Bayesian Methods for Highly Correlated Exposures: an Application to Tap Water Disinfection By-Products and Spontaneous Abortion

by
Richard F. MacLehose

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Epidemiology.

Chapel Hill
2006

Approved by:

Dr. Jay Kaufman, Advisor

Dr. David B. Dunson, Reader

Dr. Katherine E. Hartmann, Reader

Dr. Amy H. Herring, Reader

Dr. Charles Poole, Reader

Dr. David A. Savitz, Reader

© 2006
Richard F. MacLehose
ALL RIGHTS RESERVED

ABSTRACT

RICHARD F. MACLEHOSE: Bayesian Methods for Highly Correlated Exposures: an Application to Tap Water Disinfection By-Products and Spontaneous Abortion.

(Under the direction of Dr. Jay Kaufman.)

Highly correlated exposures are common in epidemiology. However, standard maximum likelihood techniques frequently fail to provide reliable estimates in the presence of highly correlated exposures. As a result, hierarchical regression methods are increasingly being used. Hierarchical regression places a prior distribution on the exposure-specific regression coefficients in order to stabilize estimates and incorporate prior knowledge. We examine three types of hierarchical models: semi-Bayes, fully-Bayes, and Dirichlet Process Priors. In the semi-Bayes approach, the prior mean and variance are treated as fixed constants chosen by the epidemiologist. An alternative is the fully-Bayes approach that places hyperprior distributions on the mean and variance of the prior distribution to allow the data to inform about their values. Both of these approaches rely on a parametric specification for the exposure-specific coefficients. As a more flexible nonparametric option, one can use a Dirichlet process prior which also serves to cluster exposures into groups, effectively reducing dimensionality. We examine the properties of these three models and compare their mean squared error in simulated datasets.

We use these hierarchical models to examine the relationship between disinfection by-products and spontaneous abortion. Spontaneous abortion is a common pregnancy outcome, although relatively little is known about its causes. Previous research has generally indicated an increased risk of spontaneous abortion among those who consume higher amounts of disinfection by-products. Right from the Start is a large multi-center cohort study of women who were followed through early pregnancy. Disinfection by-product concentrations were measured each week during the study, allowing for more precise exposure measurement than previous epidemiologic studies. We focus our attention on the concentrations of 13 constituent disinfection by-products (4 trihalomethanes and 9 haloacetic acids), some of which are so highly correlated that conventional maximum likelihood estimates are unreliable. To allow simultaneous estimation of effects, we implement 4 Bayesian hierarchical models : semi-Bayes, fully-Bayes, Dirichlet process

prior (DPP1) and Dirichlet process prior with a selection component (DPP2). Models that allowed prior parameters to be updated from the data tended to give far more precise coefficients and be more robust to prior specification. The DPP1 and DPP2 models were in close agreement in estimating no effect of any constituent disinfection by-products on spontaneous. The fully-Bayes model largely agreed with the DPP1 and DPP2 models but had less precision, while the semi-Bayes model provided the least precise estimates.

ACKNOWLEDGMENTS

This dissertation contains a great deal of work that I could not have accomplished without the help of many people. I am grateful to the help of my entire dissertation committee, from whom I have learned a great deal and who have suffered my frequent topic changes with great patience. My advisor and committee chair, Jay Kaufman, has been selfless with his time and advice. I have been the recipient of his tremendous knowledge of epidemiologic methods since I entered the department. This dissertation would not have been possible without the support of David Dunson, whose knowledge of biostatistics is only equalled by his willingness to share it. He has been a source of wonderful ideas and tremendous support.

Shalini Kulasingam has been a constant source of love and humor throughout this process ... not to mention a tremendous proofreader who has helpfully reviewed nearly every word I've written for the past year. Finally, it should go without saying, but warrants saying anyway, that none of this could have been possible without the support of my parents, Ruth and Len.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

BACKGROUND

1.1 Spontaneous Abortion

Spontaneous abortion is defined as a pregnancy loss prior to 20 weeks of completed gestation. The exact risk of spontaneous abortion is unknown, largely because of difficultly in detecting early pregnancy. However, spontaneous abortion is well known to be a common occurrence in pregnancy, with over 30% of all pregnancies ending in a loss and roughly 20% of all pregnancies ending in loss before they are clinically detectable.(Wilcox et al., 1988) Risk of spontaneous abortion remains high (roughly 1.0% each week) through the 12^{th} week of gestation and then rapidly declines.(Goldhaber and Fireman, 1991)

Given the high prevalence of spontaneous abortion, it is surprising that so little is known about its causes. Increased risk of spontaneous abortion has consistently been associated with advanced maternal age and prior spontaneous abortion.(Coste et al., 1991; Osborn et al., 2000) Smoking has also been associated with an increased risk of spontaneous abortion.(Coste et al., 1991; Harlap and Shiono, 1980; Ness et al., 1999; Windham et al., 1992) In addition, high levels of maternal lead exposure and paternal occupational exposures such as mercury, lead, and solvents have been associated with increased risk of spontaneous abortion.(Hertz-Picciotto, 2000; Lindbohm et al., 1991a,b; Savitz et al., 1994; Taskinen et al., 1989) Although a number of studies have found an association between caffeine consumption and spontaneous abortion, potential recall bias and difficulty with exposure measurement have left any conclusion uncertain.(al Ansary and Babay, 1994; Cnattingius et al., 2000; Fenster et al., 1991, 1997; Hansteen, 1990; Infante-Rivard et al., 1993; Kline et al., 1991; Mills et al., 1993; Parazzini et al., 1991, 1998; Signorello and McLaughlin, 2004; Srisuphan and Bracken,

1986; Wen et al., 2001)

1.2 Disinfection Process

One of the first uses of chlorine as a disinfectant was by Semmelweis who reduced the transmission rate of puerperal fever by hand-washing with chlorine. Following John Snow's research on the cause of cholera in London in 1850, interest was raised in finding ways to provide safe, uncontaminated drinking water. Indeed, Snow himself added chlorine to the Broad street pump in an effort to eliminate cholera. Thirty-one years later, Koch formally demonstrated the anti-microbial properties of hypochlorite. In 1902, the public water supplies in Middelkerke, Belgium began to be routinely treated with chlorine. The first municipality to adopt chlorination in the United States was Jersey City, New Jersey in 1908.(White, 1999) Since then, routine disinfection of water has become standard, although the type of disinfection varies among municipalities.

Water disinfection has become much more sophisticated over the past century, although no single approach to water disinfection is used by all municipalities in the United States. Federal law stipulates drinking water standards that must be met by all public water systems (for instance, the allowable concentration of arsenic or coliform). Each system can meet these requirements in different ways. Some systems, particularly those served by ground water (roughly 30% of the US population) may need less disinfection than public water systems that get their water from surface areas (such as lakes or reservoirs). Generally, however, water treatment proceeds through a series of three steps: removal of solids from the water, primary disinfection and residual disinfection. The removal of solids from the water may proceed by the addition of a coagulant (such as alum or iron salts) which precipitates suspended matter out of the water. Filtration can then remove the precipitates as well as smaller solids that did not precipitate out. Primary disinfection is generally accomplished through the addition of chlorine (either free chlorine or chloramines) or ozone. Chlorine is the traditional method of disinfection and has different biocidal properties for different organisms. It is generally more effective at low pH's and appears to act through disruption of nucleic acid and the cell wall.(Dennis et al., 1979; Haas and Engelbrecht, 1980a,b; Venkobachar et al., 1975) Recently its inability to eliminate all pathogens (particularly *Cryptosporidium*), as well as the potential effect of disinfection by-products, has caused concern. Ozone has more recently been used as an alternative to chlorine in the primary disinfection process due to its extreme toxicity to organisms, including *Cryptosporidium*. The chem-

ical composition of ozone, however, makes it very unstable and insoluble in water and therefore ozone provides little or no continued disinfection after the water leaves the treatment facility. To prevent contamination of the newly treated water while it flows through the pipes, water treatment facilities commonly put a small amount of chlorine (again, either free chlorine or chloramines) into the water supply before it leaves the facility. In equal concentrations, chloramines (a mixture of NH_2Cl, $NHCl_2$, and NCl_3) are less effective in killing bacteria and viruses but also less likely to combine with organic material and form disinfection by-products than free chlorine, which has led to its widespread use.(Hoff, 1986) Finally, in order to remove any residual biotic growth on pipes downstream of the treatment facility, many public water systems introduce higher concentrations of free chlorine for a short time each year.

The use of disinfectants in the water supply has led to dramatic decreases in the incidence of typhoid, paratyphoid, cholera, legionnaire's disease, and dysentery. However, the addition of these disinfectants has not been without controversy: chlorination of water supplies has led to many law suits, which ended in the courts upholding the rights of the state to disinfect the water supply by routine use of chlorination in order to better protect the public health.

1.3 Disinfection By-products

Chlorine is a halogen that, in nature, is always found in combined form. Its propensity to react with other molecules enables it to kill microbes and viruses in the water supply; it also enables it to react with inanimate organic matter. The most common source of organic matter in the water supply is decaying vegetation, but microbes and algae contribute significant amounts as well. Because of this, organic matter is common in surface water, but uncommon in ground water. In 1974, two groups of researchers identified disinfection byproducts in water treated with chlorine.(Bellar et al., 1974; Rook, 1974) It is now recognized that chlorine reacts with organic matter commonly found in surface water to produce a large number of disinfection by-products. Two classes of disinfection by-products are of interest to us; the first are halogenated methanes, or trihalomethanes (THMs): chloroform ($CHCl_3$), bromodichloromethane ($CHBrCl_2$), chlorodibromomethane ($CHBr_2Cl$) and bromoform ($CHBr_3$). The second are halogenated acetic acid, or haloacetic acids (HAAs): monochloroacetic acid (ClAA), monobromoacetic acid (BrAA), dichloroacetic acid (Cl_2AA), bromochloroacetic acid (BrClAA), dibromoacetic acid (Br_2AA), trichloroacetic acid

(Cl_3AA), bromodichloroacetic acid ($BrCl_2AA$), dibromochloroacetic acid (Br_2ClAA), and tribromoacetic acid (Br_3AA).

Following the discovery of disinfection by-products in the water supply, epidemiologic studies began to examine potential adverse outcomes associated with disinfection by-products. These studies initially focused on the effect of disinfection by-products (particularly THMs) on different types of cancer. Increased risk of bladder cancer and to a lesser extent rectal and colon cancer have been associated with increased consumption of disinfection by-products.(Crump and Guess, 1982; Mughal, 1992; Villanueva et al., 2004) Other studies of disinfection by-products and reproductive health have linked THM4, chloroform and bromodichloromethane to intrauterine death, stillbirth and miscarriage.(Aschengrau et al., 1989; Bove et al., 1995; Dodds et al., 2004, 1999; King et al., 2000; Savitz et al., 1995)

1.4 Animal Studies

Animal studies provide some insights into the potential mechanisms by which disinfection by-products cause spontaneous abortion although most of the exposures in these studies occur at doses thousands of times higher than humans could ever be exposed to. The effect of trihalomethanes on reproductive outcomes in rats has been studied the most extensively. Very high levels of chloroform have not shown teratogenic effects but have been shown to have fetotoxic effects and reduce fetal weight.(Murray et al., 1979; Palmer et al., 1979; Ruddick et al., 1983; Schwetz et al., 1974; Thompson et al., 1974) Chloroform has also been shown to have a toxic effect on the kidney, liver, sex organs and bone marrow.(Palmer et al., 1979) Chlorodibromomethane has also been shown to have a fetotoxic response in rats.(Ruddick et al., 1983) Bromodichloromethane, which had the strong effect on spontaneous abortion in Waller et al., and bromofrom have been shown by some studies to decrease the viability of offspring; however, other studies have indicated little effect of either.(Gulati et al., 1989; Narotsky et al., 1997; Ruddick et al., 1983) Bromodichloromethane has been shown to change sperm morphology.(Klinefelter and Linder, 1996) Halogenated acetic acids have been less well studied but exposure to them reduces fetal body weight and changes neural tube development in utero.(Hunter et al., 1996; Smith et al., 1992, 1988)

1.5 Previous Research on Disinfection By-products and Spontaneous Abortion

During 1980 and 1981 an industrial spill from a semiconductor manufacturer leaked solvents into the groundwater of Santa Clara County, California. An investigation into whether exposure to these solvents could explain a cluster of spontaneous abortions in the community revealed that hypothesis to be highly unlikely. Surrounding communities with much higher levels of solvent exposure did not have an increased rate of spontaneous abortions. (Deane et al., 1989; Wrensch et al., 1990) However, during these studies the investigators noted that women who drank tap water had an increased risk of spontaneous abortion, relative to women who drank bottled water.(Deane et al., 1989) In order to more thoroughly investigate this surprising finding, five studies were conducted examining the association between drinking tap-water and risk of spontaneous abortion. Two retrospective cohort studies found the strongest associations: Deane et al. reported that increased consumption of tap water was associated with an increased risk of spontaneous abortion, with an Odds Ratio (OR) of 3.4 and 95% Confidence Interval (CI) of (0.6, 19.4).(Deane et al., 1989) Wrensch et al. found that relative to not drinking tap water, drinking tap water was associated with an increased risk of spontaneous abortion (OR= 6.9, 95%CI: 2.7, 17.7).(Wrensch et al., 1990, 1992) Windham et al. conducted a case-control study and found a moderate increase in risk of spontaneous abortion among women who reported any consumption of cold tap water (vs. none) (OR=1.2, 95%CI: 1.0, 1.5).(Windham et al., 1992) Fenster et al. found a moderately decreased risk of spontaneous abortion among women who drank tap water relative to non-drinkers of tap water and also noted evidence of reporting bias among women in their study.(Fenster et al., 1992) Finally, in a case-control study, Hertz-Picciotto et al. found that the relationship between tap water consumption and spontaneous abortion depended on whether respondents were interviewed over the phone (in which case a positive association was found: OR=2.2, 95% CI: 1.4, 3.6) or through the mail (in which case a much diminished association was found: OR=1.3, 95% CI: 0.8, 2.0).(Hertz-Picciotto et al., 1992) A review of these articles noted that the effects of the two retrospective studies that showed the largest association may have been due to recall bias since both studies mentioned the well-publicized solvent spill and its potential effect on spontaneous abortion in a letter to study subjects.(Swan et al., 1992) Another study published during the same year, but conducted in a different part of California, also observed an increased risk of spontaneous abortion among tap-water

drinkers relative to bottled-water drinkers (RR=2.2, 95%CI (1.3, 3.6)). (Aschengrau et al., 1989)

All of these studies are limited in their exposure assessment; none attempted to measure the amount of disinfection by-products in the water, with most relying simply on consumption of tap-water as a surrogate. In 1995, Savitz et al. used quarterly averages of THM levels to measure the effect of THM consumption on spontaneous abortion in a case-control study in central North Carolina. They found a modest increase in the odds of spontaneous abortion (OR=1.7; 95% CI: 1.1, 2.7) for each 50 part per billion unit increase in THM level.(Savitz et al., 1995) In a prospective cohort study, Swan et al found an increased risk of spontaneous abortion among women who drank more than 5 glasses of tap water per day (OR=2.2, 95%CI: 1.2-3.9), however this result was only found in one region of their study. Waller et al. furthered these findings by assigning a THM level to each woman in the study, equal to the reported THM level from each woman's water service provider. They found that of the four trihalomethanes, $CHBrCl_2$ was associated with an increased risk of spontaneous abortion (OR=2.0, 95%CI: 1.2-3.5).(Waller et al., 1998) Waller et al. and a previous analysis of Right from the Start by Savitz et al. remain the only study that has examined constituent disinfection by-products rather than the aggregate measures of THM or glasses of water consumed.(Savitz et al., 2005; Waller et al., 1998)

While these studies generally indicate a positive association between disinfection by-products and spontaneous abortion, reaching a conclusion about whether the association is causal is hindered by limitations in each of the studies. The positive associations found in earlier studies are consistent with the hypothesis that women who experienced a spontaneous abortion may be more likely to recall (or perhaps overestimate) the amount of tap water they consumed.(Deane et al., 1989; Hertz-Picciotto et al., 1992; Neutra et al., 1992; Petitti, 1992; Swan et al., 1992; Wrensch et al., 1992; Zierler, 1992) Only two studies examined disinfection by-product levels in the water, and of those only Waller examined individual disinfection by-products.(Savitz et al., 1995; Waller et al., 1998) However, the studies that did examine disinfection by-products only determined exposure level based on the level reported from the water utility in their quarterly report, leaving the possibility of substantial misclassification.

1.6 Highly Correlated Data in Epidemiology

Because of the high proportion of pregnant women who are exposed to disinfection by-products through tapwater, any effect of disinfection by-products on spontaneous abortion could have enormous public health implications. Unfortunately, efforts to measure the effect of the 13 constituent disinfection by-products (4 THMs and 9 HAAs) on spontaneous abortion are hindered by the high correlation between the disinfection by-products. The amount of chlorine in the disinfection process, the amount of organic matter in the water supply, and the amount of bromide in the water supply all effect the concentration of the 13 disinfection by-products. These common latent factors not only cause a high correlation but also serve to confound the effect of any one of the 13 constituent disinfection by-products unless the remaining 12 are controlled for (Figure 1.1). Unfortunately, common approaches to controlling confounding, such as maximum likelihood regression, perform poorly in precisely this setting.

Highly correlated exposure data frequently arise when the multiple exposures are caused by a single, but frequently latent, factor. Such problems with high correlation are common in epidemiology. For instance, in nutritional epidemiology vitamins and nutrient levels will commonly be highly correlated because of food preferences by individuals. In epidemiologic studies of pesticides, the exposure to certain chemicals may be correlated because they are common to multiple pesticides. Occupational exposures may also be highly correlated since a person's occupation typically dictates exposure to multiple chemicals.

1.7 Common Methods for Correlated Data

The most common approach to modeling the effect of some exposures on disease in epidemiology is by using a maximum likelihood logistic regression model. Unfortunately, in the presence of highly correlated data, the maximum likelihood logistic model can produce extremely unstable estimates or even fail to converge.(Hosmer and Lemeshow, 1989) Epidemiologists have tried a variety of approaches to avoid this scenario. A common approach is to estimate the effect of one exposure at a time, leaving all other exposures out of the model. This approach produces a much more stable estimate that will be unbiased if the correlated variables are not also confounders, but it will produce biased estimates when the correlated variables are confounders. For instance, in the disinfection by-product example, if each of the 13 disinfection by-products have

an effect on spontaneous abortion and all are caused by a common unmeasured factor, then any given exposure is confounded by the remaining 12. A regression model that estimates the effect of only one disinfection by-product, while excluding the other 12, will therefore produce confounded estimates of effect. An alternative approach is to collapse the correlated exposure variables into a summary statistic, such as the mean or a weighted average. Such an approach, while generally allowing the maximum likelihood logistic regression to converge, is unappealing since it makes interpretation difficult and can mask important individual effects in the data. For instance, if only one of the 13 disinfection by-products has an effect, an exposure metric that is a weighted average of all 13 disinfection by-products will show a diluted, and possibly difficult to detect, effect. Previous analyses of disinfection by-product data have generally adopted this approach, collapsing the constituent disinfection by-products into categories such as THMs or HAAs.

Problems with collinearity have motivated a number of alternatives to maximum likelihood estimation. An early approach was *ridge regression*, which modifies maximum likelihood estimation by including a penalty, k, for large negative or positive values of the regression coefficients.(Hoerl and Kennard, 1970a,b) This penalty can be shown to correspond to the inverse of the variance of a normal prior distribution on the regression coefficients, so that ridge regression is a type of Bayesian estimator.(Lindley and Smith, 1972) When $k = 0$, there is an infinite prior variance (no penalty) and $\beta^{RG} = \beta^{MLE}$ (RG=ridge regression estimate, MLE=maximum likelihood estimate); however with $k > 0$, ridge regression coefficients will be shrunk toward zero and have smaller variance than the MLEs.(Hoerl and Kennard, 1970b) As illustration, consider a normal linear regression

$$E(Y_i|x_{i1}, x_{i2}) = \beta_1 x_{i1} + \beta_2 x_{i2} \tag{1.1}$$

with the predictors x_{i1} and x_{i2} having a bivariate normal distribution with unit variance and correlation 0.9. Supposing $\beta_1 = 2$ and $\beta_2 = 2$, we plot the distribution of the MLE and ridge regression estimator for a sample of 100 subjects in Figure 1.2. The high negative correlation of the MLEs is clear from the figure. Ridge regression coefficients exhibit far less correlation and less overall variance. The MLE are centered very close to the true value , while the ridge regression estimates are shrunk slightly towards zero. It will not always be the case that the MLE will be close to the true value; in highly correlated problems, the MLE could be far from the true values. However,

results from this example are indicative of the general improved performance for ridge regression relative to MLE: while MLEs are asymptotically unbiased, their variance can be enormous and their mean squared error (MSE) is worse than the MSE for ridge regression estimates, which are slightly biased but have a greatly decreased MSE.(Hoerl and Kennard, 1970b; Strawderman, 1978)

1.8 Hierarchical Models

Although ridge regression has only seen limited use in epidemiology, it represents a special case of a broader type of model that has seen some use: hierarchical models. Hierarchical models are those that define model parameters in an ordered structure. For instance, a basic linear regression such as that in equation 1.1 models the random variable y_i conditional on parameters β_1 and β_2. These parameters can in turn be modeled conditional on other parameters (called hyperparameters), for example $\beta_i \sim N(\mu, \phi^2)$, where N is a normal distribution with mean μ and variance ϕ^2. In the ridge regression example, $\mu = 0$ and $\phi^2 = 1/k$. Ridge regression stops at this level of the hierarchy but the hyperparameters (μ and ϕ^2) can in turn be modeled conditional on still other parameters, for example: $\mu \sim N(\psi, \zeta)$ and $\phi^2 \sim IG(\alpha_1, \alpha_2)$, where IG is the inverse gamma distribution. A parameter, conditional on the parameters one level above it in the hierarchy, is independent of other parameters. For instance, after accounting for β_1 and β_2, the parameters μ, ϕ^2, α_1 and α_2 contain no information about y_i.

Hierarchical models represent a natural way to formulate problems in epidemiology. For instance, consider the problem of estimating the effect of disinfection by-products on spontaneous abortion. A natural first step is to model the outcome, y_i, conditional on the effects of the disinfection by-products, $\beta_1 \ldots \beta_{13}$: $h(\Pr(y_i)) = \beta_0 + x_1\beta_1 + \cdots + x_{13}\beta_{13}$, where $h(\cdot)$ is a function such as the logit. In turn, $\beta_1 \ldots \beta_{13}$ can be modeled as a function of hyperparameters: $\beta_j \sim N(\mu, \phi^2)$. The parameter μ can be a function of other variables (such as an indicator for whether the chemical is brominated), to incorporate information about how the β_j varies over those variables (for instance, brominated disinfection by-products may have a different effect than non-brominated disinfection by-products). Although hierarchical models are not necessarily Bayesian, they lend themselves easily to Bayesian interpretation. For instance, the distribution placed on β_j is the prior distribution and μ incorporates our belief about the size of the effect of the j^{th} disinfection by-product and ϕ^2 is our uncertainty regarding that effect

size.

Hierarchical models are becoming more common in epidemiology. They have seen use investigating the association between occupational exposures and neuroblastoma, between pesticide exposure and neuroblastoma, between genotypes and bladder cancer, and between nutrition and breast cancer.(De Roos et al., 2001; Hung et al., 2004; Kirrane et al., 2005; Witte et al., 1994) However, these models all represent the most basic Bayesian hierarchical model: one with only two levels. Such models have been referred to as semi-Bayes models.(Greenland, 1992, 1993, 1994; Greenland and Poole, 1994) However, the hierarchical framework lends itself to being easily expanded past two levels. Specifying additional levels can allow for large gains in parameter precision and, paradoxically, can limit the reliance of model estimates on user specified parameters (such as μ and ϕ in the semi-Bayes model).

1.9 Summary

Disinfection by-products have been frequently (though not consistently) associated with spontaneous abortion in both toxicologic and epidemiologic studies. Studies with improved exposure measurement may help elucidate the possible etiologic effect disinfection by-products have on early pregnancy loss. Previous research has generally aggregated constituent disinfection by-products into categories and analyzed the effects of these categories on spontaneous abortion. Greater interest may focus instead on the effect of the constituent disinfection by-products. However, in order to estimate the effect of any single constituent disinfection by-product, the remaining 12 must be included in the model, since failure to include them could result in a confounded estimate of effect. The 13 constituent disinfection by-products are highly correlated with one another and standard epidemiologic analytic techniques perform poorly in this arena. We suggest the use of four related hierarchical models for correlated exposure data: semi-Bayes, fully-Bayes, and two semi-parametric models. We compare the properties of these four hierarchical models and implement them in a study examining the effect of disinfection by-products on spontaneous abortion.

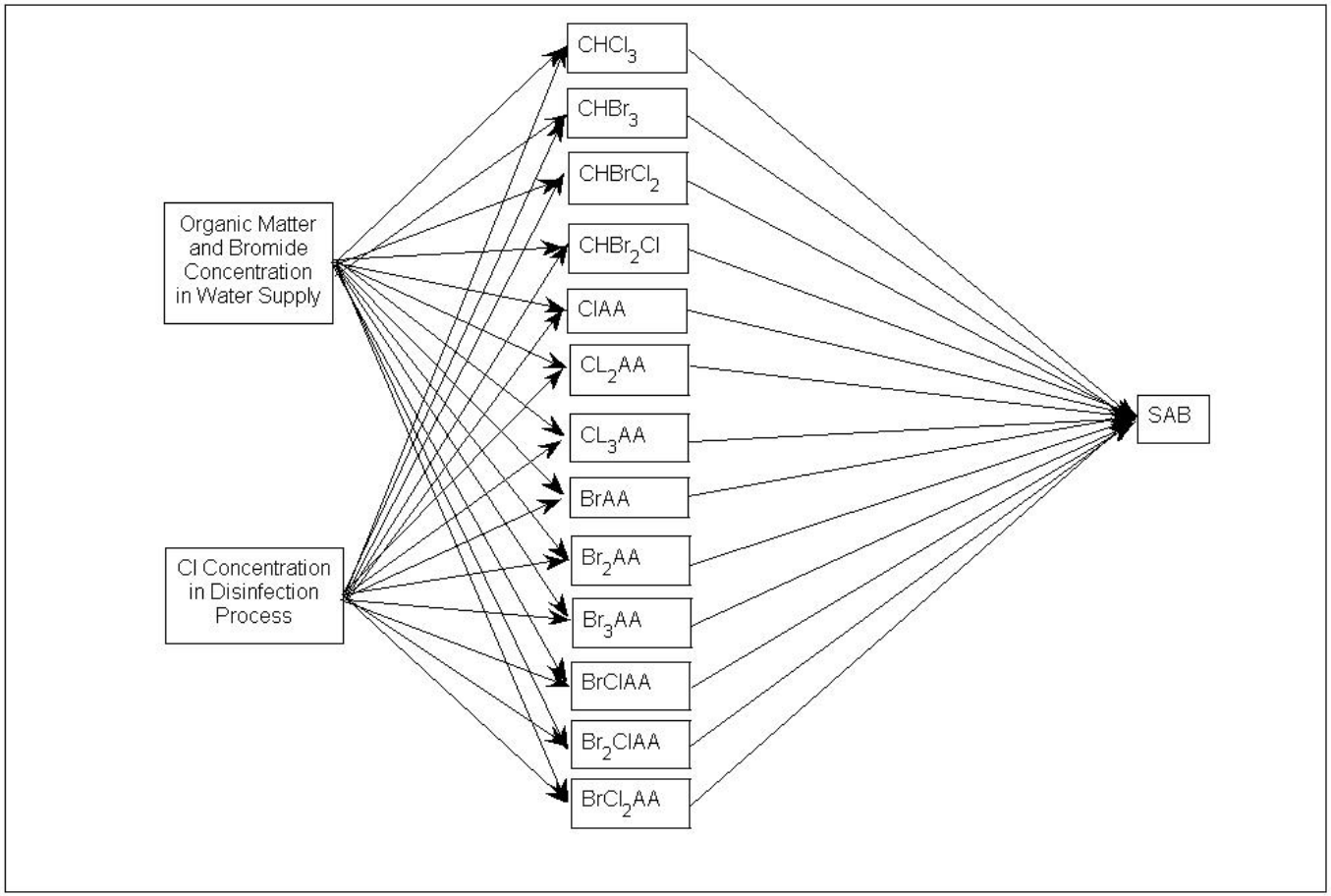

FIGURE 1.1: Directed acyclic graph depicting the relation between disinfection by-products and spontaneous abortion.

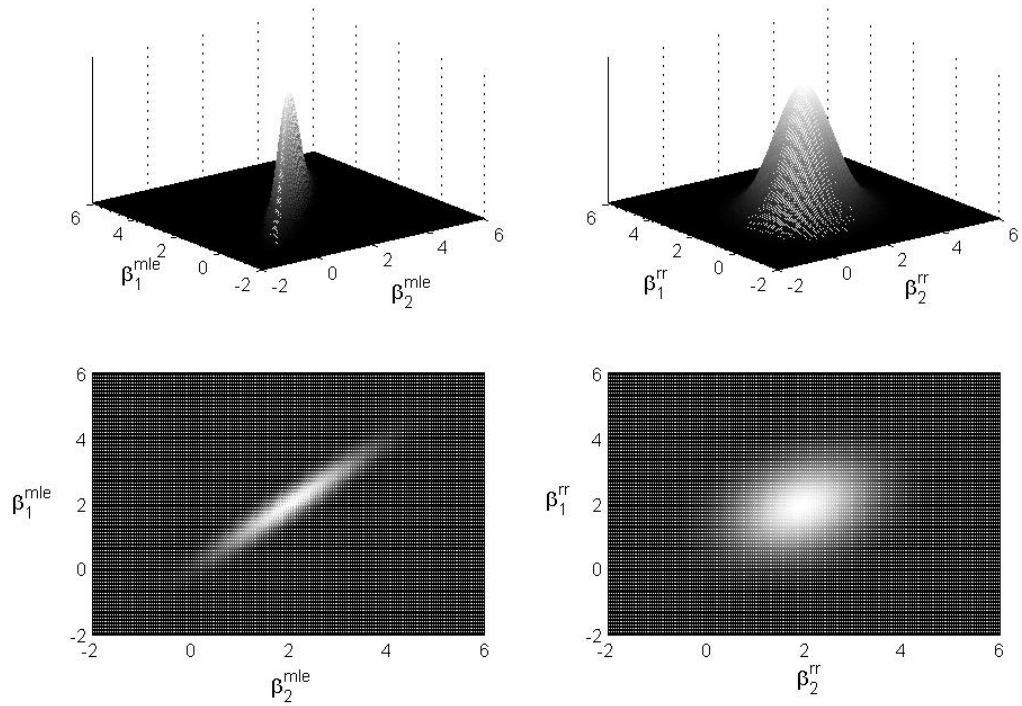

FIGURE 1.2: Distribution of MLE and ridge regression coefficients.

CHAPTER 2

METHODS

2.1 Overview of Right From the Start

This dissertation implemented four hierarchical models (semi-Bayes, fully-Bayes and two semi-parametric models) using data from a recently concluded study of pregnancy: Right from the Start(PI: Dr. David Savitz). Right from the Start was a prospective cohort study examining the effect of disinfection by-products on spontaneous abortion. The study was conducted from 2001-2004 and enrolled an ethnically diverse cohort of women over 18 from 3 study sites with different disinfection by-product distributions. Site 1 drew its water supply from a lake. The water was initially disinfected with ozone when it first reached the water treatment plant and then treated with chloramines before it left the plant. A low concentration of bromides together with a high concentration of organic matter caused the disinfection by-products in site 1's water supply to be distributed most heavily towards the chlorinated THMs and HAAs. For the entire month of March each year, site 1 added free chlorine to its water supply in order to disinfect the pipes. The second site drew its water supply from a groundwater source that had very low levels of organic matter and consequently very low levels of disinfection by-products were found in its water system. The third site also used chlorination to disinfect its water source, which had high concentrations of organic matter and bromides, leading to a higher concentration of brominated THMs and HAAs. Like site 1, site 3 used free chlorine to disinfect its water system once a year for two weeks.

Women were eligible for participation in the study if they 1) were over 18 and pregnant (or attempting to become pregnant), 2) resided in areas served by one of the three water systems, 3) were not using assisted reproductive technology, 4) had a positive pregnancy test 5) intended to carry the pregnancy to term 6) did not intend

to move out of the area before the end of the study 7) were able to read and write in English or Spanish and 8) if they had not yet conceived, they could not have been trying to conceive for greater than 6 months. Enrollment in site 1 began in 2001; sites 2 and 3 began enrollment in 2002. Women were recruited into the study through promotional information in public and private obstetric practices, community-based recruitment (child-care facilities, churches, fitness clubs, etc), and through local drug stores (where invitations to join the study were available near pregnancy test kits). After women contacted the study, an initial screening interview was performed to ensure that they met eligibility requirements.

2.1.1 Data Collection

If a woman met the eligibility criteria, informed consent was obtained and a baseline interview was conducted to collect pertinent information including: age, ethnicity, caffeine consumption, education, marital status, income, smoking status, alcohol use during pregnancy, previous pregnancy history, menstrual history, diabetes history, vitamin use and water consumption. Following the baseline interview, study participants were scheduled for an ultrasound that occurred between 6 2/7 and 7 5/7 weeks of gestation but no later than 14 0/7 weeks. The first trimester ultrasound was used to accurately determine gestational age of the fetus, fibroid status of the mother, and other physiologic information. A follow-up interview with all participants occurred between the 20th and 25th week of gestation and was used to ascertain water use, pregnancy related symptoms, and prenatal care. Following the end of the pregnancy, trained chart reviewers abstracted data from each participant's medical records for outcome ascertainment as well as additional medical information.

2.1.2 Water Sampling

Water samples from each site were extracted from the water treatment facility in each city. Samples were taken at the point of entry (POE) of the treated water into the water system. Four samples of water were obtained weekly from each location for the duration of the study. The three cities in this study were chosen partly due to their use of chemical disinfectants that minimize spacial variability of the exposure measurement over the distribution system. For instance, site 2 had very low levels of organic material in its water supply and very low levels of disinfection by-products throughout its entire distribution system. Sites 1 and 3 both used chloramine (for the majority of the year)

as the secondary disinfectant. Chloramine is less likely to combine with organic material and form disinfection by-products outside of the water treatment facility, ensuring a relatively constant concentration disinfection by-products throughout these to cities. However, for a period of 2 weeks in site 3 and one month in site 1, free chlorine was added to disinfect the pipes in the water system. The highly reactive chlorine (which makes it a particularly good disinfectant) readily combined with organic molecules, producing heterogeneity in levels of disinfection by-products throughout the water system. During the months of free chlorine use in these two cities, samples were drawn from 10 locations throughout each water distribution system, in order to reflect the potential heterogeneity of disinfection by-product concentrations. Additionally, periodically during the study, samples were drawn at locations throughout the water distribution system in order to ensure that disinfection by-product measurements calculated from samples at the point of entry correlated with measurements throughout the distribution system. THM samples were analyzed within 2 weeks of collection and HAA samples within 3 weeks. EPA standard Method 551.1 was used to analyze the concentration of THM levels in water samples and EPA standard Method 552.2 was used to analyze HAA concentrations.(EPA, 1995a,b) All samples were analyzed with a 5890 series II gas chromatograph (Agilent Technologies, Palo Alto, CA) equipped with an electron capture detector. A carrier gas of Ultra High Purity helium and a make-up gas of Ultra High Purity Nitrogen were used.

2.2 Overview of Analysis

The purpose of this dissertation was to estimate the effects of the 13 constituent disinfection by-products on spontaneous abortion. Because each of the 13 constituent disinfection by-products depend on shared factors (i.e., the concentration of bromide and organic matter in the water reservoir and the concentration of chlorine used in the disinfection process) the effect of any one of the 13 disinfection by-products may be confounded by the remaining 12, so all must be retained in any regression model to produce unbiased estimates. A standard maximum likelihood logistic regression that includes all 13 constituent disinfection by-products would result in unstable estimates because of the high correlation between the disinfection by-products. Instead, we adopted a hierarchical Bayesian approach that allowed us to stabilize parameter estimates and incorporate prior knowledge regarding the effects of the constituent disinfection by-products on spontaneous abortion. The general Bayesian approach and

four hierarchical models are described in detail below and in Chapter 3.

2.3 Bayesian Analysis

The vast majority of analytic techniques employed in epidemiology are frequentist and rely on hypothetical repeated sampling of some super-population for their interpretation. While there are many reasons to object to frequentist inference (such as violation of the likelihood principle), there are three very pragmatic reasons why epidemiologists should be skeptical of a strictly frequentist approach to data analysis.(Lindley and Phillips, 1976) First, frequentist analyses, by relying on repeated sampling, often give obtuse answers to questions. For instance, the interpretation of a 95% confidence interval for an OR is that under a very large number of samples generated in precisely the same way, 95% of the constructed intervals will contain the true OR. In most epidemiologic settings such an interval has little use: the constructed interval in one study either does or does not contain the true value and the 95% confidence interval does nothing to inform us whether that it does or does not. Second, there are broad classes of problems for which frequentist analyses have not produced useful results. Exact statistics and change-point problems are two examples where frequentist approaches are limited and/or extremely difficult to implement. Third, human beings are remarkably bad at combining evidence in a coherent fashion and frequentist approaches do not offer any way to combine prior knowledge with the current data.

The Bayesian approach offers a solution to these three limitations. In the first case, Bayesian inference provides statistics that have a clear interpretation (i.e., a 95% credible interval around an OR is the region within which we are 95% certain that the true OR lies). In the second case, Bayes theorem provides a natural and systematic way to approach complex problems (for example, Bayesian analyses naturally provide exact statistics without relying on asymptotic assumptions). In the third case, by incorporating prior knowledge in the analysis, the Bayesian approach provides a way to coherently update prior knowledge in light of newly collected data.

The essence of the Bayesian approach is that it quantifies prior information about a parameter (perhaps $\beta = \ln(\text{OR})$) through a probability distribution, $f(\beta)$. We may not (and frequently don't) believe that the parameter is a random quantity, but instead use the prior distribution to quantify our prior knowledge. Bayes theorem provides a method for combining the prior information ($f(\beta)$) with some observed data y (characterized by the likelihood function, $f(y|\beta, x)$) to generate a distribution that represents

our new state of knowledge (a posterior distribution, $f(\beta|y)$). Bayes theorem for continuous data is:

$$f(\beta|y) = \frac{f(y|\beta,x)f(\beta)}{\int_\beta f(y|\beta,x)f(\beta)\partial\beta} = \frac{f(y|\beta,x)f(\beta)}{f(y|x)}. \tag{2.1}$$

Standard epidemiologic practice is to ignore the prior distribution and base inferences only on the likelihood. For instance, the most common technique in epidemiology is the logistic regression, in which case:

$$f(\boldsymbol{y}|\beta,\boldsymbol{x}) = \prod_{i=1}^{N}\left(\frac{exp(x_i\beta)}{1-exp(x_i\beta)}\right)^{y_i}\left(1-\frac{exp(x_i\beta)}{1-exp(x_i\beta)}\right)^{1-y_i}$$

is the likelihood that is maximized in a standard frequentist logistic regression to produce a maximum likelihood estimate, $\widehat{\boldsymbol{\beta}}$. Instead, the Bayesian approach specifies a prior distribution, $f(\beta)$. For instance, we may assume that the log-odds are normally distributed with mean μ and variance ϕ^2, in which case $f(\beta) = N(\mu,\phi^2)$. The prior distribution and the likelihood are combined in equation 2.1 to give $f(\beta|\boldsymbol{y})$, the distribution of the OR that is our updated prior belief in β's effect given the observed data.

2.4 Markov Chain Monte Carlo Algorithms

There are a few special cases in which the posterior distribution from equation 2.1 is available in closed form, but these are relatively rare. For instance, let $\boldsymbol{y} = (y_1 \dots y_n)'$, $\boldsymbol{X}$ be an $n \times k$ design matrix and $\boldsymbol{\beta} = (\beta_1 \dots \beta_k)'$. Then we can define a normal linear model with normal prior:

$$f(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta}) = N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2) \tag{2.2}$$

$$f(\boldsymbol{\beta}|\boldsymbol{\beta}_0,\Sigma_0) = N_k(\boldsymbol{\beta}_0,\Sigma_0) \tag{2.3}$$

where equation 2.2 is the likelihood, σ^2 is a known variance of the outcome, equation 2.3 is the prior distribution of $\boldsymbol{\beta}$ with vector of prior means $\boldsymbol{\beta}_0$, prior covariance matrix Σ_0 and N_k is a k-dimensional normal distribution with k the number of covariates. We can combine equation 2.2 and equation 2.3 using Bayes theorem. The posterior

distribution is available, after some matrix algebra, in closed form as:

$$\begin{aligned} f(\boldsymbol{\beta}|\boldsymbol{y}) &= N_k\big(A, B\big) \\ A &= \big(\boldsymbol{X}'\boldsymbol{X}/\sigma^2 + \Sigma_0^{-1}\big)^{-1}\big(\boldsymbol{X}'\boldsymbol{y}/\sigma^2 + \Sigma_0^{-1}\boldsymbol{\beta}_0\big) \\ B &= \big(\boldsymbol{X}'\boldsymbol{X}/\sigma^2 + \Sigma_0^{-1}\big)^{-1} \end{aligned}$$

While certain conjugate prior distributions will allow the posterior distribution to be calculated in closed form, this is seldom encountered in practical applications. In situations where the posterior distribution is not available in closed form, a variety of approaches can be taken. Potentially, if the posterior distribution is of small dimension (only a few parameters) a discrete grid based approach could work well (where the grid is a set of points of the unknown parameters). Since the likelihood and prior are known, their product could be calculated for the value of the unknown parameter at every point on the grid and divided by the sum of all the products to approximate the posterior density at each grid point. Such approximations are potentially dangerous if the sample space is large (since the chosen grid may not correspond well to the sample space with highest posterior probability) and too onerous if the posterior distribution is of more than 3 or 4 dimensions.

A more fruitful approach is to abandon the task of integrating equation 2.1 and focus attention instead on drawing samples directly from the posterior distribution, $f(\boldsymbol{\beta}|y)$. If a large number of samples of $\boldsymbol{\beta}$ can be drawn from the posterior distribution, inference is trivial: we calculate whatever statistic (mean, median, variance) we are interested in from our generated samples. This also allows us to approximate the posterior distribution as closely as we like by simply generating more samples.

A widely used approach that allows samples to be drawn from the posterior distribution is Markov Chain Monte Carlo (MCMC) simulation. We focus on a particular form of MCMC sampling called Gibbs sampling in the remainder of this section.(Casella and George, 1992) Gibbs sampling is particularly useful in generalized linear models where a full conditional distribution for a parameter is typically easy to derive or sample from. A full conditional distribution is the posterior distribution of a parameter conditional on all other parameters. Gibbs sampling proceeds by repeatedly sampling parameter values from their full conditional posterior distributions. At each step of the Gibbs sampler, the conditional posterior distributions are conditioned on the value of the other parameters at the most recent iteration. Consider the following linear model with no covariates, just an intercept, β, and an unknown error term, σ^2, and prior

distributions on both of them:

$$\begin{aligned} y_i &\sim N(\beta, \sigma^2) \\ \beta &\sim N(\mu, \phi^2) \\ \sigma^2 &\sim IG(\alpha_1/2, \alpha_2/2) \end{aligned}$$

where IG is the inverse gamma distribution. The inverse gamma distribution is a common choice for the prior distribution of a variance term, since it allows for easy computation of conditional posteriors (however, it is not without controversy in some settings).(Gelman, 2005) A closed form solution is not available for the marginal distributions $f(\beta|\boldsymbol{y})$ and $f(\sigma^2|\boldsymbol{y})$, but the full conditional posterior distributions can be easily obtained:

$$\begin{aligned} f(\beta|\sigma^2, \boldsymbol{y}) &\propto \prod f(y_i|\beta, \sigma^2) f(\beta) \\ &= \prod N(\beta, \sigma^2) N(\mu, \phi^2) \\ &\propto N\left(\frac{\mu/\phi^2 + \sum y_i/\sigma^2}{1/\phi^2 + n/\sigma^2}, \frac{1}{1/\phi^2 + n/\sigma^2}\right) \\ f(\sigma^2|\beta, \boldsymbol{y}) &\propto \prod f(y_i|\beta, \sigma^2) f(\sigma^2) \\ &= \prod N(\beta, \sigma^2) IG(\alpha_1/2, \alpha_2/2) \\ &\propto IG\left(\frac{\alpha_1 + n}{2}, \frac{\alpha_2 + \sum (y_i - \beta)^2}{2}\right) \end{aligned}$$

A Gibbs sampling algorithm for this model can be implemented by specifying initial values of $\beta(0)$ and $\sigma^2(0)$ and sampling from the full conditional posterior distributions as follows:

$1_a.$ $[\beta(1)|\sigma^2(0)]$

$1_b.$ $[\sigma^2(1)|\beta(1)]$

$2_a.$ $[\beta(2)|\sigma^2(1)]$

$2_b.$ $[\sigma^2(2)|\beta(2)]$

$$3_a.\ [\beta(3)|\sigma^2(2)]$$

$$3_b.\ [\sigma^2(3)|\beta(3)]$$

$$\vdots$$

$$N_a.\ [\beta(n)|\sigma^2(n-1)]$$

$$N_b.\ [\sigma^2(n)|\beta(n)]$$

An initial k number of iterations are discarded to allow the Gibbs algorithm to achieve convergence, and samples following that burn-in are treated as random draws from $f(\beta, \sigma^2|\boldsymbol{y})$. To find the mean of β, we simply calculate the sample average of $\beta((k+1)_a) \ldots \beta(n_a)$. To find the variance of β, we simply calculate the sample variance of $\beta((k+1)_a) \ldots \beta(n_a)$. Similarly, if we wish to calculate the mean of the posterior distribution of σ^2, we can simply calculate the sample mean of $\sigma^2((k+1)_b) \ldots \sigma^2(n_b)$. Thus, as this simple example demonstrates, even in the absence of a closed form solution for the marginal posterior distribution, Gibbs sampling makes it possible to approximate that distribution by sampling from the full conditional distributions. Although the resultant samples only form an approximation to the posterior distribution, we can make our approximation arbitrarily close to the true posterior distribution by simply running the Gibbs sampler for a larger number of iterations.

2.4.1 Data Augmentation Approach

In non-linear equations however, full conditional posterior distributions can be more difficult to obtain. For example, in logistic models the full conditionals are not immediately available. Modifications to the Gibbs algorithm that use adaptive rejection sampling allow Gibbs algorithms to be generated without specifying the full conditional posterior, however such algorithms can be difficult to implement and slow to converge. As an alternative, Albert and Chib propose a data augmentation approach that is easily implemented and allows full conditional posterior distributions to be calculated for logistic and probit models.(Albert and Chib, 1993) Let y_i be a dichotomous outcomes for the i^{th} individual. We wish to model y_i as a function of predictors $\boldsymbol{x}_i$ (a $1 \times k$ vector of predictors) that have effects $\boldsymbol{\beta} = (\beta_1 \ldots \beta_k)'$. First consider modeling $\boldsymbol{y}$ using a probit model:

$$\Pr(\boldsymbol{y}) = \Phi(\boldsymbol{X}\boldsymbol{\beta}),$$

where Φ is the cumulative of the standard normal distribution. It is possible to express the probit model as a latent variable model. We assume there is a continuous latent variable $\boldsymbol{z}$ that generates $\boldsymbol{y}$ by the function:

$$\begin{aligned} y_i = 1 \quad &\text{if} \quad z_i > 0 \\ y_i = 0 \quad &\text{if} \quad z_i \leq 0 \end{aligned}$$

and model the latent variable as a function of the predictors:

$$\Pr(\boldsymbol{z}) = N(\boldsymbol{X}\boldsymbol{\beta}, 1),$$

where the variance of z is chosen as 1 to ensure identifiability. It is important to note two features of this formulation. First, introducing a latent variable $\boldsymbol{z}$ does not change our interpretation of $\boldsymbol{\beta}$ in any way. Second, it simplifies a non-linear probit model to an ordinary linear regression and makes full conditional posterior distributions easy to calculate. Let the prior distribution for $\boldsymbol{\beta}$ be $f(\boldsymbol{\beta}) = N(\boldsymbol{\beta}_0, \Sigma_0)$, then the conditional posteriors are:

$$f(\boldsymbol{z}|\boldsymbol{y} = 0, \boldsymbol{\beta}) \propto N(\boldsymbol{X}\boldsymbol{\beta}, 1) \text{ truncated to the right of 0} \tag{2.4}$$

$$f(\boldsymbol{z}|\boldsymbol{y} = 1, \boldsymbol{\beta}) \propto N(\boldsymbol{X}\boldsymbol{\beta}, 1) \text{ truncated to the left of 0} \tag{2.5}$$

$$f(\boldsymbol{\beta}|z, y) \propto N\left(\left(\Sigma_0^{-1} + \boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}'\boldsymbol{y}\right), \left(\Sigma_0^{-1} + \boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right) \tag{2.6}$$

These full conditionals make it easy to implement a Gibbs sampling algorithm to obtain the posterior distribution for $\boldsymbol{\beta}$ in a probit model. After specifying initial values of $\boldsymbol{z}$ and $\boldsymbol{\beta}$, we first sample (impute) the latent variable $\boldsymbol{z}$ given $\boldsymbol{y}$ and $\boldsymbol{\beta}$ using equations 2.4 and 2.5. Next, we sample $\boldsymbol{\beta}$ conditional on $\boldsymbol{z}$ using equation 2.6.

Extending this result to a logit model is straight forward.(Albert and Chib, 1993; O'Brien and Dunson, 2004) A t-distribution with 7 or 8 degrees of freedom is a nearly perfect representation of the logistic distribution. Because sampling from a t-distribution can be difficult, we express the t-distribution as a scale mixture of normal distributions. So rather than specifying that $\boldsymbol{z} \sim N(\boldsymbol{X}\boldsymbol{\beta}, 1)$ as in the probit model, we can specify $\boldsymbol{z} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\phi_i)$ with $\phi_i \sim G(\nu/2, \nu/2)$. The normal distribution of $\boldsymbol{z}$ is 'mixed' over the parameter ϕ_i to produce a t-distribution. As a result, if we choose ν properly we can use this parametrization to produce a logit model. Following O'Brien and Dunson we choose a value of $\nu = 7.3$ and $\sigma^2 \approx 0.87$ for a nearly exact

approximation of a logistic distribution.

2.4.2 Gibbs Algorithm for Semi-Bayes

The semi-Bayes model is a hierarchical model that places a prior distribution on effects. In the Albert and Chib data augmentation form, the semi-Bayes model is:

$$\begin{aligned}
\boldsymbol{y} &= 1 \text{ if } \boldsymbol{z} > 0 \\
&= 0 \text{ if } \boldsymbol{z} < 0 \\
\boldsymbol{z} &\sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\phi_i) \\
\boldsymbol{\beta} &\sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0) \\
\phi_i &\sim G(\nu/2, \nu/2)
\end{aligned}$$

Full conditional posterior distributions of the random variables are immediately available as:

$$\begin{aligned}
f(\boldsymbol{z}|\boldsymbol{y}=0,\boldsymbol{\beta}) &\propto N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{W}) \text{ truncated to the right of 0} \\
f(\boldsymbol{z}|\boldsymbol{y}=1,\boldsymbol{\beta}) &\propto N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{W}) \text{ truncated to the left of 0} \\
f(\boldsymbol{\beta}|\boldsymbol{z},\boldsymbol{y},\boldsymbol{\phi}) &\propto N\Big(\big(\Sigma_0^{-1} + \boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X}\big)^{-1}\big(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{y}\big), \big(\Sigma_0^{-1} + \boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X}\big)^{-1}\Big) \\
f(\phi_i) &\propto G\left(\frac{\nu+1}{2}, \frac{\nu + \sigma^{-2}(z_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2}\right)
\end{aligned}$$

where $\boldsymbol{W}$ is an $n \times n$ matrix with diagonal elements $\sigma^2 \times \phi_i$ and off diagonal elements zero. We implement this Gibbs algorithm in Matlab, however Winbugs is capable of estimating coefficients in semi-Bayes models using adaptive rejection sampling (rather than relying on the data augmentation approach). Results between the Gibbs algorithm presented above and results from Winbugs should be virtually identical, but by programming in Matlab we give ourselves greater flexibility (and speed).

2.4.3 Gibbs Algorithm for Fully-Bayes

The fully-Bayes model expands on the semi-Bayes model by allowing the prior variance, $\boldsymbol{\Sigma}_0$ to be random. For simplicity, let $\boldsymbol{\Sigma}_0 = \tau\mathrm{I}$, where I is the identity matrix and τ is a

constant prior variance for model coefficients (τ can be allowed to vary over coefficients with little increase in difficulty). The fully-Bayes model can be written as:

$$\begin{aligned}
\boldsymbol{y} &= 1 \text{ if } \boldsymbol{z} > 0 \\
&= 0 \text{ if } \boldsymbol{z} < 0 \\
\boldsymbol{z} &\sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\phi_i) \\
\boldsymbol{\beta} &\sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0) \\
\tau &\sim IG(\alpha_1/2, \alpha_2/2) \\
\phi_i &\sim G(\nu/2, \nu/2)
\end{aligned}$$

The fully-Bayes model allows the prior variance to be update based on the observed data. This is apparent from the full conditional distributions:

$$\begin{aligned}
f(\boldsymbol{z}|\boldsymbol{y}=0,\boldsymbol{\beta}) &\propto N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{W}) \text{ truncated above at 0} \\
f(\boldsymbol{z}|\boldsymbol{y}=1,\boldsymbol{\beta}) &\propto N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{W}) \text{ truncated below at 0} \\
f(\boldsymbol{\beta}|\boldsymbol{z},\boldsymbol{y},\boldsymbol{\phi}) &\propto N\Big(\big(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\big)^{-1}\big(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}\big), \big(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\big)^{-1}\Big) \\
f(\tau|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{z}) &\propto IG\left(\frac{\alpha_1+n}{2}, \frac{\alpha_2 + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)'(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{2}\right) \\
f(\phi_i) &\propto G\left(\frac{\nu+1}{2}, \frac{\nu+\sigma^{-2}(z_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2}\right)
\end{aligned}$$

Because the posterior distribution of τ is conditional on the variance in the observed data, $(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'(\boldsymbol{\beta}-\boldsymbol{\beta}_0)$, the results of the fully-Bayes analysis will be more robust to prior specification of τ than the semi-Bayes model.

2.5 Dirichlet Process Prior

The fully-Bayes and semi-Bayes approaches both assume a parametric distribution for the prior distribution of the coefficients. In the examples above, a normal prior distribution was chosen, though this is not the only possible distribution. In some settings, it may be preferable to avoid specifying a prior distribution for the coefficients and a non-parametric approach may be more appealing. The Dirichlet process prior is

a prior distribution that allows such non-parametric inference.

Parametric models dominate the epidemiologic literature, with nonparametric approaches largely limited to rank correlation methods and Kaplan-Meier curves.(Kaplan and Meier, 1958) Parametric models, as their name implies, specify the parameters that are used to index specific distributions (e.g., a normal distribution is specified when one specifies the mean and variance). Non-parametric models differ by not presuming that the specific distribution is known. Semi-parametric models represent a useful middle ground between these two classes of models, with one part of the model specified parametrically with another specified non-parametrically. Use of semi-parametric models in epidemiology is limited almost exclusively to Cox's proportional hazards model, that specifies a linear predictor and link function but not a baseline hazard function.(Cox, 1972) Although other semi-parametric models are uncommon in the epidemiologic literature, they have attractive features by "avoid[ing] restrictive assumptions about secondary aspects of a problem while preserving a tight formulation for the features of primary concern."(Oakes, 1988) For instance in a hierarchical model (see below), a first level parametric model could be specified for the effect of a covariate on the outcome while a second level non-parametric model would be specified for the distribution of the coefficient for that predictor. As many (particularly frequentist) non-parametric methods reduce assumptions about the parameters in a distribution, Bayesian non-parametric methods specify a prior that places a probability distribution over the set of all possible probability distributions. Common choices of priors include the Dirichlet process and Pólya tree, both of which can be centered on a simple parametric distribution (e.g., normal), while allowing flexible deviations.(Muller and Quintana, 2004) This approach limits sensitivity and distributional assumptions, while allowing for constraints on the unknown distributions, such as smoothness. In contrast, nonparametric maximum likelihood estimators and other frequentist methods commonly produce estimates inconsistent with prior belief - for example, such estimates commonly take the form of un-smoothed step functions.

2.5.1 The Dirichlet Distribution

Before introducing the Dirichlet process, it is necessary to briefly review the properties of Dirichlet distributions, which are commonly used in Bayesian analyses but uncommon in epidemiology. The Dirichlet distribution is a multivariate extension of the beta distribution and is a conjugate prior for the multinomial family of distributions (just

as the beta distribution is conjugate with the binomial family of distributions). Random variables drawn from a Dirichlet distribution are constrained to lie between 0 and 1. Dirichlet distributions have as many parameters as the discrete sample space over which they are placed has categories. These parameters are restricted to the set of real numbers greater than 0, and influence the relative probability of sampling from one of the discrete categories in the sample space. For instance, a prior Dirichlet distribution could be placed on the probability that a person will respond to one of three possible answers on a survey question. In this example, three parameters $(\alpha_1, \alpha_2, \alpha_3)$ need to be specified. The probability of choosing answer j (where j=1,2 or 3) will be $\alpha_j / \sum_{i=1}^{3} \alpha_i$.

2.5.2 The Dirichlet Process

Dirichlet process, as the name suggests, is a distribution that generates a Dirichlet distribution. It serves as the genesis of most Bayesian non-parametric techniques and has the important property that it places a probability distribution over the set of all possible probability distributions. Developed by Ferguson in 1973, a Dirichlet process, denoted $DP(\lambda D_0)$, serves as a way to randomly generate a distribution D.(Fabius, 1964; Ferguson, 1973; Freedman, 1963) Two parameters specify the Dirichlet process: D_0 is a specified base distribution, such as a standard normal and λ is a positive scalar precision parameter determining how close draws from $DP(\lambda D_0)$ will follow D_0. A random distribution D follows a Dirichlet process, $DP(\lambda D_0)$, if for any partition of a sample space into categories $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_r$, then $D(\mathcal{B}_1), D(\mathcal{B}_2), \ldots, D(\mathcal{B}_r)$ has a Dirichlet distribution with parameters $(\lambda D_0(\mathcal{B}_1), \lambda D_0(\mathcal{B}_2), \ldots, \lambda D_0(\mathcal{B}_r))$. As $\lambda \to \infty$, the sample distribution $D \to D_0$, and thus the Dirichlet process degenerates to the parametric distribution D_0. More intuitive definitions of the Dirichlet Process have been given; we briefly discuss two of them.

The first definition of the Dirichlet process is via the stick breaking process. Random draws from a Dirichlet process almost surely generate discrete distributions, as can be seen more easily in the stick-breaking formulation of the Dirichlet process.(Ferguson, 1973; Sethuraman, 1994) In the stick-breaking construction, we define a draw, D, from a Dirichlet process as the infinite weighted sum of degenerate point masses δ_{θ_j}, that place all their mass on point θ_j.

$$D = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$$

where

$$\begin{aligned} w_j &= z_j \prod_{s=1}^{j-1}(1 - z_s) \\ z_j &\sim beta(1, \lambda) \\ \theta_j &\sim D_0 \end{aligned}$$

That is, samples θ_j are drawn from the base distribution D_0. A "stick" that is initially of unit length is repeatedly "broken" to assign a weight, w_j, to each θ_j. Each w_j is broken from what remains of the stick following the previous $j-1$ breaks. The sum of the weighted point masses is D. Note that if λ is large, small weights will (in expectation) be given to each θ_j, so any large deviation from the baseline distribution will receive a small weight and D will tend to closely resemble D_0, as can be seen in Figure 2.1. A small λ will allow large deviations from D_0 to potentially have a large weight and D may not resemble D_0, as can be seen in Figure 2.2. The stick breaking representation of the Dirichlet process nicely represents its discrete nature. A result of this is that the probability of sampling the same θ_j more than once is non-zero. In fact, the discrete nature of the Dirichlet process allows for clustering of data which we will discuss in more detail below.

The second useful way of describing the Dirichlet process is through the Pólya Urn representation. Many statistical distributions can be derived from urn models.(Johnson and Kotz, 1977) The Pólya urn representation serves not only as a way to describe the Dirichlet process but also as a method of implementing Gibbs sampling algorithms.(Blackwell and MacQueen, 1973; Escobar, 1994; Ferguson, 1973) Consider a random variable β_i which is distributed as some unknown distribution D, which in turn has a Dirichlet process prior, $D \sim DP(\lambda D_0)$. Sampling β_i proceeds as a follows:

1. β_1 is sampled from the base distribution, D_0.
2. β_2 is set equal to β_1 with probability p_1. Otherwise it is drawn from D_0 with probability $1 - p_1$
3. β_j is set equal to β_k $(k = 1 \ldots j-1)$, with probability p_k; otherwise it is drawn from D_0 with probability $1 - \sum_{i=1}^{k} p_i$.

We define $p_j = \frac{1}{\alpha+n-1}$ for $j = (1, \ldots, n)$.

The conditional distribution of β_i given $\beta^{(i)} = (\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_n)$ is given by:

$$[\beta_i|\beta^{(i)}] \sim \big(1 - \sum_{j \neq i} p_j\big) D_0 + \sum_{j \neq i} p_j \delta_{\beta_j}$$

This representation illustrates an important property of Dirichlet processes: the grouping of observations. A series of n draws from a Dirichlet process will be clustered into k ($k \leq n$) groups. Note that if all draws of β_i have the same value, then $\beta_i \sim D_0$. We take advantage of this clustering property to both reduce the dimensionality of the data as well as to cluster effect estimates into groups of disinfection by-products that have similar effects on risk of spontaneous abortion.

2.5.3 The Dirichlet Process Prior in Practice

Although Dirichlet processes were introduced by Ferguson in 1973 and Dirichlet process mixture models were introduced by Antoniak in 1974, they were not computationally feasible until the work of Escobar and West provided MCMC techniques to obtain posterior distributions.(Escobar, 1994; Escobar and West, 1995; Ferguson, 1973; MacEachern, 1994; West et al., 1994) Since then, Dirichlet processes have seen widespread use in a variety of fields, often with the common theme of needing to reduce the dimensionality of a problem. Cao and West incorporate multiple Dirichlet process priors in a mixture model to examine neurological response data.(Cao and West, 1996) Dirichlet process mixture models have also been used in molecular biology to estimate equilibrium frequencies of gene mutations, where the number of large number of genes makes individual estimation of each frequency impossible.(Lartillot and Philippe, 2004) Gelfand and Kuo use a Dirichlet process prior to aid in the estimation of potency curves in bioassay experiments.(Gelfand and Kuo, 1991) Gopalan and Berry propose using Dirichlet process priors to make multiple comparisons.(Gopalan and Berry, 1998) Generalized linear models (GLMs) have been extended by incorporating Dirichlet priors reducing their dependence on model specification and making them semi-parametric. A Dirichlet process prior on the link function has been implemented by Newton et al.(Newton et al., 1996) Placing a Dirichlet process prior on coefficients or error terms in a generalized linear model has led to semi-parametric GLMs, generalized linear mixed models and overdispersed GLM's.(Kleinman and Ibrahim, 1998; Mukhopadhyay and Gelfand, 1997)

2.5.4 Dirichlet Process Priors for Clustering Regression Coefficients

While both semi-Bayes and fully-Bayes models are a distinct improvement over standard epidemiologic analytic techniques, they may be unsuitable in two ways. First, results may be sensitive to the assumed prior distribution of β_j and a non-parametric prior would be preferable. Second, when sufficient prior information exists the coefficients may be grouped into exchangeable categories by incorporating second level coefficients. Unfortunately, in many epidemiologic applications, prior knowledge on how to group the coefficients may be unknown and a procedure that allows them to be grouped into clusters based on similarity of effect sizes would be preferred.

An important property of the Dirichlet process prior is its ability to cluster coefficients into groups. Assuming $\beta_j \sim D$ and $D \sim DPP(\lambda D_0)$, implies the following prior distribution on β_j:(West et al., 1994)

$$[\beta_j] \sim \frac{\lambda}{\lambda + k - 1} D_0 + \frac{1}{\lambda + k - 1} \sum_{i \neq j} \delta_{\beta_i} \tag{2.7}$$

were δ_{β_i} is a point mass at β_i. Thus, β_j has a probability of being distributed as the base distribution, D_0, or being clustered with any other $\beta_i, i \neq j$. Group membership is determined by the precision parameter λ, with higher probability of clustering any two coefficients together increasing as λ decreases. At each iteration of the Gibbs sampler, a coefficient is either clustered in a group with some other coefficient(s) or occupies its own cluster. It is important to note that while coefficients will be clustered together during particular iterations of the Gibbs sampler, they will (generally) not be clustered together at every iteration of the Gibbs sampler. So posterior means of coefficients will be similar if the two are frequently clustered together, but are unlikely to be identical.

2.5.5 Gibbs Algorithm for Dirichlet Process Priors

The Gibbs sampling algorithm for the Dirichlet process prior model is more difficult to implement than either the fully-Bayes or semi-Bayes models. Also, unlike the semi-Bayes or fully-Bayes models, the Dirichlet process prior model cannot be implemented in Winbugs and requires more programming knowledge. We begin our discussion of the Gibbs algorithm by expressing the Dirichlet process prior model in hierarchical form:

$$
\begin{aligned}
\boldsymbol{y} &= 1 \text{ if } \boldsymbol{z} > 0 \\
&= 0 \text{ if } \boldsymbol{z} < 0 \\
\boldsymbol{z} &\sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\phi_i) \\
\boldsymbol{\beta} &\sim D \\
D &\sim DP(\lambda D_0) \\
\lambda &\sim G(a, b) \\
D_0 &= N(\mu, \tau^2) \\
\tau^2 &\sim IG(\alpha_1/2, \alpha_2/2) \\
\phi_i &\sim G(\nu/2, \nu/2)
\end{aligned}
$$

Unlike the semi-Bayes and fully-Bayes models which specified a particular distribution for β_j, the Dirichlet process prior model allows the distribution of β_j to be random. A precision parameter, λ, determines how closely the random distribution follows the base distribution D_0. We have placed a gamma prior distribution on λ to allow the data to inform about it. The coefficients can be clustered together into k groups that have unique values: $\gamma_1 \ldots \gamma_k$. For instance, β_1 and β_4 may have a common value γ_3, while β_2 and β_{10} have common value γ_1. We use the notation (j) to denote a parameter's value when the j^{th} element is excluded. For instance, $\boldsymbol{\beta}^{(j)} = (\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots \beta_p)$. In order to implement a Gibbs sampling algorithm, we need full conditional distributions, however they are not as easily obtained for the Dirichlet process prior model. The necessary full conditionals can be shown to be:

$$
\begin{aligned}
f(\boldsymbol{z}|\boldsymbol{y} = 0, \boldsymbol{\beta}) &\propto N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{W}) \text{ truncated above at 0} && (2.8) \\
f(\boldsymbol{z}|\boldsymbol{y} = 1, \boldsymbol{\beta}) &\propto N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{W}) \text{ truncated below at 0} && (2.9) \\
f(\beta_j|\boldsymbol{z}, \boldsymbol{y}, \phi) &\propto p_{new,j} N\big(E_j^{dpp}, V_j^{dpp}\big) + \sum_{l=1}^{k^{(j)}} p_{l,j} \delta_{\gamma_l^{(j)}} && (2.10) \\
f(\tau|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{z}) &\propto IG\left(\frac{\alpha_1 + n}{2}, \frac{\alpha_2 + \sum(\gamma_j - \mu)^2}{2}\right) && (2.11) \\
f(\phi_i) &\propto G\left(\frac{\nu + 1}{2}, \frac{\nu + \sigma^{-2}(z_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2}\right) && (2.12)
\end{aligned}
$$

where

$$E_j^{dpp} = (\tau^{-2} + \sum_i^n x_{ij}^2/\phi_i^2)^{-1}(\mu/\tau^2 + \sum_i^n x_{ij}h_i^{(j)}/\phi_i^2) \quad (2.13)$$

$$V_j^{dpp} = (\tau^{-2} + \sum_i^n x_{ij}^2/\phi_i^2)^{-1} \quad (2.14)$$

We define $h_i^{(j)} = z_i - \boldsymbol{x}_i^{(j)'}\boldsymbol{\beta}^{(j)}$. The full conditional posterior distribution of β_j contains the weights:

$$p_{new,j} = \frac{\lambda}{\lambda + p - k^{(j)} - 1} \times \frac{N(0|\mu, \tau^2)\prod N(h_i^{(j)}|0, \phi_i^2)}{N(0|E_j^{dpp}, V_j^{dpp})} \quad (2.15)$$

$$p_{l,j} = \frac{p}{\lambda + p - k^{(j)} - 1} \times \prod_{i=1}^n N(h_i^{(j)}|x_{ij}\beta_l^{(j)}, \phi_i^2) \quad (2.16)$$

The Gibbs sampling algorithm proceeds by first imputing the latent continuous variable $\boldsymbol{z}$. Second, coefficients $\beta_1 \ldots \beta_p$ are assigned to clusters $\gamma_1 \ldots \gamma_k$. Cluster allocation is determined by the weights in equation 2.15 and equation 2.16. For each coefficient, we sample from the multinomial distribution defined by equations 2.15 and 2.16. With probability $p_{new,j}$ the j^{th} coefficient is assigned to a new cluster or it is assigned to existing cluster l with probability $p_{l,j}$. After determining the cluster allocation of each coefficient, the third step is to define a new design matrix to reflect the allocation. For example, if we had 4 coefficients that were clustered into groups as follows:

$$\gamma_1 = \beta_1 = \beta_2 = \beta_4$$
$$\gamma_2 = \beta_3$$

We would then generate a new matrix, $\boldsymbol{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2)$ where $\boldsymbol{r}_1 = (\boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{x}_4)$ and $\boldsymbol{r}_2 = \boldsymbol{x}_3$. Now, the cluster-specific coefficients can be updated by sampling from $N(E_\gamma, V_\gamma)$, where $V_\gamma = (\boldsymbol{\Sigma}_\gamma^{-1} + \boldsymbol{R}'\boldsymbol{W}_\gamma^{-1}\boldsymbol{R})^{-1}$ and $E_\gamma = V_\gamma(\boldsymbol{\Sigma}_\gamma^{-1}\mu + \boldsymbol{R}\boldsymbol{W}_\gamma^{-1}\boldsymbol{z})$ and $\boldsymbol{W}_\gamma$ is a matrix with diagonal terms $\sigma^2\phi_i$.

The fourth step of the Gibbs sampler updates the precision parameter using the data augmentation technique of Escobar and West and updates the prior variance in the base distribution as in the fully-Bayes model.(Escobar and West, 1998)

2.5.6 Dirichlet Process Prior with Selection Component

Although we wish to estimate the effect of each exposure, we anticipate that in many studies some of the exposures will have no effect. If a given exposure has no effect on the outcome it cannot confound the effect of any other exposure and we would prefer to exclude it from the model. Variable selection techniques in the epidemiologic literature are limited, generally relying on backward or forward selection strategies. These strategies generally look at a large number of models to determine whether individual terms should be included or excluded. A common exclusion criterion in epidemiologic variable selection is that the OR of interest change by less than 10% when the variable is excluded (and frequently includes a component examining whether the variable is an effect modifier as well). A final model is arrived at and is treated as the only model that was examined, a strategy leads to inappropriately small reported variances.(Draper, 1995; Leamer, 1978; Raftery, 1996) However, there has been an increasing focus on variable selection methods in the statistical literature, largely motivated by gene expression applications.(Efron and Tibshirani, 2002; Newton et al., 2001) For example, Geweke proposed a mixture prior, that allows an unknown subset of the predictors to have zero coefficients ($\beta_j = 0$), while using a normal prior for the remaining coefficients.(Geweke, 1996) When using a Dirichlet process prior for the coefficients, the exposures are automatically clustered into groups. By using Geweke's mixture prior for the group specific coefficients, we allow a cluster of exposures that has coefficients equal to zero. We adopt this prior distribution in the Dirichlet process prior to perform simultaneous variable selection and clustering which is known to have excellent properties.(Ishwaran and Rao, 2005)

2.5.7 Gibbs Algorithm for Dirichlet Process Prior with Selection Component

The Gibbs algorithm for the selection component is similar to the algorithm without it. The hierarchical model can be defined as follows:

$$
\begin{aligned}
\boldsymbol{y} &= 1 \text{ if } \boldsymbol{z} > 0 \\
&= 0 \text{ if } \boldsymbol{z} < 0 \\
\boldsymbol{z} &\sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\phi_i) \\
\boldsymbol{\beta} &\sim D \\
D &\sim DP(\lambda D_0) \\
\lambda &\sim G(a, b) \\
D_0 &= \pi\delta_0 + (1-\pi)N(\mu, \tau^2) \\
\tau^2 &\sim IG(\alpha_1/2, \alpha_2/2) \\
\phi_i &\sim G(\nu/2, \nu/2) \\
\pi &\sim beta(a, b)
\end{aligned}
$$

where δ_0 is a degenerate distribution with all its mass at zero. The probability, π, that a randomly selected coefficient will be zero is given a beta prior to allow the data to help inform about its value.

The Gibbs sampler proceeds as above except the weights for assigning cluster allocation are now defined:

$$
p_{new,j} = \frac{\lambda(1-\pi)}{\lambda + p - k^{(j)} - 1} \times \frac{N(0|\mu, \tau^2)\prod N(h_i^{(j)}|0, \phi_i^2)}{N(0|E_j^{dpp}, V_j^{dpp})} \tag{2.17}
$$

$$
p_{0,j} = \pi \tag{2.18}
$$

$$
p_{l,j} = \frac{p(1-\pi)}{\lambda + p - k^{(j)} - 1} \times \prod_{i=1}^{n} N(h_i^{(j)}|x_{ij}\beta_l^{(j)}, \phi_i^2) \tag{2.19}
$$

These weights are used as parameters in the multinomial distribution as before, with the difference being that now a draw can take the value of another coefficient ($p_{l,j}$), a new value ($p_{new,j}$), or be assigned a value of zero ($p_{0,j}$). The next step is to update the cluster specific coefficients as before. The only additional step is to update the probability of assigning a coefficient a zero value, π. Its conditional posterior distribution is a function of the number of coefficients assigned a zero value in the last iteration, n_0:

$$
f(\pi|\boldsymbol{y}, \boldsymbol{\beta}) = beta(a + n_0, b + p - n_0)
$$

.

2.6 Model Specification for Analysis of Disinfection By-Products and Spontaneous Abortion

We specified a discrete time hazard model for the probability that a spontaneous abortion occurs in a given gestational week with terms for gestational week specific intercepts confounders and 13 constituent disinfection by-products. The concentrations of these by-products were categorized to allow for a more flexible relationship between the logit of the probability of spontaneous abortion and dose, we categorized constituent disinfection by-products into quartiles, when possible. We implemented the four Bayesian hierarchical models we previously discussed: semi-Bayes, fully-Bayes, Dirichlet process prior, and Dirichlet process with a selection component. We use the existing literature to specify prior distributions for these models. Because the results of any analysis depend heavily on modeling assumptions, we performed sensitivity analyses to assess how changes to our prior specifications alter our assumptions.

We programmed Gibbs sampling algorithms for each of the four models in Matlab.(Mathworks Development, 2005) All models were run for 60,000 iterations, with the initial 5,000 iterations discarded as a burn-in. The remaining iterations were examined for convergence by examining trace plots of the sample parameter values by iteration of the algorithm. Because MCMC algorithms can be sensitive to initial values, we ran our algorithms several times with different starting values.

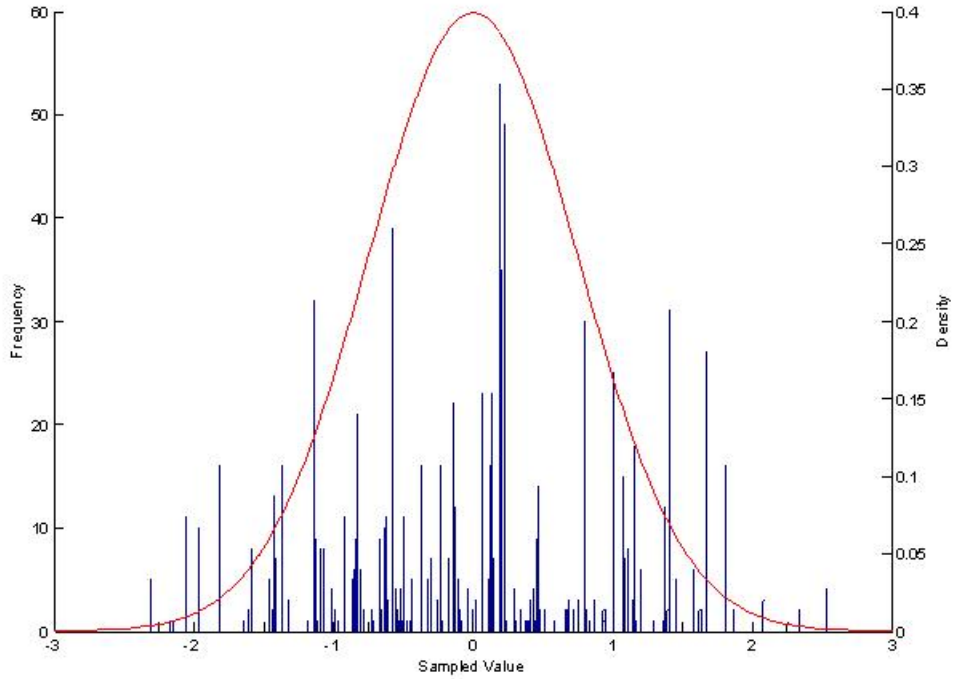

FIGURE 2.1: Histogram of 1000 samples drawn from $DP(\lambda = 50, G_0 = N(0, 1))$.

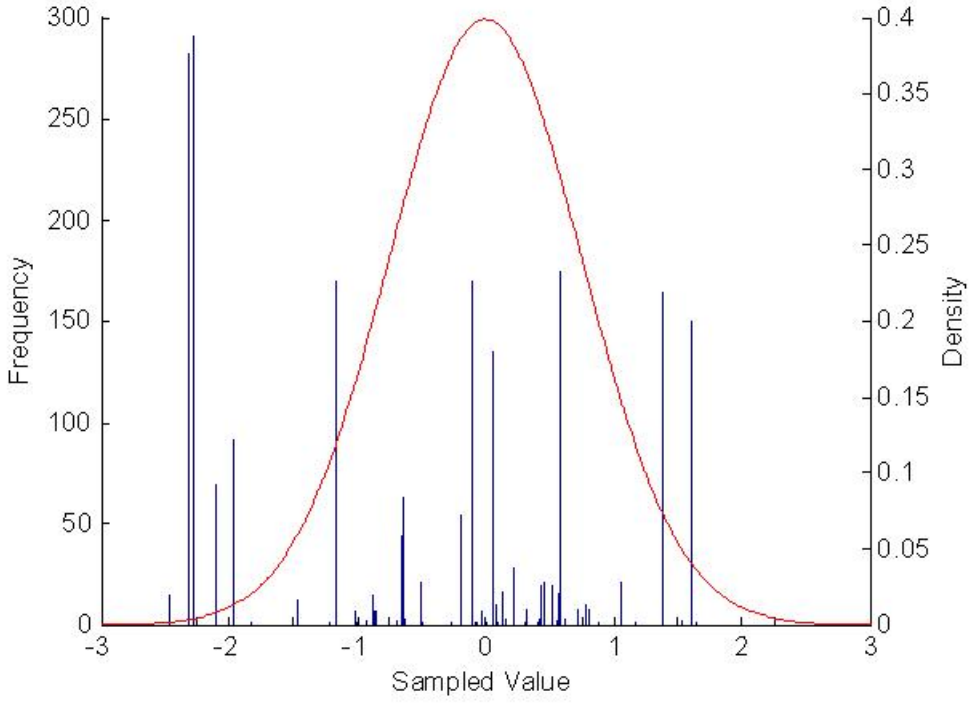

FIGURE 2.2: Histogram of 1000 samples drawn from $DP(\lambda = 5, G_0 = N(0, 1))$.

CHAPTER 3

BAYESIAN METHODS FOR HIGHLY CORRELATED EXPOSURE DATA

3.1 Abstract

Studies that include individuals with multiple highly correlated exposures are common in epidemiology. Because standard maximum likelihood techniques often fail to provide plausible estimates in such instances, hierarchical regression methods have seen increasing use. Bayesian hierarchical regression places prior distributions on exposure-specific regression coefficients to stabilize estimation and incorporate prior knowledge if available. In the semi-Bayes approach, the prior mean and variance are treated as fixed constants chosen by the researcher. An alternative is a fully-Bayes approach that places distributions on the prior mean and variance to allow the data to inform about their values. Both of these approaches typically rely on a normal prior for the exposure-specific coefficients. As a more flexible semi-parametric option, one can use a Dirichlet process prior, that clusters exposures into groups, effectively reducing dimensionality. We compare these hierarchical regression methods and demonstrate the reduced mean squared error of fully Bayes and Dirichlet process prior models in many instances.

3.2 Introduction

3.2.1 Motivation and Background

Highly correlated exposures are ubiquitous in epidemiologic research, and may arise due to an association between the measured exposures and one or more latent factors. For example, pesticide exposures for farm workers tend to be highly correlated because individuals apply multiple pesticides in a year, with choice of pesticide influenced by type of crop.(Alavanja et al., 1996; Kirrane et al., 2005) Another example is the correlation in nutrient intake that arises from an individual's food preferences. Lifestyle factors can also contribute to dependency between exposures, such as smoking, alcohol intake, and illicit drug use.

We depict this correlated exposure problem in more general fashion using the directed acyclic graph (DAG) in Figure 3.1. Let $x_1, \ldots, x_k$ denote the levels of k different exposure variables, let U denote an unmeasured variable or variables explaining the correlation in $x_1, \ldots, x_k$, and let Y denote the outcome. Researchers will generally be interested in estimating effect measures, $\beta_1, \ldots, \beta_k$, for exposures $x_1, \ldots, x_k$. Hence, a common strategy is to fit the logistic regression model:

$$\text{logit}\{\Pr(Y_i = 1 \,|\, x_{i1}, \ldots, x_{ik})\} \;=\; \alpha_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}. \tag{3.1}$$

Unfortunately, maximum likelihood estimation of the model in equation 3.1 can fail to converge when predictors are highly correlated, and estimated coefficients may be unreliable even when convergence is achieved.

This problem has led many epidemiologists to fit logistic regression models incorporating one exposure variable at a time. However, the other exposure variables may be confounders and, if so, must be included in order to assess the causal effect of any specific exposure.(Greenland et al., 1999) Another commonly-used strategy is to collapse the specific exposure information into summaries, such as a sum across chemicals in a class. Unfortunately, this results in a loss of information, does not allow inferences on effects of specific exposures, and can be sensitive to the summary chosen.

3.2.2 Hierarchical Regression

Problems with collinearity have motivated increased use of hierarchical models.(Greenland, 1992) Ordinary regression models treat the outcome as a random vari-

able, dependent on parameters. For example in equation 3.1, Y_i is a random variable that depends on the parameters α_0 and $\beta_1 \ldots \beta_k$. Hierarchical regression extends ordinary regression models by also treating parameters as random variables that depend on further coefficients through a prior distribution. Estimates obtained through hierarchical regression are shrinkage estimators in the sense that they are moved away from the unbiased maximum likelihood estimate (MLE) and toward the center of the prior distribution. The amount of shrinkage is controlled by the variance of the prior distribution. A smaller prior variance causes greater shrinkage. By changing the prior distribution, a wide variety of hierarchical regression models can be specified.

Two types of hierarchical regression models have seen wide use in epidemiologic research: empirical Bayes (EB) and semi-Bayes (SB).(De Roos et al., 2001; Engel et al., 2005a,b; Greenland, 1992, 1993, 1994; Greenland and Poole, 1994; Steenland et al., 2000) These methods vary in how they specify prior distributions on coefficients. A typical prior distribution for β_j (where j indexes the k coefficients in equation 3.1) is $N(\mu, \phi^2)$, where μ characterizes the investigator's prior knowledge about the true value of the coefficients and ϕ^2 is the uncertainty regarding that value. SB and EB procedures differ in how they treat ϕ^2. EB models use the current data to estimate ϕ^2, while SB methods offer the researcher an opportunity to specify the prior variance based on substantive knowledge.(Casella, 1985; Greenland, 1994) One process of elicitation for ϕ^2 that may be used in SB procedures is for the researcher to specify a range of values within which 95% of coefficient values are expected to fall under repeated sampling. This range can be used to calculate a value for the variance term, which is then treated as fixed and used in the hierarchical model.

Typically, in a model such as equation 3.1, a large number of coefficients will need to be estimated. Consider a model in which 20 coefficients are estimated and each has a $N(0, \phi^2)$ prior. Prior scientific knowledge may exist about the variability of the estimates, but the data also contain information about that variability, with a simplistic estimate being the variance about the prior mean of the 20 MLEs. The EB method uses the observed variability to estimate ϕ^2 but ignores prior substantive information. The SB method incorporates prior knowledge by treating ϕ^2 as known (and fixed) but ignores the information regarding the variability of the coefficients about the prior mean that is contained in the observed data. Thus SB models have a fixed amount of shrinkage regardless of the support for the prior distribution provided by the data. Alternatively, a fully-Bayes (FB) approach estimates ϕ^2 by combining prior knowledge regarding the variance of the coefficients with the observed variability in the data,

resulting in estimates that will generally be more robust than SB methods and provide a more realistic summary of the current state of knowledge than EB methods. We note that although we refer to one particular hierarchical model as a fully-Bayes model, all four hierarchical models that we present are equally Bayesian, including the SB model. Our nomenclature was chosen to be in keeping with existing naming conventions.

3.2.3 Extensions

The SB and FB models have potential disadvantages. First, results may be overly sensitive to the assumed parametric form of the prior distribution. Second, in order for SB and FB methods to shrink parameter estimates towards multiple prior means, the coefficients must be specified into classes (e.g., if the coefficients are the effects of different pesticides, they could be classified as fungicides or herbicides to allow coefficients in those classes to be shrunk toward different means). In many situations, it may be impossible to specify which effects should be grouped in to which classes, or even how many classes there should be. In this situation, a method that allows the data to guide the clustering of coefficients into classes would be preferable. For this reason, we place a Dirichlet process prior (DPP) on the distribution of the coefficients.(Ferguson, 1973, 1974; Gopalan and Berry, 1998) The DPP allows for non-parametric estimation of β_j, while simultaneously clustering the β_j into groups based on effect size.

Although we wish to estimate the effect of each exposure, we anticipate that in many studies some of the exposures will have no effect. If exposure x_j (Figure 3.1) has no effect on the outcome it cannot confound the effect of any other exposure and we would prefer to exclude it from the model. Variable selection techniques in the epidemiologic literature are limited, generally relying on backward or forward selection strategies that increase the type I error rate (Draper, 1995; Leamer, 1978; Raftery, 1996). However, there has been an increasing focus on variable selection methods in the statistical literature, largely motivated by gene expression applications.(Efron and Tibshirani, 2002; Newton et al., 2001) For example, Geweke proposed a mixture prior, that allows an unknown subset of the predictors to have zero coefficients ($\beta_j = 0$), while using a normal prior for the remaining coefficients.(Geweke, 1996) When using a DPP for the coefficients, the exposures are automatically clustered into groups. By using Geweke's mixture prior for the group specific coefficients, we allow a cluster of exposures that has coefficients equal to zero. We adopt this prior distribution in the DPP to perform simultaneous variable selection and clustering which has been shown

to have excellent properties.(Ishwaran and Rao, 2005)

3.3 Properties of SB and FB Estimators

SB and FB models have been discussed in detail elsewhere.(Greenland, 1992, 1993, 1994, 2000; Lindley and Smith, 1972) Here, we illustrate some of their properties in the simple setting of an ordinary linear regression model in which covariates $x_{i1} \ldots x_{ik}$ are regressed on an outcome y_i. For ease of presentation, we assume the linear model has a known error term, σ^2, and that the covariates are orthogonal (i.e., they are not correlated).

As mentioned above, the SB model incorporates information on β_j through a prior distribution. A typical specification for the SB ordinary linear model is:

$$\begin{aligned} [y_i|\beta_j^{sb}] &\sim N\left(\sum_{j=1}^{k} \beta_j^{sb} x_{ij}, \sigma^2\right) \\ \left[\beta_j^{sb}\right] &\sim N\left(\eta_j, \phi_j^2\right) \end{aligned} \tag{3.2}$$

where the prior mean, η_j, incorporates prior evidence regarding the size of the effect for the j^{th} coefficient and x_{ij} may be standardized so they are on the same scale. Prior scientific knowledge may indicate that the prior mean is the same for all coefficients, that it varies across the coefficients (i.e, some coefficients have one prior mean and others have a different prior mean) or that each coefficient has its own mean. For example, if $\beta_1 \ldots \beta_k$ are the effect of pesticides on retinal degeneration, one could assume that the prior knowledge of the effect of pesticides is the same for all pesticides (e.g., no effect: $\eta_j = 0$), or that the effect varies over different classes of pesticides (such as fungicide, herbicide, insecticide, etc).(Kirrane et al., 2005) In this case, indicator variables for pesticide class, z_{lj}, can be introduced into the prior distribution by allowing $\eta_j = \sum_{l=1}^{p} \theta_l^{sb} z_{lj}$. The prior variance, ϕ_j^2 represents the certainty of the prior evidence that β_j^{sb} has an effect of size η_j. The prior variance could be specified from a meta-analysis or could be calculated by choosing a range within which the researcher believes 95% of effect estimates on this topic would lie. Solving the the standard confidence interval formula for the variance term allows the researcher to specify the prior variance. The lack of a prior distribution on θ_l^{sb} or ϕ_j^2 is the distinguishing feature of SB.

The posterior distribution (i.e., the distribution that results when the prior distri-

bution is updated with the observed data) for β_j^{sb} is given by:

$$\left[\beta_j^{sb}|Data\right] \sim N\left(\frac{\eta_j/\phi_j^2 + \sum x_{ij}y_i/\sigma^2}{1/\phi_j^2 + \sum x_{ij}^2/\sigma^2}, \frac{1}{1/\phi_j^2 + \sum x_{ij}^2/\sigma^2}\right) \tag{3.3}$$

The posterior mean is an average of the prior mean (η_j) and the maximum likelihood estimate ($\sum x_{ij}y_i/\sum x_{ij}^2$), inverse weighted by their respective variances, ϕ_j^2 and $\sigma^2/\sum x_{ij}^2$. This is the essence of a shrinkage estimator: the posterior distribution of β_j^{sb} is shrunk towards its prior distribution. For concreteness, we generate a small (n=50) dataset with 5 orthogonal covariates, none of which have an effect. We assume the SB model in equation 3.2 with $\eta_j = 0$ and $k = 5$. Figure 3.2 shows the distribution of β_1^{mle} and β_1^{sb} for $\phi_j^2 = 0.5, 1.0$, and 2.0. The amount of shrinkage is a function of the prior variance: as the prior variance decreases (representing increasing certainty about the effect of β_1^{sb}), the posterior distribution shrinks towards the prior mean. Conversely, as the prior variance increases the posterior distribution converges to the distribution of the maximum likelihood estimate. Also, as can be seen from formula 3.3, as the number of observations increases, the posterior distribution is weighted more heavily toward the observed data. With orthogonal data of moderate size, the observed data will quickly overwhelm anything but the strongest priors (i.e., those with very small ϕ^2), and SB or FB results will be similar to the MLE.

In studies with a large number of covariates, SB methods have been advocated as a way to reduce problems with multiple comparisons since their shrinkage properties can decrease the probability of finding false positives.(Hung et al., 2004) We briefly comment on two troubling aspects of this approach (Appendix 1 contains more details). First, the lower type-I error rate only occurs when the prior mean is given the value of β_j under the null hypothesis (typically, $\beta_j = 0$). If a different value is chosen for the prior mean (including non-null values more consistent with scientific knowledge), the probability of rejecting the null will increase. Second, although SB methods can improve the overall error-rate, the improvement may be much less dramatic than most researchers would prefer. For example, we simulate a dataset with 20 orthogonal covariates and show (Figure 3.3) the increase in the overall type-I error rate as the number of covariates being tested increases from 1 to 20 (with prior mean: $\eta_j = 0$). The MLE exhibits an error rate of 5% when 1 covariate is tested and 64.15% when 20 covariates are tested. On the other hand the SB estimate (with prior variance: $\phi^2 = 1/2$) has an error rate of 4.29% when 1 covariate is tested and 58.43% when 20 are tested. The error rate for SB models can be reduced by assuming a smaller prior variance, however the level

of prior knowledge regarding β_j required to obtain an 'acceptable' error rate may be incommensurate with existing research.

Because ϕ^2 is so vital to SB methods, users are advised to vary it in sensitivity analyses to see how β_j^{sb} changes with different plausible values of ϕ^2. In Figure 3.2, for example, $\phi^2 = 1.0$ may have been the best guess of the variance of β_j^{sb} with sensitivity analyses conducted for $\phi^2 = 2.0$ and $\phi^2 = 0.5$. FB methods implicitly account for this uncertainty by placing a prior distribution on ϕ^2, resulting in β_j^{fb} estimates that are averaged over plausible values for ϕ^2. Unlike SB methods that have a fixed amount of shrinkage, FB models that treat ϕ^2 as random allow shrinkage of β_j^{fb} to be based not only on the specification of the prior variance but also the observed variability of β_j from the prior mean in the data. Additionally, when the prior mean is a function of covariates (e.g., $\eta_j = \sum \theta_l z_{lj}$), prior information may exist for the effect of those variables and a prior distribution can be placed on those parameters. For instance in the same scenario as above, a typical FB model is specified as:

$$\begin{aligned}
\left[y_i|\beta_j^{fb}\right] &\sim N\left(\sum_{j=1}^{k} \beta_j^{fb} x_{ij}, \sigma^2\right) \\
\left[\beta_j^{fb}|\theta, \phi_j^2\right] &\sim N\left(\sum_{l=1}^{p} \theta_l^{fb} z_{lj}, \phi_j^2\right) \\
\left[\theta_l^{fb}\right] &\sim N\left(\mu_l, \omega_l^2\right) \\
\left[\phi_j^2\right] &\sim IG\left(\alpha_1, \alpha_2\right)
\end{aligned} \tag{3.4}$$

Here, the θ_l^{fb} are the effects of the z_{lj} covariates and their prior mean, μ_l, is the prior knowledge regarding the size of that effect, while the prior variance ω_l^2 represents uncertainty in that effect. The prior distribution for the ϕ_j^2 is chosen as an inverse gamma (IG) distribution with parameters α_1 and α_2. The inverse gamma distribution is a common choice for the prior distribution of a variance term because of its flexibility and for computational convenience. The prior mean of ϕ_j^2 is $1/(\alpha_2(\alpha_1 - 1))$ and its variance is $1/(\alpha_2^2(\alpha_1 - 1)^2(\alpha_1 - 2))$. In choosing values of α_1 and α_2 for an analysis, we suggest specifying a most likely value of ϕ^2 (call this $\mathrm{E}(\phi^2)$) and a value for the variance of ϕ (call this $\mathrm{V}(\phi^2)$) such that 95% of the reasonable ϕ^2 values would fall within $\mathrm{E}(\phi^2) \pm 1.96\sqrt{\mathrm{V}(\phi^2)}$. Solving these equations for α_1 and α_2 gives: $\alpha_1 = (\mathrm{E}(\phi^2)^2/\mathrm{V}(\phi^2)) + 2$ and $\alpha_2 = (\mathrm{E}(\phi^2)^3/\mathrm{V}(\phi^2) + \mathrm{E}(\phi^2))^{-1}$.

The full conditional posterior distributions for the parameters in the FB model are:

$$\left[\beta_j^{fb}|Data,\sigma_2^2,\theta_j,\phi^2\right] \sim N\left(\frac{\sum \theta_l^{fb} z_{lj}/\phi_j^2 + \sum x_{ij}y_i/\sigma^2}{1/\phi_j^2 + \sum x_{ij}^2/\sigma^2}, \frac{1}{1/\phi^2 + \sum x_{ij}^2/\sigma^2}\right) \quad (3.5)$$

$$\left[\theta_j|Data,\beta_j^{fb},\phi_j^2\right] \sim N\left(\frac{\mu_l/\omega_l^2 + \sum z_{lj}\beta_j^{fb}/\phi^2}{1/\omega_l^2 + \sum z_{lj}^2/\phi_j^2}, \frac{1}{1/\omega_l^2 + \sum z_{ij}^2/\phi_j^2}\right) \quad (3.6)$$

$$\left[\phi^2|Data,\beta_j^{fb},\theta_j\right] \sim IG\left(\alpha_1 + p/2, (\frac{\sum(\beta_j^{fb} - z_{lj}\theta_j^{fb})^2}{2} + \frac{1}{\alpha_2})^{-1}\right) \quad (3.7)$$

The conditional distribution of ϕ^2 is of particular interest. Its adaptive shrinkage properties are apparent from the $\sum(\beta_j^{fb} - z_{lj}\theta_j)^2$ term, that is the variation in the β_j^{fb} from their prior mean. As the variance of the parameters increases, ϕ^2 also increases and when the variance decreases, ϕ^2 decreases. Thus, if there is little evidence in the data to support the prior specification for ϕ^2, the posterior estimate of ϕ^2 is increased to reflect this. Since ϕ^2 determines the amount of shrinkage, if little evidence in support of the prior specification of ϕ^2 is seen in the data, ϕ^2 will increase and less shrinkage will be observed. The converse is also true; if the there is little variability in the data of the estimates from the prior mean, the posterior estimate of ϕ^2 will decrease and cause greater shrinkage of β_j^{fb} to their prior distribution. Non-informative values of α_1 and α_2 could be chosen to allow the data to completely guide inference, however epidemiologists generally have information regarding the prior variance and should incorporate that knowledge. In cases where prior knowledge is completely lacking, the inverse gamma prior for ϕ^2 should be avoided in favor of the half-t distribution to ensure a proper posterior distribution.(Gelman, 2005)

The distribution of β_j^{fb} in equation 3.5 is very similar to the distribution of β_j^{sb} in equation 3.3. However, the distribution of the SB estimates is conditional on known values while the distribution of the FB estimates is conditional on random variables (ϕ^2 and θ_j). Thus, the distribution of β_j^{fb} should be averaged over these random variables before inferences are made (i.e., the inferences should be based on the marginal distribution of β_j^{fb} rather than its conditional distribution). Markov chain Monte Carlo (MCMC) sampling provides a way to generate the marginal posterior distribution. Gibbs sampling (a type of MCMC) proceeds by iteratively drawing random samples from the full conditional distributions in equations 3.5, 3.6 and 3.7, given the value of the other random variables in the previous iteration. After running the Gibbs sampler for a large number of iterations and discarding some initial number of iterations to allow for a burn-in period, the mean and variance of β_j^{fb} in the remaining samples are

the mean and variance of the marginal posterior distribution of interest. Although we implemented our Gibbs sampling routines in Matlab, they are also easily implemented in Winbugs, a freely downloadable program.(Spiegelhalter et al., 1999) Winbugs generates and runs Gibbs sampling algorithms based on modeling assumptions provided by the user, so little programming knowledge is required. We also note that MCMC algorithms generate the *exact* posterior distribution of the coefficient which will be useful in small datasets (which is also when SB and FB methods will be most useful). This is an improvement over previous methods for fitting SB models that rely on asymptotic assumptions.(Witte et al., 1998)

We analyze, under the FB model, the dataset we previously examined for the SB model. We assume the prior mean for β_j^{fb} is fixed at zero and assume that the parameters for the prior variance, ϕ^2, are $\alpha_1 = 1$ and $\alpha_2 = 1$. We ran a Gibbs sampling algorithm for 50000 iterations and excluded the first 5000 iterations as a burn-in period. The marginal posterior distributions of β_1^{fb} and ϕ^2 are presented in Figure 3.4. The mean of $\beta_1^{fb} = -0.51$, which is between the mean of the SB estimates under the assumption of a fixed $\phi^2 = 1$ ($\beta_1^{sb} = -0.56$) and $\phi^2 = .5$ ($\beta_1^{sb} = -0.43$). Although the mean of the prior variance was 1 in the FB model, $\beta_1^{fb} \ldots \beta_5^{fb}$ exhibited less variability than the prior indicated, and the posterior mean of ϕ^2 (0.87) decreased to reflect this additional information. Thus, by incorporating information on ϕ^2 that is contained in the data, we adaptively allow greater shrinkage of β_1^{fb} towards its prior mean.

Although we have focused on linear regression with orthogonal data, the results we have presented can be generalized to correlated data and logistic regression as well. It is only for computational convenience that we have focused on linear models here. We implement logistic hierarchical models in simulations and the applied example presented later in this paper.

3.4 Dirichlet Process Priors

As we will demonstrate through simulations, both SB and FB models are a distinct improvement over standard epidemiologic analytic techniques. However, results of either model may be sensitive to the assumed prior distribution of β_j and a non-parametric prior would be preferable. Further, although when sufficient prior information exists, coefficients may be grouped into exchangeable categories by incorporating second level coefficients, in many epidemiologic applications such prior knowledge may not exist. Instead, we explore a procedure that allows coefficients to be grouped into clusters

based on similarity of effect sizes before shrinking them toward a prior distribution.

One approach to solving these problems is to assume that the prior distribution of β_j is a mixture of 2 or more distributions. That is, rather than assuming $\beta_j \sim N(\eta_j, \phi^2)$, we could assume $\beta_j \sim \sum_{k=1}^{m} q_k N(\eta_k, \phi_k^2)$, where q_k is the probability of β_j being distributed as the k^{th} normal distribution and m is the number of components of the mixture distribution. Such models, while still parametric, allow for much more flexible modeling of β_j (e.g., allowing multiple modes). However, such mixture priors can be difficult to implement due to identifiability problems (for instance, there is ambiguity in the ordering of components) and also require the number of mixtures to be prespecified. Increased flexibility can be gained by allowing the number of mixture components, m, to be unknown. This can be accomplished using the DPP, which allows for nonparametric modeling of β_j and simultaneous clustering of the β_j into groups.(Richardson and Green, 1997)

In Bayesian nonparametric inference, a common method to limit the dependence of a parameter on a particular prior distribution is to let the prior distribution itself be random. For example, in the previous section we had $\beta_j \sim N(\mu, \phi^2)$. Instead, we could specify $\beta_j \sim D$, where D is an unspecified random distribution. Because D is random we place a prior distribution on it; in this case we choose a DPP, $D \sim DPP(\lambda D_0)$, where D_0 is a base distribution, such as a normal and λ is a precision parameter determining how closely D will follow D_0. As $\lambda \to \infty$, then $D \to D_0$, so the DPP converges to the parametric distribution D_0 and hence $\beta_j \sim D_0$. Smaller values of λ indicated less certainty that $\beta_j \sim D_0$.

An important property of the DPP is its ability to cluster coefficients into groups. Assuming $\beta_j \sim D$ and $D \sim DPP(\lambda D_0)$, implies the following conditional prior distribution on β_j:(West et al., 1994)

$$[\beta_j|\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k] \sim \left(\frac{\lambda}{\lambda + k - 1}\right) D_0 + \left(\frac{1}{\lambda + k - 1}\right) \sum_{i \neq j} \delta_{\beta_i} \quad (3.8)$$

were δ_{β_i} is a point mass at $\beta_j = \beta_i$. Thus, β_j has a probability of being distributed as the base distribution, D_0, or being clustered with any other $\beta_i, i \neq j$. Prior group membership is determined by the precision parameter λ, with higher probability of clustering any two coefficients together increasing as λ decreases.

A semi-parametric version of the FB model (semi-parametric because the distribution of y_i is parametric, while the distribution of β_j is non-parametric) can be specified

as:

$$\begin{aligned} y_i &\sim N(\sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2) \\ \beta_j &\sim D \\ D &\sim DP(\lambda D_0) \\ D_0 &= N(\mu, \phi^2) \\ \lambda &\sim G(a, b) \\ \phi^2 &\sim IG(\alpha_1, \alpha_2), \end{aligned} \tag{3.9}$$

where G is a gamma distribution with mean ab and variance ab^2. Placing a prior on the precision parameter, λ, serves the same function as placing a parameter on ϕ^2 in the FB model: it allows the data to help guide inference rather than relying solely on prior knowledge. Generally, relatively noninformative values are chosen for a and b, such as $a = 1, b = 1$ or $a = .01, b = .01$. However, empirical Bayes methods are available to estimate this parameter as well.(McAuliffe et al., 2005)

As with the FB model, marginal distributions are not available in closed form and estimating these parameters requires a Gibbs sampling algorithm. The properties of this model can be more clearly seen by briefly describing the Gibbs algorithm used for posterior computation (a modification of that proposed by Escobar and West).(Escobar and West, 1995, 1998) At each iteration of the Gibbs sampler, β_j is either sampled from the posterior base distribution (i.e. the distribution of D_0 after it has been updated based on the observed data) or is set equal to the value of one of the other $\beta_i, i \neq j$ coefficients. That is, each coefficient is either clustered with another coefficient or sampled from the posterior base distribution.

In many situations, such as the one in Figure 3.1, a variable selection technique may be beneficial. For efficiency, we may wish to exclude variables that have no effect on the outcome or there may be prior substantive knowledge that the exposure has no effect. In either case, modification of the base distribution D_0 in equation 3.9 allows a variable selection prior to be incorporated in a DP model. Following the approach of

Dunson et al., we specify a second DP model:(Dunson et al., 2005)

$$\begin{aligned} y_i &\sim N(\sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2) \\ \beta_j &\sim D \\ D &\sim DP(\lambda D_0) \\ D_0 &= \pi\delta_0 + (1-\pi)N(\mu, \phi^2) \\ \lambda &\sim G(a, b) \\ \pi &\sim beta(c, d) \\ \phi^2 &\sim IG(\alpha_1, \alpha_2) \end{aligned} \tag{3.10}$$

where δ_0 indicates a point mass at the value zero. The base distribution has a value of 0 with probability π, and distribution $N(\mu, \phi^2)$ with probability $1 - \pi$. This simple modification to the base distribution, allows β_j to be exactly equal to 0, in which case it is effectively removed from the regression equation, $100\pi\%$ of the time. When $\pi = 0$, this model reduces to the first DPP model. The coefficient π is given a $beta(c, d)$ distribution in order to allow the data to inform the probability that a coefficient is zero. Elicitation of c and d can proceed by specifying the expected probability, $\mathrm{E}(\pi)$, that a randomly selected coefficient is zero and the variance surrounding that estimate, $\mathrm{V}(\pi)$. Solving the equations for the mean and variance of the beta distribution:

$$\begin{aligned} c &= \frac{\mathrm{E}(\pi)^2 - \mathrm{E}(\pi)^3}{\mathrm{V}(\pi)} - \mathrm{E}(\pi) \\ \mathrm{d} &= \frac{\mathrm{E}(\pi)^3(1/\mathrm{E}(\pi) - 1)^2}{\mathrm{V}(\pi)} + \mathrm{E}(\pi) - 1. \end{aligned}$$

3.5 Performance of Models in Simulated Datasets

To assess these models (SB, FB, DPP, and DPP with selection prior) and their ability to estimate effects in a variety of scenarios, we examined their performance in simulated data. Data were simulated from the logistic model:

$$\mathrm{logit}\Pr(Y_i = 1 \,|x_{i1}, \ldots, x_{i10}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{10} x_{i10},$$

with 1) all $\beta_j = 0$; 2) $\beta_1 = 0.5$ and $\beta_2 \ldots \beta_{10} = 0$; 3) $\beta_1 \ldots \beta_5 = 0.5$ and $\beta_5 \ldots \beta_{10} = 0$; 4) $\beta_1 = 0.05, \beta_2 = 0.1, \beta_3 = 0.15 \ldots \beta_{10} = 0.5$. Each of the 4 models was simulated for

orthogonal data and for data with a correlation of 0.9 between each of the 10 variables. Datasets of 500 observations were generated from each model and were analyzed using a standard maximum likelihood logistic regression as well as a logistic regression with the priors specified in Table 3.1. Gibbs sampling algorithms to analyze each model were programmed and run in Matlab for 10,000 iterations. The initial 3000 iterations were discarded as a burn-in. Prior parameter values (shown in Table 3.1) were chosen to enhance comparability between the results of the models. We simulated each dataset 250 times and estimate the MSE (i.e., average squared difference between the model estimate and true parameter) for each of the estimates (the posterior means) as shown in figure 3.5. Because of the moderate size of the dataset, the ML, SB and FB methods all produced roughly equivalent MSE in orthogonal data. However, ML had a notably worse MSE than other methods, and FB methods performed somewhat better than SB in highly correlated data. The two DPP models generally had better performance than ML, SB, and FB models not only in terms of MSE, but also had lower type-I error rates and higher power. In the 2^{nd} set of simulations, in which only one coefficient had an effect, the DP model with a selection prior had poorer MSE for the coefficient with an effect (but still had far better MSE for the coefficients that had no effect). Even in the 4^{th} set of simulations, where none of the coefficients had the same effect, the two DPP models performed somewhat better than the FB and SB models.

3.6 Application to Study of Pesticides and Retinal Degeneration

The Agricultural Health Study (AHS) enrolled farmers who applied for pesticide licenses in Iowa or North Carolina between 1993 and 1997 and has been described in more detail elsewhere.(Alavanja et al., 1996) Kirrane et al. recently examined the association between pesticide exposure and retinal degeneration among the wives of AHS farmers.(Kirrane et al., 2005) A questionnaire was sent to spouses of farmers to determine if they had retinal degeneration and to determine, among other things, the types of pesticide they had used. We analyzed the same cohort Kirrane et al. used in their analysis (31,173 women,281 of whom experienced retinal degeneration), but limit our analysis only to herbicides, of which there are 18 unique chemicals. Table 3.2 shows the 4 hierarchical models used to analyze the data. Gibbs sampling algorithms were programmed in Matlab and run for 50,000 iterations with the initial 5,000

excluded as a burn-in period. The results of the models are presented in Table 3.3. The maximum likelihood model estimated a large effect of imazethapyr on macular degeneration (OR=2.6, 95% CI (1.0, 6.3)). The result is statistically significant but imprecise. The hierarchical models shrunk this result toward the prior distribution to varying degrees. The SB model produced an effect of imazethapyr that was no longer statistically significant but still markedly elevated (OR=1.7, 95% CI (0.8, 3.6)). The FB, DPP1 and DPP2 models were all in agreement, indicating little evidence of effect of imazethapyr on macular degeneration. Because little variability was observed between estimated coefficients, the posterior of the prior variance, ϕ^2, was much smaller in the FB, DPP1 and DPP2 models than its fixed value in the SB model and subsequently greater shrinkage was observed in these models.

3.7 Discussion

Highly correlated data are common in epidemiologic research, however standard analytic techniques can produce extremely imprecise confidence intervals or fail altogether in this setting. In this paper, we have examined four Bayesian models for use in this context; however, these models may have broad use beyond highly correlated data settings (for example in problems with a large number of covariates).

When deciding which of the four models to use in an analysis, consideration should be given to the properties of each model as well as the computational skill required to implement them. Of the four methods, SB and FB are the easiest approaches computationally. Either model can be easily implemented in Winbugs, using the code we provide in the appendix as a starting point. The advantages of the FB approach over the SB approach justify its use despite the (very) minor increase in computation. SB estimates assume a fixed prior variance, while FB estimators update the prior variance based on the observed data. This 'Bayesian learning' allows for adaptive shrinkage in FB models and makes their estimates more data-driven and less sensitive to prior specification than SB estimates. Further, in some epidemiologic settings the prior variance may be a parameter of scientific interest and rather than specifying its value (as in SB models), the researcher may wish to estimate it. However, as the sample size increases, the difference between FB and SB (and MLE) will tend to decrease. In orthogonal (or nearly orthogonal data), specifying a prior distribution as FB or SB may make little difference unless the prior variance is very small. In highly correlated data specification of a prior distribution can make a large difference.

Although more computationally intensive than the FB or SB models, the two DPP models discussed in section 3 have very desirable properties in many situations. In models where some coefficients have similar values, the DPP models decrease MSE by aggregating data within clusters of coefficients. Indeed, even if the clustered coefficients are not exactly identical (as in the 4^{th} set of simulations in Figure 3.5), occasionally clustering them together can still reduce MSE. However, when clustering of coefficients can occur, the DPP models perform remarkably well. The decision to implement the DPP with or without the selection component should be made on substantive grounds. When researchers have a high prior probability that many of the effects in question may be zero, the selection prior can help estimation. However, when the true value of most coefficients is zero and only a few coefficients are non-zero (but still close to zero), the selection prior could perform slightly worse than the DPP model without the selection prior.

In summary, the difficulties of analyzing highly correlated data can be greatly diminished through Bayesian methods. The SB, FB and two DPP models we examine in this paper provide useful alternatives to current ML techniques. FB models are generally superior to SB models and are easily implemented in Winbugs. DPP models, although more difficult to implement, often have better performance than other methods.

TABLE 3.1: Hierarchical models used in analysis of simulated data.

SB*			FB*		
β_j	$\sim$	$N(0,1)$	β_j	$\sim$	$N(0,\phi^2)$
			ϕ^2	$\sim$	$IG(3,1/2)$

DPP1*			DPP2*		
β_j	$\sim$	D	β_j	$\sim$	D
D	$\sim$	$DP(\lambda D_0)$	D	$\sim$	$DP(\lambda D_0)$
D_0	$=$	$N(0,\phi^2)$	D_0	$=$	$\pi\delta_0 + (1-\pi)N(0,\phi^2)$
λ	$\sim$	$G(5,1)$	λ	$\sim$	$G(5,1)$
ϕ^2	$\sim$	$IG(3,1/2)$	ϕ^2	$\sim$	$IG(3,1/2)$
			π	$\sim$	$beta(1,1)$

* SB=semi-Bayes; FB=fully-Bayes; DPP1=Dirichlet process prior; DPP2=Dirichlet process prior with selection component

TABLE 3.2: Hierarchical models used to analyze Agricultural Health Study data on herbicides and macular degeneration.

SB*			FB*		
β_j	$\sim$	$N(0, .35)$	β_j	$\sim$	$N(0, \phi^2)$
			ϕ^2	$\sim$	$IG(2.1, 2.5)$

DPP1*			DPP2*		
β_j	$\sim$	D	β_j	$\sim$	D
D	$\sim$	$DP(\lambda D_0)$	D	$\sim$	$DP(\lambda D_0)$
D_0	$=$	$N(0, \phi^2)$	D_0	$=$	$\pi\delta_0 + (1-\pi)N(0, \phi^2)$
λ	$\sim$	$G(1, 1)$	λ	$\sim$	$G(1, 1)$
ϕ^2	$\sim$	$IG(2.1, 2.5)$	ϕ^2	$\sim$	$IG(2.1, 2.5)$
			π	$\sim$	$beta(1.5, 1.5)$

* SB=semi-Bayes; FB=fully-Bayes; DPP1=Dirichlet process prior; DPP2=Dirichlet process prior with selection component

TABLE 3.3: Estimated effects of exposure to herbicides on retinal degeneration among the wives of pesticide applicators, Agricultural Health Study, North Carolina and Iowa, 1993-1997.

Herbicide	**MLE**[*] OR[*,‡] (95% CI[*])	**SB**[*] OR[*,‡] (95% CI[*])	**FB**[*] OR[*,‡] (95% CI[*])	**DPP1**[*] OR[*,‡] (95% CI[*])	**DPP2**[*] OR[*,‡] (95% CI[*])
Imazethapyr	2.6 (1.0, 6.3)	1.7 (0.8, 3.6)	1.2 (0.8, 2.1)	1.1 (0.5, 2.2)	1.1 (0.7, 1.7)
Chlorimuron ethyl	1.9 (0.7, 5.0)	1.4 (0.6, 3.1)	1.2 (0.7, 2.0)	1.0 (0.5, 1.9)	1.0 (0.7, 1.5)
Alachlor	1.4 (0.6, 3.1)	1.2 (0.6, 2.3)	1.1 (0.7, 1.7)	0.9 (0.5, 1.6)	1.0 (0.7, 1.3)
Petroleum oil	1.4 (0.7, 2.9)	1.3 (0.7, 2.3)	1.2 (0.7, 1.8)	0.9 (0.5, 1.7)	1.0 (0.7, 1.4)
2,4,5-TP[†]	1.3 (0.1, 11.2)	1.0 (0.3, 2.7)	1.0 (0.6, 1.7)	1.0 (0.5, 1.9)	1.0 (0.7, 1.4)
2,4-D[†]	1.3 (0.8, 1.9)	1.2 (0.8, 1.8)	1.1 (0.8, 1.6)	1.0 (0.6, 1.6)	1.0 (0.8, 1.3)
Butylate	1.1 (0.3, 3.9)	1.0 (0.4, 2.4)	1.0 (0.6, 1.7)	1.0 (0.5, 1.7)	1.0 (0.7, 1.4)
Gylphosate	1.1 (0.8, 1.5)	1.1 (0.8, 1.4)	1.1 (0.8, 1.4)	0.9 (0.6, 1.5)	1.0 (0.7, 1.3)
Dicamba	1.0 (0.4, 2.2)	1.0 (0.5, 1.9)	1.0 (0.6, 1.6)	0.9 (0.5, 1.8)	1.0 (0.8, 1.3)
Trifluralin	1.0 (0.5, 2.1)	1.0 (0.5, 1.9)	1.0 (0.7, 1.6)	0.9 (0.4, 1.9)	1.0 (0.7, 1.4)
Cyanazine	0.9 (0.3, 2.5)	0.9 (0.4, 1.9)	1.0 (0.6, 1.6)	0.8 (0.3, 2.1)	1.0 (0.7, 1.4)
Metribuzin	0.9 (0.3, 3.1)	0.9 (0.4, 2.2)	1.0 (0.6, 1.6)	0.8 (0.4, 2.0)	1.0 (0.7, 1.4)
EPTC[†]	0.8 (0.2, 3.4)	0.9 (0.4, 2.2)	0.9 (0.6, 1.6)	0.9 (0.5, 1.8)	1.0 (0.7, 1.4)
2,4,5-T[†]	0.7 (0.1, 3.2)	0.8 (0.3, 2.0)	0.9 (0.5, 1.6)	0.9 (0.5, 1.7)	1.0 (0.7, 1.3)
Atrazine	0.6 (0.2, 1.4)	0.7 (0.3, 1.4)	0.8 (0.5, 1.4)	0.7 (0.3, 1.9)	0.9 (0.6, 1.4)
Metolachlor	0.5 (0.2, 1.4)	0.6 (0.3, 1.4)	0.8 (0.5, 1.4)	0.7 (0.3, 2.1)	1.0 (0.6, 1.4)
Pendimethalin	0.5 (0.2, 1.6)	0.7 (0.3, 1.6)	0.9 (0.5, 1.5)	0.7 (0.3, 2.2)	0.9 (0.6, 1.5)
Paraquat	0.3 (0.0, 2.1)	0.6 (0.2, 1.6)	0.8 (0.5, 1.5)	0.8 (0.3, 2.1)	1.0 (0.6, 1.5)

* OR, odds ratio; CI, confidence interval for ML and credible interval for SB, FB, DPP1, and DPP2; MLE, maximum likelihood estimate; SB, semi-Bayes; FB, fully-Bayes; DPP1, Dirichlet process prior; DPP2, Dirichlet process prior with selection component

† 2,4,5-TP, 2,4,5-trichlorophenoxypropionic acid; 2,4,5-T, 2,4,5-trichlorophenoxyacetic acid; 2,4-D, 2,4-dichlorophenoxyacetic acid; EPTC, S-ethyl dipropylthiocarbamate

‡ All models adjusted for state and age.

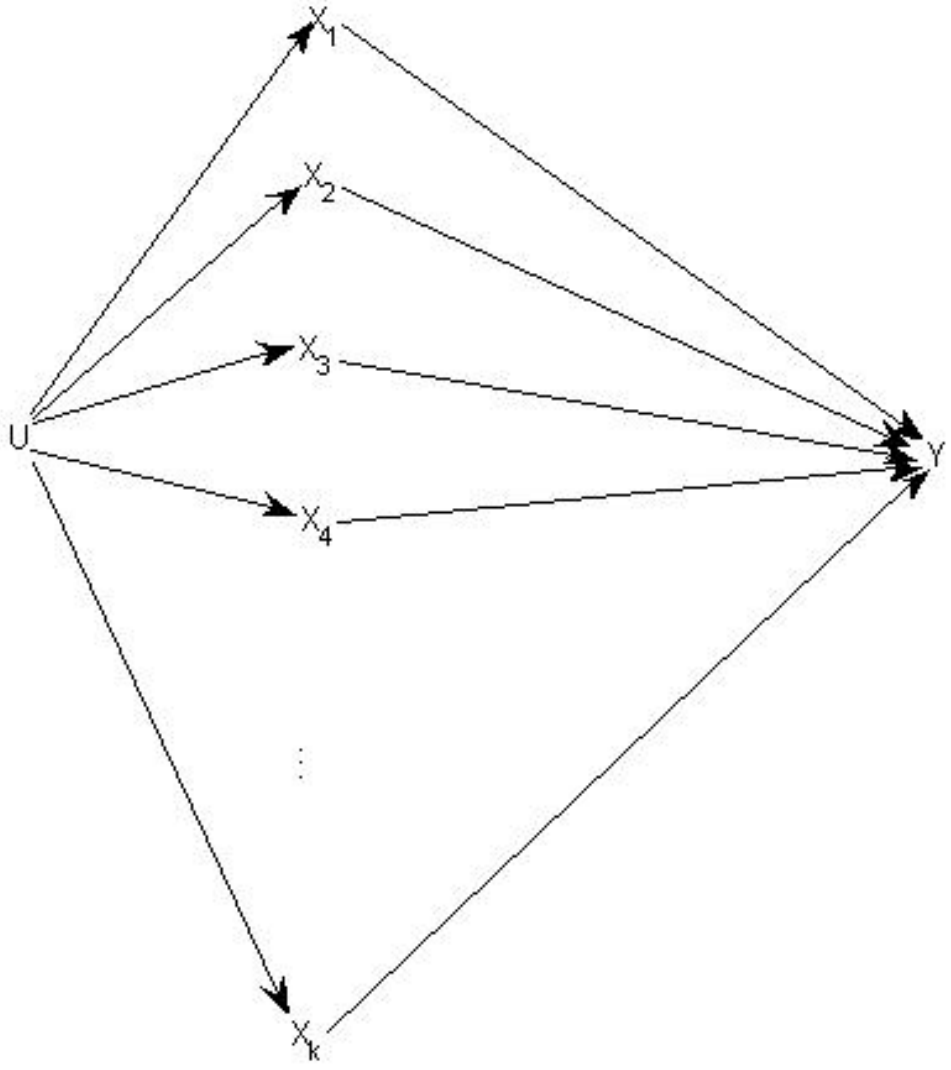

FIGURE 3.1: DAG for correlated exposure variables.

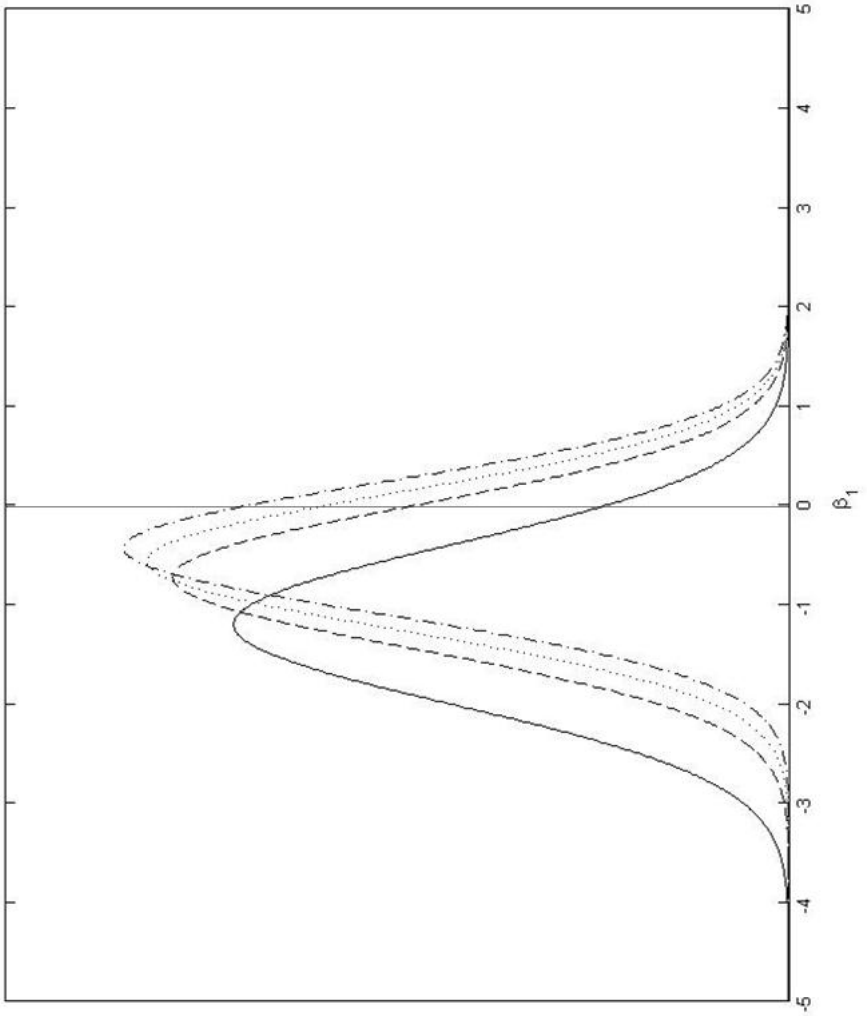

FIGURE 3.2: Distribution of SB and ML estimators.

Figure 3.2: Distribution of SB and ML estimators.
solid line: distribution of ML estimator.
dashed line: distribution of β_1^{sb} with $\phi_j^2 = 2$
dotted line: distribution of β_1^{sb} with $\phi_j^2 = 1$
dash-dot line: distribution of β_1^{sb} with $\phi_j^2 = .5$
verticle line: true value of β_1

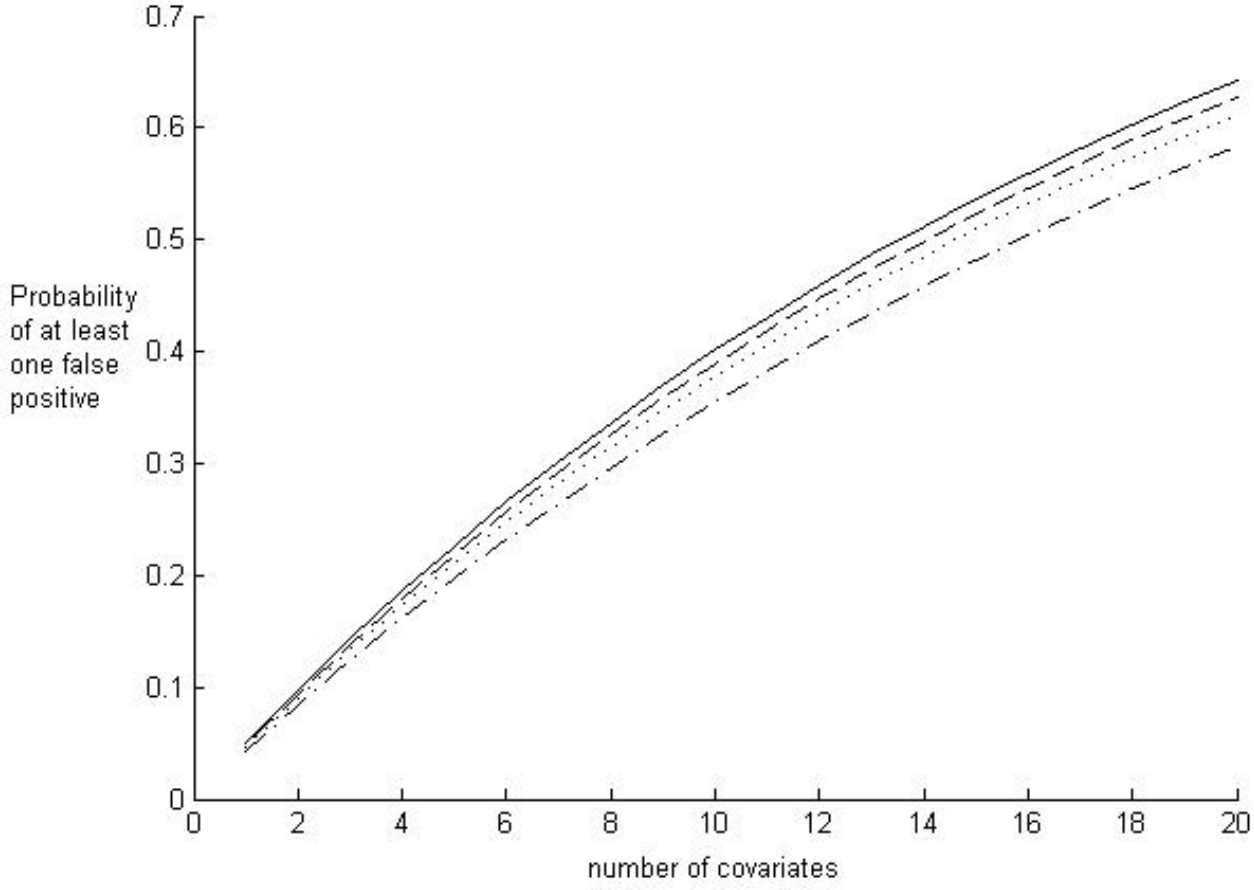

FIGURE 3.3: Probability of finding at least one false positive result in SB models as the number of covariates increases.

Figure 3.3: Probability of finding at least one false positive result in SB models as the number of covariates increases.
solid line: ML estimate.
dashed line: $\phi_j^2 = 2$
dotted line: $\phi_j^2 = 1$
dash-dot line: $\phi_j^2 = .5$

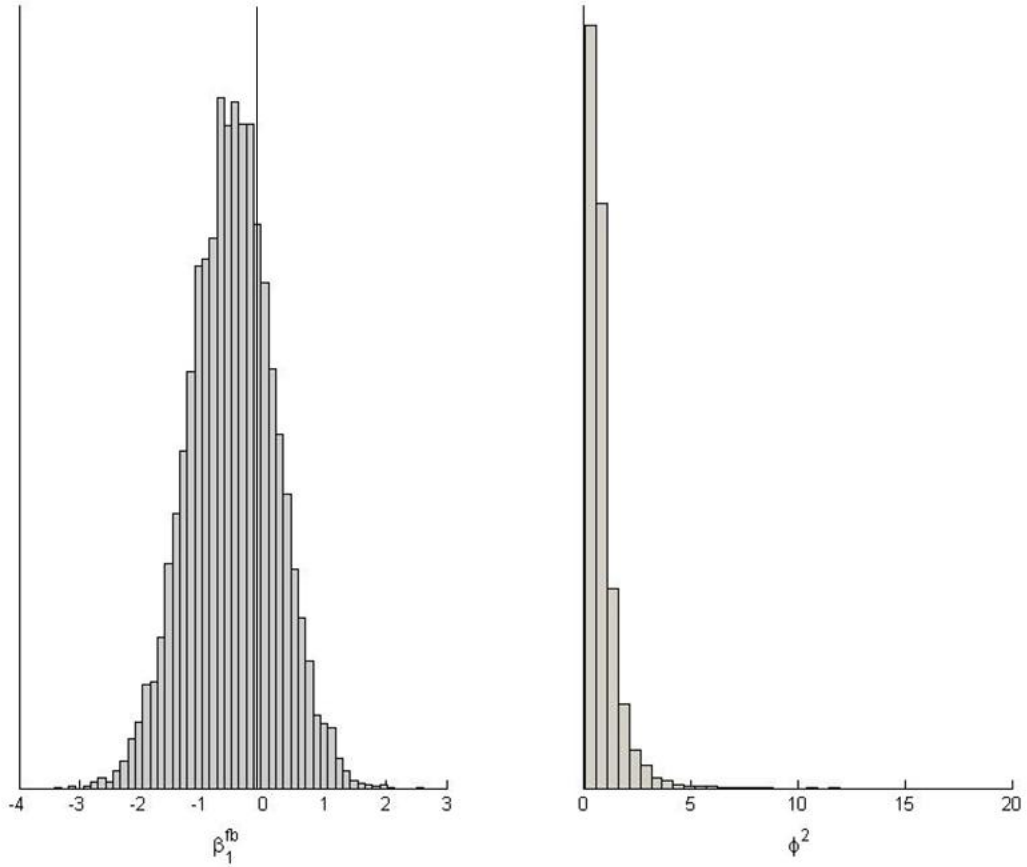

FIGURE 3.4: Distribution of β_1^{fb} and ϕ^2 in FB analysis with $\alpha_1 = 1$ and $\alpha_2 = 1$.

Figure 3.4: Distribution of β_1^{fb} and ϕ^2 in FB analysis with $\alpha_1 = 1$ and $\alpha_2 = 1$. Verticle line: true value of β_1.

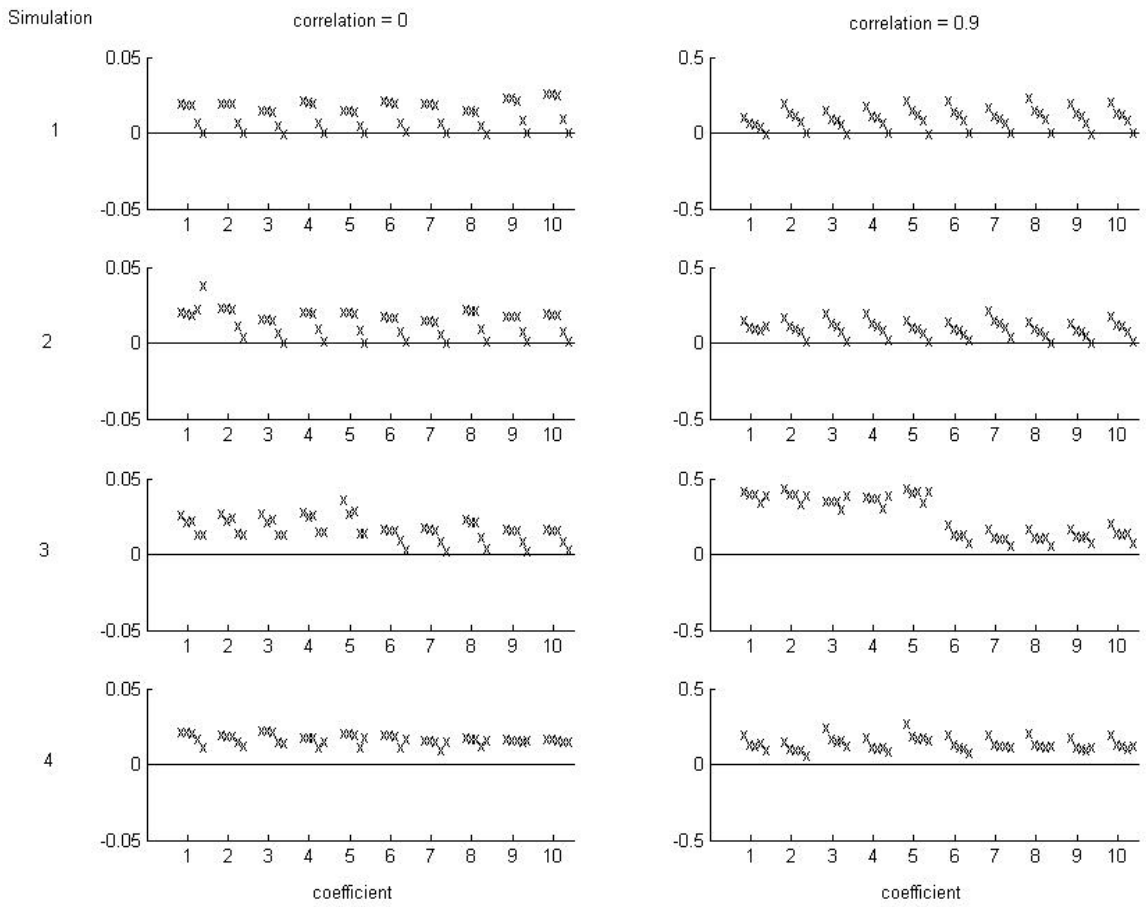

FIGURE 3.5: Mean squared error of parameter estimates under different combinations of coefficient effects and correlation. The parameter estimates from the 5 models (MLE, SB, FB, DPP, DPP with selection component) are grouped in order within each of the 10 coefficients.

Figure 3.5: Mean squared error of parameter estimates under different combinations of coefficient effects and correlation. The parameter estimates from the 5 models (MLE, SB, FB, DPP, DPP with selection component) are grouped in order within each of the 10 coefficients.
label:
1) all $\beta_j = 0$ 2) $\beta_1 = 0.5$ and $\beta_2 \dots \beta_{10} = 0$ 3) $\beta_1 \dots \beta_5 = 0.5$ and $\beta_5 \dots \beta_{10} = 0$, 4) $\beta_1 = 0.05, \beta_2 = 0.1, \beta_3 = 0.15 \dots \beta_{10} = 0.5$

3.8 Appendix 1

Bayesian credible intervals, the region within which we are $100(1-\alpha)\%$ certain β_j lies given our prior knowledge, can be calculated using the posterior mean and variance in equation 3.3:

$$E_j^{sb} - z_{\alpha/2}\sqrt{V_j^{sb}} \leq \beta_j \leq E_j^{sb} + z_{\alpha/2}\sqrt{V_j^{sb}}$$

While credible intervals give our certainty regarding the size of β_j, frequentist confidence intervals only offer $100(1-\alpha)\%$ coverage probability of the true effect over repeated studies. Credible intervals do not generally guarantee the same coverage probability as confidence intervals. Instead, the frequentist coverage probability of Bayesian credible intervals can be calculated as:

$$2 \times \Phi\left(z_{\alpha/2}\left(1 + \frac{\sigma^2}{\phi_j^2 \sum x_{ij}^2}\right)^{1/2} - \frac{\mu_j - \beta_j}{\phi_j^2}\left(\frac{\sigma^2}{\sum x_{ij}^2}\right)^{1/2}\right) \tag{3.11}$$

where Φ is the cumulative probability of the standard normal distribution. The probability of covering the true parameter is plotted in Figure 3.6 by values of the prior mean and variance in a sample dataset. It is clear from the figure that how well the credible intervals cover the true parameter value depends on the specification of the prior distribution. As the prior variance increases, the credible intervals provide nominal $100(1-\alpha)\%$ coverage of the true parameter. As the prior variance decreases (i.e., as more belief is placed in the prior mean), coverage generally decreases as the credible intervals become increasingly narrow about the prior mean. The exception occurs when the prior mean is equal to the true mean and coverage increases as the prior variance decreases. As a special case (the dotted line in Figure 3.6), consider testing the null hypothesis that $\beta_j = 0$. The increased coverage of the null hypothesis when $\eta_j = 0$ implies that the SB estimate will be less likely to flag a result as significant than the MLE.

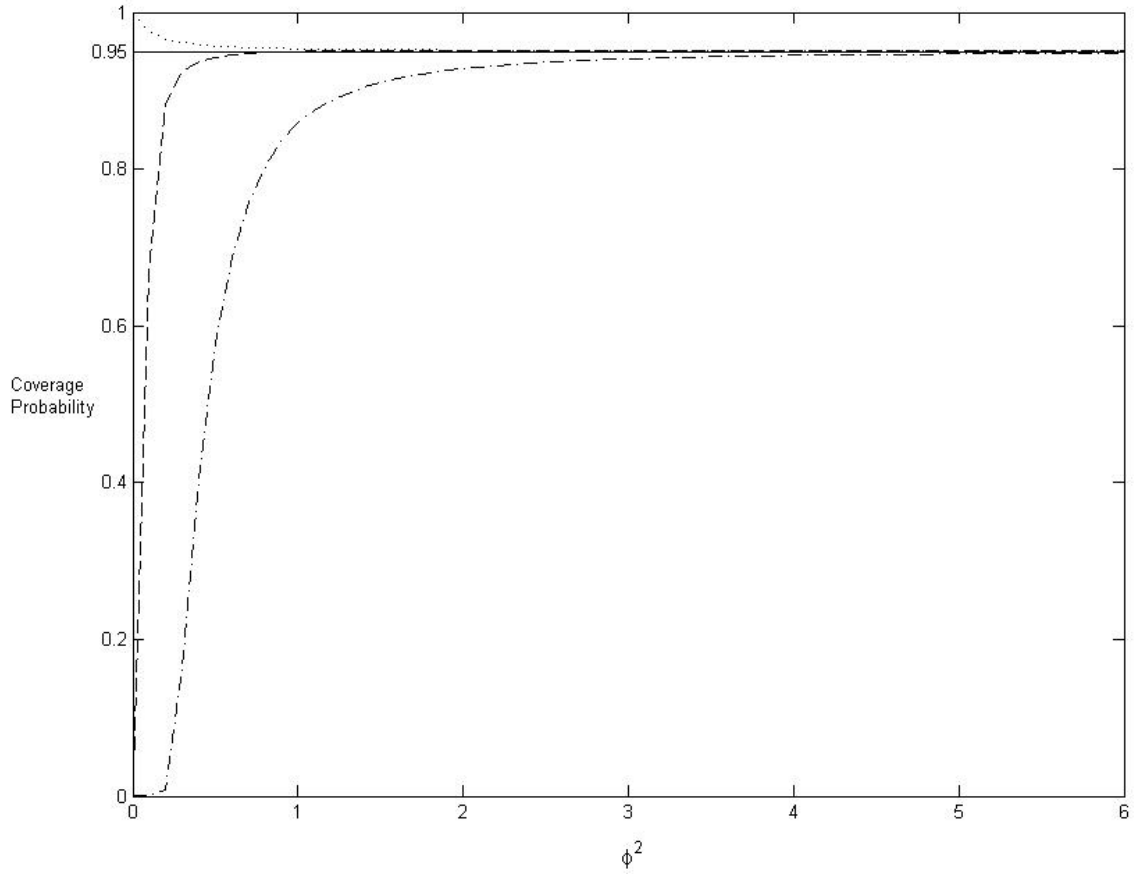

FIGURE 3.6: Coverage probability for credible intervals by prior mean and variance.

Figure 3.6: Coverage probability for credible intervals by prior mean and variance.
dotted line: prior mean=0
dashed line: prior mean=1
dash-dot line: prior mean=5

CHAPTER 4

A BAYESIAN HIERARCHICAL ANALYSIS OF DISINFECTION BY PRODUCTS AND SPONTANEOUS ABORTION

4.1 Abstract

Spontaneous abortion (SAB) is a common pregnancy outcome, with over 30% of all pregnancies ending in loss. Previous research suggests an increased risk of SAB among those who consume higher amounts of tap-water disinfection by-products (DBPs). Right from the Start is a large multi-site cohort study of women's exposure to DBPs followed through early pregnancy. We examined the effect of 13 constituent DBPs (4 trihalomethanes and 9 haloacetic acids) on SAB. Some of the constituent DBPs are highly correlated making conventional maximum likelihood regression models containing all DBPs unreliable. To allow simultaneous estimation of effects, we implemented 4 Bayesian hierarchical models : semi-Bayes (SB), fully-Bayes (FB), Dirichlet process prior (DPP1) and Dirichlet process prior with a selection component (DPP2). Models that allowed prior parameters to be updated from the data gave far more precise coefficients and were more robust to prior specification. The DPP1 and DPP2 models were in close agreement in estimating no effect of any constituent DBP on SAB. The FB model largely agreed with the DPP1 and DPP2 models but had less precision, while the SB model provided the least precise estimates. Our results suggest none of the constituent DBPs have an effect on SAB.

4.2 Introduction

Spontaneous abortion (SAB), defined as a pregnancy loss prior to 20 weeks of completed gestation, is a common occurrence, with over 30% of all pregnancies ending in a loss.(Wilcox et al., 1988) Increased risk of SAB has consistently been associated with advanced maternal age, smoking and prior spontaneous abortion.(Coste et al., 1991; Ness et al., 1999) Caffeine consumption and exposure to industrial solvents and heavy metals have also been associated with SAB, though with less consistency.(Fenster et al., 1991, 1997; Hertz-Picciotto, 2000; Savitz et al., 1994)

In the 1980's, a series of epidemiologic studies found an association between high consumption of tap water during pregnancy (relative to low consumption of tap water) and SAB.(Aschengrau et al., 1989; Fenster et al., 1992; Swan et al., 1992, 1998; Windham et al., 1992; Wrensch et al., 1992) Mechanisms through which increased consumption of tap water could increase the risk of SAB are unknown. However, certain disinfection by-products (DBPs) present in tap-water have been consistently associated with an increased risk of bladder and colorectal cancer and have been shown to have fetotoxic effects in rats.(Mughal, 1992; Nieuwenhuijsen et al., 2000) Chlorine, the most common drinking water disinfectant in the United States, combines with organic matter in the water supply to produce DBPs. Two classes of DBPs have been subject to regulation in the U.S.: trihalomethanes (THMs), consisting of $CHCl_3$, $CHBrCl_2$, $CHBr_2Cl$, and $CHBr_3$ and haloacetic acids (HAAs), consisting of ClAA, Cl_2AA, Cl_3AA, BrAA, Br_2AA, Br_3AA, BrClAA, $BrCl_2AA$, and Br_2ClAA. Little epidemiologic research exists regarding HAAs and SAB. Results of previous research are inconsistent but suggest there may be an increased risk of SAB among women with higher intake of THMs (particularly $CHBrCl_2$).(Savitz et al., 1995; Waller et al., 1998)

The purpose of this study is to estimate the effects of the tap water concentration of 13 constituent DBPs on SAB. Although, concentration of DBPs may be less biologically relevant than amount of ingested DBPs, we believe tap water concentration serves as a meaningful proxy. Because each of the DBPs depend on shared factors (i.e., the concentration of bromides and organic matter in the water reservoir and the concentration of chlorine used in the disinfection process), the effect of any one of the 13 DBPs may be confounded by the remaining 12, so all must be retained in a regression model to produce unbiased estimates. A standard maximum likelihood logistic regression that includes all 13 constituent DBPs would result in unstable estimates because of the high correlation among the DBPs. Instead, we adopt a hierarchical Bayesian approach

that allows us to stabilize parameter estimates while incorporating prior knowledge regarding the effects of the constituent DBPs on SAB.(MacLehose et al., 2005)

4.3 Methods

4.3.1 Study Design

Right from the Start (RFTS) was a prospective cohort study designed to investigate effects of DBPs on spontaneous abortion.(Promislow et al., 2004) A diverse cohort of 2483 women over 18 was enrolled from 3 metropolitan areas between 2001 and 2004. Women were eligible for enrollment if they could speak English or Spanish, had not used assisted reproductive technology to conceive, intended to carry the pregnancy to term, and did not plan to move outside of the area of study. A baseline interview was conducted to collect information on potential confounding factors and last menstrual period, which was used to date the onset of pregnancy. The time at which a pregnancy loss occurred was determined by self-report or chart abstraction. Water samples were taken at treatment facilities in the 3 metropolitan areas weekly in two sites and once every 2 weeks in the other site that had low DBP levels. EPA standard methods were used to estimate the concentration of THM and HAA, respectively.(EPA, 1995a,b) A concentration for each of the 13 constituent DBPs for each gestational week was assigned for all women in the study.

Although some women enrolled in the study for multiple pregnancies, this analysis is limited to the first pregnancy for which a woman was enrolled. We excluded 4 losses that occurred before the 5^{th} week of gestation due to inability to routinely detect pregnancies at such an early stage for a final sample size of 2407 women.

4.3.2 Analysis

We specify a discrete time hazard model for the probability that a SAB occurs in a given gestational week, which is analogous to a continuation odds ratio model.(Cole and Ananth, 2001) We included terms in the model for gestational week specific intercepts (i.e., one coefficient for each week to allow probability of SAB to vary by gestational week), potential confounders and 13 constituent DBPs. DBP concentrations change over the course of a woman's pregnancy, so the 13 constituent DBPs were included as time-varying covariates. To allow for a more flexible relationship between the logit of the probability of SAB and the concentration of the DBPs, we categorized 8 constituent

DBPs into quartiles. Five of the HAAs could not be categorized into quartiles because of their scarcity and were categorized into tertiles or dichotomized. Estimating the effects of the 32 categorized DBP coefficients ($\beta_1 \ldots \beta_{32}$) is hindered by the high correlation between many of the constituent DBPs (greater than 90% between some), a situation in which standard maximum likelihood techniques are known to perform poorly.

As an alternative, we incorporated prior knowledge about the size of $\beta_1 \ldots \beta_{32}$ using Bayesian methods. We implemented four hierarchical models (Table 4.1), which placed slightly different prior distributions on $\beta_1 \ldots \beta_{32}$: semi-Bayes (SB), fully-Bayes (FB), Dirichlet process prior (DPP1), and Dirichlet process prior with selection component (DPP2).(Dunson et al., 2005; Greenland, 1992; MacLehose et al., 2005) Coefficients from each of these models are shrinkage estimates since they are slightly biased towards the prior distribution, but have reduced variance resulting in a smaller mean squared error than maximum likelihood techniques.

4.3.3 Semi-Bayes (SB) Model

The SB model assumes the j^{th} coefficient, $\beta_j = \ln(\mathrm{OR}_j)$ with $j = 1 \ldots 32$, has a prior mean μ_j and prior variance ϕ_j^2 (Table 4.1).(Greenland, 1992) The prior mean characterizes our knowledge regarding the effect of the j^{th} category of constituent DBP on SAB, and the prior variance is our certainty in that knowledge. Ideally, we would use previous research to inform prior values for μ_j. Unfortunately, no previous epidemiologic studies specifically examined the effect of HAAs on SAB, and while some studies examined the effect of THMs on SAB, only one gives results for the effect of the four constituent THMs.(Waller et al., 1998) Further, the definition of exposure in that study was based on a woman consuming more than 5 glasses of water per day and falling in the highest exposure quartile. Their study observed a greater range of THM exposure than was seen in RFTS. Consequently, this study provided little guidance for choosing prior values for the effect of the four THMs on SAB.

Initially, we conservatively assumed that none of the constituent DBPs (HAAs and THMs) has an effect on SAB, which corresponds to assuming $\mu_j = \ln(1) = 0$ for all j. Next, we specified our certainty regarding the prior mean. Again, no previous research exists to explicitly help us quantify the uncertainty regarding the effect of DBPs on SAB, but the largest deviation from the null of any constituent DBP or summary DBP measure from any previous study was from Waller et al. who noted an OR=3.0.(Waller et al., 1998) For this initial analysis, we assumed that 95% of the

coefficients of interest have an OR between 3.0 and 1/3 and calculate ϕ^2 as: $\phi^2 = ((\ln(3) - \ln(1/3))/(2 * 1.96))^2 = 0.3142$. Because ϕ^2 plays an important role in the degree of shrinkage, it is desirable to assess the sensitivity of results to a variety of values, which we do below.

4.3.4 Fully-Bayes (FB) Model

The SB model extends the traditional frequentist regression analysis by treating $\beta_1 \ldots \beta_{32}$ as random. The FB model we propose analogously extends the SB model by treating ϕ^2 as random and placing a prior distribution on them. This allows us to incorporate substantive knowledge regarding the prior variance while also allowing the data to help inform about it. For instance, our prior guess at the variability among the ORs may be much larger (or smaller) than the observed variability. The FB model estimates ϕ^2 as a weighted average of our initial belief of the prior variance and the observed variance of the estimates, resulting in a more data-driven procedure.

To proceed with the FB analysis, we chose values of the hyperparameters α_1 and α_2 (Table 4.1) of the inverse gamma distribution for ϕ^2. First, we specified our best guess for the prior variance, $\mathrm{E}(\phi^2)$. In keeping with our reasoning for the SB analysis, it made sense to choose $\mathrm{E}(\phi^2) = 0.3142$. Next, we chose a value for the variance of ϕ^2, $\mathrm{V}(\phi^2)$, such that $\mathrm{E}(\phi^2) \pm 1.96\sqrt{\mathrm{V}(\phi^2)}$ contains 95% of reasonable ϕ^2 values. An OR$\geq$6.0 or OR$\leq$1/6 for any of the constituent DBPs would be extremely unlikely. The value of ϕ^2 if 95% of OR's fall between 6.0 and 1/6 is $\phi^2 = ((\ln(6) - \ln(1/6))/(2 \times 1.96))^2 = 0.8357$. We treated this value of ϕ^2 as the upper 95% CI and used it to calculate $\mathrm{V}(\phi^2)$. Since $0.3142 + 1.96\sqrt{\mathrm{V}(\phi^2)} = 0.8357$, then $\mathrm{V}(\phi^2) = 0.0708$. Values for α_1 and α_2 were calculated as 3.39 and 1.33, respectively, using the formulae: $\alpha_1 = \mathrm{E}(\phi^2)^2/\mathrm{V}(\phi^2) + 2$ and $\alpha_2 = (\mathrm{E}(\phi^2)^3/\mathrm{V}(\phi^2) + \mathrm{E}(\phi^2))^{-1}$.

4.3.5 Dirichlet Process Prior (DPP1) Model

The third hierarchical model allowed us to avoid specifying a particular family of distributions (such as the normal family) for $\beta_1 \ldots \beta_{32}$, while simultaneously clustering them into groups based on the magnitude of their effects. Very little prior information exists on the effects of the constituent DBPs on SAB, and there may be classes of DBPs that have similar coefficients; for example, all brominated haloacetic acids could have a similar effect. The DPP1 model automatically clusters the coefficients into groups, without any prior specification of what the groups might be. The probability that two

coefficients are clustered together depends on how similar two coefficients are and a parameter λ. The more similar two coefficients are the more likely they are to be clustered together, and the smaller λ is, the more likely they are to be clustered together. These clusters could be of great regulatory and scientific interest. In this model, along with treating $\beta_1 \ldots \beta_{32}$ and ϕ^2 as random, we treat the distribution of $\beta_1 \ldots \beta_{32}$ as random. This random distribution, D, may be similar to a base distribution, D_0, where the similarity depends on a precision parameter, λ. If λ is small, D will not resemble D_0, but if λ is large, D converges to D_0, and the DPP1 model is equivalent to the FB model. We placed a prior distribution on λ to allow the data to help determine its value, in the same way placing a prior on ϕ^2 allowed the data to help inform about its value.

To complete the DPP1 model, we specified values for $\mu, \alpha_1, \alpha_2, a$ and b (Table 4.1). Specification of μ, α_1, and α_2 in the DPP1 model was identical to that in the FB model. The parameters a and b were prior parameters for λ and determined how closely D follows D_0. We began our analysis with a fairly noninformative choice of $a = 1$ and $b = 1$.

4.3.6 Dirichlet Process Prior with Selection Component (DPP2) Model

The fourth hierarchical model modified the DPP1 model by incorporating a group of chemicals that has no effect ($\beta_j = 0$). Priors that allow zero coefficients are commonly referred to as selection priors, because if $\beta_j = 0$, then the j^{th} predictor is effectively excluded from the model.(Geweke, 1996) This conveniently serves two purposes. First, if a constituent DBP has no effect, it cannot confound the effect of any of the other DBPs, and we would prefer to remove it from the model. Second, when estimating the effects of a large number of exposures, some of them may have no effect. The variable selection prior we used assigns a prior probability, π, that a randomly selected coefficient is zero. In similar fashion to the approach we took in the FB and DPP1 model, we assigned a prior distribution to π in order to allow the data to inform the proportion of chemicals having no effect ($\beta_j = 0$).

The parameters in the DPP2 model were identical to the DPP1 model except for the addition of π, whose prior distribution required specification of c and d. We calculated these values by specifying our belief that a randomly selected coefficient was zero, $\mathrm{E}(\pi)$, and the variability of that estimate, $\mathrm{V}(\pi)$. We began our analyses by specifying that

a coefficient being null was E(π)=0.5, but with a fairly large variance, V(π)=0.0625 (corresponding to 95% confidence intervals: 0.01, 0.99). By using the equations,

$$\begin{aligned} c &= \frac{\mathrm{E}(\pi)^2 - \mathrm{E}(\pi)^3}{\mathrm{V}(\pi)} - \mathrm{E}(\pi) \\ \mathrm{d} &= \frac{\mathrm{E}(\pi)^3(1/\mathrm{E}(\pi) - 1)^2}{\mathrm{V}(\pi)} + \mathrm{E}(\pi) - 1, \end{aligned}$$

we specified $c = 1.5$ and $d = 1.5$. Alternatively, a similar approach to estimating π would be to specify the probability that none of the coefficients have an effect (π_{all}) and solve for π in the equation: $\pi_{all} = \pi^p$, where p is the number of coefficients.

4.3.7 Week Specific Risk of SAB

We completed our specification of the discrete time hazard model by placing prior distributions on the gestational-week specific probabilities of SAB. Because coefficients for the week specific probability of SAB in our model were log-odds, we calculated the log-odds of SAB and variances of these log-odds using Goldhaber and Fireman's results.(Goldhaber and Fireman, 1991) To illustrate, for the 10^{th} week of gestation, Goldhaber and Fireman reported 62 losses occurring among 4437 pregnancies at risk in that week, for a risk of 1.4%. This translates to $\ln(p/(1-p)) = -4.27$ and V($\ln(p/(1-p))$)=0.016. We used these results to place a $N(-4.27, .016)$ prior on the coefficient for the probability of a SAB in the 10^{th} week. Priors for remaining weeks were calculated in the same manner.

4.3.8 Sensitivity Analysis

The results of any analysis depend heavily on modeling assumptions. In a Bayesian analysis, there may be concern over the specification of the prior distribution. It is important to alter those specifications over a plausible range of values to assess how those changes modify our results.

Our choice of $\mu_j = 0$ in all four priors may be overly conservative in light of epidemiologic studies that have found an increased risk of SAB among women who report higher consumption of tap-water and among women who are exposed to higher levels of THMs. To address this, we chose two alternative specifications for sensitivity analyses: $\mu_j = \ln(6.9) = 1.9$ and $\mu_j = \ln(3.0) = 1.1$. The first specification was the most extreme result observed among studies examining the effect of drinking tap water on

SAB; the second was the most largest OR among studies examining the effect of THM concentration on SAB.(Savitz et al., 1995; Wrensch et al., 1992)

To assess the impact of our specification of the prior variance in the SB analysis, we varied it from 0.1 (strong prior belief in the value of μ_j) to 5.0 (weak belief in the value of μ_j). We varied the prior parameters for ϕ^2 in the FB, DPP1, and DPP2 models in a similar fashion. We chose values of $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ that correspond to E(ϕ^2)=0.1, 0.3 and 5.0 with a large variance, V(ϕ^2)=3.0: $\boldsymbol{\alpha} = (2, 10)$, $\boldsymbol{\alpha} = (2, 3.2)$ and $\boldsymbol{\alpha} = (2.1, 1.8)$; and a small variance, V(ϕ^2)=0.1: $\boldsymbol{\alpha} = (2.1, 9)$, $\boldsymbol{\alpha}(2.90, 1.8)$ and $\boldsymbol{\alpha}(252, 0.001)$.

In the DPP models, we varied prior parameters for the precision estimate, λ. We initially chose values of $a = 1$ and $b = 1$, corresponding to a mean of $\lambda = 1$. Here, we ran a sensitivity analysis with $a = .01$ and $b = .01$ representing a less informative prior, and thus allowing large deviations from the base distributions. Sensitivity analyses that increase λ were unnecessary, because as λ increases, the DPP1 becomes equivalent to the FB model, for which we already have results. Finally, we consider the prior on π and vary c and d, which determined the prior probability of a randomly selected coefficient being zero in the DPP2 model. We began with an uninformative prior, $(c = 1, d = 1)$ that implies we believe every value of π from 0 to 1 to be equally likely. We also specified two highly informative priors with E(π)=0.15 (a null result is unlikely) and 0.85 (a null result is likely) and V(π)=0.1: $(c = 0.04, d = 0.23)$ and $(c = 0.23, d = 0.04)$, respectively.

4.3.9 MCMC Sampling and Convergence Monitoring

We programmed MCMC algorithms for each of the four models using a data augmentation approach and ran them in Matlab.(Albert and Chib, 1993; Mathworks Development, 2005; O'Brien and Dunson, 2004) All models were run for 60,000 iterations, with the initial 5,000 iterations discarded as a burn-in. The remaining iterations were examined for convergence by examining trace plots of the sampled parameter values by iteration of the algorithm.

4.4 Results

Figure 4.1 shows trace plots used for monitoring convergence of one of the 32 coefficients (the 4^{th} quartile of Cl_2AA). A sequence of parameter values that has converged will show a fuzzy horizontal band; a sequence that has not converged may show an increasing or

decreasing trend. While the SB and FB trace plots are typical of MCMC algorithms, the DPP1 and DPP2 trace plots are less common. The trace plots for DPP1 indicate that the sampled coefficients are frequently near zero, which occurs when this coefficient is clustered with other coefficients. Similarly, the trace plot for DPP2 indicates that the sampled coefficient is often zero (sampled from the zero cluster). All trace plots for all coefficients from all models indicated good convergence. We tested sensitivity to starting values by initiating the MCMC algorithms from different points. All algorithms quickly converged to the same region.

The estimated effects of the 13 constituent DBPs on SAB for the 4 hierarchical models and a maximum-likelihood logistic regression are shown in Table 4.2. The maximum likelihood results showed a strong, but imprecise, negative effect of the 3^{rd} quartile of BrClAA on SAB and strong positive, but imprecise, effects of the 2^{nd} quartiles of Cl_2AA and Cl_3AA on SAB. The SB model produced ORs most similar to the maximum likelihood estimates. The SB model, however, indicated that Cl_2AA had little effect on SAB, but an elevated risk of Cl_3AA was observed (particularly, for the 2^{nd} quartile). No negative association between BrClAA and SAB was observed in the SB model. The estimates in the SB model were more precise than those in the maximum likelihood model.

The posterior mean of ϕ^2 in the FB model was 0.04, far smaller than its fixed value in the SB model ($\phi^2 = 0.3142$) or the prior mean in the FB model ($E(\phi^2) = 0.3142$). This indicates that our initial guess at the variability of the 32 DBP coefficients was far from accurate, with little variability between coefficients. The small posterior mean of ϕ^2 caused estimates from the FB model to be shrunk much further toward the prior mean ($\mu_j = 0$) than estimates from the SB model. Most ORs from the FB model indicated no association between DBPs and SAB.

The results from the DPP1 and DPP2 models were in close agreement with each other, both indicating a nearly null effect for all constituent DBPs. In each iteration of the MCMC algorithm for these 2 models, coefficients could either belong to their own cluster or be clustered with other coefficients. Both DPP1 and DPP2 models tended to strongly cluster coefficients together (with aggregate effects near zero). The clustering in these models allowed far more precise estimates than either the SB or FB models.

The results for the highest quartile of Cl_2AA were typical of the pattern of results seen in these analyses. In the maximum likelihood analysis, a woman in the highest quartile of Cl_2AA was estimated to have 2.67 times the odds of SAB as a woman in the lowest quartile (95% Confidence Interval: 0.50, 14.29). The SB model reduced the

estimated OR to 1.72 and increased the precision in the estimate as well. The 95% credible intervals from the SB model were 0.65, 2.55, indicating that if we believe the assumptions of the SB model, we are 95% certain that the true OR lies between 0.65 and 2.55. By allowing ϕ^2 to be informed by the data, the FB model shrunk the estimated effect even closer to the prior distribution (OR= 1.06, 95% credible interval: 0.77, 1.47). The DPP1 (OR=1.00, 95% credible interval: 0.88, 1.14) and DPP2 (OR=1.00, 95% credible interval: 0.89, 1.11) models produced similar estimates, but reduced the width of the credible interval. The posterior distribution for the effect of Cl_2AA is shown in Figure 4.2, which clearly illustrates the gains in precision. The posterior distribution from the DPP2 model also illustrates that this coefficient is grouped in the null cluster a large proportion of times.

We ran a number of sensitivity analyses to assess how our results would vary with different prior assumptions (see Appendix). Results from the sensitivity analyses for the SB model indicated our interpretation of the results would remain largely unchanged over a range of different prior parameters. Only Cl_2AA and Cl_3AA showed some evidence of effect if the prior mean was either $\mu_j = \ln(3.0)$ or $\mu_j = \ln(6.9)$. Results from sensitivity analyses for the FB model showed similar results: varying priors generally had little impact on interpretation, with the exception again being for Cl_2AA and Cl_3AA. The two DPP models were relatively robust to prior specification. Under a variety of different parameterizations, the DPP1 and DPP2 models indicated little or no effect of any DBPs.

4.5 Discussion

Standard maximum likelihood logistic regression results indicated several imprecise but strong positive (Cl_2AA, Cl_3AA) and negative (BrClAA) associations between DBPs and SAB. These imprecise estimates are typical of maximum likelihood in the presence of highly correlated data. We implemented four hierarchical models to allow more precise estimation of effects. The SB estimates exhibited the least shrinkage and indicated only moderate increased or decreased risk of SAB for some constituent DBP categories. The results of the FB, DPP1, and DPP2 models were all consistent and indicated none of the constituent DBPs had an effect on SAB.

Results of sensitivity analyses indicated that with the SB or FB model, our interpretation of results for Cl_2AA and Cl_3AA depended, to some extent, on our prior belief about the effect of these constituent DBPs. However, for Cl_3AA the dose response was

opposite what one would expect; the second quartile (vs. first quartile) had the highest OR and effect for subsequent quartiles diminished. The effect of Cl_2AA was more dependent on prior information, indicating a lack of information regarding this effect in the data. The DPP1 and DPP2 models provided little evidence of effect and were more robust to prior specification. The robustness of the coefficients in these models was expected, since their prior distribution was nonparametric. If the prior parameters μ and ϕ^2 specifying the base distribution, D_0, were widely inaccurate, the precision parameter, λ, allows the random distribution, D, to vary widely from D_0.

These results are consistent with the results of Savitz et al., who analyzed these data without controlling for other constituent DBPs.(Savitz et al., 2005) However, the results are at odds with the two other studies that measured DBP levels (rather than using the proxy of consumed water).(Savitz et al., 1995; Waller et al., 1998) The discrepancy could be due to more precise DBP measurement in RFTS, which measured DBP concentrations every week as opposed to other studies that relied on quarterly measurements. Cumulative exposure to DBPs may be more important, etiologically, than the dose received in a given week. Quarterly measurements may better reflect this than the week-specific concentrations we have used. Additionally, the study of Waller et al. and this study were conducted in different geographic regions with different study populations and different DBP levels.

The hierarchical models we used greatly reduced the variability of the estimates. The SB model proved to be somewhat sensitive to the prior specification of ϕ^2. The FB model allowed ϕ^2 to be data driven, which was important in this study because the prior specification of $\phi^2 = 0.3142$ was far greater than the variability noted in the data; the posterior estimate of ϕ^2 from the FB model was almost 8 times smaller. By allowing ϕ^2 to be updated based on the observed data, we were able to obtain much more precise estimates. The semi-parametric DPP models produced estimates with greater precision than the FB model; when estimates were clustered together, the model contained fewer terms and the clustered terms contained more information about the effect of those clusters.

In conclusion, the use of hierarchical models enabled us to adjust for a large number of correlated exposures while incorporating prior subject matter knowledge. Although SB models are the most commonly used Bayesian hierarchical models in epidemiology, our results suggest that more complex models that allow prior parameters to be updated based on the data can have large benefits. The FB, DPP1 and DPP2 models all produced results that suggest that none of the constituent DBPs have an effect on

SAB.

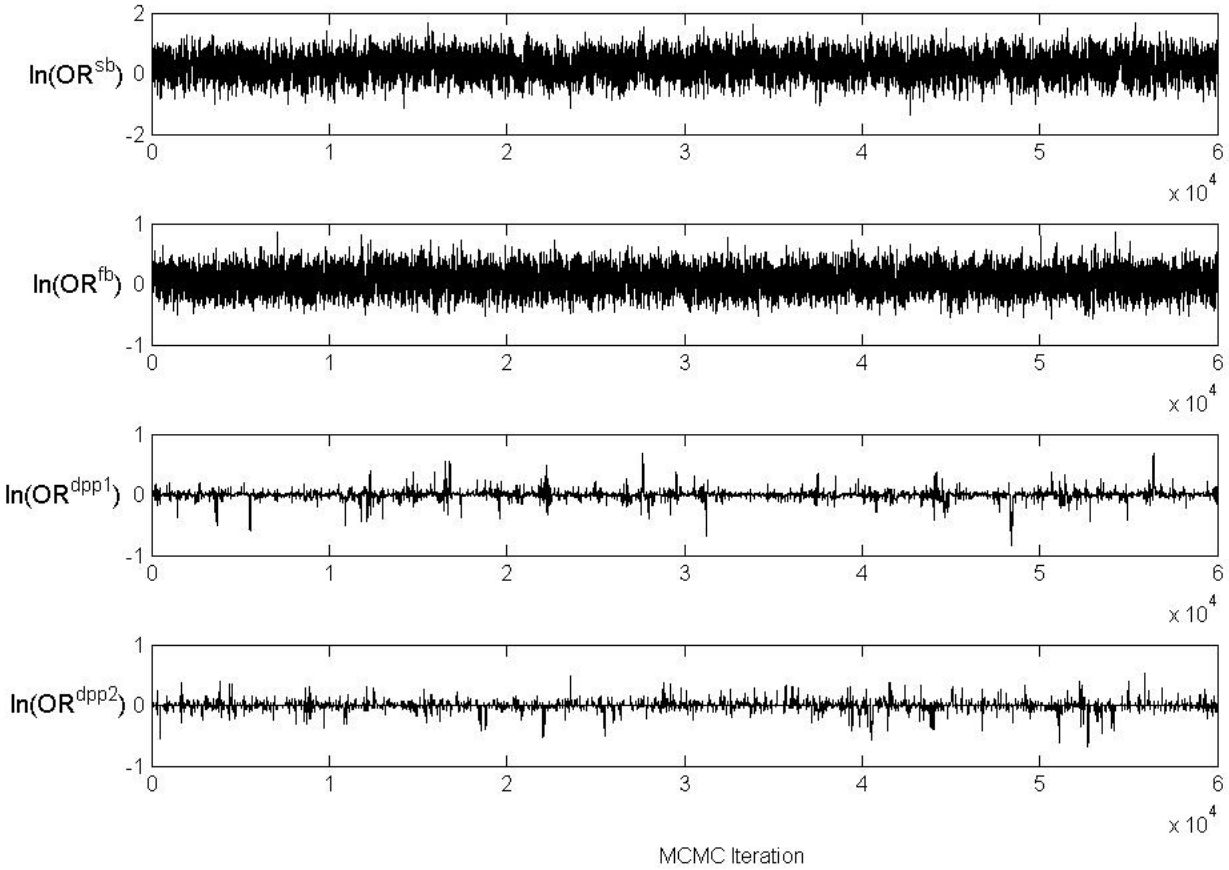

FIGURE 4.1: Convergence of the 4 hierarchical models for the effect of the 4^{th} quartile of Cl_2AA (vs the 1^{st} quartile) on SAB.

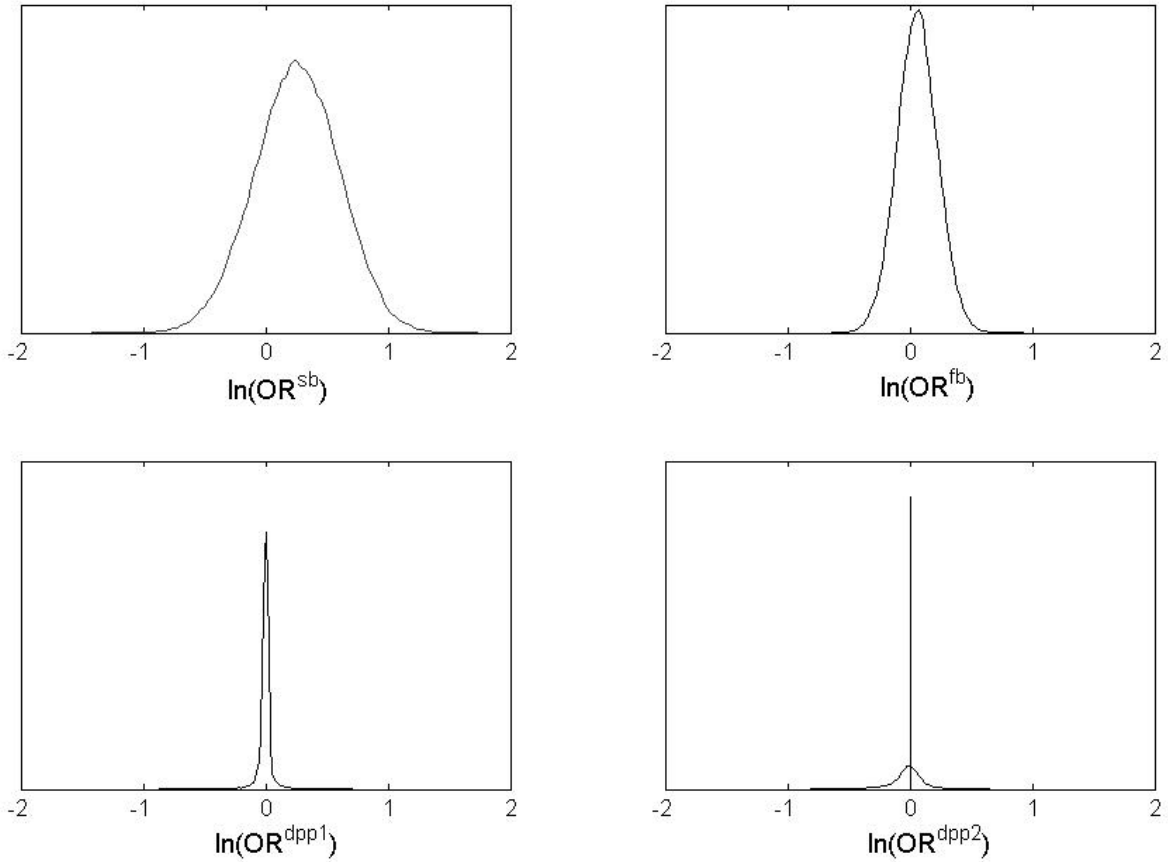

FIGURE 4.2: Posterior distribution of the effect of the highest quartile of CL_2AA (vs. the lowest quartile) for all four hierarchical models.

Figure 4.2: Posterior distribution of the effect the highest quartile of Cl_2AA (vs the lowest quartile) for all four hierarchical models
SB=semi-Bayes; FB=fully-Bayes; DPP1=Dirichlet process prior; DPP2=Dirichlet process prior with selection component

TABLE 4.1: Bayesian hierarchical models used in RFTS Analysis.

SB*			FB*		
β_j	$\sim$	$N(\mu_j, \phi^2)$	β_j	$\sim$	$N(\mu_j, \phi^2)$
			ϕ^2	$\sim$	$IG(\alpha_1, \alpha_2)$

DPP1*			DPP2*		
β_j	$\sim$	D	β_j	$\sim$	D
D	$\sim$	$DP(\lambda D_0)$	D	$\sim$	$DP(\lambda D_0)$
D_0	$\sim$	$N(\mu, \phi^2)$	D_0	$\sim$	$\pi\delta_0 + (1-\pi)N(\mu, \phi^2)$
λ	$\sim$	$G(a, b)$	λ	$\sim$	$G(a, b)$
ϕ^2	$\sim$	$IG(\alpha_1, \alpha_2)$	ϕ^2	$\sim$	$IG(\alpha_1, \alpha_2)$
			π	$\sim$	$beta(c, d)$

* SB=semi-Bayes; FB=fully-Bayes; DPP1=Dirichlet process prior; DPP2=Dirichlet process prior with selection component

TABLE 4.2: Adjusted odds ratios for the association between constituent DBPs and SAB estimated from RFTS.

DBP*	ML*	SB*	FB*	DPP1*	DPP2*
	OR†,‡ (95% CI†)	OR†,‡ (95% CrIn†)	OR†,‡ (95% CrIn†)	OR†,‡ (95% CrIn†)	OR†,‡ (95% CrIn†)
$CHCl_3$					
0-0.1	1.0	1.0	1.0	1.0	1.0
0.1-16.3	0.97 (0.43, 2.22)	0.98 (0.58, 1.65)	1.00 (0.74, 1.35)	1.02 (0.84, 1.26)	1.05 (0.77, 1.43)
16.3 - 44.3	0.75 (0.26, 2.18)	0.82 (0.45, 1.49)	0.89 (0.66, 1.20)	1.01 (0.87, 1.18)	1.02 (0.86, 1.20)
>44.3	0.99 (0.32, 3.11)	1.15 (0.61, 2.16)	1.08 (0.78, 1.48)	1.11 (0.74, 1.66)	1.14 (0.72, 1.78)
$CHBrCl_2$					
0-1.1	1.0	1.0	1.0	1.0	1.0
1.1-11.4	0.56 (0.22, 1.41)	0.73 (0.44, 1.23)	0.82 (0.61, 1.10)	1.01 (0.91, 1.12)	1.01 (0.90, 1.13)
11.4-17.5	0.69 (0.69, 2.10)	0.87 (0.49, 1.55)	0.92 (0.68, 1.24)	0.99 (0.88, 1.12)	0.99 (0.88, 1.12)
>17.5	1.11 (0.33, 3.75)	1.36 (0.72, 2.56)	1.20 (0.87, 1.66)	1.01 (0.89, 1.14)	1.00 (0.90, 1.12)
$CHBr_2Cl$					
0-1.4	1.0	1.0	1.0	1.0	1.0
1.4-2.8	1.06 (0.59, 1.89)	0.86 (0.56, 1.33)	0.84 (0.63, 1.11)	0.64 (0.27, 1.55)	0.61 (0.27, 1.38)
2.8-7.4	1.26 (0.57, 2.80)	1.05 (0.61, 1.81)	1.05 (0.78, 1.42)	0.99 (0.84, 1.16)	0.97 (0.77, 1.22)
>7.4	1.19 (0.39, 3.58)	1.06 (0.53, 2.11)	1.05 (0.75, 1.48)	1.00 (0.87, 1.15)	1.00 (0.90, 1.10)

TABLE 4.2: continued.

DBP*	ML*	SB*	FB*	DPP1*	DPP2*
	OR†,‡ (95% CI†)	OR†,‡ (95% CI†)	OR†,‡ (95% CI†)	OR†,‡ (95% CI†)	OR†,‡ (95% CI†)
$CHBr_3$					
0	1.0	1.0	1.0	1.0	
0-0.5	0.79 (0.47, 1.34)	0.82 (0.55, 1.22)	0.94 (0.71, 1.25)	0.98 (0.80, 1.20)	0.97 (0.78, 1.22)
0.5-0.9	0.81 (0.44, 1.51)	0.81 (0.56, 1.57)	0.89 (0.70, 1.15)	0.99 (0.86, 1.13)	0.99 (0.83, 1.18)
> 0.9	1.00 (0.48, 2.11)	0.96 (0.58, 1.57)	1.00 (0.74, 1.36)	1.00 (0.89, 1.13)	1.01 (0.88, 1.15)
ClAA					
0	1.0	1.0	1.0	1.0	1.0
0-2.1	0.98 (0.59, 1.62)	0.98 (0.63, 1.51)	1.02 (0.77, 1.36)	0.99 (0.90, 1.10)	1.00 (0.92, 1.08)
>2.1	0.89 (0.53, 1.49)	0.88 (0.56, 1.38)	0.92 (0.69, 1.23)	0.99 (0.88, 1.11)	0.99 (0.89, 1.11)
Cl_2AA					
0	1.0	1.0	1.0	1.0	1.0
0-8.8	1.69 (0.37, 7.83)	0.99 (0.50, 1.96)	1.04 (0.73, 1.47)	1.00 (0.87, 1.15)	1.00 (0.88, 1.15)
8.8-18.3	1.76 (0.37, 8.40)	0.97 (0.53, 1.77)	0.96 (0.71, 1.30)	1.00 (0.87, 1.14)	1.00 (0.89, 1.11)
>18.3	2.67 (0.50, 14.29)	1.29 (0.65, 2.55)	1.06 (0.77, 1.47)	1.00 (0.88, 1.14)	1.00 (0.89, 1.11)

TABLE 4.2: continued.

DBP*	ML*	SB*	FB*	DPP1*	DPP2*
	OR†,‡ (95% CI†)	OR†,‡ (95% CI†)	OR†,‡ (95% CI†)	OR†,‡ (95% CI†)	OR†,‡ (95% CI†)
Cl_3AA					
0	1.0	1.0	1.0	1.0	1.0
0-5.4	2.95 (1.18, 7.37)	1.72 (0.98, 3.04)	1.25 (0.91, 1.73)	1.06 (0.78, 1.45)	1.05 (0.79, 1.40)
5.4-11.8	2.36 (0.80, 6.99)	1.34 (0.74, 2.44)	1.11 (0.81, 1.51)	1.00 (0.89, 1.13)	1.00 (0.85, 1.17)
>11.8	1.38 (0.40, 4.80)	0.83 (0.43, 1.61)	0.86 (0.62, 1.19)	1.00 (0.85, 1.17)	0.99 (0.80, 1.23)
BrAA					
0	1.0	1.0	1.0	1.0	1.0
>0	1.03 (0.62, 1.71)	1.06 (0.66, 1.68)	1.04 (0.76, 1.43)	1.02 (0.88, 1.18)	1.04 (0.82, 1.31)
Br_2AA					
0	1.0	1.0	1.0	1.0	1.0
0-1.2	0.95 (0.53, 1.71)	0.87 (0.52, 1.45)	0.95 (0.68, 1.32)	0.98 (0.83, 1.16)	0.98 (0.79, 1.21)
>1.2	1.09 (0.63, 1.88)	0.99 (0.62, 1.59)	1.02 (0.76, 1.37)	1.00 (0.90, 1.12)	1.00 (0.92, 1.09)
Br_3AA					
0	1.0	1.0	1.0	1.0	1.0
>0	0.98 (0.62, 1.54)	0.94 (0.62, 1.44)	0.98 (0.73, 1.30)	1.01 (0.91, 1.12)	1.01 (0.90,1.13)
BrClAA					
0	1.0	1.0	1.0	1.0	1.0
0-3.8	0.34 (0.10, 1.19)	0.92 (0.51, 1.67)	1.04 (0.76, 1.43)	1.00 (0.89, 1.12)	1.00 (0.90, 1.10)
3.8-5.9	0.23 (0.06, 0.89)	0.71 (0.39, 1.31)	0.88 (0.65, 1.20)	0.99 (0.88, 1.11)	0.99 (0.89, 1.10)
>5.9	0.33 (0.07, 1.49)	1.04 (0.52, 2.06)	1.08 (0.77, 1.50)	1.02 (0.85, 1.21)	1.01 (0.88, 1.16)

TABLE 4.2: continued.

DBP*	ML*	SB*	FB*	DPP1*	DPP2*
	OR†,‡ (95% CI†)	OR†,‡ (95% CI†)	OR†,‡ (95% CI†)	OR†,‡ (95% CI†)	OR†,‡ (95% CI†)
$BrCl_2AA$					
0-1.6	1.0	1.0	1.0	1.0	1.0
1.6-3.7	1.61 (1.01, 2.56)	1.39 (0.94, 2.05)	1.17 (0.90, 1.53)	1.00 (0.91, 1.10)	1.01 (0.88, 1.17)
3.7-5.8	1.28 (0.62, 2.65)	1.03 (0.62, 1.74)	0.97 (0.72, 1.30)	0.99 (0.87, 1.12)	0.99 (0.87, 1.13)
>5.8	1.16 (0.52, 2.62)	0.98 (0.54, 1.78)	0.98 (0.72, 1.35)	1.01 (0.89, 1.13)	1.00 (0.92, 1.09)
Br_2ClAA					
0	1.0	1.0	1.0	1.0	1.0
0-2	0.58 (0.36, 0.94)	0.66 (0.43, 1.03)	0.84 (0.63, 1.11)	0.98 (0.82, 1.18)	0.98 (0.83, 1.17)
>2	0.76 (0.47, 1.24)	0.81 (0.52, 1.26)	0.91 (0.68, 1.23)	0.99 (0.89, 1.10)	0.99 (0.90, 1.10)

*DBP=disinfection byproduct; ML=maximum likelihood; SB=semi-Bayes; FB=fully-Bayes; DPP1=Dirichlet process prior; DPP2=Dirichlet process prior with selection component

† OR= odds ratio; CI=confidence interval for ML and Credible Interval for SB, FB, DPP1 and DPP2

‡ Models are adjusted for smoking, alcohol use, ethnicity, and maternal age

4.6 Appendix 1: Sensitivity Analyses

TABLE 4.3: Sensitivity analysis for semi-Bayes model.

μ	1.0	1.0	3.0	3.0	6.9	6.9
ϕ^2	0.1	5.0	0.1	5.0	0.1	5.0
$CHCl_3$						
0-0.1	1.0	1.0	1.0	1.0	1.0	1.0
0.1-16.3	0.99 (0.66, 1.47)	0.97 (0.44, 2.15)	1.02 (0.71, 1.48)	0.99 (0.47, 2.08)	1.04 (0.72, 1.50)	0.94 (0.43, 2.05)
16.3 - 44.3	0.86 (0.57, 1.30)	0.74 (0.27, 2.04)	0.89 (0.60, 1.32)	0.78 (0.30, 2.05)	0.91 (0.62, 1.34)	0.72 (0.26, 1.97)
>44.3	1.11 (0.72, 1.71)	1.02 (0.35, 3.01)	1.18 (0.78, 1.80)	1.10 (0.38, 3.13)	1.23 (0.81, 1.85)	1.01 (0.34, 3.05)
$CHBrCl_2$						
0-1.10	1.0	1.0	1.0	1.0	1.0	1.0
1.1-11.4	0.77 (0.53, 1.13)	0.62 (0.26, 1.48)	0.85 (0.58, 1.23)	0.6 (0.26, 1.40)	0.88 (0.62, 1.26)	0.60 (0.26, 1.38)
11.4-17.5	0.89 (0.59, 1.35)	0.77 (0.27, 2.21)	0.89 (0.60, 1.31)	0.76 (0.27, 2.09)	0.87 (0.60, 1.27)	0.74 (0.27, 2.02)
>17.5	1.28 (0.83, 1.99)	1.28 (0.41, 3.98)	1.21 (0.79, 1.86)	1.21 (0.40, 3.67)	1.16 (0.76, 1.78)	1.24 (0.42, 3.69)
$CHBr_2Cl$						
0-1.4	1.0	1.0	1.0	1.0	1.0	1.0
1.4-2.8	0.84 (0.58, 1.21)	0.94 (0.52, 1.72)	0.98 (0.69, 1.39)	0.97 (0.55, 1.71)	1.07 (0.76, 1.49)	1.00 (0.57, 1.74)
2.8-7.4	1.06 (0.72, 1.56)	1.16 (0.54, 2.51)	1.12 (0.77, 1.64)	1.17 (0.55, 2.49)	1.14 (0.78, 1.66)	1.18 (0.54, 2.56)
>7.4	1.07 (0.67, 1.71)	1.15 (0.39, 3.38)	0.97 (0.61, 1.54)	1.13 (0.42, 3.04)	0.92 (0.59, 1.44)	1.12 (0.40, 3.15)
$CHBr_3$						
0	1.0	1.0	1.0	1.0	1.0	1.0
0-0.5	0.89 (0.63, 1.25)	0.74 (0.47, 1.19)	1.22 (0.88, 1.69)	0.75 (0.48, 1.19)	1.48 (1.08, 2.04)	0.75 (0.48, 1.18)
0.5-0.9	0.85 (0.63, 1.13)	0.76 (0.51, 1.14)	1.37 (1.02, 1.83)	0.78 (0.52, 1.16)	1.82 (1.38, 2.41)	0.79 (0.53, 1.16)
> 0.9	0.98 (0.67, 1.45)	0.92 (0.52, 1.61)	1.33 (0.90, 1.96)	0.94 (0.53, 1.64)	1.58 (1.08, 2.31)	0.96 (0.54, 1.72)

TABLE 4.3: continued.

μ	1.0	1.0	3.0	3.0	6.9	6.9
ϕ^2	0.1	5.0	0.1	5.0	0.1	5.0
ClAA						
0	1.0	1.0	1.0	1.0	1.0	1.0
0-2.10	1.00 (0.70, 1.44)	0.96 (0.57, 1.61)	1.00 (0.72, 1.39)	0.94 (0.56, 1.58)	0.97 (0.70, 1.34)	0.95 (0.57, 1.59)
>2.10	0.89 (0.62, 1.28)	0.88 (0.51, 1.50)	0.84 (0.59, 1.18)	0.87 (0.52, 1.46)	0.81 (0.58, 1.13)	0.88 (0.53, 1.46)
Cl_2AA						
0	1.0	1.0	1.0	1.0	1.0	1.0
0-8.8	1.02 (0.63, 1.65)	1.26 (0.34, 4.71)	1.33 (0.82, 2.14)	1.3 (0.35, 4.86)	1.56 (1.00, 2.45)	1.29 (0.36, 4.60)
8.8-18.3	0.94 (0.63, 1.41)	1.34 (0.35, 5.17)	1.08 (0.72, 1.62)	1.34 (0.36, 4.94)	1.17 (0.79, 1.72)	1.37 (0.38, 4.92)
>18.3	1.15 (0.74, 1.78)	2.05 (0.50, 8.52)	1.26 (0.82, 1.92)	2.04 (0.50, 8.39)	1.35 (0.88, 2.06)	2.05 (0.51, 8.21)
Cl_3AA						
0	1.0	1.0	1.0	1.0	1.0	1.0
0-5.4	1.40 (0.93, 2.10)	2.63 (1.12, 6.19)	1.64 (1.10, 2.44)	2.55 (1.10, 5.93)	1.81 (1.21, 2.71)	2.68 (1.13, 6.38)
5.4-11.8	1.18 (0.79, 1.79)	1.96 (0.70, 5.50)	1.33 (0.89, 1.97)	1.94 (0.72, 5.21)	1.44 (0.97, 2.13)	2.04 (0.76, 5.48)
>11.8	0.83 (0.53, 1.29)	1.09 (0.36, 3.33)	0.99 (0.64, 1.53)	1.07 (0.34, 3.35)	1.10 (0.72, 1.69)	1.14 (0.35, 3.73)
BrAA						
0	1.0	1.0	1.0	1.0	1.0	1.0
>0	1.04 (0.70, 1.55)	1.05 (0.62, 1.75)	1.29 (0.89, 1.87)	1.06 (0.63, 1.80)	1.49 (1.05, 2.11)	1.07 (0.64, 1.78)
Br_2AA						
0	1.0	1.0	1.0	1.0	1.0	1.0
0-1.2	0.91 (0.60, 1.39)	0.87 (0.46, 1.67)	1.17 (0.80, 1.71)	0.88 (0.49, 1.59)	1.34 (0.93, 1.93)	0.89 (0.49, 1.62)
>1.2	1.00 (0.69, 1.45)	1.02 (0.59, 1.76)	0.89 (0.61, 1.29)	1.01 (0.58, 1.75)	0.8 (0.56, 1.14)	1.02 (0.59, 1.77)

TABLE 4.3: continued.

μ	1.0	1.0	3.0	3.0	6.9	6.9
ϕ^2	0.1	5.0	0.1	5.0	0.1	5.0
Br_3AA						
0	1.0	1.0	1.0	1.0	1.0	1.0
>0	0.96 (0.67, 1.37)	0.93 (0.57, 1.51)	0.96 (0.69, 1.34)	0.96 (0.60, 1.53)	0.97 (0.69, 1.36)	0.94 (0.59, 1.50)
BrClAA						
0	1.0	1.0	1.0	1.0	1.0	1.0
0-3.8	1.03 (0.68, 1.55)	0.47 (0.15, 1.46)	1.33 (0.89, 1.99)	0.48 (0.16, 1.50)	1.56 (1.05, 2.32)	0.48 (0.16, 1.47)
3.8-5.9	0.83 (0.55, 1.24)	0.33 (0.10, 1.13)	0.97 (0.65, 1.44)	0.35 (0.10, 1.20)	1.06 (0.72, 1.57)	0.34 (0.10, 1.16)
>5.9	1.10 (0.70, 1.74)	0.50 (0.13, 1.92)	1.16 (0.74, 1.82)	0.54 (0.13, 2.19)	1.22 (0.78, 1.92)	0.51 (0.13, 2.02)
$BrCl_2AA$						
0-1.6	1.0	1.0	1.0	1.0	1.0	1.0
1.6-3.7	1.27 (0.91, 1.77)	1.58 (0.99, 2.52)	1.42 (1.04, 1.92)	1.57 (1.00, 2.48)	1.51 (1.11, 2.05)	1.63 (1.05, 2.54)
3.7-5.8	0.99 (0.67, 1.44)	1.26 (0.60, 2.64)	1.04 (0.72, 1.50)	1.22 (0.59, 2.56)	1.09 (0.75, 1.57)	1.29 (0.62, 2.68)
>5.8	0.97 (0.63, 1.49)	1.14 (0.50, 2.6)	1.07 (0.70, 1.62)	1.12 (0.50, 2.52)	1.14 (0.76, 1.71)	1.16 (0.51, 2.62)
Br_2ClAA						
0	1.0	1.0	1.0	1.0	1.0	1.0
0-2	0.76 (0.53, 1.07)	0.58 (0.35, 0.95)	0.89 (0.65, 1.23)	0.58 (0.36, 0.95)	0.98 (0.72, 1.33)	0.59 (0.36, 0.97)
>2	0.86 (0.60, 1.23)	0.77 (0.46, 1.27)	0.87 (0.61, 1.23)	0.75 (0.45, 1.24)	0.87 (0.62, 1.21)	0.76 (0.46, 1.26)

TABLE 4.4: Sensitivity analysis for fully-Bayes model (prior mean=1.0).

μ	1.0	1.0	1.0	1.0
E(ϕ^2)	0.10	5.0	0.10	5.0
V(ϕ^2)	3.0	3.0	0.10	0.10
$CHCl_3$				
0-0.10	1.0	1.0	1.0	1.0
0.1-16.3	1.00 (0.85, 1.18)	0.99 (0.75, 1.32)	1.00 (0.85, 1.18)	0.96 (0.47, 1.99)
16.3 - 44.3	0.96 (0.82, 1.12)	0.90 (0.68, 1.19)	0.95 (0.81, 1.13)	0.74 (0.28, 1.94)
>44.3	1.02 (0.86, 1.20)	1.07 (0.80, 1.43)	1.02 (0.86, 1.21)	1.05 (0.37, 2.97)
$CHBrCl_2$				
0-1.10	1.0	1.0	1.0	1.0
1.1-11.4	0.92 (0.78, 1.08)	0.83 (0.63, 1.09)	0.91 (0.77, 1.07)	0.63 (0.28, 1.40)
11.4-17.5	0.97 (0.83, 1.14)	0.92 (0.70, 1.22)	0.97 (0.82, 1.14)	0.82 (0.31, 2.18)
>17.5	1.08 (0.92, 1.27)	1.2 (0.88, 1.62)	1.08 (0.91, 1.28)	1.36 (0.46, 3.97)
$CHBr_2Cl$				
0-1.4	1.0	1.0	1.0	1.0
1.4-2.8	0.91 (0.78, 1.07)	0.84 (0.65, 1.11)	0.90 (0.77, 1.07)	0.97 (0.55, 1.69)
2.8-7.4	1.02 (0.86, 1.19)	1.04 (0.79, 1.39)	1.01 (0.86, 1.20)	1.14 (0.54, 2.42)
>7.4	1.05 (0.89, 1.23)	1.06 (0.77, 1.46)	1.05 (0.88, 1.24)	1.12 (0.40, 3.14)

TABLE 4.4: continued.

μ	1.0	1.0	1.0	1.0
E(ϕ^2)	0.10	5.0	0.10	5.0
V(ϕ^2)	3.0	3.0	0.10	0.10
$CHBr_3$				
0	1.0	1.0	1.0	1.0
0-0.5	0.99 (0.84, 1.16)	0.95 (0.73, 1.23)	0.98 (0.83, 1.17)	0.73 (0.46, 1.17)
0.5-0.9	0.96 (0.82, 1.12)	0.9 (0.71, 1.14)	0.96 (0.82, 1.12)	0.74 (0.50, 1.10)
> 0.9	1.02 (0.87, 1.20)	1.00 (0.75, 1.34)	1.02 (0.86, 1.21)	0.90 (0.52, 1.56)
ClAA				
0	1.0	1.0	1.0	1.0
0-2.10	1.02 (0.87, 1.20)	1.03 (0.79, 1.35)	1.02 (0.86, 1.20)	0.93 (0.54, 1.58)
>2.10	0.97 (0.83, 1.13)	0.92 (0.70, 1.21)	0.96 (0.82, 1.13)	0.85 (0.50, 1.46)
Cl_2AA				
0	1.0	1.0	1.0	1.0
0-8.8	1.04 (0.88, 1.23)	1.05 (0.76, 1.45)	1.04 (0.87, 1.24)	1.22 (0.36, 4.13)
8.8-18.3	0.98 (0.84, 1.15)	0.96 (0.72, 1.27)	0.98 (0.83, 1.16)	1.27 (0.38, 4.27)
>18.3	1.00 (0.85, 1.18)	1.05 (0.77, 1.42)	1.00 (0.84, 1.18)	1.90 (0.51, 7.13)
Cl_3AA				
0	1.0	1.0	1.0	1.0
0-5.4	1.08 (0.91, 1.28)	1.23 (0.91, 1.67)	1.08 (0.91, 1.29)	2.49 (1.09, 5.68)
5.4-11.8	1.03 (0.88, 1.21)	1.09 (0.82, 1.46)	1.04 (0.88, 1.22)	1.87 (0.68, 5.11)
>11.8	0.94 (0.80, 1.11)	0.87 (0.64, 1.18)	0.94 (0.79, 1.11)	1.03 (0.32, 3.32)

TABLE 4.4: continued.

μ	1.0	1.0	1.0	1.0
E(ϕ^2)	0.10	5.0	0.10	5.0
V(ϕ^2)	3.0	3.0	0.10	0.10
BrAA				
0	1.0	1.0	1.0	1.0
>0	1.03 (0.87, 1.21)	1.04 (0.77, 1.39)	1.03 (0.86, 1.23)	1.05 (0.63, 1.77)
Br_2AA				
0	1.0	1.0	1.0	1.0
0-1.2	0.99 (0.83, 1.17)	0.96 (0.70, 1.31)	0.99 (0.83, 1.18)	0.87 (0.48, 1.59)
>1.2	1.02 (0.87, 1.19)	1.01 (0.76, 1.34)	1.02 (0.86, 1.20)	1.03 (0.60, 1.79)
Br_3AA				
0	1.0	1.0	1.0	1.0
>0	1.01 (0.86, 1.18)	0.98 (0.74, 1.30)	1.00 (0.85, 1.18)	0.93 (0.58, 1.49)
BrClAA				
0	1.0	1.0	1.0	1.0
0-3.8	1.01 (0.85, 1.19)	1.04 (0.77, 1.40)	1.01 (0.85, 1.20)	0.52 (0.18, 1.49)
3.8-5.9	0.96 (0.81, 1.12)	0.89 (0.67, 1.20)	0.95 (0.80, 1.12)	0.37 (0.12, 1.15)
>5.9	1.05 (0.89, 1.24)	1.08 (0.79, 1.47)	1.05 (0.88, 1.24)	0.56 (0.16, 2.01)
$BrCl_2$AA				
0-1.6	1.0	1.0	1.0	1.0
1.6-3.7	1.04 (0.89, 1.22)	1.16 (0.89, 1.51)	1.05 (0.89, 1.23)	1.58 (1.01, 2.45)
3.7-5.8	0.98 (0.84, 1.14)	0.96 (0.73, 1.27)	0.97 (0.82, 1.15)	1.23 (0.60, 2.54)
>5.8	1.01 (0.86, 1.19)	0.98 (0.73, 1.33)	1.01 (0.85, 1.19)	1.14 (0.49, 2.62)

TABLE 4.4: continued.

μ	1.0	1.0	1.0	1.0
E(ϕ^2)	0.10	5.0	0.10	5.0
V(ϕ^2)	3.0	3.0	0.10	0.10
Br_2ClAA				
0	1.0	1.0	1.0	1.0
0-2	0.94 (0.80, 1.11)	0.84 (0.64, 1.12)	0.94 (0.80, 1.11)	0.59 (0.36, 0.96)
>2	0.98 (0.84, 1.15)	0.93 (0.70, 1.23)	0.98 (0.83, 1.16)	0.75 (0.46, 1.22)

TABLE 4.5: Sensitivity analysis for fully-Bayes model (prior mean=3.0).

μ	3.0	3.0	3.0	3.0
E(ϕ^2)	0.10	5.0	0.10	5.0
V(ϕ^2)	3.0	3.0	0.10	0.10
$CHCl_3$				
0-0.10	1.0	1.0	1.0	1.0
0.1-16.3	0.95 (0.51, 1.75)	0.94 (0.50, 1.76)	0.95 (0.50, 1.77)	0.94 (0.43, 2.09)
16.3 - 44.3	0.77 (0.36, 1.64)	0.76 (0.36, 1.63)	0.76 (0.36, 1.62)	0.75 (0.27, 2.08)
>44.3	1.10 (0.49, 2.51)	1.08 (0.47, 2.47)	1.13 (0.50, 2.53)	1.08 (0.37, 3.14)
$CHBrCl_2$				
0-1.10	1.0	1.0	1.0	1.0
1.1-11.4	0.68 (0.36, 1.28)	0.69 (0.36, 1.35)	0.70 (0.37, 1.34)	0.63 (0.27, 1.47)
11.4-17.5	0.79 (0.38, 1.64)	0.81 (0.37, 1.78)	0.82 (0.39, 1.70)	0.79 (0.29, 2.17)
>17.5	1.26 (0.56, 2.82)	1.29 (0.55, 3.00)	1.29 (0.57, 2.87)	1.28 (0.42, 3.91)
$CHBr_2Cl$				
0-1.4	1.0	1.0	1.0	1.0
1.4-2.8	0.93 (0.57, 1.50)	0.91 (0.55, 1.51)	0.90 (0.54, 1.48)	0.93 (0.52, 1.66)
2.8-7.4	1.08 (0.58, 2.02)	1.06 (0.56, 2.01)	1.05 (0.57, 1.95)	1.13 (0.50, 2.53)
>7.4	1.01 (0.43, 2.37)	0.98 (0.41, 2.32)	0.97 (0.42, 2.24)	1.09 (0.37, 3.20)

TABLE 4.5: continued.

μ	3.0	3.0	3.0	3.0
E(ϕ^2)	0.10	5.0	0.10	5.0
V(ϕ^2)	3.0	3.0	0.10	0.10
$CHBr_3$				
0	1.0	1.0	1.0	1.0
0-0.5	0.85 (0.55, 1.32)	0.85 (0.56, 1.30)	0.86 (0.56, 1.34)	0.75 (0.48, 1.16)
0.5-0.9	0.88 (0.60, 1.28)	0.86 (0.59, 1.27)	0.89 (0.62, 1.27)	0.79 (0.53, 1.18)
> 0.9	1.04 (0.61, 1.78)	1.01 (0.58, 1.75)	1.04 (0.62, 1.74)	0.93 (0.51, 1.68)
ClAA				
0	1.0	1.0	1.0	1.0
0-2.10	0.95 (0.60, 1.50)	0.95 (0.59, 1.54)	0.95 (0.60, 1.51)	0.97 (0.60, 1.56)
>2.10	0.85 (0.53, 1.36)	0.87 (0.53, 1.43)	0.85 (0.53, 1.38)	0.89 (0.52, 1.50)
Cl_2AA				
0	1.0	1.0	1.0	1.0
0-8.8	1.14 (0.47, 2.79)	1.13 (0.46, 2.75)	1.16 (0.48, 2.80)	1.25 (0.36, 4.35)
8.8-18.3	1.13 (0.49, 2.63)	1.10 (0.47, 2.57)	1.12 (0.49, 2.57)	1.30 (0.38, 4.45)
>18.3	1.58 (0.63, 4.01)	1.57 (0.63, 3.92)	1.58 (0.64, 3.89)	1.94 (0.51, 7.44)
Cl_3AA				
0	1.0	1.0	1.0	1.0
0-5.4	2.13 (1.08, 4.21)	2.12 (1.05, 4.25)	2.16 (1.09, 4.26)	2.61 (1.12, 6.10)
5.4-11.8	1.59 (0.74, 3.39)	1.60 (0.75, 3.40)	1.64 (0.77, 3.46)	1.92 (0.74, 4.95)
>11.8	0.95 (0.41, 2.20)	0.92 (0.39, 2.18)	0.94 (0.41, 2.16)	1.06 (0.34, 3.27)

TABLE 4.5: continued.

μ	3.0	3.0	3.0	3.0
E(ϕ^2)	0.10	5.0	0.10	5.0
V(ϕ^2)	3.0	3.0	0.10	0.10
BrAA				
0	1.0	1.0	1.0	1.0
>0	1.11 (0.67, 1.83)	1.09 (0.67, 1.78)	1.10 (0.67, 1.80)	1.07 (0.64, 1.80)
Br_2AA				
0	1.0	1.0	1.0	1.0
0-1.2	0.90 (0.53, 1.55)	0.92 (0.51, 1.65)	0.89 (0.52, 1.53)	0.87 (0.47, 1.63)
>1.2	0.96 (0.58, 1.59)	0.98 (0.60, 1.60)	0.96 (0.59, 1.56)	1.01 (0.58, 1.75)
Br_3AA				
0	1.0	1.0	1.0	1.0
>0	0.91 (0.59, 1.41)	0.94 (0.61, 1.46)	0.93 (0.59, 1.48)	0.96 (0.60, 1.53)
BrClAA				
0	1.0	1.0	1.0	1.0
0-3.8	0.82 (0.38, 1.79)	0.82 (0.38, 1.79)	0.81 (0.38, 1.72)	0.51 (0.18, 1.46)
3.8-5.9	0.61 (0.27, 1.39)	0.60 (0.26, 1.39)	0.62 (0.28, 1.40)	0.36 (0.12, 1.11)
>5.9	0.92 (0.36, 2.37)	0.93 (0.37, 2.35)	0.95 (0.38, 2.38)	0.55 (0.15, 1.99)
$BrCl_2$AA				
0-1.6	1.0	1.0	1.0	1.0
1.6-3.7	1.51 (0.98, 2.31)	1.52 (1.00, 2.29)	1.51 (0.99, 2.30)	1.61 (1.02, 2.54)
3.7-5.8	1.13 (0.60, 2.11)	1.13 (0.62, 2.05)	1.08 (0.58, 2.02)	1.24 (0.61, 2.53)
>5.8	1.04 (0.52, 2.08)	1.08 (0.57, 2.04)	1.02 (0.52, 2.02)	1.15 (0.52, 2.51)

TABLE 4.5: continued.

μ	3.0	3.0	3.0	3.0
E(ϕ^2)	0.10	5.0	0.10	5.0
V(ϕ^2)	3.0	3.0	0.10	0.10
Br_2ClAA				
0	1.0	1.0	1.0	1.0
0-2	0.65 (0.41, 1.02)	0.64 (0.40, 1.03)	0.65 (0.41, 1.03)	0.59 (0.36, 0.97)
>2	0.78 (0.48, 1.26)	0.77 (0.47, 1.26)	0.77 (0.48, 1.26)	0.74 (0.45, 1.22)

TABLE 4.6: Sensitivity analysis for fully-Bayes model (prior mean=6.9).

μ	6.9	6.9	6.9	6.9
E(ϕ^2)	0.10	5.0	0.10	5.0
V(ϕ^2)	3.0	3.0	0.10	0.10
$CHCl_3$				
0-0.10	1.0	1.0	1.0	1.0
0.1-16.3	0.97 (0.47, 2.02)	0.96 (0.48, 1.94)	0.95 (0.45, 2.01)	0.99 (0.47, 2.11)
16.3 - 44.3	0.77 (0.30, 1.99)	0.73 (0.29, 1.85)	0.74 (0.28, 1.91)	0.76 (0.28, 2.06)
>44.3	1.09 (0.41, 2.92)	1.06 (0.40, 2.82)	1.06 (0.38, 2.95)	1.07 (0.38, 2.97)
$CHBrCl_2$				
0-1.10	1.0	1.0	1.0	1.0
1.1-11.4	0.62 (0.28, 1.38)	0.63 (0.29, 1.37)	0.64 (0.29, 1.40)	0.59 (0.26, 1.34)
11.4-17.5	0.75 (0.30, 1.91)	0.76 (0.30, 1.95)	0.76 (0.30, 1.94)	0.75 (0.27, 2.06)
>17.5	1.23 (0.45, 3.41)	1.26 (0.46, 3.49)	1.24 (0.45, 3.39)	1.24 (0.42, 3.69)
$CHBr_2Cl$				
0-1.4	1.0	1.0	1.0	1.0
1.4-2.8	0.94 (0.55, 1.62)	0.94 (0.55, 1.61)	0.95 (0.55, 1.63)	0.96 (0.54, 1.69)
2.8-7.4	1.11 (0.54, 2.29)	1.12 (0.54, 2.34)	1.11 (0.54, 2.29)	1.13 (0.52, 2.47)
>7.4	1.01 (0.38, 2.67)	1.04 (0.39, 2.75)	1.03 (0.39, 2.73)	1.08 (0.39, 3.00)

TABLE 4.6: continued.

μ	6.9	6.9	6.9	6.9
E(ϕ^2)	0.10	5.0	0.10	5.0
V(ϕ^2)	3.0	3.0	0.10	0.10
$CHBr_3$				
0	1.0	1.0	1.0	1.0
0-0.5	0.80 (0.51, 1.25)	0.80 (0.52, 1.24)	0.80 (0.52, 1.25)	0.78 (0.49, 1.23)
0.5-0.9	0.84 (0.57, 1.23)	0.84 (0.57, 1.24)	0.82 (0.55, 1.21)	0.80 (0.53, 1.19)
> 0.9	0.99 (0.58, 1.69)	1.01 (0.57, 1.80)	0.97 (0.56, 1.69)	0.96 (0.55, 1.69)
ClAA				
0	1.0	1.0	1.0	1.0
0-2.10	0.95 (0.58, 1.57)	0.96 (0.59, 1.58)	0.94 (0.58, 1.54)	0.94 (0.56, 1.56)
>2.10	0.85 (0.51, 1.43)	0.87 (0.52, 1.46)	0.87 (0.52, 1.47)	0.85 (0.50, 1.46)
Cl_2AA				
0	1.0	1.0	1.0	1.0
0-8.8	1.26 (0.41, 3.88)	1.25 (0.41, 3.84)	1.23 (0.40, 3.77)	1.33 (0.39, 4.49)
8.8-18.3	1.28 (0.42, 3.89)	1.30 (0.42, 3.98)	1.28 (0.41, 3.93)	1.41 (0.40, 4.93)
>18.3	1.86 (0.56, 6.19)	1.92 (0.56, 6.59)	1.89 (0.56, 6.39)	2.13 (0.55, 8.24)
Cl_3AA				
0	1.0	1.0	1.0	1.0
0-5.4	2.47 (1.11, 5.51)	2.43 (1.11, 5.32)	2.44 (1.11, 5.38)	2.48 (1.08, 5.67)
5.4-11.8	1.83 (0.72, 4.64)	1.85 (0.73, 4.66)	1.86 (0.74, 4.66)	1.85 (0.69, 4.96)
>11.8	1.06 (0.37, 3.05)	1.07 (0.37, 3.10)	1.03 (0.34, 3.06)	1.05 (0.34, 3.27)

TABLE 4.6: continued.

μ	6.9	6.9	6.9	6.9
E(ϕ^2)	0.10	5.0	0.10	5.0
V(ϕ^2)	3.0	3.0	0.10	0.10
BrAA				
0	1.0	1.0	1.0	1.0
>0	1.10 (0.66, 1.83)	1.08 (0.65, 1.81)	1.08 (0.65, 1.80)	1.07 (0.65, 1.76)
Br_2AA				
0	1.0	1.0	1.0	1.0
0-1.2	0.89 (0.50, 1.60)	0.88 (0.49, 1.59)	0.9 (0.50, 1.62)	0.90 (0.51, 1.59)
>1.2	0.99 (0.57, 1.73)	1.00 (0.58, 1.71)	1.00 (0.58, 1.75)	1.02 (0.61, 1.73)
Br_3AA				
0	1.0	1.0	1.0	1.0
>0	0.94 (0.59, 1.51)	0.95 (0.59, 1.53)	0.95 (0.60, 1.51)	0.93 (0.59, 1.48)
BrClAA				
0	1.0	1.0	1.0	1.0
0-3.8	0.60 (0.22, 1.63)	0.61 (0.23, 1.63)	0.60 (0.22, 1.66)	0.52 (0.18, 1.49)
3.8-5.9	0.44 (0.15, 1.30)	0.44 (0.15, 1.24)	0.44 (0.15, 1.32)	0.36 (0.11, 1.16)
>5.9	0.68 (0.20, 2.26)	0.65 (0.20, 2.10)	0.68 (0.20, 2.34)	0.56 (0.15, 2.06)
$BrCl_2$AA				
0-1.6	1.0	1.0	1.0	1.0
1.6-3.7	1.56 (1.01, 2.43)	1.56 (1.00, 2.43)	1.59 (1.00, 2.52)	1.60 (1.02, 2.50)
3.7-5.8	1.19 (0.59, 2.40)	1.16 (0.56, 2.40)	1.2 (0.59, 2.43)	1.25 (0.61, 2.55)
>5.8	1.10 (0.51, 2.37)	1.06 (0.48, 2.33)	1.1 (0.49, 2.47)	1.15 (0.51, 2.58)

TABLE 4.6: continued.

μ	6.9	6.9	6.9	6.9
E(ϕ^2)	0.10	5.0	0.10	5.0
V(ϕ^2)	3.0	3.0	0.10	0.10
Br_2ClAA				
0	1.0	1.0	1.0	1.0
0-2	0.61 (0.38, 0.99)	0.61 (0.38, 0.96)	0.6 (0.37, 0.98)	0.59 (0.35, 0.97)
>2	0.76 (0.46, 1.26)	0.76 (0.46, 1.26)	0.76 (0.47, 1.25)	0.75 (0.44, 1.26)

TABLE 4.7: Sensitivity analysis for DPP1 model.

μ	1.0	1.0	3.0	3.0	6.9
E(ϕ^2)	0.31	0.30	0.30	0.10	5.0
V(ϕ^2)	0.07	0.10	0.10	0.10	0.10
a	.01	1	1	1	1
b	.01	1	1	1	1
CHCl3					
0-0.1	1.0	1.0	1.0	1.0	1.0
0.1-16.3	1.01 (0.94, 1.08)	1.02 (0.85, 1.22)	1.02 (0.84, 1.23)	1.05 (0.72, 1.54)	1.01 (0.89, 1.15)
16.3 - 44.3	1.01 (0.92, 1.10)	1.01 (0.89, 1.16)	1.01 (0.91, 1.11)	1.01 (0.94, 1.08)	1.01 (0.92, 1.10)
>44.3	1.02 (0.84, 1.24)	1.08 (0.74, 1.58)	1.07 (0.73, 1.58)	1.06 (0.73, 1.54)	1.05 (0.74, 1.50)
CHBrCl2					
0-1.1	1.0	1.0	1.0	1.0	1.0
1.1-11.4	1.00 (0.94, 1.07)	1.00 (0.91, 1.11)	1.00 (0.96, 1.05)	1.00 (0.96, 1.05)	1.00 (0.96, 1.05)
11.4-17.5	1.00 (0.93, 1.08)	1.00 (0.89, 1.12)	1.00 (0.93, 1.08)	1.00 (0.96, 1.05)	1.00 (0.95, 1.05)
>17.5	1.01 (0.95, 1.07)	1.01 (0.87, 1.18)	1.00 (0.96, 1.05)	1.01 (0.90, 1.13)	1.00 (0.95, 1.06)
CHBr2Cl					
0-1.4	1.0	1.0	1.0	1.0	1.0
1.4-2.8	0.81 (0.35, 1.90)	0.59 (0.22, 1.57)	0.75 (0.27, 2.04)	1.00 (0.97, 1.04)	0.68 (0.20, 2.31)
2.8-7.4	0.99 (0.87, 1.13)	0.98 (0.81, 1.19)	1.00 (0.96, 1.04)	1.00 (0.97, 1.04)	1.00 (0.96, 1.05)
>7.4	1.00 (0.94, 1.07)	1.00 (0.89, 1.13)	1.00 (0.96, 1.05)	1.00 (0.97, 1.04)	1.00 (0.93, 1.08)

TABLE 4.7: continued.

μ	1.0	1.0	3.0	3.0	6.9
E(ϕ^2)	0.31	0.30	0.30	0.10	5.0
V(ϕ^2)	0.07	0.10	0.10	0.10	0.10
a	.01	1	1	1	1
b	.01	1	1	1	1
CHBr3					
0	1.0	1.0	1.0	1.0	1.0
0-0.5	1.00 (0.93, 1.07)	0.98 (0.81, 1.18)	0.99 (0.86, 1.15)	1.00 (0.97, 1.04)	1.00 (0.96, 1.04)
0.5-0.9	1.00 (0.93, 1.08)	0.99 (0.84, 1.15)	1.00 (0.96, 1.05)	1.00 (0.97, 1.04)	1.00 (0.96, 1.04)
> 0.9	1.00 (0.95, 1.06)	1.00 (0.88, 1.15)	1.01 (0.94, 1.07)	1.00 (0.97, 1.04)	1.00 (0.96, 1.05)
ClAA					
0	1.0	1.0	1.0	1.0	1.0
0-2.1	1.00 (0.94, 1.06)	0.99 (0.89, 1.11)	1.00 (0.94, 1.07)	1.00 (0.97, 1.04)	1.00 (0.96, 1.04)
>2.1	1.00 (0.95, 1.06)	0.99 (0.87, 1.12)	1.00 (0.96, 1.04)	1.00 (0.97, 1.04)	1.00 (0.95, 1.05)
Cl2AA					
0	1.0	1.0	1.0	1.0	1.0
0-8.8	1.00 (0.94, 1.07)	1.01 (0.86, 1.19)	1.01 (0.93, 1.08)	1.04 (0.76, 1.42)	1.01 (0.86, 1.19)
8.8-18.3	1.00 (0.95, 1.06)	1.00 (0.90, 1.11)	1.00 (0.95, 1.07)	1.00 (0.96, 1.05)	1.00 (0.96, 1.05)
>18.3	1.00 (0.94, 1.07)	1.00 (0.90, 1.12)	1.00 (0.96, 1.05)	1.00 (0.96, 1.05)	1.00 (0.94, 1.07)

TABLE 4.7: continued.

μ	1.0	1.0	3.0	3.0	6.9
E(ϕ^2)	0.31	0.30	0.30	0.10	5.0
V(ϕ^2)	0.07	0.10	0.10	0.10	0.10
a	.01	1	1	1	1
b	.01	1	1	1	1
Cl3AA					
0	1.0	1.0	1.0	1.0	1.0
0-5.4	1.02 (0.85, 1.23)	1.05 (0.80, 1.37)	1.02 (0.87, 1.19)	1.02 (0.84, 1.24)	1.07 (0.74, 1.55)
5.4-11.8	1.00 (0.94, 1.07)	1.01 (0.89, 1.14)	1.01 (0.94, 1.08)	1.01 (0.90, 1.13)	1.00 (0.92, 1.09)
>11.8	1.00 (0.92, 1.09)	0.99 (0.84, 1.18)	1.01 (0.92, 1.10)	1.00 (0.94, 1.08)	1.00 (0.92, 1.08)
BrAA					
0	1.0	1.0	1.0	1.0	1.0
>0	1.01 (0.90, 1.14)	1.04 (0.80, 1.35)	1.03 (0.81, 1.3)	1.07 (0.71, 1.61)	1.02 (0.81, 1.29)
Br2AA					
0	1.0	1.0	1.0	1.0	1.0
0-1.2	0.99 (0.84, 1.17)	0.99 (0.88, 1.12)	1.00 (0.95, 1.05)	1.00 (0.97, 1.04)	0.99 (0.89, 1.11)
>1.2	1.01 (0.95, 1.06)	1.01 (0.91, 1.12)	1.01 (0.90, 1.14)	1.01 (0.88, 1.17)	1.00 (0.94, 1.07)
Br3AA					
0	1.0	1.0	1.0	1.0	1.0
>0	1.01 (0.94, 1.07)	1.01 (0.90, 1.12)	1.01 (0.90, 1.14)	1.00 (0.97, 1.04)	1.01 (0.93, 1.09)

TABLE 4.7: continued.

μ	1.0	1.0	3.0	3.0	6.9
E(ϕ^2)	0.31	0.30	0.30	0.10	5.0
V(ϕ^2)	0.07	0.10	0.10	0.10	0.10
a	.01	1	1	1	1
b	.01	1	1	1	1
BrClAA					
0	1.0	1.0	1.0	1.0	1.0
0-3.8	1.00 (0.95, 1.06)	1.00 (0.91, 1.10)	1.00 (0.95, 1.05)	1.00 (0.97, 1.04)	1.00 (0.96, 1.05)
3.8-5.9	1.00 (0.95, 1.06)	1.00 (0.91, 1.09)	1.00 (0.96, 1.04)	1.00 (0.97, 1.04)	1.00 (0.96, 1.05)
>5.9	1.00 (0.93, 1.08)	1.00 (0.89, 1.13)	1.01 (0.90, 1.13)	1.00 (0.97, 1.04)	1.00 (0.92, 1.10)
BrCl2AA					
0-1.6	1.0	1.0	1.0	1.0	1.0
1.6-3.7	1.00 (0.95, 1.06)	1.01 (0.90, 1.13)	1.00 (0.96, 1.05)	1.00 (0.96, 1.05)	1.00 (0.93, 1.08)
3.7-5.8	1.00 (0.90, 1.10)	0.98 (0.83, 1.16)	0.99 (0.82, 1.19)	1.00 (0.97, 1.04)	0.99 (0.88, 1.12)
>5.8	1.01 (0.91, 1.12)	1.00 (0.92, 1.09)	1.00 (0.95, 1.05)	1.01 (0.92, 1.10)	1.00 (0.96, 1.05)
Br2ClAA					
0	1.0	1.0	1.0	1.0	1.0
0-2	0.99 (0.88, 1.12)	0.98 (0.80, 1.19)	1.00 (0.93, 1.07)	1.00 (0.97, 1.04)	0.98 (0.74, 1.30)
>2	1.00 (0.95, 1.05)	1.00 (0.92, 1.09)	1.00 (0.94, 1.07)	1.00 (0.97, 1.04)	1.00 (0.96, 1.05)

TABLE 4.8: Sensitivity analysis for DPP2 model.

μ	1.0	1.0	1.0	1.0	6.9
E(ϕ^2)	0.31	0.30	0.30	0.10	5.0
V(ϕ^2)	0.07	0.10	0.10	0.10	0.10
c	1	0.04	0.23	1.5	1.5
d	1	0.23	0.04	1.5	1.5
CHCl3					
0-0.1	1.0	1.0	1.0	1.0	1.0
0.1-16.3	1.04 (0.76, 1.43)	1.04 (0.78, 1.41)	1.01 (0.84, 1.23)	1.03 (0.84, 1.25)	1.02 (0.79, 1.31)
16.3 - 44.3	1.01 (0.88, 1.17)	1.01 (0.83, 1.22)	1.00 (0.95, 1.06)	1.02 (0.87, 1.18)	1.00 (0.97, 1.03)
>44.3	1.11 (0.72, 1.72)	1.1 (0.74, 1.64)	1.03 (0.80, 1.33)	1.10 (0.78, 1.56)	1.05 (0.76, 1.44)
CHBrCl2					
0-1.1	1.0	1.0	1.0	1.0	1.0
1.1-11.4	1.00 (0.92, 1.09)	1.00 (0.85, 1.18)	1.00 (0.94, 1.07)	1.01 (0.89, 1.13)	1.00 (0.98, 1.02)
11.4-17.5	0.99 (0.88, 1.11)	0.98 (0.81, 1.18)	1.00 (0.96, 1.04)	0.99 (0.85, 1.14)	1.00 (0.97, 1.03)
>17.5	1.00 (0.93, 1.07)	1.02 (0.84, 1.24)	1.00 (0.94, 1.07)	1.01 (0.87, 1.16)	1.00 (0.98, 1.02)
CHBr2Cl					
0-1.4	1.0	1.0	1.0)	1.0	1.0
1.4-2.8	0.62 (0.22, 1.73)	0.60 (0.23, 1.54)	0.58 (0.21, 1.66)	0.86 (0.54, 1.37)	0.64 (0.18, 2.27)
2.8-7.4	0.99 (0.85, 1.14)	0.98 (0.80, 1.2)	1.00 (0.95, 1.05)	0.98 (0.83, 1.15)	1.00 (0.94, 1.06)
>7.4	1.00 (0.90, 1.11)	1.00 (0.82, 1.21)	1.00 (0.91, 1.10)	0.99 (0.87, 1.13)	1.00 (0.98, 1.02)

TABLE 4.8: continued.

μ	1.0	1.0	1.0	1.0	6.9
E(ϕ^2)	0.31	0.30	0.30	0.10	5.0
V(ϕ^2)	0.07	0.10	0.10	0.10	0.10
c	1	0.04	0.23	1.5	1.5
d	1	0.23	0.04	1.5	1.5
CHBr3					
0	1.0	1.0	1.0	1.0	1.0
0-0.5	0.98 (0.79, 1.2)	0.95 (0.71, 1.27)	0.99 (0.86, 1.14)	0.97 (0.80, 1.18)	0.99 (0.88, 1.12)
0.5-0.9	0.98 (0.83, 1.17)	0.97 (0.80, 1.18)	0.99 (0.85, 1.15)	0.98 (0.84, 1.15)	1.00 (0.96, 1.04)
> 0.9	1.00 (0.91, 1.10)	1.00 (0.84, 1.2)	1.00 (0.95, 1.06)	1.00 (0.88, 1.15)	1.00 (0.96, 1.04)
ClAA					
0	1.0	1.0	1.0	1.0	1.0
0-2.1	0.99 (0.89, 1.10)	1.00 (0.86, 1.15)	1.00 (0.95, 1.05)	0.98 (0.86, 1.12)	1.00 (0.96, 1.04)
>2.1	1.00 (0.93, 1.07)	0.98 (0.85, 1.14)	1.00 (0.98, 1.02)	0.98 (0.87, 1.12)	1.00 (0.98, 1.02)
Cl2AA					
0	1.0	1.0	1.0	1.0	1.0
0-8.8	1.01 (0.87, 1.17)	1.00 (0.82, 1.22)	1.00 (0.92, 1.09)	1.00 (0.87, 1.15)	1.00 (0.95, 1.06)
8.8-18.3	0.99 (0.88, 1.12)	0.99 (0.83, 1.19)	0.99 (0.89, 1.11)	0.99 (0.86, 1.14)	1.00 (0.93, 1.08)
>18.3	1.00 (0.90, 1.10)	1.01 (0.84, 1.23)	1.00 (0.94, 1.06)	1.00 (0.88, 1.13)	1.00 (0.96, 1.04)

TABLE 4.8: continued.

μ	1.0	1.0	1.0	1.0	6.9
E(ϕ^2)	0.31	0.30	0.30	0.10	5.0
V(ϕ^2)	0.07	0.10	0.10	0.10	0.10
c	1	0.04	0.23	1.5	1.5
d	1	0.23	0.04	1.5	1.5
Cl3AA					
0	1.0	1.0	1.0	1.0	1.0
0-5.4	1.04 (0.80, 1.34)	1.10 (0.75, 1.6)	1.02 (0.83, 1.27)	1.05 (0.81, 1.37)	1.00 (0.95, 1.05)
5.4-11.8	1.00 (0.87, 1.15)	1.01 (0.83, 1.23)	1.00 (0.94, 1.07)	1.00 (0.87, 1.16)	1.00 (0.92, 1.08)
>11.8	0.99 (0.81, 1.21)	0.99 (0.77, 1.27)	0.99 (0.78, 1.25)	1.00 (0.83, 1.19)	1.00 (0.96, 1.05)
BrAA					
0	1.0	1.0	1.0	1.0	1.0
>0	1.05 (0.78, 1.43)	1.02 (0.83, 1.26)	1.01 (0.87, 1.17)	1.03 (0.84, 1.26)	1.00 (0.93, 1.08)
Br2AA					
0	1.0	1.0	1.0	1.0	1.0
0-1.2	0.99 (0.83, 1.17)	0.97 (0.77, 1.24)	0.99 (0.82, 1.19)	0.98 (0.84, 1.14)	1.00 (0.94, 1.06)
>1.2	1.00 (0.94, 1.07)	1.00 (0.87, 1.15)	1.00 (0.96, 1.04)	1.00 (0.91, 1.10)	1.00 (0.98, 1.02)
Br3AA					
0	1.0	1.0	1.0	1.0	1.0
>0	1.01 (0.92, 1.10)	1.01 (0.87, 1.17)	1.01 (0.92, 1.09)	1.02 (0.89, 1.16)	1.00 (0.99, 1.02)

TABLE 4.8: continued.

μ	1.0	1.0	1.0	1.0	6.9
E(ϕ^2)	0.31	0.30	0.30	0.10	5.0
V(ϕ^2)	0.07	0.10	0.10	0.10	0.10
c	1	0.04	0.23	1.5	1.5
d	1	0.23	0.04	1.5	1.5
BrClAA					
0	1.0	1.0	1.0	1.0	1.0
0-3.8	1.00 (0.94, 1.06)	1.00 (0.86, 1.18)	1.00 (0.94, 1.07)	1.00 (0.90, 1.12)	1.00 (0.97, 1.04)
3.8-5.9	0.99 (0.87, 1.12)	0.97 (0.80, 1.18)	1.00 (0.96, 1.04)	0.98 (0.85, 1.14)	1.00 (0.94, 1.05)
>5.9	1.00 (0.92, 1.09)	1.02 (0.83, 1.24)	1.00 (0.96, 1.04)	1.01 (0.85, 1.20)	1.00 (0.95, 1.05)
BrCl2AA					
0-1.6	1.0	1.0	1.0	1.0	
1.6-3.7	1.00 (0.93, 1.09)	1.03 (0.85, 1.24)	1.00 (0.96, 1.05)	1.01 (0.89, 1.15)	1.00 (0.97, 1.04)
3.7-5.8	0.99 (0.91, 1.08)	0.98 (0.83, 1.17)	0.99 (0.88, 1.12)	0.98 (0.85, 1.13)	1.00 (0.94, 1.06)
>5.8	1.00 (0.94, 1.08)	1.00 (0.85, 1.17)	1.00 (0.95, 1.05)	1.01 (0.89, 1.14)	1.00 (0.99, 1.01)
Br2ClAA					
0	1.0	1.0	1.0	1.0	1.0
0-2	0.98 (0.77, 1.23)	0.97 (0.78, 1.20)	0.99 (0.90, 1.10)	0.98 (0.84, 1.14)	1.00 (0.97, 1.03)
>2	1.00 (0.94, 1.06)	0.98 (0.82, 1.17)	1.00 (0.96, 1.04)	0.99 (0.90, 1.10)	1.00 (0.99, 1.01)

4.7 Appendix 2: Winbugs Code for Semi-Bayes and Fully-Bayes Models

We provide a generic template of Winbugs code that can be used to implement either SB and FB models. We present code for a hypothetical dataset with a binary outcome, y, and 7 dichotomous covariates $x_1 \ldots x_7$. Information on how to read data into Winbugs can be found in the Winbugs manual.(Spiegelhalter et al., 1999) We use the following data:

list(x1=c(0,1,0,0,0,0,0,0), x2=c(0,0,1,0,0,0,0,0), x3=c(0,0,0,1,0,0,0,0),
x4=c(0,0,0,0,1,0,0,0), x5=c(0,0,0,0,0,1,0,0), x6=c(0,0,0,0,0,0,1,0),
x7=c(0,0,0,0,0,0,0,1),
n = c(100,100,100,100,100,100,100,100), y = c(10, 8, 11, 12, 9, 13, 11, 14), N = 8,
J=7)

The data are in aggregate form (i.e., there 100 people who are unexposed to $x_1 \ldots x_7$ and 10 of them have the outcome. There are 100 people who are exposed to x_1 and 8 of them have the outcome, etc). The following Winbugs code can be used to analyze this dataset using a SB model:

4.7.1 Winbugs Code for SB Model

model {
for(i in 1 : N) {
y[i] ~ dbin(p[i],n[i])
logit(p[i]) <- alpha + bsb[1]*x1[i]+bsb[2]*x2[i]+bsb[3]*x3[i]
+ bsb[4]*x4[i]+bsb[5]*x5[i]+bsb[6]*x6[i]+bsb[7]*x7[i] }
for(j in 1:J) {
bsb[j] ~ dnorm(0,.3) }
alpha ~ dnorm(0.0,0.01) }

We note that dnorm(a,b) is a normal distribution with mean a and variance 1/b. Therefore, in the FB model a gamma prior is place on the inverse of the variance (as opposed to our approach earlier which placed an inverse gamma prior on the variance).

4.7.2 Winbugs Code for FB Model

model {
for(i in 1 : N) {
y[i] ~ dbin(p[i],n[i])
logit(p[i]) <- alpha + bfb[1]*x1[i]+bfb[2]*x2[i]+bfb[3]*x3[i]
+ bfb[4]*x4[i]+bfb[5]*x5[i]+bfb[6]*x6[i]+bfb[7]*x7[i] }
for(j in 1:J) {
bfb[j] ~ dnorm(0,phi) }
alpha ~ dnorm(0.0,0.01)
phi ~ dgamma(0.075,4)}

In the FB model, dgamma is a Gamma(α, β) distribution with mean= $\alpha\beta$ and variance=$\alpha\beta^2$. So our above specification gives a prior mean of 0.3 and prior variance of 1.2.

The code in sections A.2.1 and A.2.2 can be run for 50,000 iterations in Winbugs in a matter of seconds.

CHAPTER 5

DISCUSSION

5.1 The Use of Bayesian Methods for Correlated Data

Although highly correlated data are common in epidemiologic research. Standard analytic techniques, such as maximum likelihood regression may provide very unstable estimates or even fail to converge. We have presented four Bayesian hierarchical models that have superior performance when compared to standard techniques. Although we have presented these models in the scope of correlated data, these hierarchical models could also prove useful in regressions with a large number of variables.(Dunson et al., 2005)

5.1.1 The Semi-Bayes Model

The semi-Bayes model was introduced over 10 years ago and has seen periodic use. Researchers have used semi-Bayes models in occupational, genetic, nutritional and cancer epidemiology.(De Roos et al., 2001; Greenland, 1992; Hung et al., 2004; Witte et al., 1994) By placing a prior distribution on model coefficients, the semi-Bayes model not only allows the researcher to incorporate prior knowledge but also shrinks coefficients toward that prior distribution. The amount of shrinkage in the semi-Bayes model depends on the prior variance. Smaller prior variances (indicating more prior knowledge) cause greater shrinkage to the prior mean while larger prior variances (indicating less prior knowledge) cause less shrinkage. In datasets of moderate size, the impact of the prior distribution is likely to be minimal. Previous studies that have used semi-Bayes models frequently specify relatively large prior variances. For instance, Kirrane et al.

specify a prior variance equivalent to 95% of possible ORs falling in a ten-fold range. There are two problems with such large prior variances. First they are almost guaranteed to cause little shrinkage and be dominated by the observed data. Second, they are frequently incommensurate with prior knowledge. In the study by Kirrane et al., the authors indicate prior research showed a small increased risk of macular degeneration among users of pesticides (with the OR observed in a previous study of 2.0). It is unlikely the investigators would truly assign any prior probability to an OR=5, let alone OR=10. Users of semi-Bayes models should consider specifying more substantively realistic prior variances (ORs of 10 could be ruled out a priori in most studies) to reap more benefits from the Bayesian model.

A further troubling aspect of the use of semi-Bayes models is their role in reducing the type-I error rate in hypothesis testing.(Hung et al., 2004; Steenland et al., 2000) As we have demonstrated, there are two problems with this approach. First, semi-Bayes credible intervals only have increased frequentist coverage (i.e., they cover the true parameter estimate $\geq (1-\alpha)\%$ of the time and so are less likely to incorrectly reject the null) when the prior mean is zero. While such a prior mean may sometimes be justifiable, it will frequently be incommensurate with existing research. Second, even if setting the prior mean to zero is reasonable, the increased coverage probability will generally be minimal. Since this method requires assumptions that will frequently be untenable and even when tenable, will produce little gain in coverage, we suggest against using semi-Bayes methods for reducing type-I error rates.

Our simulation results generally demonstrate that the semi-Bayes model has somewhat worse properties than the other three Bayesian hierarchical models that we examine. This is not a surprising result. The semi-Bayes models suffers, to paraphrase Jimmie Savage, from breaking the Bayesian egg without making a Bayesian omelet.(Savage, 1954) That is, the researcher who uses semi-Bayes models allows some amount of Bayesian learning by updating the prior distribution about the effects with the observed data, but doesn't allow the prior variance to be updated with the observed data. It stands to reason that methods that do allow the prior variance to be updated will outperform the semi-Bayes method simply because they make use of more available data.

This result is also somewhat misleading: it is possible to generate scenarios in which the semi-Bayes model outperforms (in terms of mean squared error) the fully-Bayes model. The scenarios in which the fully-Bayes model will most radically outperform the semi-Bayes model will be ones in which the semi-Bayes model has specified a prior

variance that is completely incompatible with the data. On the other hand, were we to generate a dataset and fit a semi-Bayes model with a prior variance equal to the variance observed in the dataset, the semi-Bayes model could perform somewhat better than the fully-Bayes model (since the prior variance was correctly specified to begin with). However, as we are never likely to know what the true variance is, we view the fully-Bayes approach as superior to the semi-Bayes approach. Indeed, in the applied example on disinfection by-products and spontaneous abortion we found that our entirely plausible prior variance in the semi-Bayes model was completely incompatible with the small amount of variability between estimates in the observed data.

When semi-Bayes models were introduced, presumably, it was because they were easier to fit than fully-Bayes models given the limitations of existing software at that time. However, in presenting methods to fit semi-Bayes models, these authors relied on asymptotic properties.(Witte et al., 1998) It is important for researchers to recognize that it is precisely those situations where asymptotics will hold (i.e., with large datasets) that semi-Bayes methods will be least useful. It is in those datasets where asymptotic assumptions are most tenuous that Bayesian methods will be most useful. Many of the recent articles using semi-Bayes techniques are frequently implemented in studies where asympototic assumptions may be tenuous, at best (for instance Kirrane et al. observe cell sizes of zero and De Roos et al. observe cell sizes of two).(De Roos et al., 2001; Kirrane et al., 2005) We have given templates of semi-Bayes and fully-Bayes code in Winbugs to alleviate the need to rely on asymptotic normality in fitting Bayesian models. We have also presented the basics of the Gibbs sampling routines we programmed in Matlab.

5.1.2 The Fully-Bayes Model

While the semi-Bayes model is a large improvement over standard techniques, it too is easy to improve upon. The fully-Bayes model is the most straightforward improvement. Rather than treating the prior mean and variance as fixed, it places distributions on them and allows them to be updated using data in the study. These models are common in other disciplines, but lacking in epidemiology. The simulation results we presented showed the fully-Bayes model typically having smaller mean squared error than the semi-Bayes model. It generally did not perform as well as the more complicated Dirichlet process models. Our applied analysis of disinfection by-products and spontaneous abortion showed a situation in which the fully-Bayes model offered

profound benefits, compared to the semi-Bayes model. Very little variation was seen between estimated effects in the Right from the Start data. By updating the prior variance with the observed data, the fully-Bayes model generate estimates with greater shrinkage and much greater precision than the semi-Bayes model.

The fully-Bayes applications in this dissertation have generally treated the prior mean as fixed but allowed the prior variance to be random by placing a prior distribution on it. The reason we have treated the prior mean as fixed is because our models have not specified more complex formulations of the prior mean (for instance, the prior mean could be a linear combination of other covariates: $\mu_j = \alpha_0 + \alpha_1 z_{1j} + \alpha_2 z_{2j}$). If we had specified the prior mean as a function of other covariates, we could easily place a prior distribution on the effects of those covariates (say, $\alpha_k \sim N(\mu_2, \phi_2^2)$). However, when the prior mean is a constant, placing a prior distribution on it is redundant and only serves to increase the prior variance. If uncertainty exists concerning the prior mean, it should be incorporated directly into the prior variance rather than through a hyperprior distribution on the prior mean.

More generally, some may be concerned with how many hierarchies (prior distributions) are sufficient in a hierarchical model. We suggest that prior distributions should be used when there is important information to incorporate through that prior distribution and when the use of the prior distribution has practical advantages. For instance, placing a prior distribution on a main effect (such as the effect of $CHCl_3$ on spontaneous abortion) allows incorporation of prior knowledge and, in practical terms, allows shrinkage of estimates and decreased mean squared error. Placing a hyperprior distribution on the prior variance has similar advantage: it allows us to use the data to help update our prior knowledge of the variance. Placing a hyperprior distribution on a constant mean, however, has no practical advantage and only serves to increase the prior variance.

Presumably, the lack of fully-Bayes models in the epidemiologic literature is partially due to the lack of a SAS procedure to fit them (as there is for semi-Bayes). The Winbugs code we provide for the fully-Bayes model contains only a few more lines than the Winbugs code for the semi-Bayes model and we feel certain that if researchers invest a few hours learning Winbugs they will find these programs easier to run than the SAS code for the semi-Bayes model.

5.1.3 The Dirichlet Process Models

We relaxed the parametric assumptions of the semi-Bayes and fully-Bayes model by introducing a Dirichlet process prior. Rather than assuming that the coefficients had a normal distribution, we allowed their distribution to be unknown and assigned that unknown distribution a Dirichlet process prior, both without (DPP1) and with (DPP2) a selection component. The Dirichlet process models allowed the regression coefficients to be clustered into groups at each iteration of the Gibbs sampler. The clustering process served to increase precision in the estimates. The two Dirichlet process models generally had the most precise estimates and smallest MSE of the four models we examined.

A possible concern of the semi-Bayes and fully-Bayes models is that they serve to shrink all estimates toward the same common mean. For instance, the effects of all 13 constituent disinfection by-products were shrunk toward the prior mean of zero. This is a good property to have when all estimates are believed to have the same effect, however consider the situation in which one of the 13 constituents has an effect but the other 12 do not. The by-product that has an effect is still shrunk toward the prior mean of zero, making its effect less apparent. The Dirichlet process prior rectifies this problem by allowing coefficients to occupy their own cluster. In the disinfection by-product example, the 12 disinfection by-products will be shrunk toward zero while the one by-product that does have an effect will have its own cluster that will not be shrunk toward zero.

The increased precision of the Dirichlet process models can be viewed in two ways. First, by clustering coefficients we are essentially including fewer terms in the regression model. With fewer terms comes increased precision. Second, clustering coefficients inherently decreases the variance. Consider a linear regression with orthogonal data: $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$. Estimates of $\widehat{\beta_1}$ and $\widehat{\beta_2}$ are given by $\sum x_{i1} y_i / \sum x_{i1}^2$ and $\sum x_{i2} y_i / \sum x_{i2}^2$ with variance $\sigma^2 / \sum x_{i1}^2$ and $\sigma^2 / \sum x_{i2}^2$, respectively. If we cluster β_1 and β_2 together, we are assuming $\beta_1 = \beta_2 = \beta_{cl}$. We can rewrite the regression as $y_i = \beta_{cl}(x_{i1} + x_{i2}) + \epsilon_i$, in which case $\widehat{\beta_{cl}} = \sum (x_{i1} + x_{i2}) y_i / \sum (x_{i1} + x_{i2})^2$ with variance $\sigma^2 / \sum (x_{i1} + x_{i2})^2$. Taking expectations of $\widehat{\beta_1}$, $\widehat{\beta_2}$ and $\widehat{\beta_{cl}}$ we see that they are all unbiased and equal (assuming $\beta_1 = \beta_2$). However, we see that the variance of $\widehat{\beta_{cl}}$ is smaller than the variance of either $\widehat{\beta_1}$ or $\widehat{\beta_2}$. Thus, when variables are clustered variability decreases. This is intuitively appealing, since by combining coefficients in clusters we are indicating that we have additional information about the size of the cluster's effect (that is, two

covariates worth of information rather than one). Because of the decreased variance associated with clustering coefficients, this procedure can reduce overall mean squared error even when coefficients that do not have precisely equal coefficient estimates are clustered together.

The applied examples in this dissertation show Dirichlet process prior models that estimate routinely null, but very precise, effects for all parameters. A potential concern is that these models might prove ineffective at estimating non-null effects. The results of our simulations in Chapter 3 indicate this is not the case. The same properties have been noted by Dunson et al. in genetic applications.(Dunson et al., 2005) Indeed, by allowing the unknown distribution of the coefficients to differ from the base distribution, the Dirichlet process model may be better at detecting coefficients that have an effect than the semi-Bayes or fully-Bayes models (that shrink all estimates toward zero).

5.2 Disinfection By-products and Spontaneous Abortion

Previous epidemiologic studies have found somewhat discrepant results, but generally indicated an increased risk of spontaneous abortion among women who consume more disinfection by-products. Our conclusion is in keeping with a previous analysis of this data by Savitz et al.: in contrast to previous studies, there is little evidence of any effect of any constituent disinfection by-product on spontaneous abortion in Right from the Start.

The maximum likelihood logistic regression produced results that showed quite a few large associations between constituent disinfection by-products and spontaneous abortion. In particular, the 2^{nd} through 4^{th} quartiles of Cl_2AA and Cl_3AA showed between 1.5 times and 3.0 times the risk of spontaneous abortion as those in the first quartile, while those exposed to concentrations of BrClAA in the 2^{nd} through 4^{th} quartiles had roughly 1/3 the risk of spontaneous abortion as those in the first quartile. The maximum likelihood estimates were often characterized by their extreme imprecision, with the most obvious example being the OR for the 4^{th} quartile (vs. the 1^{st} quartile) of CL_2AA which had a 95% CI of (0.50, 14.29). The extremely imprecise nature of the maximum likelihood estimates made these results virtually impossible to interpret in many cases.

The semi-Bayes model produced effect estimates that were much more precise than

the maximum likelihood estimates, however they were less precise than the other three hierarchical models. The semi-Bayes model indicated generally null results of all constituent disinfection by-products, however there were a few categories of some constituent disinfection by-products that suggested an increased risk of spontaneous abortion, such as the 4^{th} quartile of $CHBrCl_2$ and the 2^{nd} quartile of Cl_3AA. However, the lack of any systematic patterns or dose response relationships make these scattered results seem biologically implausible. Indeed, under the fully-Bayes model, these results were shrunk much further back toward the null. None of the constituent disinfection by-products seemed to have an effect when examined using the fully-Bayes model. The two Dirichlet process models produced results that were quite similar and by far the most precise of all 4 hierarchical models. These semi-parametric models produced results providing no evidence of effect for any constituent disinfection by-product.

Results of sensitivity analyses indicate that with the SB or FB model, our interpretation of results for Cl_2AA and Cl_3AA depends, to some extent, on our prior belief about the effect of these constituent DBPs. However, for Cl_3AA the dose response is opposite what one would expect, with the second quartile (vs. first quartile) having the highest OR and effect for subsequent quartiles diminishing. The effect of Cl_2AA is somewhat more dependent on prior information, indicating a lack of information regarding this effect in the data. The DPP1 and DPP2 models provided little evidence of effect and were much more robust to prior specification.

Our results are consistent with a previous analysis of these data by Savitz et al. However, Savitz et al. generally analyzed these data by aggregating over groups of disinfection by-products (such as the four trihalomethanes and nine haloacetic acids), an approach not sensitive to detecting effects of individual disinfection by-products. They also did not attempt to control for multiple disinfection by-products in their analyses of constituent disinfection by-products, so results could have been confounded.

Previous research on disinfection by-products and spontaneous abortion is limited. Early studies used crude proxies of disinfection by-product consumption, such as the number of glasses of tap-water consumed per day. The two studies that have specifically looked at disinfection by-products are the study by Savitz et al. and by Waller et al.(Savitz et al., 1995; Waller et al., 1998) The study by Savitz found a relationship between total THM consumption and spontaneous abortion, but only in the highest sextile. No association was seen when exposure was classified in tertiles. Savitz et al. measured exposure using quarterly reports from water suppliers. The Savitz et al. study is particularly interesting because it draws participants from the same geographical

location as Right from the Start. The discrepancy between these results and Savitz et al.'s could be due to a number of factors. First, Savitz et al. reported a relatively low response rate in their study so selection bias could account for their positive finding. Second, their measurement of exposure was based on quarterly reports from the water facilities, while Right from the Start had weekly data on disinfection by-products. Possibly, quarterly water data were gathered at a moment unrepresentative of recent disinfection by-product levels.

Waller et al. examined the relation between the four THMs and spontaneous abortion and are the only previous study to look at constituent disinfection by-products.(Waller et al., 1998) They found an increased risk (OR=2.0 95% CI: 1.2, 3.5) of spontaneous abortion among those in the highest exposure quartile of $CHBrCl_2$ (vs the other three quartiles combined). The three other THMs show no effect on spontaneous abortion. Interestingly, when Waller et al. combine all THMs in a single maximum likelihood logistic model their estimates become imprecise. Our study found no association between $CHBrCl_2$ and spontaneous abortion, despite a similar cutpoint for the highest quartile. We were in agreement with their null findings regarding the concentration of $ChCl_3$, $ChBr_3$, and $ChBr_2Cl$, however. It is unlikely that controlling for all disinfection by-products simultaneously accounted for the discrepancy in results since we observed similar results when we analyzed only one by-product at a time. Waller et al. conducted a prospective cohort study, limiting the chance that recall bias was responsible for their findings. The most obvious difference between the this study and theirs is the exposure assessment, with Waller et al. using quarterly data on disinfection by-product concentrations as opposed to our weekly data. Again, the quarterly data in Waller, could have been gathered at a moment unrepresentative of recent disinfection by-product levels.

Our study is not without its limitations. We have studied the relationship between disinfection by-product concentration in a given week and the probability of spontaneous abortion. The week-specific concentration of disinfection by-products may be less important than the accumulated dosage. However, examining the effect of cumulative dose on week specific probability of loss is a more difficult problem that we will examine in future research. We have not accounted for the amount of disinfection by-product actually consumed through ingestion or through other routes (such as inhalation). Although such measures are available, we believe the crude measure of disinfection by-product concentration in the water supply is a good proxy for these measures. There is concern that the time at which a pregnancy is no longer viable does

not always correlate well with the time at which the products of conception are lost from the uterus. Preliminary results adjusting for this misclassification indicate it does not change our interpretation of the results. Although Right from the Start enrolled women in early pregnancy, very early losses (those before 5 weeks) were impossible for us to detect. It is possible that disinfection by-products have an effect on very early losses, and we were unable to detect it.

5.3 Summary

The hierarchical models we used greatly reduced the variability of the estimates. The semi-Bayes model proved to be overly dependent on the specification of the prior variance. The fully-Bayes model allowed the prior variance to be update based on the data, which was important in Right from the Start since the specification of the prior variance ($\phi^2 = 0.3142$) was far greater than the variability noted in the data. The posterior estimate of the prior variance from the fully-Bayes model was almost 8 times smaller. By allowing the prior variance to be update based on the observed data, we were able to obtain much more precise estimates. The two semi-parametric Dirichlet process prior models we implemented provided the most precise results and generally had the smallest mean squared error in simulations. These models were also very robust to prior specification.

Our results suggest that disinfection by-products may not have an effect on spontaneous abortion. The use of hierarchical models enabled us to adjust for a large number of correlated exposures while also incorporating subject matter knowledge. Although semi-Bayes models are the most commonly used Bayesian hierarchical models in epidemiology, our results suggest that more complex models that allow prior parameters to be updated based on the data can have large benefits.

REFERENCES

al Ansary, L. and Babay, Z. (1994). Risk factors for spontaneous abortion: a preliminary study on saudi women. *J R Soc Health*, 114(4):188–93.

Alavanja, M., Sandler, D., McMaster, S., Zahm, S., McDonnell, C., Lynch, C., Pennybacker, M., Rothman, N., Dosemeci, M., Bond, A., and Blair, A. (1996). The agricultural health study. *Environ Health Perspect*, 104(4):362–9.

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *JASA*, 88(422):669–79.

Aschengrau, A., Zierler, S., and Cohen, A. (1989). Quality of community drinking water and the occurrence of spontaneous abortion. *Arch Environ Health*, 44(5):283–90.

Bellar, T., Lichtenberg, J., and Kroner, R. (1974). The occurrence of organohalides in chlorinated drinking waters. *Journal of the American Water Works Association*, 66(12):703.

Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1:353–355.

Bove, F. J., Fulcomer, M. C., Klotz, J. B., Esmart, J., Dufficy, E. M., and Savrin, J. E. (1995). Public drinking water contamination and birth outcomes. *Am J Epidemiol*, 141(9):850–62.

Cao, G. and West, M. (1996). Practical bayesian inference using mixtures of mixtures. *Biometrics*, 52:1334–1341.

Casella, G. (1985). An introduction to empirical bayes data analysis. *The American Statistician*, 39(2):83–7.

Casella, G. and George, E. (1992). Explaining the gibbs sampler. *American Statistician*, 46:167–74.

Cnattingius, S., Signorello, L. B., Anneren, G., Clausson, B., Ekbom, A., Ljunger, E., Blot, W. J., McLaughlin, J. K., Petersson, G., Rane, A., and Granath, F. (2000). Caffeine intake and the risk of first-trimester spontaneous abortion. *N Engl J Med*, 343(25):1839–45.

Cole, S. and Ananth, C. (2001). Regression models for unconstrained, partially or fully constrained continuation odds ratios. *International Journal of Epidemiology*, 30(6):1379–82.

Coste, J., Job-Spira, N., and Fernandez, H. (1991). Risk factors for spontaneous abortion: a case-control study in france. *Hum Reprod*, 6(9):1332–7.

Cox, D. (1972). Regression models and life tables. *Journal of the Royal Statistical Society B*, 34:187–330.

Crump, K. S. and Guess, H. A. (1982). Drinking water and cancer: review of recent epidemiological findings and assessment of risks. *Annu Rev Public Health*, 3:339–57.

De Roos, A., Poole, C., Teschke, K., and Olshan, A. (2001). An application of hierarchical regression in the investigation of multiple paternal occupational exposures and neuroblastoma in offspring. *Am J Ind Med*, 39(5):477–86.

Deane, M., Swan, S., Harris, J., Epstein, D., and Neutra, R. (1989). Adverse pregnancy outcomes in relation to water contamination, santa clara county, california, 1980-1981. *Am J Epidemiol*, 129(5):894–904.

Dennis, W., Olivieri, V., and Kruse, C. (1979). Mechanism of disinfection: Incorporation of cl-36 into f2 virus. *Water Research*, 13(4):363–369.

Dodds, L., King, W., Allen, A. C., Armson, B. A., Fell, D. B., and Nimrod, C. (2004). Trihalomethanes in public water supplies and risk of stillbirth. *Epidemiology*, 15(2):179–86.

Dodds, L., King, W., Woolcott, C., and Pole, J. (1999). Trihalomethanes in public water supplies and adverse birth outcomes. *Epidemiology*, 10(3):233–7.

Draper, D. (1995). Assessment and propagation of model uncertainty. *J. Roy. Statist. Soc. B*, 57:45–70.

Dunson, D., Herring, A., and Mulherin-Engel, S. (2005). Bayesian selection and clustering of polymorphisms in functionally related genes. *ISDS Tech Report*.

Efron, B. and Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23:70–86.

Engel, S., Erichsen, H., Savitz, D., Thorp, J., Chanock, S., and Olshan, A. (2005a). Risk of spontaneous preterm birth is associated with common proinflammatory cytokine polymorphisms. *Epidemiology*, 16(4):469–77.

Engel, S., Olshan, A., Savitz, D., Thorp, J., Erichsen, H., and Chanock, S. (2005b). Risk of small-for-gestational age is associated with common anti-inflammatory cytokine polymorphisms. *Epidemiology*, 16(4):478–86.

EPA (1995a). Method 551.1: Determination of chlorination disinfection byproducts, chlorinated solvents, and halogenated pesticides/herbicides in chromatography with electron-capture detection. Technical report, Environmental Protection Agency.

EPA (1995b). Method 552.2: determination of haloacetic acids and dlapon in drinking water by liquid-liquid extraction, derivatization and gas chromatography with electron catpure detection.

Escobar, M. (1994). Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277.

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588.

Escobar, M. and West, M. (1998). Computing nonparametric hierarchical models. In Dey, D., Muller, P., and Sinha, D., editors, *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 1–22. Springer-Verlag, New York.

Fabius, J. (1964). Asymptotic behavior of bayes estimates. *The Annals of Mathematical Statistics*, 35:846–856.

Fenster, L., Eskenazi, B., Windham, G., and Swan, S. (1991). Caffeine consumption during pregnancy and spontaneous abortion. *Epidemiology*, 2(3):168–74.

Fenster, L., Hubbard, A., Swan, S., Windham, G., Waller, K., Hiatt, R., and Benowitz, N. (1997). Caffeinated beverages, decaffeinated coffee, and spontaneous abortion. *Epidemiology*, 8(5):515–23.

Fenster, L., Windham, G., Swan, S., Epstein, D., and Neutra, R. (1992). Tap or bottled water consumption and spontaneous abortion in a case-control study of reporting consistency. *Epidemiology*, 3(2):120–4.

Ferguson, T. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230.

Ferguson, T. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2:615–29.

Freedman, D. (1963). On the asymptotic behavior of bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34:1386–1403.

Gelfand, A. and Kuo, L. (1991). Nonparametric bayesian bioassay including ordered polytomous response. *Biometrika*, 78(3):657–66.

Gelman, A. (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(2):1–19.

Geweke, J. (1996). Variable selection and model comparison in regression. In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics 5*, pages 609–620. Oxford Press.

Goldhaber, M. and Fireman, B. (1991). The fetal life table revisited: spontaneous abortion rates in three kaiser permanente cohorts. *Epidemiology*, 2(1):33–9.

Gopalan, R. and Berry, D. (1998). Bayesian multiple comparisons using dirichlet process priors. *Journal of the American Statistical Association*, 93(443):1130–1139.

Greenland, S. (1992). A semi-bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. *Stat Med*, 11(2):219–30.

Greenland, S. (1993). Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-bayes regression. *Stat Med*, 12(8):717–36.

Greenland, S. (1994). Hierarchical regression for epidemiologic analyses of multiple exposures. *Environ Health Perspect*, 102 Suppl 8:33–9.

Greenland, S. (2000). Principles of multilevel modelling. *Int. J. Epid.*, 29:158–67.

Greenland, S., Pearl, J., and Robins, J. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.

Greenland, S. and Poole, C. (1994). Empirical-bayes and semi-bayes approaches to occupational and environmental hazard surveillance. *Arch Environ Health*, 49(1):9–16.

Gulati, D., Hope, E., and Barnes, L. (1989). Bromoform: reproductive and fertility assessment in swiss cd-1 mice when administered by gavage. *NTP-86-FACB-053*.

Haas, C. and Engelbrecht, R. (1980a). Chlorine dynamics during inactivation of coliforms, acid-fast bacteria and yeasts. *Water Research*, 14(12):1749–1757.

Haas, C. and Engelbrecht, R. (1980b). Physiological alterations of vegetative microorganisms resulting from aqueous chlorination. *Journal of the Water Pollution Control Federation*, 52:1976–1989.

Hansteen, I. L. (1990). Occupational and lifestyle factors and chromosomal aberrations of spontaneous abortions. *Prog Clin Biol Res*, 340B:467–75.

Harlap, S. and Shiono, P. H. (1980). Alcohol, smoking, and incidence of spontaneous abortions in the first and second trimester. *Lancet*, 2(8187):173–6.

Hertz-Picciotto, I. (2000). The evidence that lead increases the risk for spontaneous abortion. *Am J Ind Med*, 38(3):300–9.

Hertz-Picciotto, I., Swan, S., and Neutra, R. (1992). Reporting bias and mode of interview in a study of adverse pregnancy outcomes and water consumption. *Epidemiology*, 3(2):104–12.

Hoerl, A. and Kennard, R. (1970a). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Hoerl, A. and Kennard, R. (1970b). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoff, J. (1986). Inactivation of microbial agents by chemical disinfectants.

Hosmer, D. and Lemeshow, S. (1989). *Applied Logistic Regression*. Jon Wiley & Sons, New York.

Hung, R., Brennan, P., Malaveille, C., Porru, S., Donato, F., Boffetta, P., and Witte, J. (2004). Using hierarchical modeling in genetic association studies with multiple

markers: application to a case-control study of bladder cancer. *Cancer Epidemiol Biomarkers Prev*, 13(6):1013–21.

Hunter, E. S., r., Rogers, E. H., Schmid, J. E., and Richard, A. (1996). Comparative effects of haloacetic acids in whole embryo culture. *Teratology*, 54(2):57–64.

Infante-Rivard, C., Fernandez, A., Gauthier, R., David, M., and Rivard, G. E. (1993). Fetal loss associated with caffeine intake before and during pregnancy. *Jama*, 270(24):2940–3.

Ishwaran, H. and Rao, J. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–73.

Johnson, N. L. and Kotz, S. (1977). *Urn models and their application: an approach to modern discrete probability theory*. Wiley, New York.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481.

King, W. D., Dodds, L., and Allen, A. C. (2000). Relation between stillbirth and specific chlorination by-products in public water supplies. *Environ Health Perspect*, 108(9):883–6.

Kirrane, E., Hoppin, J., Kamel, F., Umbach, D., Boyes, W., DeRoos, A., Alavanja, M., and Sandler, D. (2005). Retinal degeneration and other eye disorders in wives of farmer pesticide applicators enrolled in the agricultrual health study. *Am J Epidemiol*, 161(11):1020–9.

Kleinman, K. and Ibrahim, J. (1998). A semi-parametric bayesian approach to genearlized linear mixed models. *Statistics in Medicine*, 17:2579–2596.

Kline, J., Levin, B., Silverman, J., Kinney, A., Stein, Z., Susser, M., and Warburton, D. (1991). Caffeine and spontaneous abortion of known karyotype. *Epidemiology*, 2(6):409–17.

Klinefelter, G. R. and Linder, R. E. (1996). Recent reproductive effects associated with disinfection by-products. In *Disinfection By-Products in Drinking Water: Critical Issues in Health Effects Research-Workshop Report*. International Life Sciences Institute Press, Washington, DC.

Lartillot, N. and Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109.

Leamer, E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley, New York.

Lindbohm, M. L., Hemminki, K., Bonhomme, M. G., Anttila, A., Rantala, K., Heikkila, P., and Rosenberg, M. J. (1991a). Effects of paternal occupational exposure on spontaneous abortions. *Am J Public Health*, 81(8):1029–33.

Lindbohm, M. L., Sallmen, M., Anttila, A., Taskinen, H., and Hemminki, K. (1991b). Paternal occupational lead exposure and spontaneous abortion. *Scand J Work Environ Health*, 17(2):95–103.

Lindley, D. and Phillips, L. (1976). Inference for a bernoulli process (a bayesian view). *American Statistician*, 30:112–9.

Lindley, D. and Smith, A. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc (Ser. B)*, 34:1–41.

MacEachern, S. (1994). Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics - Simulation and Computation*, 23:727–741.

MacLehose, R., Dunson, D., and Herring, A. (2005). Bayesian methods for highly correlated exposure data. *submitted to Epidemiology*.

Mathworks Development, I. (2005). *Matlab v. 7.0.4*. Natick, Massachusetts.

McAuliffe, J., Blei, D., and Jordan, M. (2005). Nonparametric empirical bayes for the dirichlet process mixture model. *Statistics and Computing, to appear*.

Mills, J. L., Holmes, L. B., Aarons, J. H., Simpson, J. L., Brown, Z. A., Jovanovic-Peterson, L. G., Conley, M. R., Graubard, B. I., Knopp, R. H., and Metzger, B. E. (1993). Moderate caffeine use and the risk of spontaneous abortion and intrauterine growth retardation. *Jama*, 269(5):593–7.

Mughal, F. (1992). Chlorination of drinking water and cancer: a review. *J Environ Pathol Toxicol Oncol*, 11(5-6):287–92.

Mukhopadhyay, S. and Gelfand, A. (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association*, 92(438):633–639.

Muller, P. and Quintana, F. (2004). Nonparametric bayesian data analysis. *Statistical Science*, 19(95-110).

Murray, F. J., Schwetz, B. A., McBride, J. G., and Staples, R. E. (1979). Toxicity of inhaled chloroform in pregnant mice and their offspring. *Toxicol Appl Pharmacol*, 50(3):515–22.

Narotsky, M. G., Pegram, R. A., and Kavlock, R. J. (1997). Effect of dosing vehicle on the developmental toxicity of bromodichloromethane and carbon tetrachloride in rats. *Fundam Appl Toxicol*, 40(1):30–6.

Ness, R., Grisso, J., Hirschinger, N., Markovic, N., Shaw, L., Day, N., and Kline, J. (1999). Cocaine and tobacco use and the risk of spontaneous abortion. *N Engl J Med*, 340(5):333–9.

Neutra, R., Swan, S., Hertz-Picciotto, I., Windham, G., Shaw, G., Fenster, L., and Deane, M. (1992). Potential sources of bias and confounding in environmental epidemiologic studies of pregnancy outcomes. *Epidemiology*, 3(2):134–42.

Newton, M., Czado, C., and Chappell, R. (1996). Bayesian inference for semiparametric binary regression. *Journal of the American Statistical Association*, 91(433):142–153.

Newton, M., Kendziorski, C., Blattner, F., and Tsui, K. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8:37–52.

Nieuwenhuijsen, M., Toledano, M., Eaton, N., Fawell, J., and Elliott, P. (2000). Chlorination disinfection byproducts in water and their association with adverse reproductive outcomes: a review. *Occup Environ Med*, 57(2):73–85.

Oakes, D. (1988). Semi-parametric models. In Kotz, S. and Johnson, N., editors, *Encyclopedia of Statistics*, volume 8, pages 367–369. Wiley, New York.

O'Brien, S. and Dunson, D. (2004). Bayesian multivariate logistic regression. *Biometrics*, 60(3):739–46.

Osborn, J. F., Cattaruzza, M. S., and Spinelli, A. (2000). Risk of spontaneous abortion in italy, 1978-1995, and the effect of maternal age, gravidity, marital status, and education. *Am J Epidemiol*, 151(1):98–105.

Palmer, A. K., Street, A. E., Roe, F. J., Worden, A. N., and Van Abbe, N. J. (1979). Safety evaluation of toothpaste containing chloroform. ii. long term studies in rats. *J Environ Pathol Toxicol*, 2(3):821–33.

Parazzini, F., Bocciolone, L., Fedele, L., Negri, E., La Vecchia, C., and Acaia, B. (1991). Risk factors for spontaneous abortion. *Int J Epidemiol*, 20(1):157–61.

Parazzini, F., Chatenoud, L., Di Cintio, E., Mezzopane, R., Surace, M., Zanconato, G., Fedele, L., and Benzi, G. (1998). Coffee consumption and risk of hospitalized miscarriage before 12 weeks of gestation. *Hum Reprod*, 13(8):2286–91.

Petitti, D. (1992). Opening pandora's box. *Epidemiology*, 3(2):78–81.

Promislow, J., Makarushka, C., Gorman, J., Howards, P., Savitz, D., and Hartmann, K. (2004). Recruitment for a community-based study of early pregnancy: the right from the start study. *Paediatr Perinat Epidemiol*, 18(2):143–52.

Raftery, A. (1996). Approximate bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83(2):251–66.

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

Rook, J. (1974). Formation of haloforms during chlorination of natural water, water treatment and examination. *J Am Water Works Assoc*, 23:234–243.

Ruddick, J. A., Villeneuve, D. C., Chu, I., and Valli, V. E. (1983). A teratological assessment of four trihalomethanes in the rat. *J Environ Sci Health B*, 18(3):333–49.

Savage, J. (1954). *The Foundations of Statistics*. Wiley.

Savitz, D., Andrews, K., and Pastore, L. (1995). Drinking water and pregnancy outcome in central north carolina: source, amount, and trihalomethane levels. *Environ Health Perspect*, 103(6):592–6.

Savitz, D., Singer, P., Herring, A., Hartmann, K., Weinberg, H., and Makarushka, C. (2005). Drinking water disinfection by-product exposure and pregnancy loss. *American Journal of Epidemiology(in submission)*.

Savitz, D., Sonnenfeld, N., and Olshan, A. (1994). Review of epidemiologic studies of paternal occupational exposure and spontaneous abortion. *Am J Ind Med*, 25(3):361–83.

Schwetz, B. A., Leong, B. K., and Gehring, P. J. (1974). Embryo- and fetotoxicity of inhaled chloroform in rats. *Toxicol Appl Pharmacol*, 28(3):442–51.

Sethuraman, J. (1994). A constructive definition of the dirichlet process prior. *Statistica Sinica*, 2:639–650.

Signorello, L. B. and McLaughlin, J. K. (2004). Maternal caffeine consumption and spontaneous abortion: a review of the epidemiologic evidence. *Epidemiology*, 15(2):229–39.

Smith, M. K., Randall, J. L., Read, E. J., and Stober, J. A. (1992). Developmental toxicity of dichloroacetate in the rat. *Teratology*, 46(3):217–23.

Smith, M. K., Randall, J. L., Tocco, D. R., York, R. G., Stober, J. A., and Read, E. J. (1988). Teratogenic effects of trichloroacetonitrile in the long-evans rat. *Teratology*, 38(2):113–20.

Spiegelhalter, D., Thomas, A., and NG, B. (1999). Winbugs version 1.2 user manual. Technical report, MRC Biostatistics Unit.

Srisuphan, W. and Bracken, M. B. (1986). Caffeine consumption during pregnancy and association with late spontaneous abortion. *Am J Obstet Gynecol*, 154(1):14–20.

Steenland, K., Bray, I., Greenland, S., and Boffetta, P. (2000). Empirical bayes adjustments for multiple results in hypothesis-generating or surveillance studies. *Cancer Epidemiolog, Biomarkers, and Prevention*, 9:895–903.

Strawderman, W. (1978). Minimax adaptive generalized ridge regression estimators. *JASA*, 73(363):623–7.

Swan, S., Neutra, R., Wrensch, M., Hertz-Picciotto, I., Windham, G., Fenster, L., Epstein, D., and Deane, M. (1992). Is drinking water related to spontaneous abortion?

reviewing the evidence from the california department of health services studies. *Epidemiology*, 3(2):83–93.

Swan, S., Waller, K., Hopkins, B., Windham, G., Fenster, L., Schaefer, C., and Neutra, R. (1998). A prospective study of spontaneous abortion: relation to amount and source of drinking water consumed in early pregnancy. *Epidemiology*, 9(2):126–33.

Taskinen, H., Anttila, A., Lindbohm, M. L., Sallmen, M., and Hemminki, K. (1989). Spontaneous abortions and congenital malformations among the wives of men occupationally exposed to organic solvents. *Scand J Work Environ Health*, 15(5):345–52.

Thompson, D. J., Warner, S. D., and Robinson, V. B. (1974). Teratology studies on orally administered chloroform in the rat and rabbit. *Toxicol Appl Pharmacol*, 29(3):348–57.

Venkobachar, C., Iyengar, L., and Prabhakara Rao, A. (1975). Mechanisms of disinfection. *Water Research*, 9(1):119–124.

Villanueva, C. M., Cantor, K. P., Cordier, S., Jaakkola, J. J., King, W. D., Lynch, C. F., Porru, S., and Kogevinas, M. (2004). Disinfection byproducts and bladder cancer: a pooled analysis. *Epidemiology*, 15(3):357–67.

Waller, K., Swan, S., DeLorenze, G., and Hopkins, B. (1998). Trihalomethanes in drinking water and spontaneous abortion. *Epidemiology*, 9(2):134–40.

Wen, W., Shu, X. O., Jacobs, D. R., J., and Brown, J. E. (2001). The associations of maternal caffeine consumption and nausea with spontaneous abortion. *Epidemiology*, 12(1):38–42.

West, M., Mueller, P., and Escobar, M. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In Smith, A. and Freeman, P., editors, *Aspects of Uncertainty: A Tribute to DV Lindley*. Wiley.

White, G. C. (1999). *The handbook of chlorination and alternative disinfectants*. J. Wiley, New York, 4th edition.

Wilcox, A., Weinberg, C., O'Connor, J., Baird, D., Schlatterer, J., Canfield, R., Armstrong, E., and Nisula, B. (1988). Incidence of early loss of pregnancy. *N Engl J Med*, 319(4):189–94.

Windham, G., Swan, S., Fenster, L., and Neutra, R. (1992). Tap or bottled water consumption and spontaneous abortion: a 1986 case-control study in california. *Epidemiology*, 3(2):113–9.

Witte, J., Greenland, S., Haile, R., and Bird, C. (1994). Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. *Epidemiology*, 5(6):612–21.

Witte, J., Greenland, S., and Kim, L. (1998). Software for hierarchical modeling of epidemiologic data. *Epidemiology*, 9(5):563–6.

Wrensch, M., Swan, S., Lipscomb, J., Epstein, D., Fenster, L., Claxton, K., Murphy, P., Shusterman, D., and Neutra, R. (1990). Pregnancy outcomes in women potentially exposed to solvent-contaminated drinking water in san jose, california. *Am J Epidemiol*, 131(2):283–300.

Wrensch, M., Swan, S., Lipscomb, J., Epstein, D., Neutra, R., and Fenster, L. (1992). Spontaneous abortions and birth defects related to tap and bottled water use, san jose, california, 1980-1985. *Epidemiology*, 3(2):98–103.

Zierler, S. (1992). Drinking water and reproductive health. *Epidemiology*, 3(2):77–8.