widehat{PPV}_i\right)}{\hat{r}_i^2(1-\hat{r}_i)^2} + \frac{\left(1-\hat{r}_i^*\right)^2 \widehat{var}\left(\widehat{NPV}_i\right)}{\hat{r}_i^2(1-\hat{r}_i)^2} \tag{3.5}$$

using multivariate Delta method.

$$\widehat{var}\left(\hat{\phi}\right) = \mathbf{W}^{\mathbf{T}} \boldsymbol{\Sigma}_{\hat{\theta}^*} \mathbf{W}, \tag{3.6}$$

whereby

$$\mathbf{W} = \begin{pmatrix} \frac{\partial}{\partial r_0^*}\phi_{|\theta^*=\hat{\theta}^*} \\ \frac{\partial}{\partial r_1^*}\phi_{|\theta^*=\hat{\theta}^*} \\ \frac{\partial}{\partial PPV_0}\phi_{|\theta^*=\hat{\theta}^*} \\ \frac{\partial}{\partial PPV_1}\phi_{|\theta^*=\hat{\theta}^*} \\ \frac{\partial}{\partial NPV_0}\phi_{|\theta^*=\hat{\theta}^*} \\ \frac{\partial}{\partial NPV_1}\phi_{|\theta^*=\hat{\theta}^*} \end{pmatrix},$$

is a column vector composed of first derivatives evaluated at the MLE $\hat{\theta}^*$. Similar ideas can be adapted to obtain the asymptotic variances for MLEs of the original parameters $\hat{\theta} = \left(\hat{r}_0, \hat{r}_1, \widehat{SN}_0, \widehat{SN}_1, \widehat{SP}_0, \widehat{SP}_1\right)$.

$$\widehat{var}\left(r_i\right) = \left(\widehat{PPV}_i + \widehat{NPV}_i - 1\right)^2 \widehat{var}\left(\hat{r}_i^*\right) + \left(\hat{r}_i^*\right)^2 \widehat{var}\left(\widehat{PPV}_i\right) + \left(\hat{r}_i^* - 1\right)^2 \widehat{var}\left(\widehat{NPV}_i\right) \tag{3.7}$$

$$\begin{aligned}\widehat{var}\left(SN_i\right) = & \left[\frac{\widehat{PPV}_i}{\hat{r}_i} - \frac{\widehat{PPV}_i\hat{r}_i^*\left(\widehat{PPV}_i + \widehat{NPV}_i - 1\right)}{\hat{r}_i^2}\right]^2 \widehat{var}\left(\hat{r}_i^*\right) + \\ & \left[\frac{\hat{r}_i^*\hat{r}_i - \widehat{PPV}_i\left(\hat{r}_i^*\right)^2}{\hat{r}_i^2}\right]^2 \widehat{var}\left(\widehat{PPV}_i\right) + \\ & \left[\frac{\left(1-\hat{r}_i^*\right)\hat{r}_i^*\widehat{PPV}_i}{\hat{r}_i^2}\right]^2 \widehat{var}\left(\widehat{NPV}_i\right) \end{aligned} \tag{3.8}$$

$$\begin{aligned}\widehat{var}\left(SP_i\right) = & \left[\frac{-\widehat{NPV}_i}{1-\hat{r}_i} + \frac{\widehat{NPV}_i\left(1-\hat{r}_i^*\right)\left(\widehat{PPV}_i + \widehat{NPV}_i - 1\right)}{\left(1-\hat{r}_i\right)^2}\right]^2 \widehat{var}\left(\hat{r}_i^*\right) + \\ & \left[\frac{\hat{r}_i^*\left(1-\hat{r}_i^*\right)\widehat{NPV}_i}{\left(1-\hat{r}_i\right)^2}\right]^2 \widehat{var}\left(\widehat{PPV}_i\right) + \\ & \left[\frac{1-\hat{r}_i^*}{1-\hat{r}_i} - \frac{\widehat{NPV}_i\left(1-\hat{r}_i^*\right)^2}{\left(1-\hat{r}_i\right)^2}\right]^2 \widehat{var}\left(\widehat{NPV}_i\right) \end{aligned} \tag{3.9}$$

3.2 MLEs under nondifferential misclassification

Neither parameterization mentioned above facilitates closed-form MLEs under the circumstances of nondifferential misclassification. Although explicit formulas that are computationally less intensive are provided in approaches such as the matrix method (Barron, 1977) and inverse matrix method (Marshall, 1990), the ML estimation is in general preferable due to its efficiency. We therefore proceed by employing a quasi-Newton method ("L-BFGS-B") by Byrd et al. (1995) that is implemented in the "optim()" function in statistical software **R** to achieve likelihood optimization in original parameter space of $\theta = (r_0, r_1, SN_0, SN_1, SP_0, SP_1)$. Function values, gradients (first order derivatives) of the log likelihood function and the initial estimates are supplied to the routine in order to build up a picture of the surface to be maximized (**R** 2.4.0).

A numerically differentiated Hessian matrix at the best set of estimates in form of a matrix of mixed second derivatives,

$$\mathbf{H}(\hat{\theta}) = \left[\frac{\partial^2}{\partial\theta^2} logL\left(\hat{\theta}|\mathbf{y}\right)\right],$$

will be returned if the option *hessian* is set **TRUE** in "optim()" function. The estimate of observed Fisher information is received immediately

$$\mathbf{i}(\hat{\theta}) = -\mathbf{H}(\hat{\theta}),$$

and the asymptotic covariance matrix of $\hat{\theta}$ in the Cramér-Rao lower bound is approximated by the inverse of the estimated Fisher information, denoted by $\mathbf{i}(\hat{\theta})^{-1}$. The asymptotic variance of the log odds ratio estimate is attainable by the multivariate Delta method,

$$\widehat{var}\left(\hat{\phi}\right) = \begin{pmatrix} \frac{\partial}{\partial r_0}\phi_{\theta|\hat{\theta}} \\ \frac{\partial}{\partial r_1}\phi_{\theta|\hat{\theta}} \end{pmatrix}^T \begin{pmatrix} \hat{\sigma}^2_{\hat{r}_0} & \hat{\sigma}_{\hat{r}_0\hat{r}_1} \\ \hat{\sigma}_{\hat{r}_0\hat{r}_1} & \hat{\sigma}^2_{\hat{r}_1} \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial r_0}\phi_{\theta|\hat{\theta}} \\ \frac{\partial}{\partial r_1}\phi_{\theta|\hat{\theta}} \end{pmatrix}, \tag{3.10}$$

whereby the matrix in the middle of the right hand is the 2x2 component of the estimated asymptotic covariance matrix corresponding to the $\hat{r}_0$ and $\hat{r}_1$.

Chapter 4

Simulation Extrapolation Approach

Simulation extrapolation (SIMEX) method was initially proposed to correct additive measurement error in general regression problems by Cook and Stefanski (1994) and further extended to handle other measurement error models (Cook and Stefansk, 1994; Küchenhoff and Carroll, 1997; Eckert, Carroll and Wang, 1997). SIMEX exploits the relationship between the size of measurement error in covariate or response and the bias in parameter estimation. It utilizes a self-contained simulation component to establish a trend of measurement error-induced bias versus the variance of the extra added measurement error, and extrapolate this trend to the case of no measurement error (Carroll, Ruppert, Stefanski and Crainiceanu, 2006).

The misclassification SIMEX (MC-SIMEX) method, proposed by Küchenhoff, Mwalili and Lesaffre (2006) extends SIMEX to handle misclassification problems with discrete covariate or response. Let us describe MC-SIMEX in a logistic regression setting with a binary response variable Y, and a nondifferential misclassification prone binary covariate X that is associated with an unobservable independent variable T.

$$logit\,(Y = 1|t) = \beta_0 + \beta_1 t$$

It follows immediately after simple arithmetic that β_1 is the effect parameter of interest in this context, which is the log odds ratio of having a positive response, i.e. $\phi = \beta_1$. The

nondifferential misclassification mechanism is characterized by a misclassification matrix,

$$\begin{aligned}\Pi &= \begin{pmatrix} SP & 1-SN \\ 1-SP & SN \end{pmatrix} \\ &= \begin{pmatrix} P(X=0|T=0) & P(X=0|T=1) \\ P(X=1|T=0) & P(X=1|T=1) \end{pmatrix}.\end{aligned}$$

Assuming Π is positive definite, the spectral decomposition can be written as $\Pi = E\Lambda E^{-1}$, where Λ is a diagonal matrix of eigenvalues and E is composed of eigenvectors. Let us define a misclassification operation $MC[\Pi](Z)$, indicating the procedure to simulate a misspecified random variable related to Z with measurement error probabilities SN, SP. It is then natural to express the relationship between X and T as $X = MC[\Pi](T)$. The limit to which the *naive* estimator of β_1 converges ignoring measurement error on the exposure, is denoted by $\beta_1^*(\Pi)$, as the sample size approaches infinity.

At the simulation stage of MC-SIMEX algorithm, extra misclassification $\Pi^\lambda = E\Lambda^\lambda E^{-1}$, $(\lambda > 0)$ is attached to the error prone variable X to generate "reclassified" pseudo data,

$$X(\lambda) = MC\left[\Pi^\lambda\right](X).$$

Under the assumption that the measurement error mechanisms affecting T and X are independent, one can write

$$X(\lambda) = MC\left[\Pi^{1+\lambda}\right](T),$$

and the naive estimator based on the simulated dataset $(X_i(\lambda), Y_i)$, $i = 1, \ldots, N$, bears the form $\hat{\beta}_1^*\left(\Pi^{1+\lambda}\right)$, which is achievable by least squares method. Hence for a fixed set

of λ values, $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_M$, the MC-SIMEX reclassification step is repeated for B times at every λ_k to simulate pseudo data $(X_i(\lambda_k), Y_i)$ and to estimate $\beta^*_{1,b}\left(\Pi^{1+\lambda_k}\right)$. The situation of no misclassification is obtained at $\lambda = -1$, as $\Pi^{1+\lambda} = \Pi^0 = I$. The log odds ratio at λ_k is then estimated as

$$\hat{\beta}_{1,k} = \frac{\sum_{b=1}^{B} \hat{\beta}^*_{1,b}\left(\Pi^{1+\lambda_k}\right)}{B},$$

for $k = 1, \ldots, M$.

At the extrapolation step, the MC-SIMEX estimator $\hat{\beta}_{MC}$ is thus acquired by

1. fitting an parametric extrapolation function $\hat{L}$ on data $\left(1 + \lambda_k, \hat{\beta}_{1,k}\right)$, for instance by least squares, for $k = 1, \ldots, M$;
2. extrapolating the parametric model back to the origin crossing $\left(0, \hat{\beta}_{MC}\right)$.

Küchenhoff et al. (2006) believe that the MC-SIMEX estimator is at least approximately consistent when the extrapolation function is correctly specified or sufficiently close to $\beta^*_1\left(\Pi_k^{1+\lambda}\right)$. They conclude that a quadratic or loglinear extrapolation function works well in general for various models. Examples of MC-SIMEX plots illustrating the effect of increasing measurement error on $logOR$ via different extrapolation functions are provided in Figures 6.2 and 6.4.

The MC-SIMEX method can also be applied to differential misclassification problems. In the context to model relationship between binary covariate and response, a 4x4 misclassification matrix with separate matrices characterizing measurement error in case Π_1 and control Π_0 populations should be supplied to initiate the simulation and

extrapolation process,

$$\mathbf{\Pi} = \begin{pmatrix} \Pi_0 & \mathbf{0} \\ \mathbf{0} & \Pi_1 \end{pmatrix}. \tag{4.1}$$

The misclassification matrices are usually either directly obtainable or can be estimated from the validation data.

There are several candidates for the MC-SIMEX variance estimation $var(\hat{\beta}_{MC})$. The ease of parameter estimation in general regression context using SIMEX is somehow counterbalanced by the complex calculation of the standard errors. A bootstrap approach is outlined by Küchenhoff et al. (2006), though the intensive computation required in simulation step brings difficulty to the implementation. Two other methods avoiding nested resampling are established by Stefanski and Cook (1995) and Carroll et al. (1996). The first one goes closely with Tukey's jackknife standard error calculation, when the misclassification matrix is known or adequately estimated. The latter employs the asymptotic normal distribution and the asymptotic covariance matrix in the setting of M-estimation, and provides a more flexible estimation (Carroll et al., 2006). These two variance estimation methods are adapted in this thesis.

The MC-SIMEX method is available in R after loading the "SIMEX" package that was developed by Lederer. The main function "mcsimex()" is modified to meet our needs to automate the estimation process of log odds ratio repeatedly under the differential misclassification.

Chapter 5

Simulation Studies

5.1 Data Simulation

In order to demonstrate the performance of Bayesian adjustment to misclassification and make comparison with other statistical approaches, we conduct simulation study under two cases. Two sets of fixed simulation parameters are assigned to cases 1 and 2 respectively. Four misclassification scenarios concerning different levels of misclassification differentiality are built to contrast statistical methods introduced in Chapter 2, 3 and 4. We generate 400 datasets (NREP=400) regarding each of the scenarios in each case. Data in scenario 1 are simulated under the nondifferential misclassification, with increasing degree of differentiality in scenarios 2, 3, and 4.

Case 1: Simulation with equal numbers of cases and controls ($N_{ca} = N_{cnt} = 800$), whereby each group consists of 25 percent of validation data and 75 percent of main (or *misclassification prone*) study data ($N^v_{ca} = N^v_{cnt} = 200$, $N^m_{ca} = N^m_{cnt} = 600$). To achieve the property that the odds ratio $\Phi = \frac{\frac{r_1}{1-r_1}}{\frac{r_0}{1-r_0}}$ for the true exposure T on the response Y is 1.5, we have

- *Scenario 1*: (r_0, r_1)=(0.075, 0.1084), (SN_0, SN_1)=(0.6, 0.6), (SP_0, SP_1)=(0.9, 0.9)
- *Scenario 2*: (r_0, r_1)=(0.075, 0.1084), (SN_0, SN_1)=(0.6, 0.65), (SP_0, SP_1)=(0.9, 0.85)

- *Scenario 3*: (r_0, r_1)=(0.075, 0.1084), (SN_0, SN_1)=(0.6, 0.7), (SP_0, SP_1)=(0.9, 0.8)

- *Scenario 4*: (r_0, r_1)=(0.075, 0.1084), (SN_0, SN_1)=(0.6, 0.75), (SP_0, SP_1)=(0.9, 0.75)

Case 2: Simulation with larger percentage of controls (N_{cnt} = 1200) than cases (N_{ca} = 400), while the ratio of validation data against main data remains 1:3 (N^v_{ca} = 100, N^v_{cnt} = 300, N^m_{ca} = 300, N^m_{cnt} = 900). To achieve the property that the odds ratio Φ for the true exposure T on the response Y is 2, we have

- *Scenario 1*: (r_0, r_1)=(0.1, 0.1818), (SN_0, SN_1)=(0.6, 0.6), (SP_0, SP_1)=(0.85, 0.85)

- *Scenario 2*: (r_0, r_1)=(0.1, 0.1818), (SN_0, SN_1)=(0.65, 0.6), (SP_0, SP_1)=(0.8, 0.85)

- *Scenario 3*: (r_0, r_1)=(0.1, 0.1818), (SN_0, SN_1)=(0.7, 0.6), (SP_0, SP_1)=(0.75, 0.85)

- *Scenario 4*: (r_0, r_1)=(0.1, 0.1818), (SN_0, SN_1)=(0.75, 0.6), (SP_0, SP_1)=(0.7, 0.85)

To generate data in either case, we first simulate the true exposure status T (=0, 1) at $Y = i$ from *Bernoulli*(r_i), for i=0, 1, two binary quantities indicating the occurrence of misclassification in terms of sensitivity and specificity, of sample size N_{ca} or N_{cnt}. Using Equation 1.1, we then simulate the apparent (or surrogate) exposure measurements in two groups. Finally we cross tabulate the true and apparent exposure quantities of size N^v_{ca} or N^v_{cnt} to construct the internal validation tables, and classify the remaining apparent exposure measurements to acquire the main study table (Table 1.1).

Three Bayesian methods adopting nondifferential, nearly nondifferential and differential prior distributions respectively, are performed at each scenario within each case to adjust for possible misclassifications and assess the association between the true exposure and outcome.

5.2 Choice of Hyperparameters

Methodology about the MCMC method is described in details in Chapter 2. In simulation studies, we first specify the hyperparameters appearing in prior distributions of the effect parameter of interest $\theta = (r_0, r_1, SN_i, SP_i)$, $i = 0, 1$. According to Section 2.2, under the assumptions that $\mu_1 = \mu_2$, $\sigma_1 = \sigma_2$, $\nu_1 = \nu_2$, $\gamma_1 = \gamma_2$, $\tau_1 = \tau_2$ and $\delta_1 = \delta_2$, we assign $\mu = -2$, $\sigma = 1$ to model the prior information that the logit true exposures are normally distributed with 95% probability between logit(0.02) and logit(0.5). Mild correlation between r_0 and r_1 ($\rho_1 = 0.3$) is selected to allow relatively large standard deviation of $logOR$ around mean 0. Similarly, we set $\nu = \tau = 1.7$, $\gamma = \delta = 0.65$ to represent the prior knowledge that the logit sensitivity and logit specificity are normally distributed within logit(0.6) and logit(0.95) with 95% probability. As discussed in Chapter 2, we set $\rho_2 = \rho_3 = 1$ to reflect nondifferential misclassification; $\rho_2 = \rho_3 = 0$ to express prior belief in differential misclassification. The choice of ρ_2, ρ_3 for nearly nondifferential misclassification requires extra work. To reflect the priori that, the probability of having similar proportions of correctly measured exposure variable among cases and controls is about one quarter, i.e.

$$P\{|SN_1 - SN_0| < 0.01\} = P\{|SP_1 - SP_0| < 0.01\} = 0.24,$$

a simulation study is conducted, and 0.9 is assigned to ρ_2 and ρ_3. Alternatively, closed-form solutions $\rho_2 = \rho_3 = 0.9$ can be obtained by solving the equation,

$$P\{|logit(SN_1) - logit(SN_0)| < 0.1\} = P\{|logit(SP_1) - logit(SP_0)| < 0.1\} = 0.1,$$

using expressions (2.3), (2.4) and the standard normal table.

5.3 Convergence of MCMC Simulation

The convergence and mixing of MCMC simulation should be checked before we make inference about the posterior distributions. Decisions are made based on visual inspection of simulation plots. After discarding the first 1000 simulations to diminish the effect of initial distributions, we draw samples from 10000 iterations, based on which statistical inferences are conducted. The last 2000 iterations of Markov chains for r_i, SN_i, SP_i and $logOR$ at different scenarios (each represented by one out of 400 simulated datasets) in Case 1 are displayed in Figure 5.1 and 5.2. To be more specific, Figure 5.1 depicts the 8000-10000 realizations of a Markov chain simulated under the nondifferential misclassification assumption regarding a dataset in Case 1, scenario 1; Figure 5.2 depicts the 8000-10000 realizations of a Markov chain simulated under the differential misclassification assumption regarding a dataset in Case 1, scenario 4.

It is observed that Markov chains generated by the "nondifferential", "nearly nondifferential" and "differential" Bayesian methods, under different prior distributions display satisfactory mixing and convergence. The Markov chains are moving thoroughly within the target range. and no chain is stuck at the similar values for successive iterations. Table 5.1 reports the acceptance rates based on 10000 posterior realizations over 400 replicated datasets at each scenario in Case 1. As described in section 2.6, the univariate proposal (or jumping) distributions of model parameters embedded in the Gibbs sampler, do not directly depend on preceding states, for information from the immediate previous states facilitates update of the unobserved cell counts in the main data $\mathbf{y}_{unobs} = \{b_{ij}\}$ and, the univariate candidate parameter value is simulated solely based on updated data and previous values of other parameters. Therefore, high acceptance rates displayed in Table 5.1 demonstrate, posterior Markov chain is moving around and adequacy of mixing is achieved. It is worth pointing out that, the jumping distributions

proposed here do not involve tuning parameters as required in the random-walk MH algorithm. The typical challenge of adjusting variations (σ_k's) in the proposal distributions $\theta_k^* \sim N(\theta_k^t, \sigma_k^2)$ $(k = 1, \ldots, m)$, to maintain a moderate acceptance rate, is therefore avoided. Meanwhile the Markov chains converge roughly to the target distributions af-

Table 5.1: Average acceptance rates over 400 replicates in Case 1

Nondifferential Bayesian method - scenario 1						
	r_0	r_1	SP_0	SP_1	SN_0	SN_1
Acceptance	0.973	0.950	0.728	0.728	0.977	0.977
Nearly nondifferential Bayesian method - scenario 2						
	r_0	r_1	SP_0	SP_1	SN_0	SN_1
Acceptance	0.976	0.947	0.540	0.668	0.742	0.741
Nearly nondifferential Bayesian method - scenario 3						
	r_0	r_1	SP_0	SP_1	SN_0	SN_1
Acceptance	0.977	0.944	0.485	0.665	0.572	0.602
Differential Bayesian method - scenario 4						
	r_0	r_1	SP_0	SP_1	SN_0	SN_1
Acceptance	0.973	0.949	0.641	0.786	0.959	0.909

ter 1000 burn-in iterations. Posterior distributions of the exposure prevalences amongst diseased and non-diseased participants, and log odds ratio describing the association between the response and exposure, are illustrated by histograms of posterior samples (Figure 5.3 and 5.4). We hereby emphasize plots considering scenarios where models are consistent with the data-generation, though the simulations include inconsistent cases as well (e.g. differential model on nondifferentially misclassified data, or nondifferential model on differentially misspecified data). The pre-specified true exposure prevalence are 0.075 (control group), 0.1081 (case group) in Case 1, and 0.1 (control group), 0.1818 (case group) in Case 2. The simulation parameters along with the true log odds ratios (log(1.5) in Case 1 and log(2) in Case 2) are marked by vertical bars in the plots. It is shown in the plots that, posterior samples capture the true model parameters with high

accuracy. It is also noticeable that r_0 and r_1 are often estimated with bias towards the same direction (i.e. whose medians located both below or above the vertical bars). This however leads to a more accurate estimate of $logOR$.

Table 5.2: Posterior distributions at various scenarios in Case 1 with different prior information

Nondifferential prior at Case 1, scenario 1							
	r_0	r_1	SP_0	SP_1	SN_0	SN_1	$logOR$
Mean	0.075	0.102	0.901	0.901	0.706	0.706	0.353
SD	0.015	0.0169	0.011	0.011	0.063	0.063	0.265
5^{th} %tile	0.052	0.076	0.883	0.883	0.599	0.599	-0.070
95^{th} %tile	0.101	0.131	0.918	0.918	0.805	0.805	0.798
Nearly nondifferential prior at Case 1, scenario 2							
	r_0	r_1	SP_0	SP_1	SN_0	SN_1	$logOR$
Mean	0.053	0.098	0.884	0.847	0.777	0.785	0.686
SD	0.013	0.018	0.013	0.016	0.069	0.063	0.339
5^{th} %tile	0.032	0.070	0.862	0.820	0.658	0.677	0.142
95^{th} %tile	0.076	0.129	0.906	0.873	0.877	0.880	1.262
Nearly nondifferential prior at Case 1, scenario 3							
	r_0	r_1	SP_0	SP_1	SN_0	SN_1	$logOR$
Mean	0.060	0.091	0.884	0.798	0.715	0.750	0.463
SD	0.014	0.018	0.013	0.017	0.074	0.064	0.320
5^{th} %tile	0.039	0.064	0.863	0.769	0.588	0.637	-0.049
95^{th} %tile	0.084	0.123	0.905	0.826	0.828	0.846	0.995
Differential prior at Case 1, scenario 4							
	r_0	r_1	SP_0	SP_1	SN_0	SN_1	$logOR$
Mean	0.055	0.089	0.898	0.762	0.645	0.7520	0.529
SD	0.014	0.019	0.013	0.020	0.103	0.077	0.360
5^{th} %tile	0.034	0.060	0.876	0.730	0.468	0.614	-0.050
95^{th} %tile	0.080	0.123	0.920	0.795	0.805	0.864	1.128

The sample mean, standard deviation and 90% credible interval (regarding 10000 post burn-in iterations) for each model parameter at different scenarios of Case 1 and Case 2 (each corresponding to one simulated dataset) are calculated under the appropriate prior assumptions regarding differentiality of misclassification. Results presented in Tables 5.2

Table 5.3: Posterior distributions at various scenarios in Case 2 with different prior information

Nondifferential prior at Case 2, scenario 1							
	r_0	r_1	SP_0	SP_1	SN_0	SN_1	$logOR$
Mean	0.104	0.162	0.828	0.828	0.629	0.629	0.511
SD	0.017	0.030	0.013	0.013	0.060	0.060	0.280
5^{th} %tile	0.078	0.115	0.806	0.806	0.529	0.529	0.049
95^{th} %tile	0.132	0.215	0.849	0.849	0.727	0.727	0.975
Nearly nondifferential prior at Case 2, scenario 2							
	r_0	r_1	SP_0	SP_1	SN_0	SN_1	$logOR$
Mean	0.072	0.152	0.804	0.861	0.627	0.608	0.840
SD	0.014	0.029	0.013	0.019	0.071	0.073	0.310
5^{th} %tile	0.050	0.106	0.783	0.828	0.506	0.486	0.323
95^{th} %tile	0.095	0.202	0.826	0.892	0.742	0.727	1.344
Nearly nondifferential prior at Case 2, scenario 3							
	r_0	r_1	SP_0	SP_1	SN_0	SN_1	$logOR$
Mean	0.103	0.150	0.775	0.820	0.770	0.751	0.426
SD	0.017	0.028	0.016	0.023	0.054	0.063	0.292
5^{th} %tile	0.076	0.106	0.749	0.782	0.676	0.642	-0.053
95^{th} %tile	0.132	0.199	0.801	0.856	0.851	0.845	0.910
Differential prior at Case 2, scenario 4							
	r_0	r_1	SP_0	SP_1	SN_0	SN_1	$logOR$
Mean	0.110	0.196	0.697	0.861	0.765	0.758	0.674
SD	0.017	0.032	0.017	0.027	0.058	0.069	0.274
5^{th} %tile	0.084	0.145	0.669	0.815	0.663	0.636	0.221
95^{th} %tile	0.140	0.250	0.725	0.904	0.855	0.863	1.116

and 5.3. Most sample means are sufficiently close to the true values of the parameters at different levels of exposure misclassification, except for the sensitives. The sensitivities in case and control groups are found larger than the true values in Cases 1 and 2. This should not downgrade the performance of our algorithm, because most 90% credible intervals cover the true values, except for SN_1 at scenario 2, Case 1 and a few exceptional observations in Case 2. To conclude, the Bayesian misclassification methods utilizing MCMC algorithm facilitate convergent posterior Markov chains with adequate mixing.

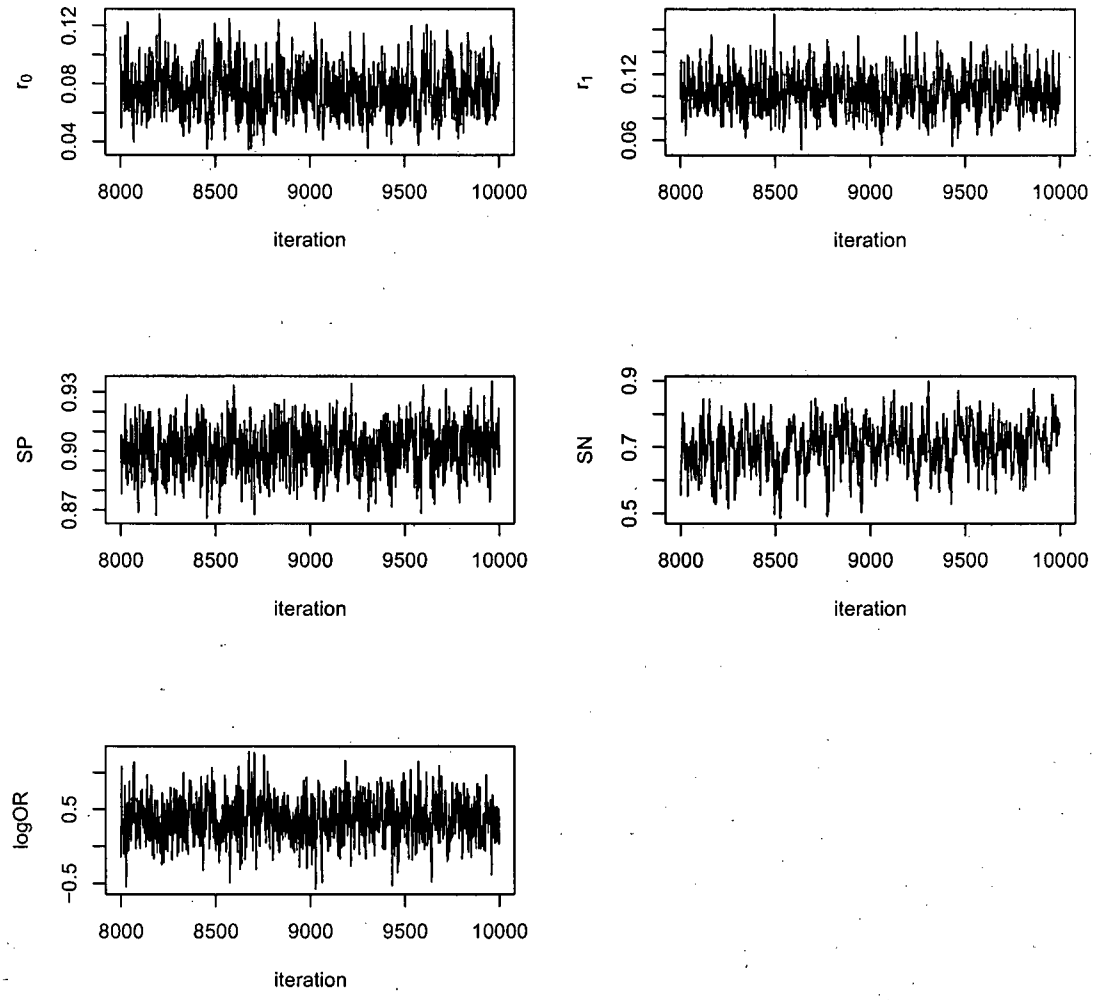

Figure 5.1: *MCMC mixing based on iterations 8000-10000 regarding a dataset in Case 1, Scenario 1, using the nondifferential Bayesian method*

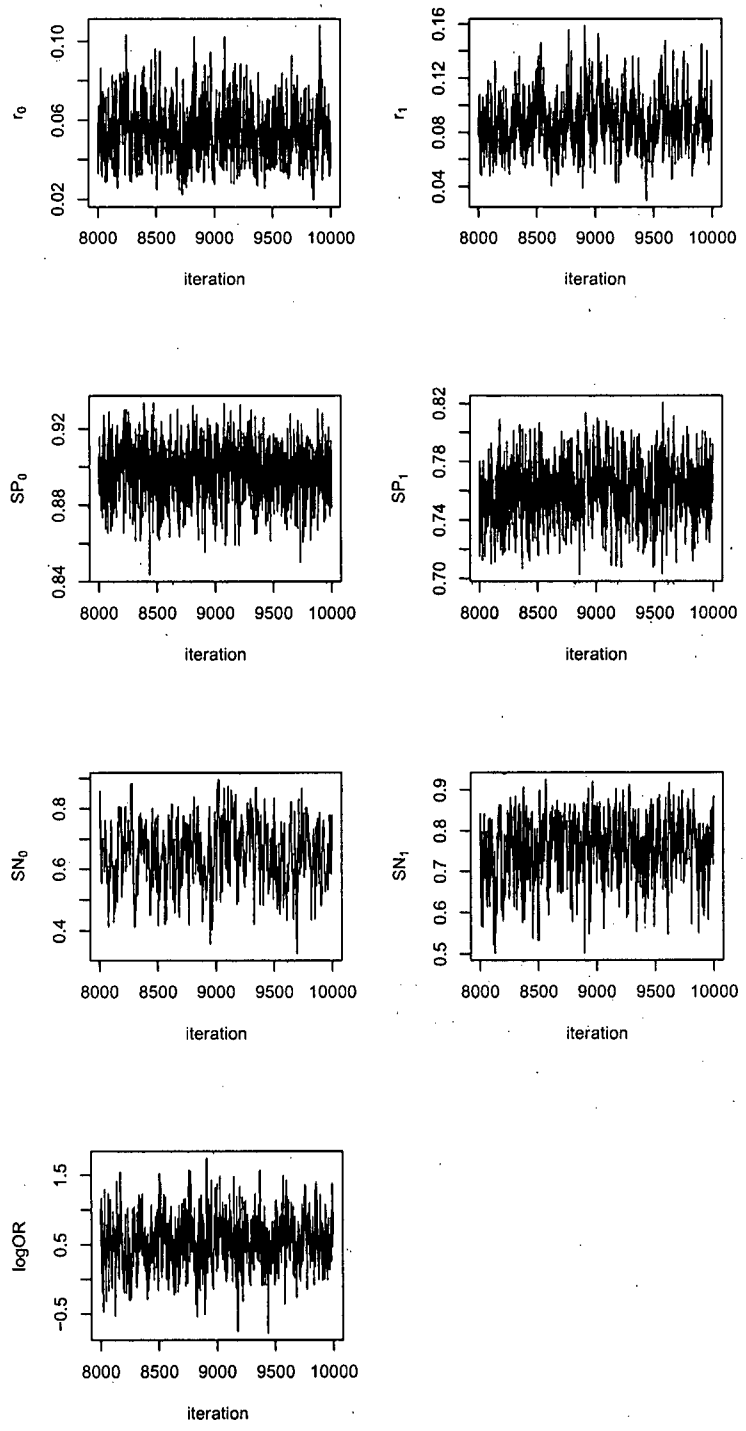

Figure 5.2: *MCMC mixing based on iterations 8000-10000 regarding a dataset in Case 1, Scenario 4, using the differential Bayesian method*

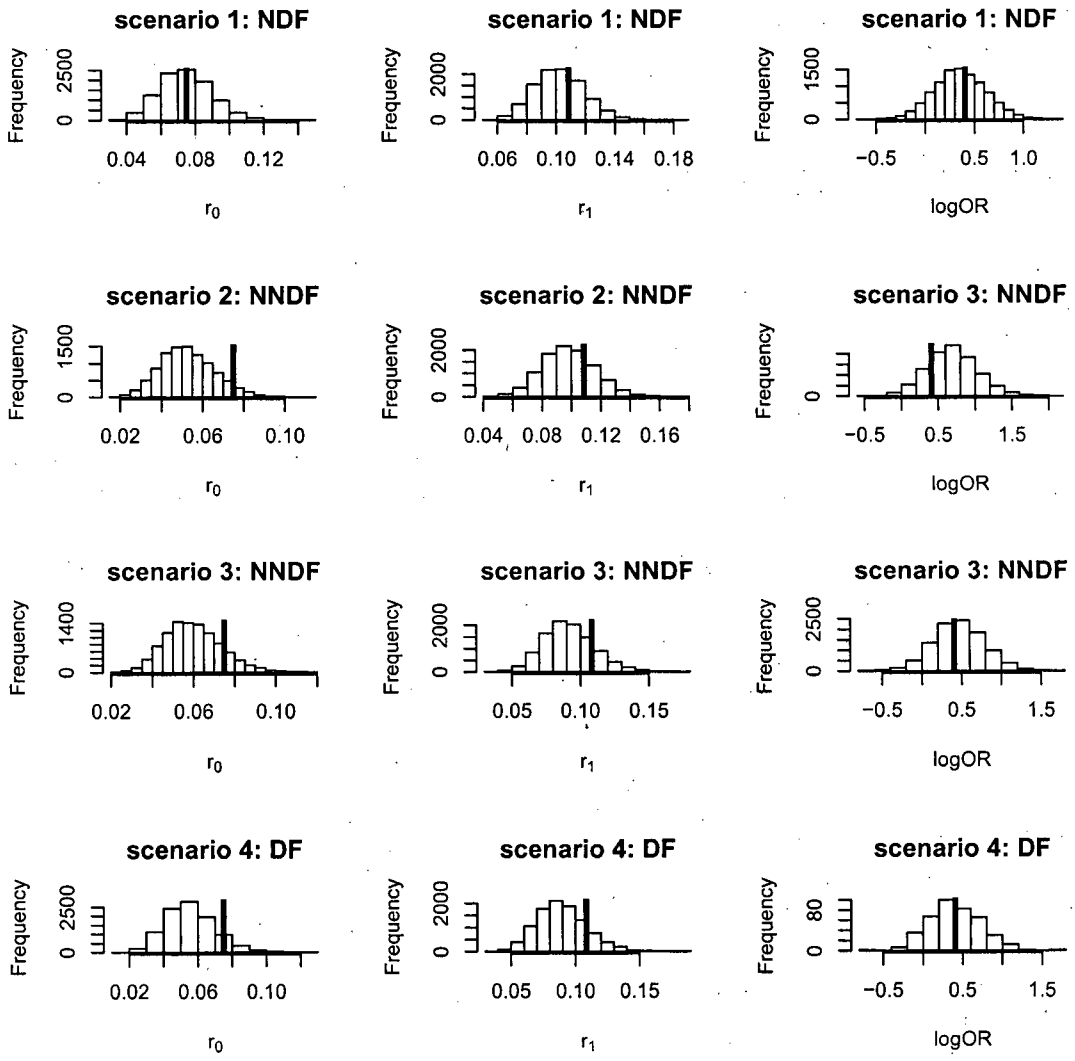

Figure 5.3: *Posterior histogram of r_0, r_1 and logOR based on 10000 iterations of Case 1*

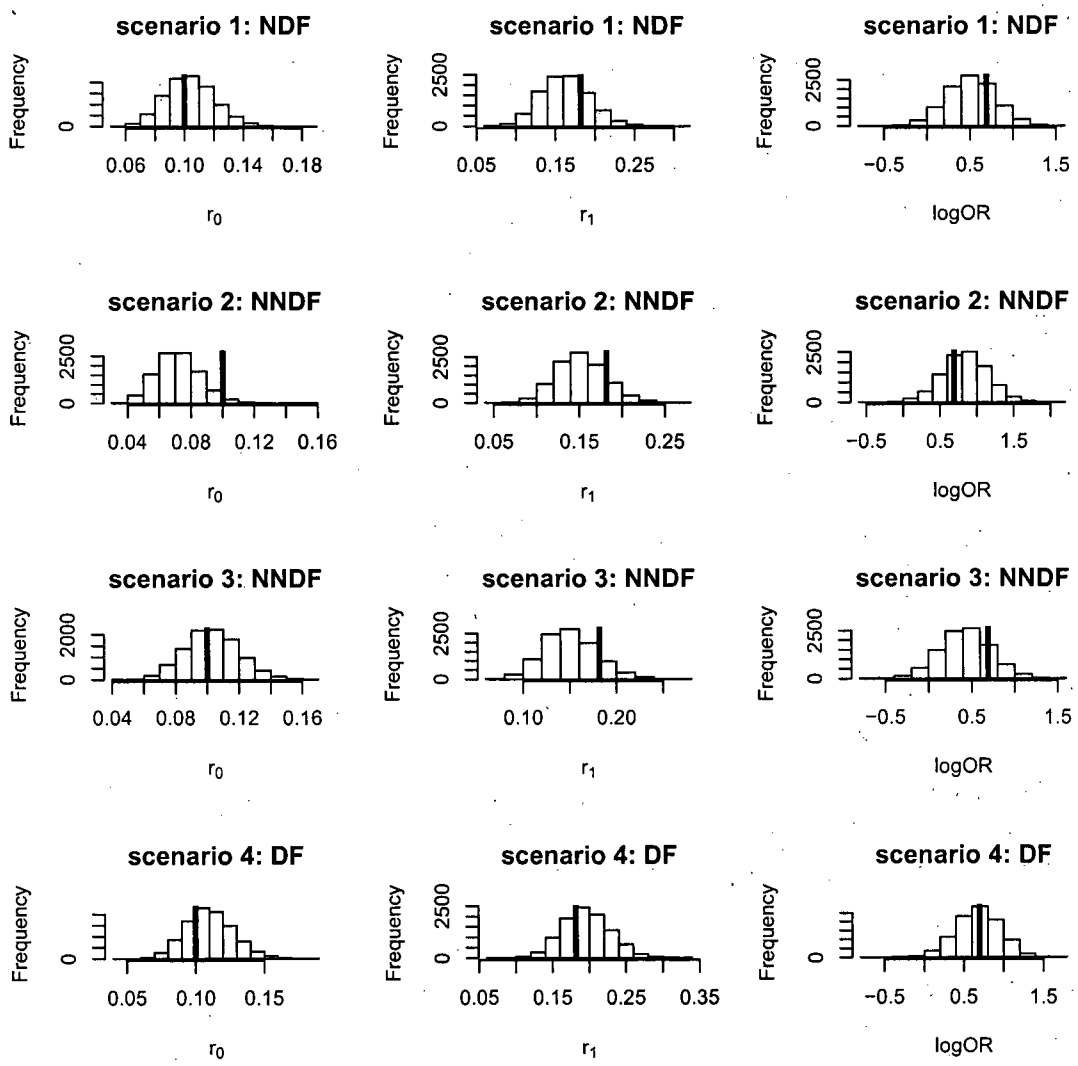

Figure 5.4: *Posterior histogram of r_0, r_1 and logOR based on 10000 iterations of Case 2*

Tables 5.2 and 5.3 also suggests that standard deviations increase with the inflation of differential misclassification. This is not unexpected because under nondifferential misclassification, data concerning cases and controls are pooled together to estimate the misclassification parameters. Whereas under differential misclassification, sensitivities and specificities are estimated separately according to data in each arm, hence resulting in less precise estimates (i.e., with larger variance).

5.4 Comparison with results obtained using MLE and MC-SIMEX

In this section, we study the comparative performance of Bayesian methods, maximum likelihood estimation methods and simulation extrapolation methods in several misclassification situations. Three methods under the assumptions of nondifferential misclassification (NDF), nearly nondifferential misclassification (Nearly NDF, only applicable to Bayesian method), and differential misclassification (DF) are applied simultaneously to data simulated in section 5.1 (4 scenarios in two cases). The estimated mean squared error (MSE), defined as

$$\widehat{MSE}(\hat{\phi}) = \frac{1}{NREP} \sum_{i=1}^{NREP} \left(\hat{\phi}_i - \phi\right)^2$$

is calculated to assess the goodness of the point estimator for $\phi \equiv logOR$, by incorporating both the precision (via variance) and accuracy (via bias) of $\hat{\phi}$. $NREP$=400 represents the number of repeated datasets generated in each scenario. The coverage proportion of the $logOR$ and average width of the 90% credible interval using Bayesian methods and the 90% confidence intervals using MLE or MC-SIMEX are also reported to evaluate the accuracy and precision of the effect estimator $\hat{\phi}$ (Table 5.4, 5.5 , 5.6 and 5.7 for Case 1

and Table 5.8, 5.9 and 5.10 for Case 2).

Table 5.4 shows the nondifferential Bayesian method has the best performance when the data is truly nondifferentially misclassified. It produces estimator with smallest bias, variation and overall error rate compared to other Bayesian methods in scenario 1. However, the related coverage proportion drops dramatically as the exposure misclassification gets more differential. The fact that the empirical coverage probability drops to 0.06 in scenario 4 illustrates the poor reliability of the nondifferential Bayesian method, for the level of differentiality involved in the misclassification is seldom well understood in practice. The differential misclassification Bayesian method behaves better than the others in terms of smaller estimated MSEs and larger coverage proportions, when data are highly differentially misclassified (scenario 3 and 4), which agrees to our expectation. In general, the differential Bayesian method is not as efficient under nondifferential misclassification, because data are split to two groups and estimations are accomplished on separate subsets. Nevertheless, it generates reliable estimations as misclassification deviates from complete nondifferential condition without losing much efficiency. The nearly nondifferential Bayesian method performs well in scenario 2 where departure from the nondifferential misclassification is mild. Separate estimation of sensitivities or specificities improves the accuracy of $\hat{\phi}$. Meanwhile by posing large positive correlation between SN_is and SP_is, it "borrows strength" from measurements in both groups and produces small variance and MSE. The MSE, coverage proportion and average width of credible interval associated with the nearly nondifferential method are always in between the other two Bayesian methods. It is a method worth exploring in practice when the mechanism of misclassification is not fully specified.

In Table 5.5, we observe analogous phenomena comparing the performances of nondifferential (NDF) MLE and differential (DF) MLE. More accurate and precise estimation of $logOR$ with smaller MSE, greater empirical coverage rate and shorter confidence in-

terval is supplied by NDF ML method in scenario 1. On the other hand, the estimates are far away from the target when the actual misclassification rates differ between cases and controls. The DF ML method yields consistent estimates with satisfactory coverage proportions ($> 88\%$) over all situations, and is especially superior at more differential scenarios such as 3 and 4. It is also found that MLEs are systematically associated with slightly greater MSE and variation compared with the Bayesian estimates in Table 5.4, which may result from the utilization of appropriate prior information.

Results obtained using differential (DF) and nondifferential (NDF) MC-SIMEX methods are summarized in Table 5.6 and 5.7. For each simulated dataset, the estimated effect parameter $\hat{\phi}$ is acquired via a quadratic extrapolation function and a loglinear extrapolation function,

$$\begin{aligned} L_{quad}(\lambda) &= \alpha_0 + \alpha_1\lambda + \alpha_2\lambda^2, \\ L_{log}(\lambda) &= \mathrm{e}^{\alpha_0+\alpha_1\lambda}. \end{aligned}$$

The variance for each $\hat{\phi}$ is estimated using the asymptotic theory and Jackknife method. Tables 5.6 and 5.7 hence summarize the performance of MC-SIMEX methods at these four combinations. Similar to Bayesian and MLE methods, the NDF MC-SIMEX works better for the nondifferential misclassification scenario, but is off-target with differential measurement errors. In Case 1, $\hat{\phi}$s achieved during reclassification at various values of λ are on average better fit by the quadratic parametric function, leading to relatively more accurate estimates of ϕ without measurement error. This is reflected by greater coverage proportions in Table 5.6 and 5.7. We also observe that estimates using either extrapolation functions bear same MSEs in each scenario, while Tukey's Jackknife variance is consistently smaller than asymptotic variance when quadratic function is applied. As the size of measurement error represented by SN_i and SP_i is not totally specified

in this study, Tukey's Jackknife method is preferred to accommodate situations where variation of $\widehat{SN_i}$, $\widehat{SP_i}$ is not negligible (Carroll et al., 2006). Compared with intervals by first two methods, the 90% confidence intervals generated by MC-SIMEX methods have apparently smaller probabilities to cover the true effect parameter $logOR$.

Table 5.4: MSEs, empirical coverages and average widths of 90% credible intervals for $logOR$ using Bayesian methods based on datasets simulated in Case 1

		NDF	Nearly NDF	DF
scenario 1	MSE	0.0773	0.0869	0.0993
	Coverage	0.9075	0.8975	0.8875
	Width	0.9247	0.9947	1.0279
scenario 2	MSE	0.1801	0.0959	0.0986
	Coverage	0.6825	0.9000	0.9075
	Width	0.9325	1.01327	1.0380
scenario 3	MSE	0.4798	0.1230	0.0946
	Coverage	0.2400	0.8550	0.8950
	Width	0.9295	1.0308	1.0501
scenario 4	MSE	0.8759	0.1545	0.1071
	Coverage	0.0600	0.8000	0.8875
	Width	0.9249	1.0478	1.0564

Similar findings present in Case 2. For in each simulated dataset, the number of controls are three time of the number of cases, data from the control group have more impact on the estimation of model parameters such as the exposure prevalences r_0, r_1 and the effect parameter $logOR$. This results in greater coverage proportions than those in Case 1 when NDF methods (Bayesian or MLE or MC-SIMEX) are incorrectly applied to the differentially misclassified data (scenario 2, 3 and 4). It is also observed that the DF MC-SIMEX method fails at some replicative datasets in scenario 2. Two possibilities may cause this problem. First, the estimated misclassification matrix $\hat{\mathbf{\Pi}}^{\lambda}$ (Equation (4.1) in Chapter 4) does not exit. This may be caused by estimation of the original misclassification matrix $\mathbf{\Pi}$ using the validation data. Küchenhoff et al. (2006) suggested use a refined approximation method (Israel et al., 2001) to estimate $\mathbf{\Pi}^{\lambda}$. Second, a few

Table 5.5: MSEs, empirical coverages and average widths of 90% confidence intervals for $logOR$ using MLE methods based on datasets simulated in Case 1

		NDF	DF
scenario 1	MSE	0.0924	0.1224
	Coverage	0.9100	0.8875
	Width	0.97706	1.1062
scenario 2	MSE	0.2258	0.1226
	Coverage	0.6650	0.9075
	Width	0.9849	1.1150
scenario 3	MSE	0.5776	0.1176
	Coverage	0.2125	0.8975
	Width	0.9800	1.1235
scenario 4	MSE	1.0361	0.1338
	Coverage	0.0575	0.8875
	Width	0.9765	1.1301

modifications are made on function "mcsimex()" to fulfil our needs to automate the estimation process of log odds ratio repeatedly under the differential misclassification. Although the performance of the modified function under boundary conditions (large misclassification rates) are not completely tested, we believe the chance that it causes a problem is very tiny.

Table 5.6: MSEs, empirical coverages and average widths of 90% confidence intervals for *logOR* using MC-SIMEX under nondifferencial misclassification based on datasets simulated in Case 1

		Quadratic extrapolation		Loglinear extrapolation	
		Asymptotic variance	Jackknife variance	Asymptotic variance	Jackknife variance
scenario 1	MSE	0.0810	0.0810	0.1078	0.1078
	Coverage	0.8875	0.8025	0.7550	0.7350
	Width	0.8929	0.7300	0.9317	0.7300
scenario 2	MSE	0.3414	0.3414	0.6136	0.6136
	Coverage	0.3600	0.2450	0.3175	0.1650
	Width	0.8704	0.6999	1.0617	0.6999
scenario 3	MSE	1.1960	1.1960	1.7361	1.7361
	Coverage	0.0100	0.0025	0.005	0
	Width	0.8561	0.6788	1.0415	0.6788
scenario 4	MSE	2.4902	2.4902	3.5249	3.5249
	Coverage	0	0	0	0
	Width	0.8607	0.6718	1.0683	0.6718

Table 5.7: MSEs, empirical coverages and average widths of 90% confidence intervals for *logOR* using MC-SIMEX under differencial misclassification based on datasets simulated in Case 1

		Quadratic extrapolation		Loglinear extrapolation	
		Asymptotic variance	Jackknife variance	Asymptotic variance	Jackknife variance
scenario 1	MSE	0.1502	0.1502	0.0829	0.0829
	Coverage	0.7700	0.6500	0.6600	0.8550
	Width	0.8950	0.7278	0.6227	0.7278
scenario 2	MSE	0.1226	0.1226	0.0877	0.0877
	Coverage	0.7875	0.6800	0.8325	0.8625
	Width	0.8637	0.6937	0.5673	0.6937
scenario 3	MSE	0.1423	0.1423	0.1992	0.1992
	Coverage	0.7175	0.6300	0.2800	0.5800
	Width	0.8319	0.6746	0.5204	0.6746
scenario 4	MSE	0.1920	0.1920	0.2995	0.2995
	Coverage	0.6075	0.5200	0.0025	0.0900
	Width	0.8194	0.6590	0.4734	0.6590

Table 5.8: MSEs, empirical coverages and average widths of 90% credible intervals for $logOR$ using Bayesian methods based on datasets simulated in Case 2

		NDF	Nearly NDF	DF
scenario 1	MSE	0.0703	0.0832	0.0965
	Coverage	0.8925	0.9000	0.8900
	Width	0.8734	0.9393	0.9627
scenario 2	MSE	0.1624	0.1049	0.0977
	Coverage	0.7300	0.8650	0.8950
	Width	0.9157	0.9536	0.9643
scenario 3	MSE	0.3317	0.1318	0.0969
	Coverage	0.4400	0.8100	0.8900
	Width	0.9471	0.9624	0.9661
scenario 4	MSE	0.5852	0.1792	0.1035
	Coverage	0.2375	0.7125	0.9025
	Width	0.9675	0.9722	0.9712

Table 5.9: MSEs, empirical coverages and average widths of 90% confidence intervals for $logOR$ using MLE methods based on datasets simulated in Case 2

		NDF	DF
scenario 1	MSE	0.0771	0.1055
	Coverage	0.8925	0.8850
	Width	0.9079	1.0215
scenario 2	MSE	0.1452	0.1080
	Coverage	0.8050	0.8950
	Width	0.9595	1.0211
scenario 3	MSE	0.2945	0.1025
	Coverage	0.5500	0.9050
	Width	0.9956	1.0211
scenario 4	MSE	0.5429	0.1006
	Coverage	0.3425	0.9000
	Width	1.0222	1.0255

Table 5.10: MSEs, empirical coverages and average widths of 90% confidence intervals for $logOR$ using MC-SIMEX under nondifferencial misclassification based on datasets simulated in Case 2

		Quadratic extrapolation		Loglinear extrapolation	
		Asymptotic variance	Jackknife variance	Asymptotic variance	Jackknife variance
scenario 1	MSE	0.1247	0.1247	0.1725	0.1725
	Coverage	0.7950	0.6650	0.7800	0.6575
	Width	0.8890	0.7193	1.0223	0.7198
scenario 2	MSE	0.7746	0.7746	0.6240	0.6240
	Coverage	0.0550	0.0200	0.0950	0.0575
	Width	0.9058	0.7172	0.6418	0.7172
scenario 3	MSE	1.9450	1.9450	1.2375	1.2375
	Coverage	0	0	0	0
	Width	0.9165	0.7145	0.4893	0.7145
scenario 4	MSE	3.4673	3.4673	1.8875	1.8875
	Coverage	0	0	0	0
	Width	0.9236	0.7099	0.4382	0.7099

Chapter 6

Application in Epidemiological Studies

6.1 The study of sudden infant death syndrome

To further illustrate how Bayesian methods, MLE methods and MC-SIMEX methods work in practice, we consider a case-control study on sudden infant death syndrome (SIDS) (Kraus, Greenland and Bulterys, 1989). It is of interest to investigate the association between the use of antibiotics during pregnancy and the occurrence of SIDS. The surrogate exposure quantity about maternal use of antibiotic (X) was measured by interview question (yes=1, no=0). Information from medical records (T) was extracted to conduct an internal validation study to assess the misclassification probabilities. The complete data is shown in Table 6.1. Ignoring possible measurement errors, the $X - Y$ log odds ratio is estimated as 0.352 with 90% confidence interval (0.141, 0.563).

Table 6.1: Validation study and main study of SIDS

	Validation data				*Main data*		
	Y=1		Y=0				
T	X=1	X=0	X=1	X=0	Y	X=1	X=0
T=1	29	17	21	16	Y=1	122	442
T=0	22	143	12	168	Y=0	101	479

Table 6.2: Estimates of model parameters in SIDS study

	NDF Bayesian			Nearly NDF Bayesian			DF Bayesian		
	Mean	SD	Rejection	Mean	SD	Rejection	Mean	SD	Rejection
r_0	0.1593	0.0200	0.0419	0.1675	0.0215	0.0481	0.1701	0.0222	0.0496
r_1	0.2194	0.0232	0.0588	0.2098	0.0234	0.0562	0.2013	0.0236	0.0589
SP_0	0.9017	0.0122	0.1987	0.9134	0.0140	0.2817	0.9216	0.0155	0.2794
SP_1	0.9017	0.0122	0.1987	0.8889	0.01661	0.1959	0.8803	0.0184	0.2717
SN_0	0.6255	0.0450	0.0241	0.6181	0.0559	0.2678	0.6378	0.0632	0.0914
SN_1	0.6255	0.0450	0.0241	0.6333	0.0515	0.2529	0.6528	0.0588	0.0199

	NDF MLE		DF MLE	
	Mean	SD	Mean	SD
r_0	0.1634	0.1981	0.1793	0.0234
r_1	0.2253	0.2649	0.2095	0.0254
SP_0	0.6031	0.6810	0.5966	0.0695
SP_1	0.6031	0.6810	0.6060	0.0653
SN_0	0.9028	0.9241	0.9255	0.0172
SN_1	0.9028	0.9241	0.8782	0.0197

Table 6.3: $\widehat{logOR}$, SD and 90% intervals for $logOR$ in SIDS study

	$\widehat{log(OR)}$	SD	90% intervals
NDF Bayesian	0.3967	0.1875	(0.0887, 0.7079)
Nearly NDF Bayesian	0.2791	0.2047	(-0.0576, 0.6135)
DF Bayesian	0.2086	0.2154	(-0.1450, 0.5650)
NDF MLE	0.3983	0.1909	(0.08442, 0.7123)
DF MLE	0.1927	0.2212	(-0.1711, 0.5566)
NDF MC-SIMEX			
quadratic, asymptotic	0.6131	0.2265	(0.2406, 0.9856)
quadratic, Jackknife		0.1772	(0.3217, 0.9045)
loglinear, asymptotic	0.7896	0.2579	(0.3654, 1.2138)
loglinear, Jackknife		0.1772	(0.4981, 1.0810)
DF MC-SIMEX			
quadratic, asymptotic	0.0365	0.2221	(-0.3288, 0.4019)
quadratic, Jackknife		0.1794	(-0.2585, 0.3316)
loglinear, asymptotic	0.2701	0.1090	(0.0908, 0.4494)
loglinear, Jackknife		0.1794	(-0.0249, 0.5651)

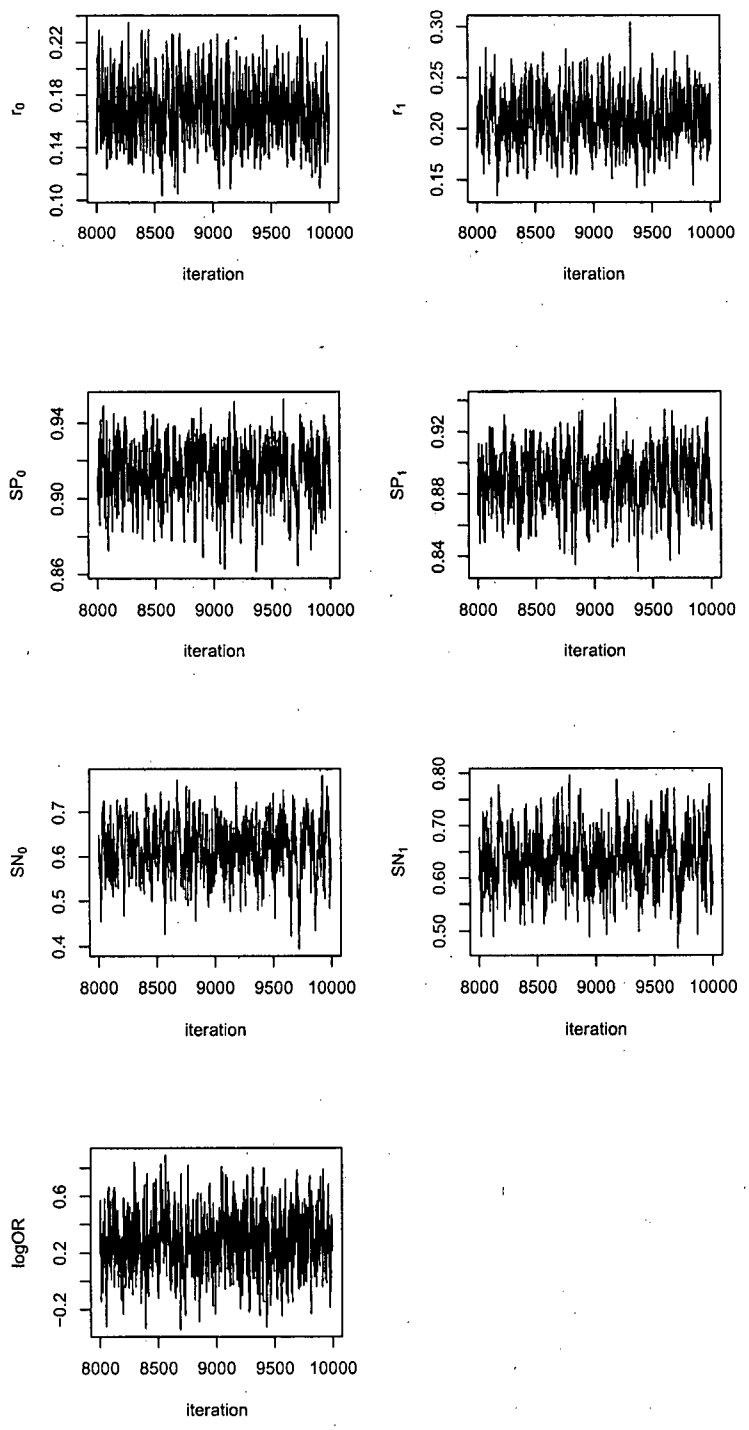

Figure 6.1: *MCMC mixing based on iterations 8000-10000 using Nearly NDF Bayesian method (SIDS study)*

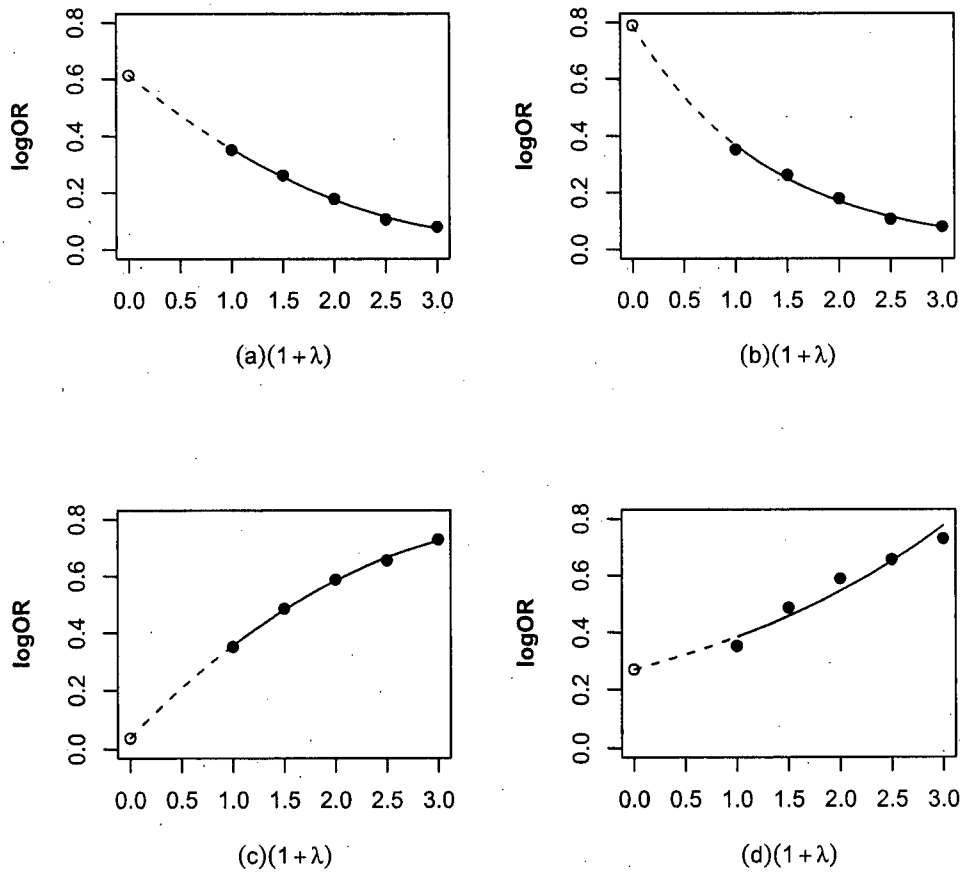

Figure 6.2: *Plots of the estimated logOR as a function of misclassification size λ in SIDS study. The upper-left panel is based on a quadratic extrapolation subject to NDF MC-SIMEX. The upper-right panel is based on a loglinear extrapolation subject to NDF MC-SIMEX. The lower-left panel is based on a quadratic extrapolation subject to DF MC-SIMEX. The lower-right panel is based on a loglinear extrapolation subject to DF MC-SIMEX.*

The Markov chains generated by Bayesian methods present adequate mixing and convergence. Figure 6.1 displays posterior simulations over the last 2000 iterations when the Nearly ND method is applied. Study results after adjustment for misclassification are presented in Tables 6.2 and 6.3. The sample means and 90% intervals regarding model parameters (r_0, r_1, SN_i, SP_i) obtained by NDF or DF Bayesian approach are close to those estimated by corresponding MLE method. Standard errors generated from Bayesian and MLE methods are similar to one another. Parameters estimated under DF misclassification assumptions are found more variable than those derived under NDF or nearly NDF situations, which is consistent with findings from simulation studies.

The point estimates of $logOR$ in form of posterior mean are greater than zero, when the Bayesian, MLE or NDF MC-SIMEX methods are applied. It appears that $log(\widehat{OR})$ by NDF Bayesian or NDF MLE is twice greater than that obtained by DF Bayesian or DF MLE algorithm. Given the equality of misclassification rates in case and control groups, data in Table 6.1 suggest a positive association between the use of antibiotic and subsequence incidence of SIDS, for 0 is not contained in the 90% credible or confidence intervals generated by NDF methods.

The estimates of $log(OR)$ returned by MC-SIMEX are somehow distinguished from those returned from Bayesian or MLE method. It is depicted in Figure 6.2 that, log odds ratio $\hat{\beta}_{1,k}$ decreases with λ_k if the measurement of drug use is nondifferentially misclassified, but increases with greater size of misclassification when SN_i or SP_i differentiates between two groups. As a result, combining the trajectory of the points, a quadratic extrapolation function produces a smaller positive log odds ratio at the origin under NDF and a smaller negative log odds ratio crossing the origin under DF. The existence of misclassification matrix $\mathbf{\Pi}^{\lambda}$ is fulfilled at $\lambda = 0.5, 1, 1.5, 2$, and two variance calculations are accomplished at each misclassification level. Because SN_i and SP_i required in MC-SIMEX are estimated from the validation study, and their variabilities are not

negligible, we consider the Jackknife variance as a more appropriate method to apply (Küchenhoff et al., 2006). One drawback of MC-SIMEX method is, it does not provide approximation of other parameters such as the exposure prevalence or misclassification rates directly. In addition, a known or adequately estimated misclassification matrix Π is required to ensure the performance of the algorithm.

6.2 The study of invasive cervical cancer

A brief profile of the case-control study investigating the impact of herpes simplex virus type 2 (HSV-2) in the progression of invasive cervical cancer is provided in Chapter 1. We adopt this study as an example to demonstrate the effectiveness of Bayesian adjustments when the differentiality of misclassification is borderline. Data for the cervical cancer study is shown in Table 6.4. It is noticeable from the validation and main data that the exposure prevalence of HSV-2 is high in both cases and controls.

Table 6.4: Validation data and main data for cervical cancer study

	Validation data				*Main data*		
	Y=1		Y=0				
T	X=1	X=0	X=1	X=0	Y	X=1	X=0
T=1	18	5	16	16	Y=1	375	318
T=0	3	13	11	33	Y=0	535	701

Ignoring measurement error arising from the inaccurate western blot procedure, the naive log odds ratio is estimated as 0.4529 (standard error = 0.0928), with 95% confidence interval (0.300, 0.606), indicating HSV-2 is positively correlated with the occurrence of invasive cervical cancer. We conduct Bayesian adjustment subject to three misclassification situations (NDF, nearly NDF and DF) via different prior distributions. Mild prior association ($\rho_1 = 0.3$) is imposed for exposure prevalences in cases and controls. The randomness and differentiality of misclassification parameters are reflected by hyperparameters, the same way as in simulation studies: $\rho_2 = \rho_3 = 1$ for NDF misclassification; $\rho_2 = \rho_3 = 0$ for DF misclassification; and $\rho_2 = \rho_3 = 0.9$ for nearly NDF misclassification.

The mixing and convergence of posterior Markov chains are satisfactory. Figure 6.3 depicts prior densities and posterior simulations regarding r_0, r_1 and $logOR$ based on 10000 posterior realizations (after 1000 burn-in iterations), under different misclassifi-

cation assumptions. The same prior distributions of r_0, r_1 and $logOR$ are applied to various differentiality conditions. Although exposure prevalences seem to be underestimated in prior densities, the posterior distributions incorporate information from the data and prior beliefs and lead to greater, distinct exposure prevalences and positive log odds ratio. Figure 6.3 illustrates that, the resulting estimates obtained using Bayesian adjustment are reasonably robust to prior knowledge, for posterior distributions generated across three conditions are sufficiently close to one another. Yet the analogous shapes of the prior and posterior densities suggest that, the prior knowledge still plays a non-negligible role in generating the resulting models. Furthermore, it is noticeable that, as the measurement error becomes more differential, $\hat{r}_0$ increases and $\hat{r}_1$ decreases slightly, resulting in small decrement of $\widehat{logOR}$, which is consistent with findings from simulation studies.

Table 6.5 presents posterior means, posterior standard deviation and 90% credible intervals (or 90% confidence intervals for non Bayesian models) for the log odds ratio describing how herpes simplex virus type 2 affect development of cervical cancer. We observe that, exposure of HSV-2 is positively associated with a growing risk of developing cervical cancer. As expected, the Bayesian and MLE adjustments in general produce similar $\widehat{logOR}$ under various misclassification assumptions. Effect estimates derived assuming more differential misclassifications move towards the unity of odds ratio, with larger variability. MC-SIMEX on the other hand generates bigger correlation, if the inaccuracy arising in western blot procedure does not depend on the incidence of cervical cancer; and smaller association if the measurement error is nondifferential. Extrapolations based on the quadratic and log linear models are displayed in Figure 6.4.

As Carroll, Gail and Lubin (1993) pointed out, there is moderate evidence to show measurement error is differential across cases and controls. Sensitivities estimated from validation data alone are 0.78 for cases and 0.5 for controls (two sided Fisher's exact

producing p-value=0.047). Nevertheless, if both the complete and incomplete data are considered, a likelihood ratio test for the nondifferentiality of misclassification with 2 degrees of freedom, generates a p-value at 0.073, indicating lack of evidence to reject the null at 5% significance level. Hence, it is more appropriate to interpret the differentiality of measurement as borderline. One advantage of Bayesian adjustment emerges in this context, as it incorporates the "in-between" scenario into consideration via nearly non-differential prior densities. By imposing reasonably large correlation coefficients (ρ_2, $\rho_3 < 1$) between sensitivities or specificities in cases and controls, the nearly ND Bayesian method compromises between loss of efficiency and accuracy associated with the "fully" differential and nondifferential adjustment methods. Furthermore, the sensitivities and specificities estimated via MLE or Bayesian methods are found consistent with the conditional probabilities of measurement errors calculated using the generalized latent variable modeling technique of Skrondal and Rabe-Hesketh (2004).

Table 6.5: $log\widehat{OR}$, SD and 90% intervals for $logOR$ in cervical cancer study

	$log(\widehat{OR})$	SD	90% interval
NDF Bayesian	0.8665	0.2047	(0.5434, 1.2175)
Nearly NDF Bayesian	0.6527	0.2700	(0.2303, 1.1066)
DF Bayesian	0.5417	0.2938	(0.0462, 1.0200)
NDF MLE	0.9579	0.2366	(0.5688, 1.3471)
DF MLE	0.6081	0.3503	(0.0318, 1.1843)
NDF MC-SIMEX			
quadratic, asymptotic	0.9106	0.1845	(0.6071, 1.2141)
quadratic, Jackknife		0.1545	(0.6566, 1.1647)
loglinear, asymptotic	1.1130	0.2426	(0.7139, 1.5120)
loglinear, Jackknife		0.1545	(0.8589, 1.3670)
DF MC-SIMEX			
quadratic, asymptotic	0.1093	0.1721	(-0.1737, 0.3923)
quadratic, Jackknife		0.1327	(-0.1090, 0.3276)
loglinear, asymptotic	0.434	0.1010	(0.2678, 0.5999)
loglinear, Jackknife		0.1327	(0.2156, 0.6521)

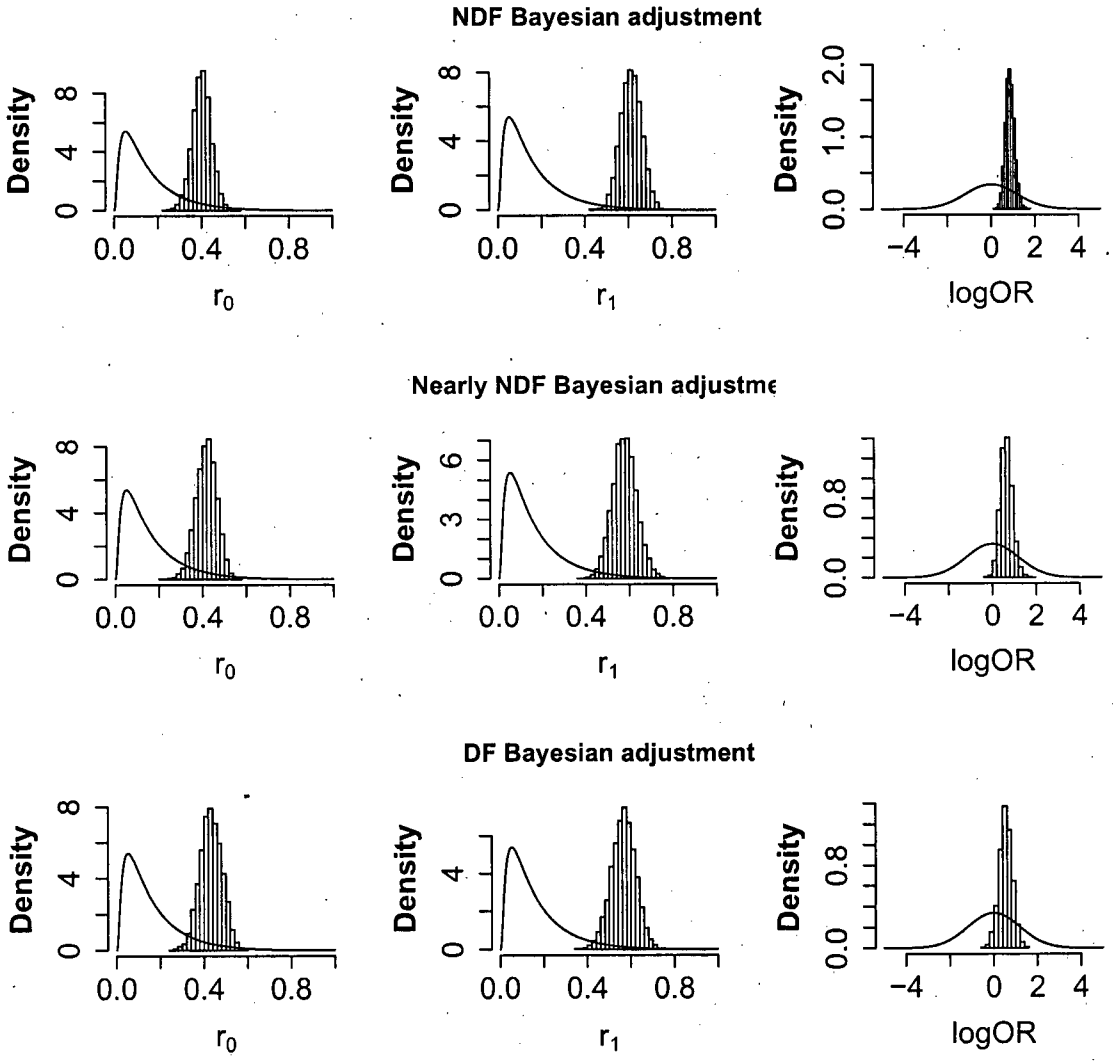

Figure 6.3: *Prior and posterior distributions of* r_0, r_1 *and* $logOR$ *subject to three misclassifications.*

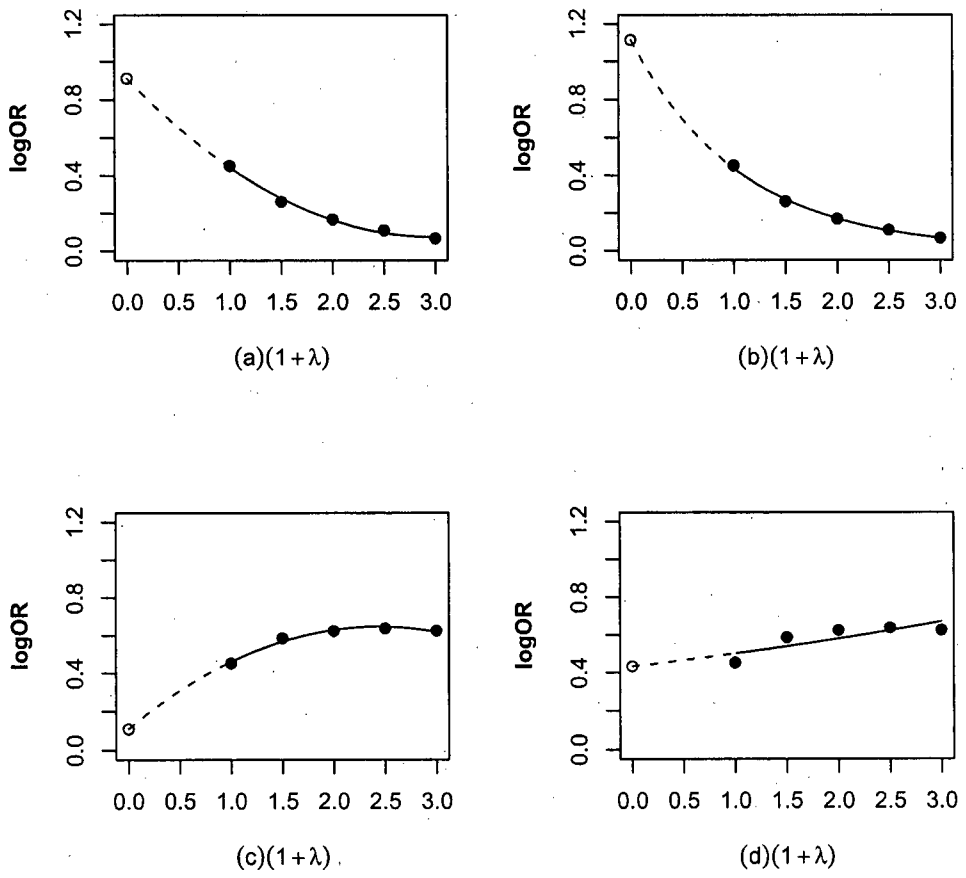

Figure 6.4: *Plots of the estimated logOR as a function of misclassification size λ in cervical cancer study. The upper-left panel is based on a quadratic extrapolation subject to NDF MC-SIMEX. The upper-right panel is based on a loglinear extrapolation subject to NDF MC-SIMEX. The lower-left panel is based on a quadratic extrapolation subject to DF MC-SIMEX. The lower-right panel is based on a loglinear extrapolation subject to DF MC-SIMEX.*

Chapter 7

Conclusion and Future Work

In this dissertation, we concentrate on comparing three adjustment models for binary exposure measurement error in case-control studies. When potential measurement error on the exposure is "ignored", statistical assessment of the impact of the exposure variable on a dichotomous health related outcome is misleading. The direction towards which the association between the actual but unobservable explanatory variable and the response is biased, is unpredictable in some circumstances.

A Bayesian model is developed to account for various types of measurement error on the covariate (i.e., nondifferential misclassification, nearly nondifferential misclassification and differential misclassification), by incorporating randomness of the exposure prevalences, sensitivities and specificities amongst cases and controls via prior distributions. The concepts and fundamental techniques of Bayesian paradigm are reviewed in Chapter 2. A hybrid algorithm alternating the Gibbs sampler and Metropolis-Hastings algorithm is proposed to sample the parameters of interest (r_i, SN_i, SP_i) i=0, 1 from the posterior distributions. Statistical inference about the odds ratio describing the correlation between the exposure and response, is made based on posterior realizations of model parameters after burn-in iterations. Other models previously developed to account for misclassification problems include the maximum likelihood estimation and simulation extrapolation method (Chapters 3 and 4). In connection to our research, we contrast model parameters and log odds ratio estimated in the Bayesian model against those generated by other methods under different misclassification scenarios, to investigate the overall

performances of the adjustment models.

Two simulation studies composed of four scenarios with gradually increasing levels of differentiality of misclassification are conducted in Chapter 5. Markov chains simulated using the Bayesian model display adequate convergence and mixing, when the prior distributions are consistent with data-generation. Our choice of the univariate proposal distribution in the MH algorithm is proportional to the likelihood function and does not depend on the current parameter value. High acceptance rates are observed in both cases, indicating the proposed hybrid algorithm is efficient. When data are simulated using equal sensitivities and specificities in diseased and non-diseased populations, the nondifferential (ND) prior behaves the best, for $log\widehat{OR}$ returned by the corresponding posterior model is associated with the smallest MSE, least variance and greatest coverage proportion. The differential (DF) Bayesian adjustment model performs better when the measurement error on the exposure is highly differential. The nearly DF prior performs well when the influence of disease status on the exposure misclassification is mild. Although separate estimation of the sensitivities and specificities under the DF assumption improves accuracy of the estimates, the DF Bayesian model has a trade-off of loss of efficiency when the exposure is actually nondifferentially misclassified. The nearly DF Bayesian model to some extent, balances the loss of efficiency and accuracy, by posing large yet less than one, positive correlation between the SN_i's and SP_i's. It is found that, the MSE, coverage proportion and average width of the 90% credible intervals obtained using the nearly ND prior distributions, are always between those associated with the completely differential or nondifferential prior densities. Hence in situations where the mechanism of misclassification is uncertain or the differentiality is borderline, the nearly ND Bayesian adjustment should be adopted to reveal the actual relationship between the disease and actual exposure variable.

Analogous phenomena regarding the performances of NDF vs. DF are detected in the

maximum likelihood estimates (MLE) and MC-SIMEX estimates. As expected, coverage proportions of the true log odds ratio associated with 90% confidence or credible intervals using the appropriate MLE or Bayesian models, are satisfactory (89%) and similar to one another. On the other hand, we notice that MLEs are associated with systematically bigger MSE and interval width, compared with the Bayesian estimates. One possible explanation is that, the properly specified prior distribution helps to reduce the variability of the effect estimator and improves the efficiency of Bayesian model. As no middle-ground scenario exists in the MLE approach, $logOR$ is estimated either through a NDF stream or a DF stream. A two-step MLE procedure, sometimes can be adapted to recover the potentially biased effect estimator. A hypothesis test for the nondifferential misclassification (the null) can first be conducted using some standard procedures, for instance, the likelihood ratio test with two degrees of freedom, based on the validation data and main data. Under guidance of the testing result, the ML estimate of $logOR$ subject to DF or NDF measurement error, will be reported as the final answer. Unfortunately, for situations where the evidence to reject the null is merely mild, say p-value is around 0.05, the two-stage procedure is incapable of providing a reliable estimate.

The extrapolated log odds ratio at zero measurement error varies with the choice of extrapolation function in a MC-SIMEX model. Although Küchenhoff, Mwalili and Lesaffre (2006) showed through some concrete examples that, the quadratic and log linear functions provided good approximation, in extrapolating $logOR$ back to the origin of no measurement error, we observe these functions sometimes return substantially distinct $\widehat{logOR}$ with variable coverage proportions. The subjective choice of an extrapolation function brings difficulty in determining an unique effect estimate in epidemiological applications and makes result interpretation ambiguous. Moreover, even if the differentiality of exposure misclassification is correctly specified, log odds ratios estimated via MC-SIMEX models tend to be more variable, having lower coverage rates and bigger

MSEs, compared with the alternatives.

The Bayesian adjustment model proposed in this dissertation can be naturally extended to other related areas. One immediate extension arises in the appearance of additional correctly recorded explanatory variables, in addition to the binary response and error-prone exposure discussed in this thesis. Previous studies show that, the impact of measurement error is more difficult to predict and the adjustment protocol becomes more complex in this case. Gustafson (2004) investigated other situation where two surrogates are employed to measure the unobserved true exposure. The likelihood functions utilized in the MLE and Bayesian models need to be reconstructed to capture information emerging from the dual error-prone covariates. Special attention should be paid to the sensitivities and specificities associated with each surrogate variable, conditioning on the binary disease status. Another interesting example is in investigation of misclassification on an exposure with more than 2 categories, for example, the exposure evaluation among three populations. It is then more appropriate to construct a posterior distribution based on a *Multinomial* likelihood instead of a *Binomial* sampling distribution. Eventually, the measurement error problems arising from combination of the above scenarios are worth exploring. Further research should be conducted to improve the validity of scientific findings in epidemiological studies.

Bibliography

Barron, B. A. (1977). The effects of misclassification on the estimation of relative risk, *Biometrics* **33**: 414–418.

Bashir, S. A. and Duffy, S. W. (1997). The correction of risk esitmates for measurement error, *Annals of Epidemiology* **7**: 154–164.

Beaton, G. H., Milner, J. and Little, J. A. (1979). Sources of variation in 24-hour dietary recall data: implications for nutrition study design and interpretation, *American Journal of Clinical Nutrition* **32**: 2546–2559.

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization, *Scientific Computing* **16**: 1190–1208.

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

Carroll, R. J., Küchenhoff, H., Lombard, F. and Stefanski, L. A. (1996). Asymptotics for the simex estimator in nonlinear measurement error models, *Journal of the American Statistical Association* **91**: 242–250.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*, Vol. 105 of *Monographs on Statistics and Applied Probability*, second edn, Chapman & Hall/CRC, Boca Raton.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, Vol. 6 of *Duxbury advanced series*, second edn, Duxbury.

Cook, J. R. and Stefansk, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error eodels, *Journal of the American Statistical Association* **89**: 1314–1328.

Eckert, R. S., Carroll, R. J. and Wang, N. (1997). Transformations to additivity in measurement error models, *Biometrics* **53**: 262–272.

Greenland, S. and Gustafson, P. (2006). Accounting for independent nondifferential misclassification does not increase certainty that an observed association is in the correct direction, *American Journal of Epidemiology* **164**: 63–68.

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*, Vol. 13 of *Interdisciplinary Statistics*, Chapman & Hall/CRC, Boca Raton.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications, *Biometrika* **57**: 97–109.

Hildesheim, A., Mann, V., Brinton, L. A., Szklo, M., Reeves, W. C. and Rawls, W. E. (1991). Herpes simplex virus type 2: a possible interaction with human papillomavirus type 16/18 in the development of invasive cervical cancer, *international Journal of Cancer* **49**: 335–340.

Israel, R., Rosenthal, J. and Wei, J. (2001). Finding generators for markov chains via empirical transition matrices, with applications to credit ratings, *Mathematical Finance* **11**: 245–265.

Jones, D. Y., Schatzkin, A., Green, S. B., Block, G., Brinton, L. A., Ziegler, R. G., Hoover, R. and Taylor, P. R. (1987). Dietary fat and breast cancer in the national health and nutrition survey 1: epidemiologic follow-up study, *Journal of the National Cancer Institute* **79**: 465–471.

Kraus, J. F., Greenland, S. and Bulterys, M. (1989). Risk factors for sudden infant death syndrome in the us collaborative penrinatal project, *International Journal of Epidemiology* **18**: 113–120.

Küchenhoff, H. and Carroll, R. J. (1997). Segmented regression with errors in predictors: semiparametric and parametric methods, *Statistics in Medicine* **16**: 169–188.

Küchenhoff, H., Mwalili, S. M. and Lesaffre, E. (2006). A general method for dealing with misclassification in regression: the misclassification simex, *Biometrics* **62**: 85–96.

Lyles, R. H. (2002). A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure, *Biometrics* **58**: 1034–1037.

Marshall, R. J. (1990). Validation study methods for estimating exposure proportions and odds ratios with misclassified data, *Journal of Clinical Epidemiology* **43**: 941–947.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines, *Mournal of Chemical Physics* **21**: 1087–1092.

Morrissey, M. J. and Spiegelman, D. (1999). Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons, *Biometrics* **55**: 338–344.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: multilevel, longitudinal, and structural equation models*, Chapman & Hall/CRC, Boca Raton.

Stefanski, L. A. and Cook, J. R. (1995). Simulation-extrapolation: the measurement error jackknife, *Journal of the American Statistical Association* **90**: 1247–1256.

Walter, S. D. and Irwig, L. M. (1987). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review, *Journal of Clinical Epidemiology* **41**: 923–937.

Wu, M. L., Whittemore, A. S. and Jung, D. L. (1986). Errors in reported dietary intakes, *American Journal of Epidemiology* **124**: 826–835.