Bayesian modeling and inference for asymmetric responses with applications

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Xiaoyue Zhao

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dipankar Bandyopadhyay and Lynn Eberly

July, 2017

Acknowledgements

There are many people that have earned my gratitude for their contributions to my time in graduate school at the University of Minnesota. Firstly, I would like to express my sincere gratitude to my advisor Prof. Dipankar Bandyopadhyay for his advising of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His constant guidance kept me on track throughout the entire process, and eventually submitting this thesis. Also, I would like to thank my thesis committee members: Prof. Susan Arnold, Prof. Lynn Eberly, and Prof. Lin Zhang for their insightful comments and encouragements which inspired me to widen my research from various perspectives. My thanks also go to Prof. Sudipto Banerjee and Prof. Gurumurthy Ramachandran, who provided me an opportunity to work on their research project, and enlightened me with the first principles of Bayesian statistics. I thank my fellow classmates in the Division of Biostatistics for their companionship and for all the fun we had studying together for homework and exams during the past five years. Finally, I want to thank my parents for their unfailing support and continuous encouragements throughout the years of my education.

Abstract

Analysis of asymmetric data poses several unique challenges. In this thesis, we propose a series of parametric models under the Bayesian hierarchical framework to account for asymmetry (arising from non-Gaussianity, tail behavior, etc) in both continuous and discrete response data. First, we model continuous asymmetric responses assuming normal random errors by using a dynamic linear model discretized from a differential equation which absorbs the asymmetry from the data generation mechanism. We then extend the skew-normal/independent parametric family to accommodate spatial clustering and non-random missingness observed in asymmetric continuous responses, and demonstrate its utility in obtaining precise parameter estimates and prediction in presence of skewness and thick-tails. Finally, under a latent variable formulation, we use a generalized extreme value (GEV) link to model multivariate asymmetric spatially-correlated binary responses that also exhibit non-random missingness, and show how this proposal improves inference over other popular alternative link functions in terms of bias and prediction. We assess our proposed method via simulation studies and two real data analyses on public health. Using simulated data, we investigate the performance of the proposed method to accurately accommodate asymmetry along with other data features such as spatial dependency and non-random missingness simultaneously, leading to precise posterior parameter estimates. Regarding data illustrations, we first validate the efficiency in using differential equations to handle skewed exposure assessment responses derived from an occupational hygiene study. Furthermore, we also conduct efficient risk evaluation of various covariates on periodontal disease responses from a dataset on oral epidemiology. The results from our investigation re-establishes the significance of moving away from the normality assumption and instead consider pragmatic distributional assumptions on the random model terms for efficient Bayesian parameter estimation under a unified framework with a variety of data complexities not earlier considered in the two aforementioned areas of public health research.

Contents

List of Tables

List of Figures

Chapter 1

Introduction: works inspired by asymmetric data

Florence Nightingale $(1820-1910)$ once said: "Statistics . . . the most important science in the whole world: for upon it depends the practical application of every other science and of every art; the one science essential to all political and social administration, all education, all organizations based upon experience for it only gives the results of our experience." She promoted to improve health care by using the art and science of statistics and to help people learn from the data (Agresti and Franklin, 2007). As a pioneer in visual presentation of information and statistical graphics, her work has inspired generations of people to advocate and apply statistics to improve people's lives in public health. She was not alone on the way to promote applied statistics in heath science. The use of models in health science has a long history. Researchers and scientists from a wide variety of disciplines are enthusiastic to develop complex and interpretable stochastic models to derive inference from data generated in real life.

The most widely used, also the simplest model, is the general linear model, which forms a linear relationship between continuous response and covariates while also accounts for normally distributed random errors. This linear model is easy to use and the coefficient estimates give straightforward interpretations. Whereas there are several assumptions under the general linear model framework, the normality assumption is non-ignorable. When this assumption is not met, we can use the generalized linear model (GLM). Nelder and Wedderburn (1972) formalized the GLM and it allows the response to follow other parametric distributions (instead of only the Normal distribution as in the general linear model) by choosing different link functions from the exponential family accordingly. For example, when response follows a Binomial distribution, we can use a logit link and do a logistic regression and it works for both continuous and discrete responses. The general linear model is a special case of the GLM where an identical link is used. The development from the general linear model to the GLM is a breakthrough with advantages. The GLM allows to relax the normality assumption on responses. The link function in the GLM can be different from the distribution assumption for random components so that there's more flexibility in modeling. Also in the GLM, estimates can be obtained by using either the maximum likelihood or drawing inference from posterior distributions under the Bayesian framework and we can choose the method accordingly. For example, Zeger and Karim (1991) showed an example using the GLM to model clustered Gaussian responses, where they assumed an independent Gaussian random error with mean 0 and variance σ_ϵ^2. This model was casted under a Bayesian framework using the Gibbs sampling to draw inference from posterior samples.

In scientific research, we do not always get symmetric data. The study of asymmetric data is not rare in research literatures across different disciplines. For examples, researchers in marketing have been using the multidimensional scaling (MDS) but there are no suitable models for analyzing asymmetric data relationships (Harshman et al., 1982). Harshman et al. (1982) developed a linear model to rewrite the original asymmetric matrix into a weight matrix, an asymmetric matrix that specifies the directional relationships, and a matrix of error terms. This method has a better representation of asymmetric data and the transformation matrices give useful marketing implications (Harshman et al., 1982). A common approach adopted for data analysis in these situations is reverting to the original multivariate normality assumptions after applying suitable data transformations of responses. Osborne (2010) reviewed the Box-Cox transformation (Box and Cox, 1964) which has been a great improvement on traditional approaches (such as square root transformation and log transformation) for normalization, where the Box-Cox transformation gave the most symmetrical results compared to other methods. The application of data transformation can be found in numerous areas. For example, in the study of environmental properties, the distributions of elements found in soil and rocks are often asymmetric. Kerry and Oliver (2007) uses data transformation to reduce asymmetry, and the paper shows how different transformation methods can be suitable to various situations. Although data transformation can lead to reasonable empirical results, they may be avoided when a suitable alternative theoretical model is available because transformation often hinders the underlying data generation mechanisms due to reduced information, and component-wise transformation (such as the multivariate cases) does not always lead to joint normality (Jara et al., 2008). In addition, interpretations on transformed and original planes may not be similar. Besides, transformations are often not universal, i.e., a transformation used for one dataset might not work for another. Furthermore, standard log-transformation might be infeasible (only works for non-negative values).

A remedy of data transformation is to use more flexible parametric families to model asymmetric data. A considerable amount of research has been done by introducing more flexible parametric families that can accommodate normality departures (skewness and kurtosis), and hence eliminate the need of data transformations. When data exhibit non-normal behavior, the fidelity of the Gaussian assumption has always been questioned (Verbeke and Lesaffre, 1996; Ghidey et al., 2004; Lin, 2010). In the context of linear mixed models (LMMs), the random effects distribution was relaxed using finite normal mixtures (Verbeke and Lesaffre, 1996), smoothing (Ghidey et al., 2004), a semi-nonparametric density (Zhang and Davidian, 2001). Much of recent frequentist and Bayesian advances in regression problems revolve around the attractive and popular skew-normal (elliptical) distributions (Azzalini and Capitanio, 2003; Sahu et al., 2003). Related literature in this context is very rich (Lin, 2010; Azzalini, 2005; Arellano-Valle et al., 2006), and the entire monograph of Genton (2004) is dedicated to discuss recent developments.

Besides data transformation and the use of flexible parametric families, we can also handle asymmetry using specified link functions under the GLM framework. When a link is misspecified, it can lead to bias and inefficiency of the estimates of covariate coefficients. Thus, it is of significant importance to use the correct link. There are substantial works done in studying different link functions. Chen et al. (1999) introduced a latent variable approach using a class of asymmetric link models for binary response data. Kim (2008) introduced a link based on the flexible generalized t-distribution, which controls the tails and the scale of the link. The generalized extreme value (GEV) distribution is also a parametric approach to model skew data, which has been widely used in a variety of different discipline, such as risk management on financial markets (Coles et al., 2001; Diebold et al., 2000). Wang and Dey (2010) presents a flexible skewed link function for modeling binary response data based on the generalized extreme value (GEV) distribution. Wang and Dey (2011) also uses the GEV link to model ordinal response data in ecological research.

In this dissertation, I try to throw some light on the adequacy of normality assumption on random errors / measurement uncertainties and the consequences on parameter estimation. The research goal is to develop and extend current parametric models via Bayesian approach to handle skewed datasets with both continuous and binary responses in applied public health research which are inadequately modeled with Gaussian assumption. Our data-driven models are built on the unique features of the individual dataset. The first dataset is about contaminant exposure assessment in evaluating occupational hygiene, and the second one is about periodontal disease (PD) progression in the context of dental epidemiology.

1.1 An overview of the motivating datasets that inspire this dissertation

I applied our models to datasets from two fields in public health: occupational hygiene and dental epidemiology. The dental epidemiology data have been successfully analyzed prior to our research (Reich and Bandyopadhyay, 2010; Reich et al., 2013; Bandyopadhyay et al., 2009, 2010a,a, 2012; Bandyopadhyay and Canale, 2016), but we emphasis on modeling the skewness (Chapter 3 and Chapter 4) with different data setups and providing alternative perspectives to modeling and analyzing this dataset.

1.1.1 Field experiment data in contaminant exposure assessment for model evaluation in occupational hygiene

The data in section 2.4.2 and section 2.4.3 are collected in an experimental study (Arnold et al., 2017) when evaluating the well-mixed room model and the two-zone model systematically to understand contaminant exposure in working environment. Because the research goal in occupational hygiene is to protect workers' health and well-being, it's crucial to use appropriate exposure models which can contribute to accurate decision making. Arnold et al. (2017) pointed out that the accuracy of using subjective opinions such as personal exposure data for estimation is no better than random chance. Therefore, objective inputs such as systematic model evaluation are needed. Based on the widespread industrial applications and range of physical chemical properties, we choose to use toluene data in our analysis. These experiments are conducted under highly controlled conditions in exposure chambers, such as well-mixed chamber and near-field far-field (two-zone) chamber.

In the well-mixed room model, concentrations were measured at six different locations around the source to obtain spatial profiles of toluene concentrations. The source was located at the center of the chamber. Measurements were taken until the concentration level in the chamber has reached a steady state. Three levels of ventilation rates were used in the experiment (details in section 2.4). Since the air in the chamber is completely well-mixed, measurements at 6 different locations were almost the same. Hence, without any spatial association we only use measurements at 1 location for analysis. In the two-zone model experiment, concentrations of toluene were measured at 4 different locations in the chamber, one at near field and other three at far field. Concentrations were measured every 90 seconds for 120 minutes until they reach a steady state. For both well-mixed chamber and two-zone chamber experiments, measurements of the concentration level were recorded and transferred into a spreadsheet. Primary analysis of the contaminant concentrate level of toluene shows that the continuous response data exhibit skewness and possibly some non-Gaussian tail behavior, which inspired us to use a flexible parametric linear model under the Bayesian framework when analyzing the data.

1.1.2 GAAD dataset

The second application to handle asymmetric data is in dental epidemiology, especially for periodontal disease (PD). PD is not uncommon in the United States. Oliver et al. (1998) pointed out that about 50% of U.S. adults over the age of 35 experience early stages of periodontal disease. If left untreated, it can progress to moderate or severe PD, which leads to tooth shifting, eventual loss, and other serious health problems. For instance, if inflammatory proteins and bacteria in the periodontal tissue enter the blood stream, it can cause coronary disease and stroke (Beck et al., 1996). Besides effects on the cardiovascular system, there are studies show a positive association between glycemic control of type 2 diabetes and severe periodontal disease (Tsai et al., 2002; Bandyopadhyay et al., 2010b; Teeuw et al., 2010). In Chapter 3 and Chapter 4, we analyze data collected from a clinical study in dental epidemiology to evaluate the association between glycemic control and periodontitis progression among the Gullah-speaking African Americans (GAAD) with type-2 diabetes mellitus (T2DM). Studies show that the African Americans' community has a higher risk of diabetes compared to other races in the United States, where possible contributors are genetic traits, prevalence of obesity, and insulin resistance(Marshall, 2005). Within the African American community, the Gullah African Americans are the most genetically homogeneous population of the African descent in the United States (Bandyopadhyay et al., 2010b), which is a suitable and informative population to study the association between PD and type-2 diabetes.

The most important biomarker to assess periodontal disease is the clinical attachment level (CAL). In the GAAD data, CAL was measured for each of the 6 sites of a tooth, nested within a subject, including various subject-level covariates such as age, gender, body mass index (indicating obesity status), glycemic level (indicating diabetic status), and tooth-site level covariates such as site in upper/lower jaw, site in tooth type, etc. With this multivariate response vector, the underlying statistical question was to investigate and estimate the functions determining the covariate-response relationships. However, note that this is complicated due to several reasons. First, the dataset exhibits a large volume of missing responses. Second, PD progression is also considered to be spatially-clustered, i.e., diseased status for a set of closely located tooth-sites are similar. Furthermore, primary analysis of CAL data exhibits skewness and (possible) thick tails (Figure 3.1). Therefore, we develop models in Chapter 3 and Chapter 4 that expand the estimation framework to accommodate all these difficulties, and produce robust parameter estimates.

1.2 Dissertation overview

A common feature from the 2 datasets in section 1.1 is asymmetric behavior, mostly skewness and possible non-Gaussian tails in the response. However, we cannot apply suitable data transformations of the response to revert to the original multivariate normality assumptions. In our case, standard log-transformation is not feasible for either of the two datasets. For the occupational hygiene data, any forms of data transformation for the responses may disrupt the underlying differential equations (DE) framework, which is the cornerstone of the stochastic model characterizing the complex dynamic system of industrial hygiene. As for the GAAD data, CAL can have zero values. Therefore, we need other approaches to accommodate asymmetry.

In this dissertation, I first develop a likelihood-based dynamic linear model using a discrete version of the underlying DEs to estimate parameters in two chamber experiments that mimic the real-life working environment for occupational hygiene. We compare model fit under Gaussian assumption to other available parametric non-Gaussian propositions. Quite interestingly, I find that although the responses look skewed, a Gaussian fit is comparable to other competing non-Gaussian models (such as a skew-normal, or a skew-t density for the random terms). However, this is not the same case for CAL in the GAAD data where I deal with spatially-referenced responses [note that proximally located sites inside the mouth usually have similar CAL values]. In addition, substantial missing responses are observed. Given that PD is the major cause of tooth loss, this missingness is often attributed to be non-random. To mitigate this, in Chapter 3 we develop a new family of parametric statistical model that incorporates skewness, kurtosis and spatial clustering within a unified framework. With simulation studies and data analyses, we show that our model produces a better fit and improves parameter estimation over other Gaussian-based models. We can also model the extreme responses of CAL, i.e., when the patient has moderate to severe PD. This setup gives us a skewed multivariate binary response because the number of observations with moderate or severe PD is much smaller compare to the number of observations without that condition. To accommodate this, in Chapter 4 we applied the GEV link using a multivariate latent random variable jointly modeled with spatial-clustering and non-random missingness. Through simulation studies and analysis of the GAAD data, we show that the GEV link model with informative missingness has the best fit over other symmetric links.

Given the complexities involved in these datasets, a classical inferential framework mostly relying on maximum-likelihood estimation techniques, although feasible, might appear to be daunting. The associated asymptotic normality results, which are quintessential to any classical advancements, are not straightforward under complex dynamical models, spatial referencing, non-random missingness, and other data features tackled in my dissertation. One difficulty in handling these datasets is the computation intensive analyses. Historically speaking, after computers became more available in the 1950's, there were great advances in computation techniques which made complex models and large datasets applicable. For example, with the help of large electronic computers, Elveback and Varma (1965) showed how computers can help with stochastic models to obtain information. They use simulated mechanical model to illustrate stochastic characters of the epidemic model for closed and randomly mixing populations (Abbey, 1952). Hence, I decisively consider the hierarchical Bayesian formulation, with the ability to incorporate expert background (prior) information about the unknown parameters, and relying on the relevant Markov chain Monte Carlo (MCMC) steps for parameter estimation. For analyses in Chapter 2 and Chapter 3, prediction is almost automatic relying on posterior predictive distributions (Carlin and Louis, 2008). However, the generalized extreme value link model in Chapter 4 requires a bit more work to get posterior predictive samples of the multivariate latent variable because of its spatially-clustered nature and the closed form full-conditional is unavailable. I approach this problem by applying Hamiltonian Monte Carlo within Gibbs sampling, which makes the GEV link model accessible to spatially-clustered data. In addition, computational codes were developed using a combination of `R` packages, such as `R2WinBUGS` (in Chapter 3) and `rjags` (in Chapter 2), that connects the popular freeware `R` to other freely downloadable softwares such as `WinBUGS` and `JAGS`. Hence, the methods are appealing to any practitioner with some basic knowledge about the software.

The rest of this thesis proceeds as follows. Chapter 2 talks about the Bayesian dynamic modeling and inference for the industrial hygiene dataset. Chapter 3 and Chapter 4 both accommodate various aforementioned features exhibited by the clinical study on PD, while Chapter 3 applies flexible parametric models and Chapter 4 develops joint modeling using the GEV link. Finally, Chapter 5 is the discussion and conclusion followed by Reference and Appendix.

Chapter 2

Dynamic Bayesian physical models for occupational exposure assessment

2.1 Introduction

A primary concern in occupational hygiene (OH) is the estimation of a worker's exposure to chemical agents. Prediction of exposure through mathematical modeling is gaining popularity, especially with the advent of the European Community Regulation (REACH) that requires assessing exposures in a variety of exposure scenarios where monitoring may not be feasible (Ramachandran, 2008). This is usually achieved by modeling the physical processes generating chemical concentrations in the workplace. An accurate representation will deliver better concentration estimates and facilitate subsequent decision-making in exposure management. Nicas (2002) points out that mathematical modeling can provide a more precise estimate than estimates derived using only a few data points collected through monitoring. However, the workplace is usually too complex for a physical model, or even a class of models. One novel approach been increasingly studied is a synergy of physical and statistical models to better estimate the underlying physical processes in the workplace (Hornung et al., 1994; Keil, 2000; Chen et al., 1999). In this sense, statistical modeling, combined with computational methods and software, can be an important tool in assessing exposure in the field of occupational hygiene.

Formal modeling in OH typically proceeds from a deterministic physical model for pollutant transportation using mass balance to predict the concentration of the contaminant, usually a function of time. For example, the so called well-mixed room (WMR) model (Ramachandran, 2008) assumes a single compartment with uniform pollutant concentration everywhere in the compartment. Here, the model has two main terms: a source (or generation) term and a dispersion term. The box has a continuous pollutant source which releases into a ventilated air space, where the ventilation rate is preassigned. The WMR model assumes a constant air flow rate through the well-mixed room (see Figure 2.1).

A second example is the two-zone model, also known as the near-field far-field model. It is first introduced by Hemeon(1963), which assumes two well-mixed compartments and a single source of emission. A compartment near the source of the emission is called the near-field while a larger chamber that encloses the first compartment but represents concentration farther away from the source is called the far-field.

Unfortunately, predicting exposure for well-mixed room model and two-zone model is challenging because (a) the physical assumptions are typically never completely met in practice, and (b) there is a lack of quantitative knowledge of exposure determinants for the statistical exposure models that can fit the scenario perfectly (Zhang et al., 2009). Since most companies do not collect such determinant information routinely, data on crucial determinants such as ventilation rates and pollutant generation rates are difficult to obtain (Kauppinen, 1994). There has been limited research in evaluating model parameters and assessing model performance. Uncertainty quantification is essential for estimation and prediction since it is inconceivable to experimentally account for all sources of variation in the data. For example, a physical parameter of primary interest in the two-zone model is the rate of air exchange between near-field and far-field. This parameter is affected by a number of factors such as the presence of a human body, body movement, and body temperature (Melikov, 1996; Nicas, 1999). It is, however, difficult and expensive to design experiments that can control for all these factors. Instead, statistical models that account for uncertainties can make this problem more tractable.

Recent work in melding statistical and physical models in OH include Zhang et al. (2009) and Monteiro et al. (2014). The former proposed a Bayesian approach regressing the mean concentration toward the nonlinear solution of differential equations representing the two-zone model. However, regressing on the solution of the physical model may prove ineffective due to biases and extraneous variation. Monteiro et al. (2014) attempted to enrich this work by synthesizing the physical model with experimental data using a stochastic process. While this enhances inferential performance with respect to predictive coverage, it considerably enhances the uncertainty in estimation of the model parameters and makes such inference less useful in practice. Neither of the aforementioned articles explored errors that violated Gaussian assumptions.

This chapter aims to contribute in the following manner. From a statistical modeling perspective, we propose a Bayesian dynamic linear model by discretizing the actual physical model. Rather than working with the non-linear solution to the differential equations, as was done by Zhang et al. (2009) and Monteiro et al. (2014), discretizing the differential equations produces a dynamic linear model. This avoids the exact nonlinear solutions, is easier to compute and, perhaps more pertinently for the practicing industrial hygienist, can be implemented easily in standard Bayesian software environments such as WinBUGS and rjags. Furthermore, we model the measurement errors on the same scale as that of the concentration and not in a logarithmic or other transformed scale. Transforming the outcome to model the errors would violate the underlying physical model, which we find undesirable. Our model comparison results in fact demonstrate that there's no substantive differences between assuming Gaussian errors and assuming more sophisticated skewed error distributions. This is a special case to model asymmetric data using dynamic linear model under Gaussian assumption, which is quite different from what we do in Chapter 3 and Chapter 4.

Another intended contribution concerns "validation". Validation of any inferential framework synthesizing physical and statistical models requires experimentation at two levels. First, we simulate computer experiments that generate synthetic data by adding statistical noise to the assumed true states of the well-mixed room and two-zone physical models. Validation proceeds by ascertaining if a statistical model is able to estimate the true state of the physical model (i.e., all model parameters) and also by assessing its predictive capabilities. At a second level, we generate concentration data from a real laboratory by carefully designing chamber experiments, where we attempt to meet the assumptions of the models and then monitor the change in exposure levels using both well-mixed room model and two-zone model (Arnold et al., 2017). These data are then analyzed using our proposed modeling framework and evaluated in terms of its inferential capabilities for the true process parameters.

The remainder of Chapter 2 is organized as follows. Section 2.2 offers details pertaining to the physical models we explore here. Section 2.3 develops the Bayesian dynamic linear model, different stochastic specifications for the measurement error and our approach to compare with competing statistical specifications. Section 2.5 presents simulation studies for the two-zone model. Implementation details for data simulating priors information and results are also presented in these sections. Section 2.6 summarizes our findings and discusses directions for future research in this area.

2.2 Physical models

2.2.1 Well-mixed room (WMR) model

The well-mixed room (WMR) model incorporates information about the air flow rate and the generation rate of contaminant in the workroom with volume Vm^3. We assume a balance in the volumetric rate m^3/min of air entering the workroom as Q_{in} and leaving the workroom as Q_{out}. We drop the subscripts and simply refer to Q as the supply or exhaust air flow rate of the workroom that might remove some of the contaminants from the workroom. In addition, there are some other factors that can reduce the contaminant. For example, the contaminant may get absorbed onto workroom surfaces (e.g., walls and floor coverings) or chemical reactions (Melikov, 1996; Nicas, 1999). We denote all these loss mechanisms as a loss rate factor K_L (1/min). So the amount of contaminant in the workroom can be reduced by various loss mechanisms and airflow rate. The loss rate coefficient, K_L, governs the mechanism by which the pollutant is removed from the room (other than ventilation). Examples of such mechanisms include adsorption of gases and vapors onto various surfaces (here K_L is an adsorption rate for the particular vapor and surface type) and particle deposition on surfaces by gravitational settling (K_L is now a function of terminal settling velocity for particles of a given diameter and density), impaction and Brownian diffusion. Thus, K_L helps generalize this model to gaseous as well as particulate air contaminants.

We are concerned with the concentration of a specific contaminant in the workroom in Figure 2.1. Air enters the room with a supply airflow rate Q and a contaminant concentration level G. The box is perfectly mixed which creates a uniform contaminant concentration throughout the room. The exhaust rate equals the supply airflow rate Q. To develop the dynamic model, we start with getting the differential equation. Using the principle of mass conservation, we can derives the first-order differential equation:

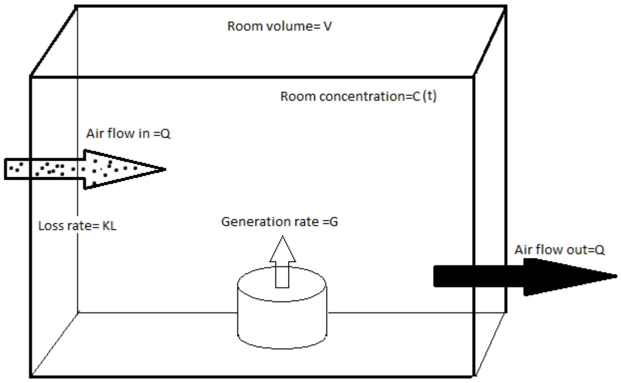


Figure 2.1: Dynamics of the one-zone model.

$$\frac{d}{dt}C(t) + \frac{Q + K_L V}{V}C(t) = \frac{G + C_{IN}Q}{V} \ , \tag{2.1}$$

where C_{IN} (e.g., in units of mg/m^3) is the concentration of the contaminant entering the room (at a flow rate of Q). Replacing the derivative in (2.1) with finite difference $C(t+1) - C(t)$ yields the discretized model (see detail in A.1),

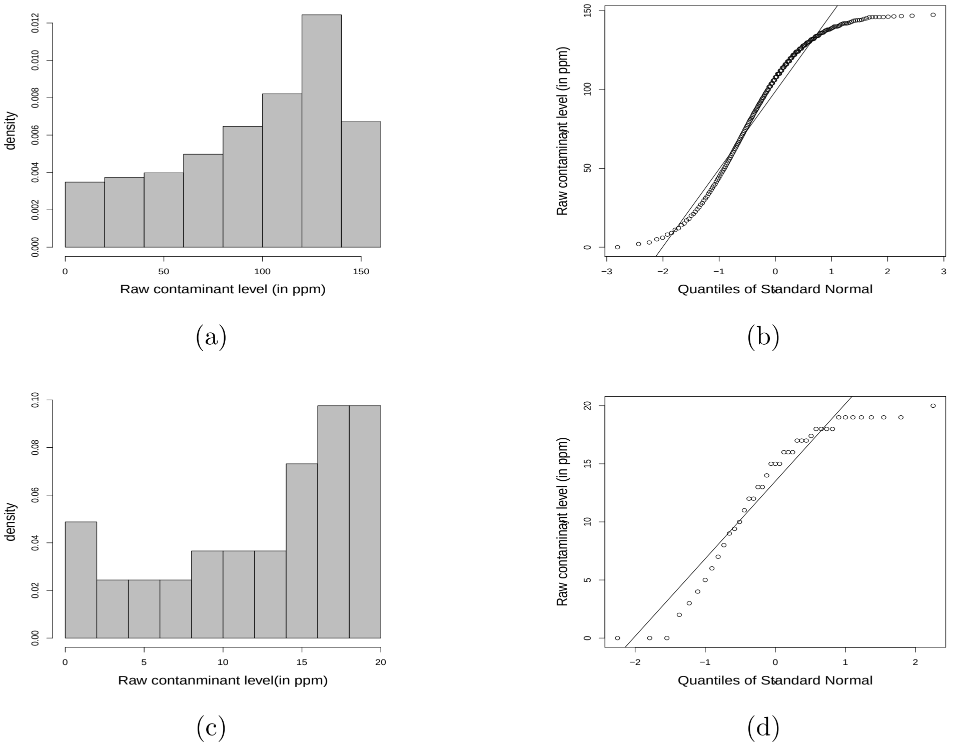
$$\begin{aligned} C(t+1) &= \left(1 - \frac{Q + K_L V}{V}\right) C(t) + \frac{G + C_{IN}Q}{V} \\ &= \left(1 - \frac{Q + K_L V}{V}\right)^{t+1} C(0) + \frac{G + C_{IN}Q}{V} \sum_{i=0}^{t} \left(1 - \frac{Q + K_L V}{V}\right)^{i} . \end{aligned} \tag{2.2}$$

We will extend (2.2) to a Bayesian dynamic linear model in section 2.3 and the resulting model will not explicitly require solving (2.1). Equilibrium of the physical state is obtained in the limit as $t \to \infty$ and given by $\lim_{t\to\infty} C(t) = (G + C_{\text{IN}}Q)/(Q + K_L V)$ in mg/m^3.

While C_{IN} may not be equal to $C(0)$ in general, in the experiment discussed here Arnold et al. (2017), we have $C_{\text{IN}} = C(0)$. Then the loss factor K_L is identifiable from Q only if the initial concentration $C(0)$ is not equal to 0. If we assume $C(0) = 0$, then the term $Q + K_L V$ is still identifiable but not the individual terms Q or K_L. These need to be identified from their prior distributions and will provide estimates sensitive to the specific choice of priors. Therefore, informative priors are crucial in obtaining solid estimates. In our subsequent simulations and experimental studies, we have explored with different choices for $C(0)$. Determined by the chamber condition at $t = 0$, $C(0)$ can be either greater than or less than the equilibrium and the asymptotic solution (2.2) is correct for both conditions. We assume $C(0) = 0$ as we are entering a clean chamber with ideal conditions. We explored the idea of a slightly contaminated chamber where $C(0)$ is about 5% to 10% of the equilibrium. We also use different sets of priors in the analysis. Further details are available in sections 2.4.2.

Figure 2.2 [panels (a) - (d)] shows the raw density histograms and the associated Q-Q plots for the raw contaminant concentration level data from the well-mixed workroom with low generation rate and high generation rate. Those plots present evidence of departures from the Normality assumptions (i.e., asymmetry). Therefore, the Gaussian distribution assumption does not seem to be congruous with the response data which exhibits strong left-skewness. One purpose of this paper is to show that when one converts the physical model into a dynamic framework, the results from assuming a Gaussian error would not be different from what you get by using some skewed distribution.

The underlying model assumes that the air in the workroom is completely well-mixed. Therefore, the concentration level at any point in the workroom is the same as at any other point in time. In other words, there is no spatial variability within the workroom. The contaminant is dispersed instantaneously throughout the volume of



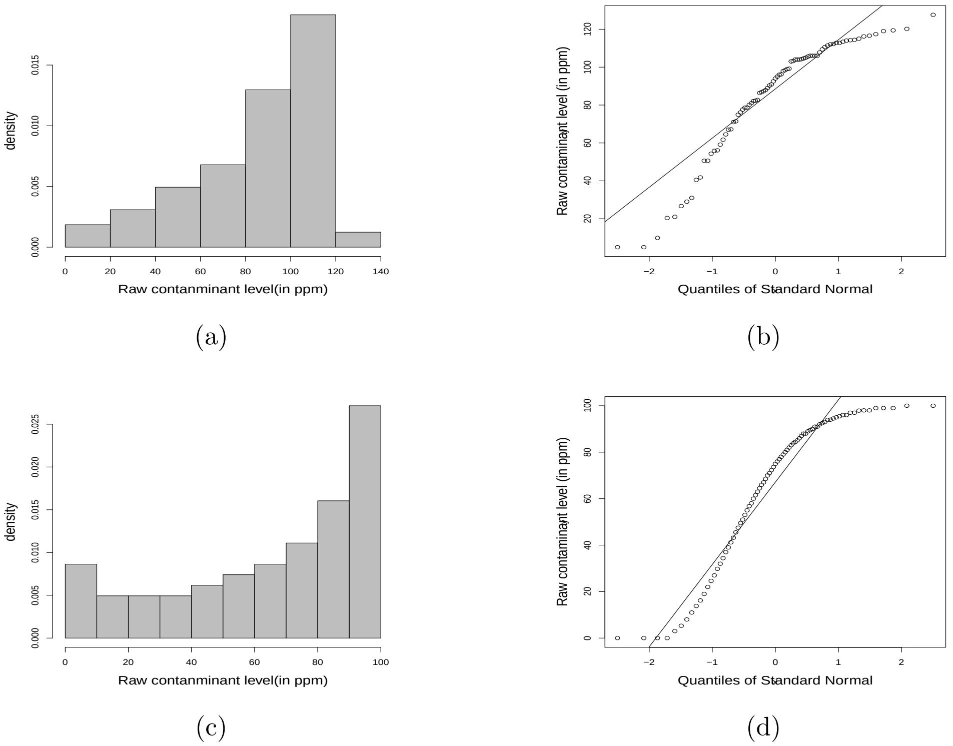
(a)

(b)

(c)

(d)

Figure 2.2: Well-mixed data: Plot of the density histogram of the raw contaminant concentration level for WMR model with high generation rate (panel a), WMR model with low generation rate (panel c), And the corresponding Q-Q plots are presented in panels (b) and (d).



(a) (b)

(c) (d)

Figure 2.3: Two-zone data: Plot of the density histogram of the raw contaminant concentration level for near field contaminant concentration level (panel a), and for the far field contaminant concentration level (panel c). The corresponding Q-Q plots are presented in panels (b) and (d).

the workroom, which leads to some simplification of the mathematical solution. This simplification comes at a cost: the model tends to underestimate exposures in situations where the workers are very close to the source, and the process is a continuous one once it has reached a quasi-steady state (i.e, there are no large variations in the process variables over time). A two-compartment (two-zone) model can compensate for some of these deficiencies, which we discuss this next.

2.2.2 Two-zone model

The two-zone model assumes the presence of a contamination source in the workplace and that the region is composed of two well-mixed fields. The near-field is the zone very near and around the source with volume V_N, while the far-field refers to the rest of the room and has volume V_F. The far field completely encloses the near field and we assume there is some air exchange between the two zones. The contaminant concentration for the near and far fields are given by $C_N(t)$ and $C_F(t)$, respectively. The supply and exhaust flow rates are the same and equal to Q (in units of m^3/min). The airflow rate between the far and near field is β (in units of m^3/min). Also, we assume that the contaminant's total mass is emitted at a constant rate G (in units of mg/min). We assume that the initial concentration level in both fields are zero, the supply airflow is free of contaminants and has rate Q and the only removal mechanism for the contaminant is ventilation. Figure 2.4 schematically represents the dynamics of the two-zone system (Zhou and Schmidler, 2009).

Based on the above assumptions, Monteiro et al. (2014) shows the rates of change in concentrations can be expressed as

$$\frac{d}{dt}\mathbf{C}(t) = \mathbf{W}(\boldsymbol{\theta}_1; \mathbf{x})\mathbf{C}(t) + \mathbf{g}(\boldsymbol{\theta}_1; \mathbf{x}) \ , \tag{2.3}$$

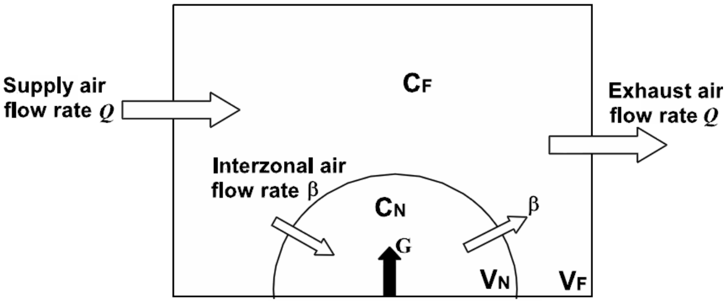


Figure 2.4: Dynamics of the two-zone model.

where $\mathbf{C}(t) = \begin{bmatrix} C_N(t) \\ C_F(t) \end{bmatrix}$, $\mathbf{W}(\boldsymbol{\theta}_1;\mathbf{x}) = \begin{bmatrix} -\frac{\beta}{V_N} & \frac{\beta}{V_N} \\ \frac{\beta}{V_F} & -\frac{(\beta+Q)}{V_F} \end{bmatrix}$, $\mathbf{g}(\boldsymbol{\theta}_1;\mathbf{x}) = \begin{bmatrix} \frac{G}{V_N} \\ 0 \end{bmatrix}$, $\boldsymbol{\theta}_1 = \{\beta, Q, G\}$ and $\mathbf{x} = \{V_N, V_F\}$. The functions $C_N(t)$ and $C_F(t)$ are the exposure concentrations at time t in the near and far fields, respectively.

Zhang et al. (2009) and Monteiro et al. (2014) provide explicit solutions to (2.3) in terms of matrix exponentials when the eigenvalues of $\mathbf{W}(\boldsymbol{\theta};\mathbf{x})$ are real and distinct. They regress the observed (bivariate) concentrations toward the nonlinear solution and estimate the parameters in $\boldsymbol{\theta}$. Here, we depart from this approach and construct a dynamic probability model based upon (2.3). This obviates dealing with the nonlinear solution and provides a more numerically stable template for inference. Replacing the derivative in (2.3) with the finite difference, we obtain

$$\mathbf{C}(t+1) = [\mathbf{I} + \mathbf{W}(\boldsymbol{\theta}_1;\mathbf{x})]\,\mathbf{C}(t) + \mathbf{g}(\boldsymbol{\theta}_1;\mathbf{x}) = \mathbf{G}(\boldsymbol{\theta}_1;\mathbf{x})\mathbf{C}(t) + \mathbf{g}(\boldsymbol{\theta}_1;\mathbf{x})\ , \tag{2.4}$$

where $\mathbf{G}(\boldsymbol{\theta}_1;\mathbf{x}) = \mathbf{I} + \mathbf{W}(\boldsymbol{\theta}_1;\mathbf{x})$. Section 2.3 extends section 2.4 into a full Bayesian model.

Figure 2.3 panels (a)-(d) show raw data density histograms and the associated Q-Q plots from the raw contaminant concentration level data collected from the two-zone model for both the near-field and far-field. Panel (a) and panel (c) presents histograms of the raw concentrations from the near and far fields, while panel (b) and panel (d) show the corresponding Q-Q plots. The two-zone experimental data also exhibit left-skewness and warrant investigations with non-Gaussian errors.

2.3 Bayesian dynamic linear model and estimation

We will describe how to build the dynamic Bayesian model using the two-zone model as an example. Equation (2.4), which is discretized from the two-zone differential equation that represents the actual physical model, and describes the relationship between $\mathbf{C}(t+1)$ and $\mathbf{C}(t)$ and they are estimated recursively in time. $\mathbf{C}(t)$ is the mean contaminant concentration level at time t which is not observed directly. As Doucet (2000) has mentioned, we can treat the mean of the contaminant concentration level as a Markovian hidden signal of interest $\{\mathbf{C}(t), t \in T\}$ and call it the hidden process. The observed contaminant concentration level $\{\mathbf{Y}(t), t \in T\}$ is implemented in the observation (measurement) equation. The relevant hidden information on $\{\mathbf{C}(t), t \in T\}$ at time T and the unknown parameters that we are interested in estimating are included in the posterior distribution $p(\{\mathbf{C}(t), t \in T\}, \boldsymbol{\theta} \mid \{\mathbf{Y}(t), t \in T\})$. Because of the Gaussian assumption we have for the observations, this linear Gaussian model is known as the Kalman filter process (Hosseini et al., 2011). Harrison (1976) defined a class of dynamic linear models, which is system equations that describe how the parameters get estimated and updated through time and how the observations are dependent on the current process. The model usually has the following components: (1) time index $(t = 1, 2, \ldots, T)$; (2) a vector of the observed data at time t, $\mathbf{Y}(t) = \begin{pmatrix} Y_N(t) \\ Y_F(t) \end{pmatrix}$; (3)

a vector of the hidden (unobserved) data at time t, $\mathbf{C}(t) = \begin{pmatrix} C_N(t) \\ C_F(t) \end{pmatrix}$; (4) a matrix of independent variables(predictors), here it's the identity matrix $\mathbf{I}$; (5) a system matrix $\mathbf{G}(\boldsymbol{\theta}; \mathbf{x})$; (6) random effects $\boldsymbol{\nu}_t$ and $\boldsymbol{\omega}_t$, in the Gaussian model we assume they are multivariate Normal with mean 0 and covariance Σ_ν and Σ_ω respectively.

We construct a Bayesian dynamic model with two components: (i) a measurement equation and (ii) a transition equation. The measurement equation, which is also known as the observation equation (Harrison, 1976), describes the relationship between the mean concentration level and the observations at any given time and for any given physical state. It specifies the stochastic process of the observation $\mathbf{Y}(t)$ on the current parameter $\{\boldsymbol{\theta}, \mathbf{C}(t)\}$ at time t. The distribution of $\mathbf{Y}(t)$ is completely defined by the parameters at the current process. The transition equation, which is also known as the system equation, describes how the mean concentration level is updated from the state at time t to the next state at time $t + 1$. The fixed system matrix $\mathbf{G}$, specifies the deterministic derivation of the unknown parameters from time t to time $t + 1$. Harrison (1976) has pointed out that dynamic linear model can be generalized by making matrix $\mathbf{G}$ indeterministic. In our analysis, we keep the matrix $\mathbf{G}$ fixed since the experiment settings don't vary with time.

We can use the dynamic linear model to do parameter estimation. Kalman (1963) showed that the dynamic linear model can be used in estimating parameters recursively as well as updating and revising parameters. In our case, the Markovian hidden value $\mathbf{C}(t-1)$ can be calculated from the most recent observation values $\mathbf{Y}(t)$, the posterior distribution $p(\mathbf{C}(t-1) \,|\, \mathbf{Y}(t))$, and the current random effects ν_t and ω_t.

2.3.1 Well-mixed room Bayesian Dynamic Model

Let $Y(t)$ be the observed concentration levels from a well-mixed chamber experiment and let $C(t)$ denote the mean concentration level at time t. The measurement equation assumes that $Y(t)$ is a noisy version of $C(t)$. The transition equation constructs the dynamic update of $C(t)$ based upon the physical model in (2.1) and, more specifically, in (2.2). Under normality assumption, this yields

$$\begin{aligned}&\text{Measurement equation: } Y(t) = C(t) + \nu_t \ , \quad \nu_t \overset{iid}{\sim} N(0, \sigma^2) \ ; \\ &\text{Transition equation: } C(t+1) = \left(1 - \frac{Q + K_L V}{V}\right) C(t) + \frac{G + C(0)Q}{V} + \omega_t \ , \\ &\qquad\qquad \omega_t \overset{iid}{\sim} N(0, \tau^2) \ .\end{aligned} \tag{2.5}$$

The collection of model parameters in (2.5) is $\boldsymbol{\theta} = \{K_L, G, Q, \sigma^2, \tau^2\}$, where the first three elements retain their interpretations from (2.1), while σ^2 and τ^2 are variance terms accounting for uncertainties in the measurements and in the physical model itself. $C(0)$ is fixed at either 0 or a small value which is about 5% to 10% of the equilibrium concentration level.

To construct a full Bayesian model from (2.5), we derive the likelihood function from the measurement equation and a distribution on the underlying state $C(t)$ from the transition equation. The specification is completed by assigning prior distributions to the parameters in $\boldsymbol{\theta}$. The corresponding posterior is proportional to

$$\begin{aligned}&IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times \text{Unif}(K_L \,|\, a_{K_L}, b_{K_L}) \times \text{Unif}(G \,|\, a_G, b_G) \times \\ &\text{Unif}(Q \,|\, a_Q, b_Q) \times \prod_{t=0}^{n-1} N\left(C(t+1) \,|\, \left(1 - \frac{Q + K_L V}{V}\right) C(t), \tau^2\right) \times \\ &\prod_{t=1}^{n} N(Y(t) \,|\, C(t), \sigma^2) \ ,\end{aligned} \tag{2.6}$$

where IG$(\cdot,\cdot)$ and Unif$(\cdot,\cdot)$ represent inverse-Gamma and Uniform densities and the measurements $Y(t)$ are assumed to have been taken at time points $t = 1, 2, \ldots, n$.

2.3.2 Two-zone Bayesian Dynamic Model

Turning to the two-zone model, the analogues of (2.5) are

$$\begin{aligned}&\text{Measurement equation: } \mathbf{Y}(t) = \mathbf{C}(t) + \boldsymbol{\nu}_t \ , \quad \boldsymbol{\nu}_t \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_\nu) \ ; \\ &\text{Transition equation: } \mathbf{C}(t+1) = \mathbf{G}(\boldsymbol{\theta}_1;\mathbf{x})\mathbf{C}(t) + \mathbf{g}(\boldsymbol{\theta}_1;\mathbf{x}) + \boldsymbol{\omega}_t \ , \quad \boldsymbol{\omega}_t \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_\omega) \ ,\end{aligned} \tag{2.7}$$

where $\mathbf{Y}(t) = \begin{bmatrix} Y_N(t) \\ Y_F(t) \end{bmatrix}$ is a 2×1 vector of the observed concentration levels in the near and far-fields from a two-zone well-mixed chamber experiment, and $\mathbf{C}(t)$, $\mathbf{G}(\boldsymbol{\theta}_1;\mathbf{x})$ and $g(\boldsymbol{\theta}_1;\mathbf{x})$ are as defined in (2.3). The corresponding posterior distribution is proportional to

$$\begin{aligned}&IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times \text{Unif}(K_L \,|\, a_{K_L}, b_{K_L}) \times \\ &\quad \text{Unif}(G \,|\, a_G, b_G) \times \text{Unif}(Q \,|\, a_Q, b_Q) \times \\ &\quad \prod_{t=0}^{n-1} N\left(\mathbf{C}(t+1) \,|\, \mathbf{G}(\boldsymbol{\theta}_1;\mathbf{x})\mathbf{C}(t) + \mathbf{g}(\boldsymbol{\theta}_1;\mathbf{x}), \boldsymbol{\Sigma}_\omega\right) \times \prod_{t=1}^{n} N(\mathbf{Y}(t) \,|\, \mathbf{C}(t), \boldsymbol{\Sigma}_\nu) \ .\end{aligned} \tag{2.8}$$

The posterior distributions in (2.6) and (2.8) are analytically intractable and are evaluated by drawing samples from the posterior distribution using Markov chain Monte Carlo (MCMC) algorithms, such as the Gibbs sampler and Metropolis random walk algorithms. MCMC yields samples from a Markov chain that requires an initial "burn-in" period to gradually converge to its stationary distribution, i.e., the posterior distribution. Convergence can be diagnosed by running multiple chains with different starting values and assess mixing using, e.g., the Gelman Rubin statistics (Gelman and Rubin, 1992). MCMC algorithms along with functions for assessing satisfactory mixing behavior are now automated in modeling languages such as `BUGS` and `rjags`, both of which have interfaces within the `R` statistical computing environment (`http://www.r-project.org/`).

2.4 Experimental Chamber Study

2.4.1 Chamber Study Design for WMR Model Evaluation

We focus upon an experiment conducted in a full size exposure chamber (2.0 m × 2.8 m × 2.1 m). A detailed description of the chamber construction is presented in Arnold et al. (2017). A factorial study design was used to evaluate the WMR model across a range of emission and ventilation rates. Three industrial solvents, acetone, 2-butanone, and toluene were selected due to their widespread industrial application and range of physical chemical properties. As shown in Table 2.1, emission rates were selected to accommodate instrument sensitivity, delivery mechanism capacity and time required to approach steady state concentrations. Three ventilation rates, 0.3, 1.3 and 3 ACH, representing ranges relevant to residential and industrial operations were used. Precise generation rates, G (mg/min), were achieved by releasing a solvent into the chamber using a Harvard Apparatus Pump, Series 11 Elite, (Harvard Apparatus, Holliston, MA) equipped with a Becton Dickenson 30 ml or 50 ml glass syringe (East Rutherford, NJ). Because of the relatively high vapor pressures, the solvents evaporated almost immediately upon delivery, emitting the solvent vapor at a known and consistent generation rate. Each set of conditions was evaluated 3 times. Thus, for each solvent, 27 studies = 3 emission rates × 3 ventilation rates × 3 repetitions were conducted.

Solvent	Molar Mass (g/mol)	Vaper Pressure (mm Hg) at 25 degree Celsius	Generation rate (mg/min)		
			Low	Medium	High
Acetone	58.08	200	39.5	79.1	116.7
2-Butanone	84.93	71	40.25	80.5	120.75
Toluene	92.14	28.4	43.8	86.5	129.5

Table 2.1: Chamber studies' solvents' physical chemical properties

Two Drger X-am 7000 Multi-Gas Monitors (MGM) equipped with Smart PID sensors were used to measure the solvent vapor concentrations in real-time. Each instrument was calibrated by the manufacturer's instructions using a standard calibration gas of Isobutylene (IBUT), 100 ppm. To ensure the most accurate results, additional calibration studies were conducted with each X-am 7000 MGM, quantifying the response factor to the specific solvent. For toluene, two standard calibration gases of 20 ppm and 200 ppm were used according to the MGM Technical Manual instructions. Since standard calibration gases for the other two solvents were not available, sorbent tubes were used to collect area time-integrated air samples and these were compared with the time weighted average MGM reported concentrations. These studies were conducted at three air exchange rates, 0.3, 1.3 and 3.0 ACH to generate a calibration curve. TWA samples were collected following NIOSH method 2500 using Anasorb 747 sorbent tubes (SKC model 226-81A, SKC, Inc. Pittsburgh, PA). Sample analysis was conducted by an AIHA Accredited laboratory following NIOSH method 2500. Response factor was determined by (2.9) and the reported and observed Response Factor (RF) are reported in Table 2.2.

$$\text{Response Factor} = \frac{\text{Desired concentration}}{\text{observed concentration}} \tag{2.9}$$

As we study the concentration of toluene, the source of toluene was located at the center of the chamber. Measurements were taken until the concentration level in the

Chemical	Response Factor(RF)	
	Reported RF	Measured RF
Toluene	0.7	0.7
2 Butanone	0.64	0.91
Acetone	1.15	1.5

Table 2.2: Reported and Measured Response Factors

chamber has reached a steady state. For the low ventilation rate ($Q = 0.067$ m^3/min) experiment set up, concentrations were measured every 90 seconds for 300 minutes in each location. For the medium ventilation rate ($Q = 0.244$ m^3/min) experiment set up, concentrations were measured every 90 seconds for 120 minutes in each location. As for the high ventilation rate ($Q = 0.563$ m^3/min), measurements were taken for 60 minutes in each location. Since the air in the chamber is completely well-mixed, the measurements at 6 different locations were almost the same. Hence, we will use measurements at one location to do the analysis. The chamber volume is 11.9 m^3 and the contaminant generation rate is 43.8 mg/min. The ventilation rate varies in three different levels: (I) low ventilation rate at 0.067 m^3/min (II) medium ventilation rate at 0.244 m^3/ min (III) high ventilation rate at 0.552 m^3/min.

2.4.2 Analysis of the Experimental Well-mixed room model Data

Model comparison between normal, skew-normal, and skew-*t*

Because the raw CAL data exhibit skewness and possible tail behavior (see Figure 2.2), we want to start with model comparison to show the dynamic linear model discretized from the differential equation can absorb the asymmetry and it does not rely on the parametric assumption of the measurement errors. We show that by using a skew-normal (SN) and a skew-*t* (ST) assumptions for the measurement error and compare with the normality assumption.

In total, we have 5 competing models with different assumptions for the measurement errors :

- N : Normal model, where $\nu_t \overset{iid}{\sim} N(0, \sigma^2)$, $\omega_t \overset{iid}{\sim} N(0, \tau^2)$.
- 1SN : Skew-normal model, where $\nu_t \overset{iid}{\sim}$ Skew-normal $(\eta, \sigma^2, \lambda)$, $\omega_t \overset{iid}{\sim} N(0, \tau^2)$.
- 2SN: Skew-normal model, where $\nu_t \overset{iid}{\sim}$ Skew-normal $(\eta_1, \sigma^2, \lambda_1)$, $\nu_t \overset{iid}{\sim}$ Skew-normal $(\eta_2, \tau^2, \lambda_2)$.
- 1ST: Skew-t model, where $\nu_t \overset{iid}{\sim}$ Skew-$t(\xi, \sigma^2, \alpha, \nu)$, $\omega_t \overset{iid}{\sim} N(0, \tau^2)$.
- 2ST: Skew-t model, where $\nu_t \overset{iid}{\sim}$ Skew-$t(\xi_1, \sigma^2, \alpha_1, \nu_1)$, $\omega_t \overset{iid}{\sim}$ Skew-$t(\xi_2, \tau^2, \alpha_2, \nu_2)$

The definitions of the skew-normal (SN) and skew-t (ST) distributions are proposed in Sahu et al. (2003). All the 5 models share the same measurement and transition equations as in (2.5), but we have different distribution assumptions for the error terms. In the SN distribution, η is the location parameter, σ^2 is the scale parameter, and λ is the shape parameter. In the ST distribution, α is the skewness parameter. Since our data is left-skewed (see Figure(2.2)), to make sure that α is negative, we use a truncated normal prior on α where $\alpha \sim N(0, 1000)\ T\ (, 0)$. And ν is the degree of freedom. The prior on ν is Uniform, $\nu \sim$ Uniform $(1, 100)$. Table A.10 presents the degree of freedom estimates for the skew-t distribution, which are all greater than 50. This shows the skew-t density converges to the skew-normal density. Besides flexible parametric assumption, we also want to see if the initial states and the choice of priors have an effect. To study the effect of initial states, there are two different initial states for the concentration: (I) $C(0) = 0$ (II) $C(0) \neq 0$ and is fixed at 5% to 10% of the equilibrium. Here, we set it to 7.5% of the equilibrium. To model with informative priors, we assume a Uniform distribution with lower/upper bounds be 5% less/more of the true value if available:

$$\begin{aligned}
K_L &\sim U(10 \times 10^{-5}\ , 0.02)\ , \\
G &\sim U(41\ , 45)\ , \\
\frac{1}{\sigma^2} &\sim U(20\ , 100)\ , \\
\frac{1}{\tau^2} &\sim U(20\ , 100)\ .
\end{aligned}$$

For the non-informative priors :

$$\begin{aligned}
&K_L \sim U(10 \times 10^{-5}\ , 0.06)\ , \\
&G \sim U(34\ , 100)\ , \\
&\frac{1}{\sigma^2} \text{ and } \frac{1}{\tau^2} \text{ have the same priors and in the informative case.}
\end{aligned}$$

As for the high ventilation rate setting, the ventilation rate is $G = 0.552$. Therefore, the informative prior for Q is Uniform (0.524, 0.580). But three digits priors are too precise, so we use $U \ \sim (0.5, 0.6)$ instead. We developed 4 model setups:

M1: $C(0) = 0$ with informative priors,

M2: $C(0) = 0$ with non-informative priors,

M3: $C(0) \neq 0$ with informative priors,

M4: $C(0) \neq 0$ non-informative priors,

and for each initial state and prior each setup, we fit 5 competing models with different measurement error assumptions.

To select the best model from various competing models such as skew-t and skew-normal using Bayesian model selection tools, we consider both deviance based criterion (Spiegelhalter et al., 2002) and the posterior predictive performance (Gelman et al., 2014a). First, we used the popular Deviance Information Criterion (DIC), and it involves a measurement of fit, usually a deviance statistic, as well as complexity which is the number of free parameters in the model. It's defined as DIC $= \overline{D(\boldsymbol{\theta})} + p_D$. $\overline{D(\boldsymbol{\theta})} = -2\mathrm{E}\{\log[f(\mathbf{y}|\boldsymbol{\theta})]|\mathbf{y}\}$, $f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\theta})$ is the likelihood function, $\mathrm{E}\{\log[f(\mathbf{y}|\boldsymbol{\theta})]|\mathbf{y}\}$ is the posterior expectation of $\log[f(\mathbf{y}|\boldsymbol{\theta})]$ and p_D is the effective number of parameters in the model, given by $p_D = \overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}})$. Since the desired model should exhibit good fit with a reasonable number of parameters, smaller values of DIC show preferred models. Besides DIC, we also use the conditional predictive ordinate (CPO) statistic (Gelfand et al., 1992), derived from the posterior predictive distribution (ppd), for model selection. Let $\boldsymbol{\mathcal{D}}$ be the full data, $\boldsymbol{\mathcal{D}}^{(-i)}$ is the data with the ith observation deleted, and Ω, be our parameter vector. We denote the posterior density of Ω, given $\boldsymbol{\mathcal{D}}^{(-i)}$ by $\pi(\Omega|\boldsymbol{\mathcal{D}}^{(-i)})$. For the i-th observation, the CPO$_i$ can be written as CPO$_i$ $= \int f(\mathbf{y}_i|\Omega)\pi(\Omega|\boldsymbol{\mathcal{D}}^{(-i)})d\Omega = \left\{\int \frac{\pi(\Omega|\boldsymbol{\mathcal{D}})}{f(\mathbf{y}_i|\Omega)} d\Omega\right\}^{-1}, i = 1, \ldots, n$. The CPO_i can be interpreted as the height of the marginal density of the time to event at $\mathbf{y}_i$. In absence of a closed form, a Monte Carlo estimate of CPO$_i$ can be obtained using a harmonic-mean approximation (Dipak, 1997) as $\widehat{\mathrm{CPO}}_i = \left\{\frac{1}{Q}\sum\limits_{q=1}^{Q} \frac{1}{f(\mathbf{y}_i|\Omega_q)}\right\}^{-1}$, where $\Omega_1, \ldots, \Omega_Q$ is a post burn-in sample of size Q from $\pi(\Omega|\boldsymbol{\mathcal{D}})$. A summary statistic of the CPO$_i$'s is the log pseudo-marginal likelihood (LPML), defined by LPML $= \sum\limits_{i=1}^{n} \log(\widehat{\mathrm{CPO}}_i)$. Larger values of LPML indicates better fit.

Table 2.3 shows model comparison results. In the DIC column, we note that the differences in DIC between a Normal model and a skew-normal model are not bigger than 10, which is considered to be small enough that the two models are almost equal in terms of goodness of fit. By the posterior predictive model choice criterion, we get similar results with Bayesian p-value and LPML. The degree of freedom estimates for ST model are greater than 30, suggesting the ST model converges to the SN model (Table A.10). More details for model comparison for low and medium ventilation rates are in Table A.9, Table A.11, and Table A.12 .

Model	Assumption	DIC	LPML	Bayesian p-value
M1	N	144	19.267	0.434
	1SN	145.6	19.253	0.437
	2SN	144.6	19.253	0.436
	1ST	154.3	20.309	0.526
	2ST	148.6	20.291	0.524
M2	N	141	19.267	0.436
	1SN	142.6	19.251	0.434
	2SN	144.6	19.252	0.437
	1ST	139.7	20.285	0.529
	2ST	145.4	20.365	0.526
M3	N	143.9	19.267	0.437
	1SN	143.6	19.252	0.439
	2SN	145.3	19.252	0.435
	1ST	142.3	20.228	0.528
	2ST	146.4	20.269	0.527
M4	N	140.1	19.267	0.463
	1SN	142.6	19.250	0.439
	2SN	141.7	19.253	0.434
	1ST	141.7	20.233	0.528
	2ST	144.8	20.263	0.525

Table 2.3: Values of posterior predictive model choice criterion for WMR model with high ventilation rate. M1-M4 are 4 model setups that specify the initial concentration level and priors. For each setup, DIC, LPML, and Bayesian p-value are calculated from the posterior samples from 5 competing models with different assumptions for the measurement errors.

We conclude that the underlying differential equation that describes the physical WMR model is adequate in explaining the skewness in the data. We stick with the Normal model assumption for our further work such as experimental data analysis and simulation studies.

We fit model in (2.5) to the WMR model chamber experimental data using two parallel MCMC chains of 20000 iterations each. Convergence diagnostics showed that convergence happened after about 2000 iterations. In this data analysis part, we want to see if the initial values of $C(0)$ matter to the estimation and we also want to see if the parameters are estimable from data.

Analysis results for data from the low and medium ventilation rates experimental setups are shown in Table A.2 and Table A.3 in the Appendix. Because the ventilation rates vary, we use $Q \sim U\ (0.06, 0.07)$ and $Q \sim\ U(0.2, 0.3)$ as informative priors for the low and medium rate models respectively. We use $Q \sim U\ (0.001, 0.1)$ and $Q \sim U\ (0.001, 0.5)$ as non-informative priors for the low and medium rate models respectively. Analysis results for data from a high ventilation experimental setup are shown in Table 2.4. The concentration at steady state is 80 mg/m^3. Therefore, the initial contaminant concentration level is $C(0) = 0$ or $C(0) = 6$. With informative priors, we have the same results regardless of the initial concentration level and the true values for G (43.18) and Q (0.552) are included in the 95% CI. As for the non-informative scenarios, Q is underestimated and the true parameter value is included in the 95% CI. Also G becomes estimable where the 95% CI captures the true value (43.18) even without informative priors.

2.4.3 The Two-zone Chamber Experimental Design

In the two-zone model setup, the volume of the near field is 0.104 m^3 and the volume of the far field is 11.796 m^3. A floor-based mixing fan was positioned outside the near-field. The measured average airflow rate through chamber was $Q = 0.298\ m^3$/min. Toluene was vaporized into the chamber using the same Harvard Syringe pump as in the WMR model and the pump was placed inside a wire mesh box. The generation rate was calculated to be $G = 129.54$ mg/min and it's constant over time. From the solution to (2.3) (see details in the Appendix A.1), we see that $C_N \to \frac{G}{Q} + \frac{G}{\beta}$ and $C_F \to \frac{G}{Q}$ as $t \to \infty$. Therefore $C_N - C_F \to \frac{G}{\beta}$ and we find the steady state solution

	$C(0) = 0$, informative priors		$C(0) = 45$, informative priors	
	Estimate $(2.5\%, 97.5\%)$	MCSE	Estimate $(2.5\%, 97.5\%)$	MCSE
G	44.03 $(41.52, 45.30)$	7.40×10^{-3}	44.02 $(41.51, 45.30)$	7.50×10^{-3}
K_L	0.00216 $(6.86 \times 10^{-5}, 0.00746)$	1.40×10^{-5}	0.00214 $(6.36 \times 10^{-5}, 0.00753)$	1.41×10^{-5}
Q	0.54 $(0.525, 0.575)$	1.02×10^{-4}	0.54 $(0.525, 0.575)$	1.02×10^{-4}
σ^2	0.025 $(0.016, 0.038)$	3.61×10^{-5}	0.022 $(0.0144, 0.034)$	3.62×10^{-5}
τ^2	4.52 $(2.89, 6.96)$	7.44×10^{-3}	4.53 $(2.90, 6.99)$	7.28×10^{-3}
	$C(0) = 0$, non-formative priors		$C(0) = 45$, non-formative priors	
	Estimate $(2.5\%, 97.5\%)$	MCSE	Estimate $(2.5\%, 97.5\%)$	MCSE
G	49.56 $(37.26, 64.30)$	4.96×10^{-2}	47.94 $(36.1, 61.88)$	4.81×10^{-2}
K_L	0.013 $(5.18 \times 10^{-4}, 0.0352)$	6.72×10^{-5}	0.0195 $(9.35 \times 10^{-4}, 0.0463)$	8.89×10^{-5}
Q	0.35 $(0.21, 0.61)$	7.73×10^{-4}	0.23 $(0.013, 0.56)$	1.08×10^{-3}
σ^2	0.022 $(0.0144, 0.034)$	3.61×10^{-5}	0.022 $(0.014, 0.034)$	3.54×10^{-5}
τ^2	4.04 $((2.29, 5.26))$	6.71×10^{-3}	3.75 $(2.33, 5.04)$	4.41×10^{-3}

Table 2.4: Well-mixed model analysis results from experimental data. Ventilation rate is High. The true values are $G = 43.18$ and $Q = 0.552$. We fit the data to 4 competing models with different initial contaminant concentration and priors. The estimate columns give posterior mean and the 95% are computed from the posterior sample. MCSE stands for Monte Carlo standard error for each model.

for β is $\frac{G}{C_N - C_F}$. Theoretically speaking, the measured and modeled concentrations agree reasonably well on steady state. We can calculate β based on the steady state solution, i.e. $\beta \to \frac{G}{C_N - C_F} = \frac{G}{C_N - \frac{G}{Q}}$. Therefore, the calculated β is $\sim \frac{129.54}{525.178 - \frac{129.54}{0.298}} = 1.43\ m^3/min$.

One innovation in this experiment is the choice of boundary between the near and far fields. Previously, the boundary has been arbitrarily selected. We used a different approach here. The far and near fields are supposed to have distinctly different concentrations and it's unlikely to see a sharp discontinuity at the boundary between the two zones. We can use the rate of change of concentration to determine the boundary between the two zones since the rate is not uniform. We used this to define the near field as a 10 cm high cylinder with a radius of 10 *cm* with its base on the plane of the source.

The two-zone model treats the area around the source as a well-mixed box and identified as the near field, and the far field encompasses the near field and the rest of the room. There is some amount of air exchange between the two boxes, which is referred to as the interzonal air flow rate and denoted as β. This model requires knowledge of room ventilation (Q) and contaminant generation rates (G) in addition to requiring a non-trivial investment in obtaining this information. This model is useful for accounting for point-sources of emission that can result in a spatial difference in the magnitude of exposure due to the concentration close to the source being greater.

2.4.4 Estimation results from the experimental two-zone study data

We fit three models in this section for the experimental data. First, we assume that $\mathbf{C}(0) = (0,\ 0)^\top$. In the second model, because the steady state concentrations are 525 and 411 for the near-field and far-field, and we set the initials at 7.5% of the equilibrium which is $\mathbf{C}(0) = (39.375,\ 30.825)^\top$. In the third model, we assume $\mathbf{C}(0)$ is random, where $\mathbf{C}(0) \sim \text{MVN}(\mathbf{s}, \mathbf{S})$. We have $\mathbf{s} = (0, 0)^\top$ and $\mathbf{S} = \left(\begin{smallmatrix} 0.001 & 0 \\ 0 & 0.001 \end{smallmatrix}\right)$. We assume that concentrations at the near and far fields were measured simultaneously. Below we present our results using informative priors on the parameters. Details about the different priors we experimented with are available in the Appendix TableA.1.

With the two-zone experimental data set, we fit the following models:

D1: $\mathbf{C}(\mathbf{0}) \sim \text{MVN}(\mathbf{s}, \mathbf{S})$, where $\mathbf{s} = (0, 0)^\top$ and $\mathbf{S} = \left(\begin{smallmatrix} 0.01 & 0 \\ 0 & 0.01 \end{smallmatrix}\right)$, D2: $\mathbf{C}(\mathbf{0}) = (0, 0)$,

D3: $\mathbf{C}(\mathbf{0}) = (5, 4)$,

and the results are shown in Table 2.5. The estimate and $(2.5\%, 97.5\%)$ are computed from the posterior sample. MCSE stands for Monte Carlo standard error for each model. The estimates and MCSE for G and Q are very similar among those 3 models. Both G and Q are estimable with the true values included in the 95% CI. Estimates for the inter-zonal airflow β is also similar. Model comparison results for these 3 models are shown in Table 2.6. One can observe that when $\mathbf{C}(0)$ is assumed to follow a multivariate Normal prior, the model has the smallest DIC, i.e., the best fit among all.

	D1		D2		D3	
Param.	Estimate	MCSE	Estimate	MCSE	Estimate	MCSE
G	131.9 $(131.9, 125.3)$	2.0×10^{-2}	129.787 $(123.7, 135.5)$	1.95×10^{-2}	131.984 $(135.8, 135.8)$	1.96×10^{-2}
Q	0.286 $(0.283, 0.288)$	2.28×10^{-5}	0.287 $(0.283, 0.297)$	2.17×10^{-5}	0.286 $(0.283, 0.294)$	2.06×10^{-5}
β	1.45 $(1.35, 1.55)$	3.70×10^{-4}	1.44 $(1.33, 1.56)$	3.83×10^{-4}	1.48 $(1.37, 1.58)$	3.76×10^{-4}

Table 2.5: Two-zone model data analysis results from experimental data with informative priors. The true values are G= 129.54, Q= 0.294, and $\beta = 1.43$. We fit the data to 3 competing models with different initial contaminant concentration and priors. The estimate columns show the posterior mean and 95% credible intervals which are computed from the posterior sample. MCSE stands for Monte Carlo standard error for each model.

Model	DIC	$\bar{D}(\boldsymbol{\theta})$	$D(\bar{\boldsymbol{\theta}})$	p_D
D1	1212	1204	1195.818	8.182
D2	5602	5593	5584.889	8.111
D3	3954	3946	3937.911	8.089

Table 2.6: Model comparison results using DIC for two-zone model with experimental data. We compare 3 competing models with different initial contaminant concentrations and priors. Model with the smallest DIC has the best fit.

2.5 Simulation studies

We conducted simulation studies using computer-simulated concentration distributions for the WMR model and the two-zone model. The analysis and results for the former are presented in the Appendix A.6, while the two-zone model is presented below.

2.5.1 Data Generation and Methods

The simulated two-zone data is based on the solutions to the deterministic equations with 200 time points; see, e.g., Zhang et al. (2009) and Monteiro et al. (2014) for the expressions for the exact solutions to (2.3). We generated observations over a period of time. Based on the knowledge of industrial hygienists, β, Q, and G are fixed at 7.25 m^3/min, 15 m^3/min, and 105 mg/min respectively. And V_N and V_F are fixed at 1.1 m^3 and 240 m^3 respectively. Unlike Monteiro et al. (2014), who assumed temporally misaligned data where the measurements at several time points are available in only one of the fields, we assume that measurements from both fields are obtained simultaneously.

The Bayesian hierarchical model that we used is in (2.7), where we take the symmetric covariance matrices as $\Sigma_\nu = \left(\begin{smallmatrix} \nu_1 & \nu_{12} \\ \nu_{12} & \nu_2 \end{smallmatrix}\right)$ and $\Sigma_\omega = \left(\begin{smallmatrix} \omega_1 & \omega_{12} \\ \omega_{12} & \omega_2 \end{smallmatrix}\right)$. We implemented our model using the `rjags` package available in **R**, where the true values for the parameters are $\beta = 7.25$, $G = 105$, $Q = 15$, and $\Sigma_\nu = \Sigma_\omega = \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right)$. We assume that the contaminant concentrations from near-field ($Y_N(t)$) and far-field ($Y_F(t)$) are dependent of each other and the correlation is fairly strong between $Y_N(t)$ and $Y_F(t)$. So we set the correlation between $Y_N(t)$ and $Y_F(t)$ as $0.8/1 = 0.8$. For better convergence results, we ran two chains with different initial values for three different scenarios:

S1: Data simulated from ODE solution with $\mathbf{C}(0) = (0, 0)^\top$,

S2: Data simulated from ODE solution with $\mathbf{C}(0) = (1.5, 0.5)^\top$, S3: Data simulated from ODE solution with $\mathbf{C}(0) \sim MVN(\mathbf{s}, \mathbf{S})$, where $\mathbf{s} = (0, 0)^\top$ and $\mathbf{S} = \left(\begin{smallmatrix} 0.01 & 0 \\ 0 & 0.01 \end{smallmatrix}\right)$.

For each model, we generated 100 samples, and for each sample we ran 10000 iterations.

2.5.2 Prior Settings

In the model where $Y_N(t)$ and $Y_F(t)$ are dependent, we have 9 unknowns which are $\boldsymbol{\theta} = \{\beta, G, Q\}$, Σ_ν and Σ_ω. Based on practical experience and physical principles, we know that β cannot be huge. The prior for β can be uniformly distributed with lower bound 0 and upper bound 14.5. Since β is strictly positive, the prior can be log-normal with mean 0 and a large variance (Zhang et al., 2009). We adopted the uniform prior in this paper. The priors for G and Q are chosen similarly. The prior for G is uniform with lower bound 73.5 and upper bound 136.5. The prior for Q is uniform with lower bound 12 and upper bound 18. For the variance-covariance matrix Σ_ν and Σ_ω, we assume an inverse-Wishart prior (Carlin and Louis, 2008) with the scale matrix $\mathbf{S} = \left(\begin{smallmatrix} 10 & 0 \\ 0 & 10 \end{smallmatrix}\right)$, and the degrees of freedom $\nu = 4$.

2.5.3 Simulation Results

This section shows simulation results from the two-zone model. For each parameter, we compute the (overall) Mean Squared Error (MSE) and Relative Bias (RB), which are defined respectively as (use parameter G for illustration) MSE $= \frac{1}{p \times M} \sum_{i=1}^{M} \sum_{j=1}^{p} (\hat{G}_j^{(i)} - G_j)^2$, and RB$_j = \frac{1}{M} \sum_{i=1}^{M} \frac{\hat{G}_j^{(i)} - G_j}{G_j}$, where $\hat{G}_j^{(i)}$ is the posterior mean of G_j from the ith simulated data set and G_j is the true value. To get posterior sample inference, we randomly pick a sample and calculate the mean and 95% CI. For S1, the true values for the parameters are all included in the 95% CI. All the estimates are very close to the true values with good coverage probabilities. For S2, we observe that the true value for G, β, Q and Σ_ν are all included in the 95% CI. The estimates are very close to the true values and the coverage probabilities are good. S1 and S2 have very similar results. Given that S1 and S2 share the same priors and the initial states of $\mathbf{C}(0)$ doesn't matter much, it's not surprising that we draw this conclusion. For S3, we can observe that the true value for parameters G, β, Q and Σ_ν are all included in the 95% credible interval. The CI is wider compared to S1 and S2 given that it has a larger MCSE. To compare results between the 3 simulation schemes, S1 and Table S2 show very similar results. For S3, G and β are underestimated and all the estimates have larger MCSE as well as larger MSE compared to results for S1 and S2. The reason is that because $\mathbf{C}(0)$ is not fixed in S3, we are adding more variations in this model which resulting in larger MCSE. DIC results from the three simulation schemes are show in Table 2.8. S3 has the smallest DIC, thus the best fit.

2.6 Discussion

In this chapter, we introduced a Bayesian linear model, combining prior information on the physical model and discretizing the differential equations with the observed data and also accounts for measurement uncertainties by adding dynamic components. In occupational hygiene, Bayesian approach has gained its popularity by utilizing subjective information (i.e., informative priors). This chapter shows a statistical modeling from practicing occupational hygienists' stand points. This dynamic model approach provides a narrower CI compared to a straightforward Bayesian nonlinear regression (Zhang et al., 2009) and as well as to a Gaussian process based on Bayesian melding (Monteiro et al., 2014). If we compare the width of the CI of the key parameters such as G, Q, and β, obtained from our method to that from Monteiro et al. (2014) and Zhang et al. (2009), we notice a decrease in the width of the CI. The reason is that with the dynamic model, we don't need to introduce random effects to measure the errors.

Instead we use variances to account for uncertainties in both the measurements and the physical model itself.

One advantage of our implementation is that there's no need to solve the DE, simply discretizing the DE will do the work. Furthermore, we can take the experimental data as it is even though it presents some level of skewness, because the discretized model can capture the skewness in the data set. However, this advantage may not be applicable to other skewed data, such as the skewness in CAL from the periodontal data set. The model comparison results in Table 3.2 of Chapter 3 and Table 4.2 of Chapter 4 show inadequacy under normality assumption compared to non-Gaussian assumption, where model/link under a skewed assumption fit the data much better. This shows that the method works for one particular data set may not be the best approach for others. Hence, it's very practical to develop data-driven models.

Our data analysis part validates the efficiency in using the underlying differential equations. Even with a left-skewed data set, the differential equation is able to capture the data information and incorporate it into a normal framework. We conducted our data analysis under Gaussian assumption for WMR model and two-zone model. Additionally, besides the two-zone model presented in Monteiro et al. (2014) and Zhang et al. (2009), we also investigated the WMR model case and estimates the loss rate K_L. In this sense, this dynamic approach can be easily applied to multi-zone cases.

Our findings reveal that it's crucial to set informative priors in order to get precise estimates of the parameters of interest. Our approach allows us to combine priors on the actual physical model with the data likelihood. In our experimental data analysis, our estimates of the air flow rate G, the generation rate Q, the air flow exchange rate between far-field and near-field β in the two-zone model, and the estimate of K_L in the WMR model are close to the true values showing that our method works efficiently for both models. It also shows that our model assumptions, both WMR model and two-zone model, agree with the reality and are good for prediction. Because industrials don't usually collect data that can be used in exposure assessment, we are not able to test our model using a real-life example. Whereas in Chapter 3 and Chapter 4, we are able to apply the proposed methods to periodontal data collected from clinical research.

Param.	Model	Estimate $(2.5\%, 97.5\%)$	MCSE	RB	MSE	Cover.
G (105)	S1	104.90 $(99.41, 109.98)$	5.31×10^{-2}	-7.64×10^{-4}	8.09	96
	S2	105.81 $(103.63, 106.33)$	4.83×10^{-3}	-8.60×10^{-5}	97.98	94
	S3	98.41 $(84.50, 110.83)$	4.99×10^{-2}	1.34×10^{-2}	4615	99
$\beta(7.25)$	S1	7.2 $(7.16, 7.23)$	3.66×10^{-4}	-7.47×10^{-4}	0.39	99
	S2	7.26 $(7.16, 7.34)$	3.33×10^{-4}	-6.93×10^{-5}	0.47	99
	S3	6.8 $(5.83, 7.65)$	3.45×10^{-3}	1.34×10^{-2}	21.99	99
Q (15)	S1	14.89 $(13.66, 16.04)$	1.32×10^{-2}	6.11×10^{-4}	1.79	100
	S2	15.04 $(13.91, 16.19)$	4.14×10^{-3}	1.22×10^{-3}	2.22	100
	S3	14.46 $(12.14, 17.51)$	1.06×10^{-2}	9.68×10^{-3}	31.87	100
$\Sigma_{\nu inv}[1,1]$ (1.0)	S1	8.07 $(0.71, 1.05)$	1.93×10^{-3}	5.51×10^{-2}	1.18	96
	S2	1.09 $(0.90, 1.33)$	7.77×10^{-4}	0.03	0.99	97
	S3	0.88 $(0.64, 1.22)$	1.07×10^{-3}	0.13	4.63	93
$\Sigma_{\nu inv}[1,2](0.8)$	S1	0.64 $(0.51, 0.81)$	1.74×10^{-3}	1.53×10^{-3}	0.69	98
	S2	0.85 $(0.67, 1.06)$	$6.96 \times 10-4$	-0.02	0.88	94
	S3	0.64 $(0.40, 0.97)$	1.03×10^{-3}	-0.014	2.25	97
$\Sigma_{\nu inv}[2,2](1.0)$	S1	0.92 $(0.76, 1.13)$	2.09×10^{-3}	5.03×10^{-2}	1.23	95
	S2	1.10 $(0.90, 1.34)$	7.81×10^{-4}	0.03	1.23	94
	S3	1.15 $(0.82, 1.60)$	1.41×10^{-3}	0.134	4.46	93

Table 2.7: Two-zone model simulation results, summaries from 3 model setups with differnt initial contaminant concentraion values. Param. is the parameter and the true values are shown inside the parenthesis. Posterior mean and 95% credible interval are computed from the posterior samples. MCSE is the Monto Carlo standard error. RB stands for relative bias , MSE stands for mean square error, cover. is the coverage probability.

Model	DIC	$\overline{D}(\boldsymbol{\theta})$	$D(\bar{\boldsymbol{\theta}})$	p_D
S1	1148	1140	1131.937	8.063
S2	576.07	562	547.93	14.07
S3	110.1	99.02	87.98	11.04

Table 2.8: Model comparison results using DIC for two-zone model with simulated data. We compare 3 competing models with different initial contaminant concentrations and priors. Model with the smallest DIC has the best fit.

Chapter 3

Spatial skew-normal/independent models for clustered periodontal data with non-random missingness

3.1 Introduction

Periodontal disease (PD) usually refers to a collection of inflammatory disease affecting tissues called periodontium that surround and support the tooth and maintains them in the maxillary (upper jaw) and mandibular (lower jaw) bones. If left untreated, it can cause progressive bone loss around the tooth with loosening and eventual loss. It is well documented that some 5% to 15% of any population is susceptible to severe generalized periodontitis worldwide (Pourabbas et al., 2005). Being the primary cause of adult tooth loss, it has been estimated that about 50% of U.S. adults over the age of 35 experience early stages of periodontal disease (Oliver et al., 1998). The most important biomarker for assessing PD status is the clinical attachment level (CAL) (Darby and Walsh, 2014), measured in mm (whole numbers) by a periodontal probe. The study aims at quantifying the disease status of this population as well as the associations between disease status and patient-level covariates such as age, BMI, gender, HbA1c and smoking status (Reich and Bandyopadhyay, 2010). The motivating data example for this chapter and Chapter 4 comes from a clinical study conducted at the Medical University of South Carolina (MUSC) to determine the PD status of Type-2 diabetic Gullah-speaking African-Americans (henceforth, GAAD data). The word "Gullah" represents unique cultural and linguistic patterns to the African Americans living on the Sea Island of South Caroline (Johnson-Spruill et al., 2009). CAL was measured for each of the 6 sites of a tooth, nested within a subject, including various subject-level covariates such as age, gender, body mass index (indicating obesity status), glycemic level (indicating diabetic status), and tooth-site level covariates such as site in upper/lower jaw, site in tooth type, etc. With this multivariate response vector, the underlying statistical question was to investigate and estimate the functions determining the covariate-response relationships. However, note that this is complicated due to a few reasons. First, the dataset exhibits a large volume of missing responses (around 27% of these data), typical of any PD dataset, given that PD is the major cause of tooth loss in adults. This missingness pattern is monotone and non-ignorable, falls under the not-missing-at-random (NMAR) category for studying missing data patterns, and earlier modeled using the shared parameter framework by Reich and Bandyopadhyay (2010) and Reich et al. (2013). Second, PD progression is also hypothesized to be spatially clustered, i.e., diseased status for a set of closely located tooth-sites are similar. Spatial modeling for PD data is not new; see Reich and Bandyopadhyay (2010), Reich et al. (2013) and Boehm et al. (2013) for a variety of contexts in this vein. Furthermore, a plot of the CAL data exhibits

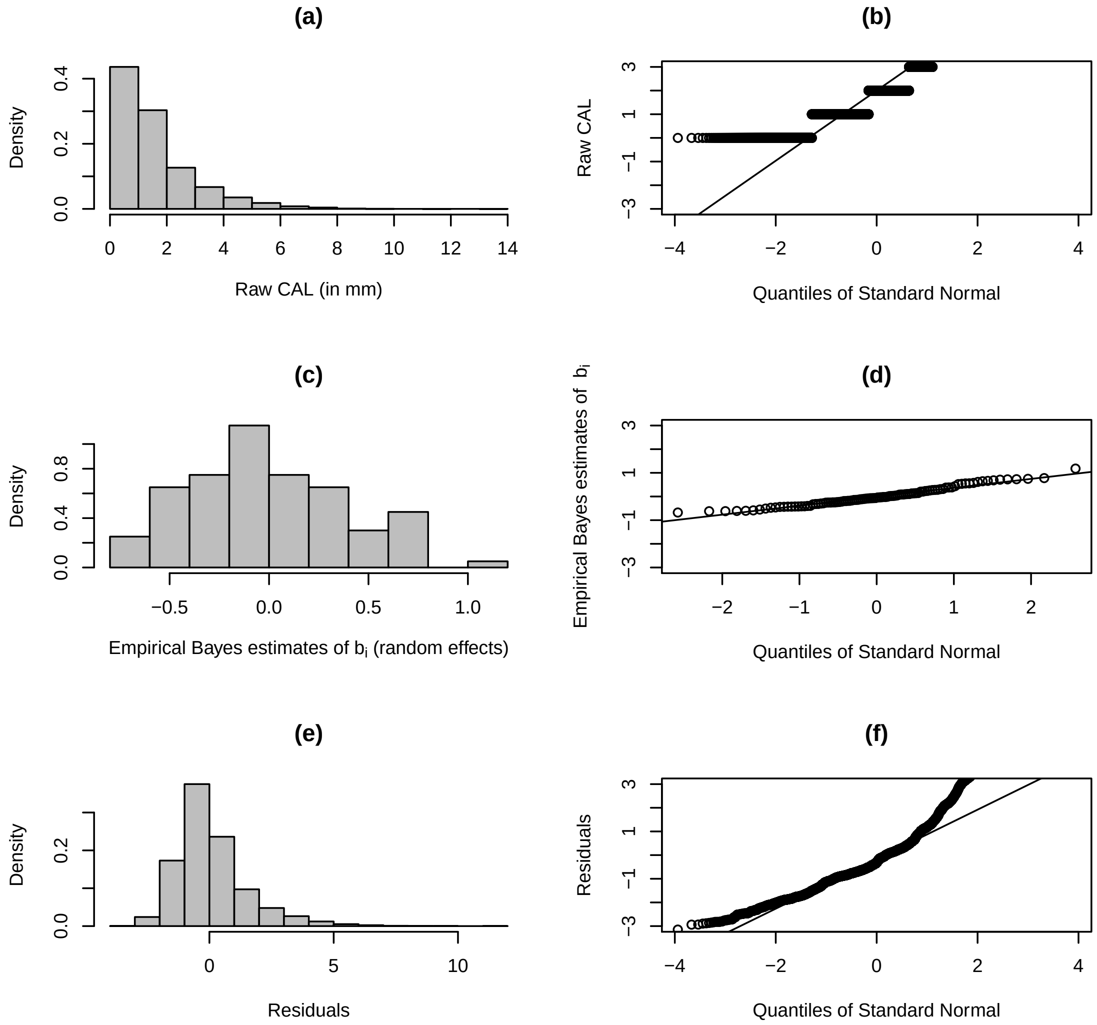


Figure 3.1: GAAD Data: Plots of the density histogram of the raw CAL responses (panel a), the empirical Bayes' estimates of random effects (panel c), and the model residuals (panel e), obtained after fitting a linear mixed model to the dataset. The corresponding Q-Q plots are presented in panels (b), (d) and (f), respectively.

possible skewness and (possible) thick tails. For example, Figure 3.1 [panels (a)-(f)] presents the raw density histogram and associated Q-Q plots for the raw CAL data, the empirical Bayes estimates of the random effects, and the model residuals, after fitting a linear mixed model (LMM) to the GAAD data using the `lme` function in `nlme` package in `R`. These plots clearly reveal evidence of asymmetry (i.e., departures from the normality assumptions) and presence of possible outliers, which cannot be explained by a standard LMM fit with Gaussian assumptions. In addition to this, ignoring the aforementioned features of non-ignorable missingness and spatial clustering can bias parameter estimates and inference. Hence, we set forward to developing a model that expand the estimation framework to accommodate all these shortcomings, and produces robust parameter estimates.

Gaussian assumption is not the best choice when data exhibit non-normal behavior. Considerable research has been done by introducing more flexible parametric families that can accommodate skewness and heavy tails, and hence eliminate the need of data transformations. In the context of LMMs, the random effects distribution was relaxed using finite normal mixtures (Verbeke and Lesaffre, 1996), smoothing (Ghidey et al., 2004), a semi-nonparametric density (Zhang and Davidian, 2001). Much of recent frequentist and Bayesian advances in regression problems revolve around the attractive and popular skew-normal (elliptical) distributions (Azzalini and Capitanio, 2003; Sahu et al., 2003). Motivated by the thick-tailed normal/independent (NI) densities of Rosa et al. (2003), Bandyopadhyay et al. (2010a) developed a robust skew-normal/independent (SNI) framework for bivariate PD data to accommodate both skewness and kurtosis within the same paradigm, and extended it to tackle censoring in the context of HIV viral load modeling in Bandyopadhyay et al. (2012).

In the same vein, the literature that accommodates asymmetry (mostly via skew-elliptical densities) in the context of spatially correlated geostatistical or areal data models is also extensive. For geostatistical modeling in point-referenced data, Kim and Mallick (2004) proposed a skew-normal spatial process and related kriging techniques, while Palacios and Steel (2006) advocated scale mixing of a stationary Gaussian process for non-Gaussian data. Zhang and El-Shaarawi (2010) developed spatial interpolation methods for a class of stationary processes with skewed marginal densities. Incorporation of asymmetry in the context of spatial GLM and latent variable models appear in Hosseini et al. (2011) and Irincheeva et al. (2012), respectively. Various non/semi-parametric (Gelfand et al., 2005; Reich and Fuentes, 2007) and point processes extensions (Ji et al., 2009) are also available. For areal data (which is our case), interest lies in smoothing and spatial dependence is typically introduced using conditional specifications and related Markov random field assumptions. Here, Nathoo and Ghosh (2013) extended the signature (Gaussian) conditionally autoregressive (CAR) setup (Banerjee et al., 2014) typical to areal data models to non-Gaussian cases via the parametric skew-t density and a nonparametric Dirichlet process prior. In addition, Bayesian formulations for multivariate finite mixture models for continuous areal-referenced standardized test scores appear in Neelon et al. (2014), and two-part spatial models for semi-continuous emergency expenditure data in Neelon et al. (2015).

Our setup differs from the above in a variety of contexts. None of the formulations above accommodates non-ignorable missing data in their estimation setup. In addition, the spatial lattice we want to incorporate involves replication (i.e., separate spatial lattices for each subject), and is fundamentally different from the traditional disease mapping setting where multiple subjects are observed at each spatial location, but the spatial lattice is not replicated. In this sense, some identifiability issues due to a single realization in the spatial skew-Gaussian formulation were pointed out by Kim and Mallick (2004), and remedied by Genton and Zhang (2012). Fortunately, we do not suffer from this inconsistency. To alleviate the inadequacies involved due to unrealistic Gaussian assumptions, data transformations and non-random missingness, we extend the multivariate parametric SNI formulation in Bandyopadhyay et al. (2010a) to asymmetric spatial data. We achieve this by embedding the CAR covariance matrix in the SNI covariance specification for the random effects, and hence develop a new class of densities called SNI-CAR. Our approach follows the multivariate skew-normal development in Sahu et al. (2003), which is readily amenable to the Bayesian regression problems. Starting with a marginal stochastic representation as in Arellano-Valle et al. (2007) and Lin (2010), our SNI-CAR formulation provides a unified class of skew-heavy-tailed densities, particularly attractive for robust parametric inference.

The remainder of this chapter unfolds as follows. Section 3.2 illustrates the motivating GAAD behind this research. In section 3.3, we develop the modeling framework, with some introductory background on the SNI class of densities. Section 3.4 presents an outline of the Bayesian estimation scheme, while section 3.5 presents a simulation study to compare the finite sample performance of the various subclasses from our proposed model. Finally, section 3.6 presents a model comparison table for data analysis followed by conclusions.

3.2 GAAD Data

The motivating data was collected from a clinical study (Fernandes et al., 2009) conducted at MUSC. The study was primarily aimed to explore the relationship between periodontal disease and diabetes level (determined by Hba1c, or ‘glycosylated hemoglobin’) in Type-2 diabetic Gullah-speaking (or simply Gullah) African-Americans (13 years or older) residing in the coastal sea-islands of South Carolina. The substantial evidence of adverse effects of diabetes on periodontal health (Taylor and Borgnakke, 2008) has

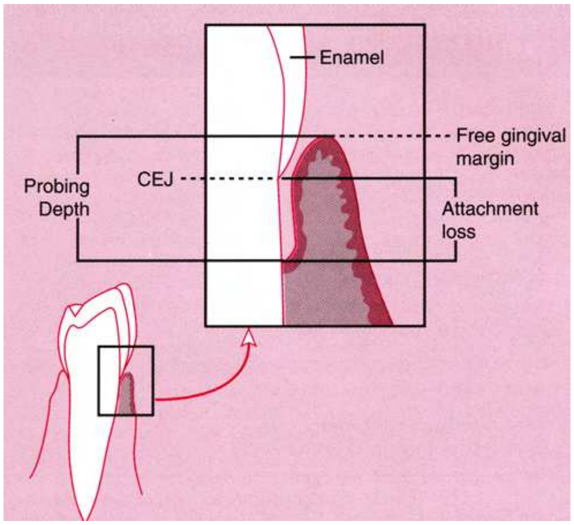


Figure 3.2: Graphical illustration of the CAL measures for a tooth. This figure was published in '*Dental Hygiene: Theory and Practice*', 1st edition, Michele L. Darby and Margaret M. Walsh, Chapter 17 Page 471, Copyright W.B.Saunders Company (1995)

been extensively explored in dental research. The 2006 American Diabetes Association (ADA) Standards of Medical Care recommend diabetic patients strive to maintain the HbA1c < 7, ideally between 4-6 (Control et al., 1993). For this current analysis, we selected 100 patients with complete covariate information. The primary measure of periodontal status/progression is defined as the clinical attachment level (CAL), measured in mm using a manual probe for 6 surfaces per tooth (disto-buccal, mid-buccal, mesio-buccal, disto-lingual, mid-lingual and mesio-lingual) for all 28 teeth per subject, except the third molars. Figure 3.2 presents a pictoral description of CAL, in addition to two other measures, pocket depth (PD) and gingival recession (CEJ-GM). PD is defined as the distance from the gingival margin to the base of the sulcus/pocket, while CEJ-GM is the distance between the free gingival margin and the cemento-enamel junction (Darby and Walsh, 2014). Next, CAL is defined as CAL = PPD – (CEJ-GM).

In addition, several subject level covariates were also collected in the study, namely Age (in years), Gender (1=Female, 0=Male), Body Mass Index or BMI (in kg/m^2), smoking status (1 = smoker, 0 = never smoker), glycemic status or Hba1c (1= High-/uncontrolled, 0 = controlled), etc. About 26% of the subjects are smokers. The mean age of the subjects is about 55 years with a range from 26-87 years. Female subjects seem to be predominant (about 73%) in our data, which is not uncommon in Gullah population (Johnson-Spruill et al., 2009). About 74% of subjects are obese (BMI >= 30) and 64% are with Hba1c = 1 (for subjects with blood sugar level higher than 7 percent), an indicator of high glycemic level. Furthermore, other site-level covariates include 'site in gap' (1 = in gap, 0 = otherwise), site in tooth-type (incisor=1, canine=2, premolar=3 and molar=4) and site in maxilla (1 = maxilla, 0 = otherwise). Inspired by Bandyopadhyay and Canale (2016) , we use Figure 3.3 to show various tooth number, site locations, and maxilla which are the tooth-level covariates.

3.3 Statistical Model

3.3.1 Skew-normal/independent Densities

We start with the definition of the SN distribution proposed in Sahu et al. (2003) as an alternative to Azzalini and Valle (1996) for straightforward Bayesian inference. A $p \times 1$ random vector $\mathbf{Y}$ follows a SN distribution with $p \times 1$ location vector $\boldsymbol{\mu}$, $p \times p$ positive definite dispersion matrix $\boldsymbol{\Sigma}$ and $p \times p$ asymmetry matrix $\boldsymbol{\Lambda} = \text{Diag}(\boldsymbol{\lambda})$, where $\text{Diag}(\cdot)$ is a diagonal matrix, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^\top$, written as $\mathbf{Y} \sim \text{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$, if its pdf is given by

$$f(\mathbf{y}) \quad = \quad 2^p \phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Omega}) \Phi_p(\boldsymbol{\Lambda}^\top \boldsymbol{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\Delta}), \tag{3.1}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$, $\boldsymbol{\Delta} = (\mathbf{I}_p + \boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda})^{-1} = \mathbf{I}_p - \boldsymbol{\Lambda}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{\Lambda}$, $\mathbf{I}_p$ is the $p \times p$ identity matrix, $\phi_p(.; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\Phi_p(.; \boldsymbol{\Sigma})$ are, respectively, the p-variate probability density

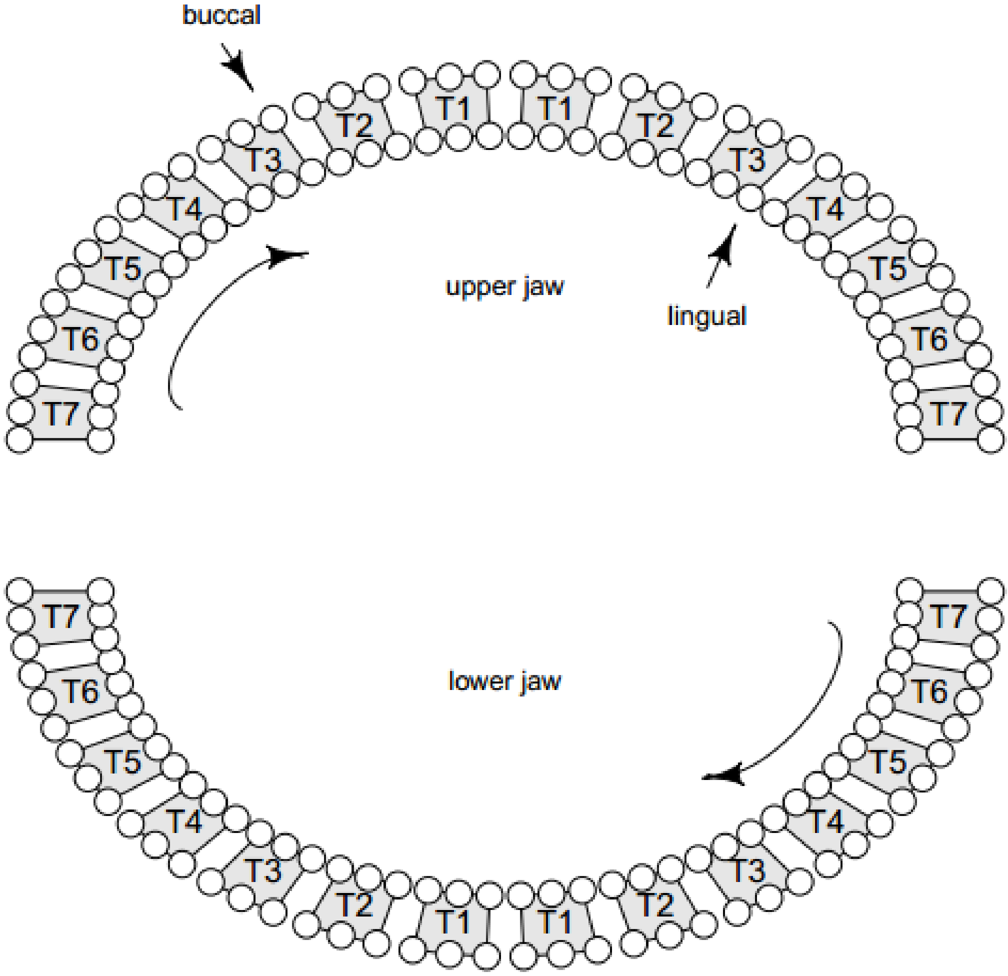


Figure 3.3: Tooth number (T1 - T7), site locations (buccal versus lingual), maxilla(upper jaw versus lower jaw), and tooth types (T7 and T6 are molars; T5 and T4 are premolars; T3 is canine; T1 and T2 are incisors)

function (pdf) of $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the p-variate cumulative distribution function (cdf) of $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. Note that for $\boldsymbol{\Lambda} = \mathbf{0}_{p\times p}$ (or $\boldsymbol{\lambda} = \mathbf{0}_{p\times 1}$) where $\mathbf{0}_{p\times p}$ and $\mathbf{0}_{p\times 1}$ are respectively a $p \times p$ matrix and a $p \times 1$ vector of zeros, (3.1) reduces to the symmetric $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-pdf, while for non zero values of $\boldsymbol{\Lambda}$, it produces an asymmetric family of $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-pdf's.

Following Bandyopadhyay et al. (2010a), we define a SNI distribution as a process of the p-dimensional random vector

$$\mathbf{Y} = \boldsymbol{\mu} + U^{-1/2}\mathbf{Z}, \tag{3.2}$$

where $\boldsymbol{\mu}$ is a location vector, U is a positive random variable with cdf $H(u|\boldsymbol{\nu})$ and pdf $h(u|\boldsymbol{\nu})$, independent of the $SN_p(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$ random vector $\mathbf{Z}$ (Arellano-Valle et al., 2007). Here, the parameter $\boldsymbol{\nu}$ is a scalar or vector indexing the distribution of U. Given $U = u$, $\mathbf{Y}$ follows a multivariate skew–normal distribution with location vector $\boldsymbol{\mu}$, scale matrix $u^{-1}\boldsymbol{\Sigma}$ and asymmetry matrix $u^{-1/2}\boldsymbol{\Lambda}$, i.e., $\mathbf{Y}|U = u \sim \mathrm{SN}_p(\boldsymbol{\mu}, u^{-1}\boldsymbol{\Sigma}, u^{-1/2}\boldsymbol{\Lambda})$. Thus, U affects both $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda}$. From (3.1), the marginal pdf of $\mathbf{Y}$ is:

$$f(\mathbf{y}) \;=\; 2^p \int_0^\infty \phi_p(\mathbf{y}; \boldsymbol{\mu}, u^{-1}\boldsymbol{\Omega})\Phi_p(u^{1/2}\boldsymbol{\Lambda}^\top\boldsymbol{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\Delta})dH(u|\boldsymbol{\nu}). \tag{3.3}$$

The notation $\mathbf{Y} \sim \mathrm{SNI}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, H)$ will be used when $\mathbf{Y}$ has pdf (3.3). When $\boldsymbol{\Lambda} = \mathbf{0}$, the SNI distributions reduces to the respective normal-independent (NI) density (Lange and Sinsheimer, 1993), represented by the pdf $f_0(\mathbf{y}) = \int_0^\infty \phi_p(\mathbf{y}; \boldsymbol{\mu}, u^{-1}\boldsymbol{\Sigma})dH(u; \boldsymbol{\nu})$. We will use the notation $\mathbf{Y} \sim \mathrm{NI}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, H)$ when $\mathbf{Y}$ has distribution in the NI class. This asymmetrical class of SNI distributions includes the skew-t (ST), the skew-slash (SSL) and the skew contaminated normal (SCN) distributions, all of which accommodates heavier tails than the SN and they all handle more skewness than the Normal distribution.

From (3.2), and the expressions for the expectation and covariance matrices of $\mathbf{Y}$ (Sahu et al., 2003), it follows that

$$E[\mathbf{Y}] = \boldsymbol{\mu}_{SNI} = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\kappa_1(\nu)\boldsymbol{\lambda}$$

and

$$Var[\mathbf{Y}] = \mathbf{\Sigma}_{SNI} \quad = \quad \kappa_2(\nu)(\mathbf{\Sigma} + \mathbf{\Lambda}\mathbf{\Lambda}^\top) + (\kappa_2(\nu) - \kappa_1^2(\nu))\frac{2}{\pi}\boldsymbol{\lambda}\boldsymbol{\lambda}^\top$$

where $\kappa_\alpha(\nu) = E[U^{-\alpha/2}]$, $\alpha \in \{1, 2\}$ and the moments are well defined. As in Bandyopadhyay et al. (2010a), this class of asymmetrical SNI density contains a variety of skewed densities as its members with various choices of the mixing variable U (Bandyopadhyay et al., 2010a), such as:

1. Multivariate skew-normal (SN): $H = 1$;
2. Multivariate skew-t (ST): $H = \Gamma(\nu/2, \nu/2)$, when $\nu \to \infty$ we get the SN distribution as the limiting case;
3. Multivariate skew-slash (SSL): $H = \text{Beta}(\nu, 1)$, the SSL distribution reduces to the SN distribution when $\nu \to \infty$;
4. Multivariate contaminated normal (SCN): $H = \begin{cases} \nu_2, & \text{with prob} \quad \nu_1, \\ 1, & \text{with prob} \quad 1 - \nu_1. \end{cases}$

The normal, Student-t, slash and contaminated normal distributions are retrieved by setting $\mathbf{\Lambda} = 0$. All these distributions have heavier tails than that of the SN, and can handle thick tails (kurtosis).

3.3.2 The SNI-CAR Model with Non-random Missingness

Let $y_i(s)$ be the CAL response for subject $i = 1, \ldots, N$ at spatial location $s = 1, \ldots, S$. For each subject, there are $S = 168$ potential measurement locations. Denote $\mathbf{y}_i = [y_i(1), \ldots, y_i(S)]^T$, the response vector for subject i. Typical for any PD data, either all 6 measurements from a tooth are observed, or all observations are missing (given that probing doesn't happen for missing tooth). We first develop the SNI-CAR model assuming all observations are present, and then extend it to include missingness. Under a standard linear mixed model (LMM) setup, the observed CAL for subject i can be written as:

$$\begin{aligned} \mathbf{y}_i &= \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i, \qquad (3.4) \\ \boldsymbol{\mu}_i &= \mathbf{X}_i^\top \boldsymbol{\beta} + \boldsymbol{\theta}_i, \end{aligned}$$

where $\boldsymbol{\mu}_i = [\mu_i(1), \ldots, \mu_i(S)]^\top$ is the vector of true CAL values for subject i for all the available observations, $\boldsymbol{\varepsilon}_i \sim \mathrm{N}(0, \sigma^2 \mathbf{I}_S)$ is the vector of random errors $\varepsilon_i(s)$, $\mathbf{X}_i^\top$ is p-vector of subject-level (say, age) and site-level (say, site in gap) covariates, $\boldsymbol{\beta}$ are the corresponding regression parameters of dimensions $p \times 1$ and $\boldsymbol{\theta}_i$ is the vector of random effects. Now, to accommodate possible spatial referencing, the latent vector $\boldsymbol{\theta}_i = [\theta_i(1), \ldots, \theta_i(S)]^\top$ can follow a S-dimensional multivariate normal distribution with mean zero ($E[\boldsymbol{\theta}_i] = 0$) and a CAR (Besag, 1974) covariance matrix. The CAR covariance of $\boldsymbol{\theta}_i$, denoted by $\boldsymbol{\Sigma}_i$, is given by $\tau^2 Q(\rho_i)^{-1}$, where $Q(\rho_i) = \mathbf{M} - \rho_i \mathbf{D}$, where $\mathbf{D}$ is $S \times S$ the adjacency matrix of the underlying graph whose elements $D_{ss'}$ equals 1 if locations s and s' are adjacent, and 0 otherwise; $\mathbf{M}$ is a $S \times S$ diagonal matrix with diagonal elements $M_{ss} = \sum_{s'} D_{ss'}$ representing the number of neighbors for site s; $\rho_i \in [0, 1]$ is the smoothing parameter controlling the degree of spatial association and $\tau^2 > 0$ controls the magnitude of spatial variation. In the adjacent matrix, we model the adjacent sites on the same tooth and sites that share a gap between teeth as 'neighbors'. For issues with identifiability, henceforth, we assume $\rho_i = \rho$ for all i, i.e., all subjects have the same spatial variation. This assumption is not unrealistic from a clinical standpoint, given that the set of subjects from the GAAD data are all Type-2 diabetic with extremely homogeneous socio-economic features (Johnson-Spruill et al., 2009). Now, due to the presence of possible asymmetry (skewness and thick-tails), we assume $\boldsymbol{\theta}_i$ follows a SNI-CAR density of dimension S, which we write as $\boldsymbol{\theta}_i \sim \text{SNI-CAR}_S(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, H(\cdot;\nu))$, where $\boldsymbol{\Sigma}$ is the CAR covariance, $\boldsymbol{\Lambda}$ is a diagonal matrix associated with the skewness parameter, and $H(\cdot;\nu)$ denotes one of the distributions presented in Subsection 3.3.1. Centering $\boldsymbol{\theta}_i$ to have zero mean, we assume the location parameter $\boldsymbol{\mu} = -\sqrt{\frac{2}{\pi}}\kappa_1\boldsymbol{\lambda}$. Thus, we have $\boldsymbol{\theta}_i \sim \text{SNI-CAR}_S(-\sqrt{\dfrac{2}{\pi}}\kappa_1(\boldsymbol{\nu})\lambda\mathbf{1}_S, \boldsymbol{\Sigma}, \lambda\boldsymbol{I}, H(\cdot;\nu))$, where the skewness parameter λ is chosen to be a scalar to avoid over-parametrization and identifiability problems. This representation partitions the skewness component and the spatial component, and hence provides a flexible way to incorporate multivariate asymmetric spatial random effects into our modeling.

However, in reality, substantial proportion of missing data is observed from PD studies, and the GAAD dataset is no exception (Reich and Bandyopadhyay, 2010). Hence, the complete response vector $\mathbf{y}_i$ for subject i is incomplete, and can be decomposed into $\mathbf{y}_i^o$ and $\mathbf{y}_i^m$, the observed and missing components respectively, according to the missingness process Δ, which is assumed non-ignorable, i.e., missingness is induced due to unobserved responses. For instance, in the GAAD study, subjects with higher level of PD tend to have teeth that had fallen out due to previous incidence of PD. Furthermore, the missingness is monotone, i.e., a missing tooth is never going to come back, and is different from the non-monotone assumption typical in longitudinal studies. In this situation, it has been shown that ignoring the missingness process and analyzing ‘only available’ data can lead to biased parameter estimates (Follmann and Wu, 1995; Reich and Bandyopadhyay, 2010). Hence, joint modeling of the observed CAL data and the missingness process is indicated, and we achieve this via the popular shared parameter models (SPM), where a set of (spatial) random effects induces the interdependence of the two processes (Follmann and Wu, 1995; Tsonaka et al., 2009).

Note that in our setup, we define the missing process at the tooth level, given that we cannot have an observed site and a missing site from the same tooth. Let $\delta_i(t) = 1$ if tooth t is missing for subject i, and 0 otherwise, and Δ_i denotes the corresponding vector for subject i. Under this SPM framework, the joint density of $\mathbf{y}_i$ and Δ (suppressing i) can be factored as:

$$f(\mathbf{y}^o, \mathbf{y}^m, \Delta|\Omega) = \int f_1(\mathbf{y}^o, \mathbf{y}^m|\boldsymbol{\theta}, \Omega) f_2(\Delta|\boldsymbol{\theta}, \Omega) g(\boldsymbol{\theta}|\Omega) d\boldsymbol{\theta}$$

where f, f_1 and f_2 are the respective probability density functions, $\boldsymbol{\theta}$ is the vector of SNI-CAR random effects, and Ω is the parameter vector. From this factorization, it follows that given $\boldsymbol{\theta}$, the processes $\mathbf{y}$ and Δ are independent. Now, the missing tooth locations are not random, but are related to the periodontal health of that region inside the mouth. Hence, for subject i, we allow the missing tooth indicator $\delta_i(t) \sim \text{Bernoulli}(p_{it})$, such that

$$\text{logit}(p_{it}) = a_0 + b_0 \mathbf{Z}_t^\top \boldsymbol{\theta}_i \tag{3.5}$$

where $\mathbf{Z}_t^\top \boldsymbol{\theta}_i$ is the mean of $\boldsymbol{\theta}_i$ at the six observations on tooth t, $t = 1, \ldots, 28$, with $\mathbf{Z}_t(s)$ equal 1/6 if site s is on tooth t, and 0 otherwise, and a_0 and b_0 relate the latent process to the missing tooth indicator (Reich and Bandyopadhyay, 2010). Note that since $\theta_i(s)$ is included in both the model for presence of and value of the responses, both presence and value of the data contribute to the posterior of $\theta_i(s)$, and thus the posterior of Ω, the full parameter vector under consideration. Also note that $b_0 = 0$ corresponds to independence between the latent true CAL and the location of missing teeth, in which case the location of missing teeth does not contribute to estimating Ω. Note that in our current formulation, we assume the missingness process is dependent only on the (latent) spatial random effects, and not on any covariates. This was assumed for simplicity of interpretation, and also to avoid identifiability issues; however this can certainly be relaxed in our estimation framework. Assuming $\boldsymbol{\theta}_i \sim G$ (the distribution function), the joint density of the observed data vector $(\mathbf{y}_i,\ \Delta_i)$ for the ith subject is obtained from the following marginalization:

$$f(\mathbf{y}_i, \Delta_i|G, \Omega) = \int f_1(\mathbf{y}_i|\boldsymbol{\theta}_i, \Omega) f_2(\Delta_i|\boldsymbol{\theta}_i, \Omega) dG(\boldsymbol{\theta}_i)$$

3.4 Bayesian Inference

3.4.1 Likelihood, Priors and Posteriors

In this section, we describe our choice of priors and associated posterior distributions of model parameters to implement Bayesian inference for our SNI-CAR setup. A key feature of this model is its flexible hierarchical representation. From (3.2) and the marginal stochastic representation of a SN random vector, it follows that the SNI-CAR model defined in (3.4) has the following hierarchical representation:

$$\begin{aligned}
\mathbf{Y}_i|\ \boldsymbol{\theta}\ _i, \mathbf{T}_i = \mathbf{t}_i, U_i = u_i &\overset{\text{ind}}{\sim} N_S(\mathbf{X}_i\boldsymbol{\beta} +\ \boldsymbol{\theta}\ _i, u_i^{-1}\sigma^2\mathbf{I}_S), && (3.6)\\
\boldsymbol{\theta}\ _i|\mathbf{T}_i = \mathbf{t}_i, U_i = u_i &\overset{\text{ind}}{\sim} N_S(-\sqrt{\frac{2}{\pi}}\kappa_1(\boldsymbol{\nu})\lambda\mathbf{1}_S + u_i^{-1/2}\lambda\mathbf{t}_i, u_i^{-1}\boldsymbol{\Sigma}), && (3.7)\\
\mathbf{T}_i|U_i = u_i &\overset{\text{ind}}{\sim} TN_S(\mathbf{0}, u_i^{-1}\mathbf{I}_S; \mathbb{R}_+^S), && (3.8)\\
U_i &\overset{\text{iid}}{\sim} H(u_i|\boldsymbol{\nu}), && (3.9)
\end{aligned}$$

$i = 1, \ldots, n$, where $\mathbb{R}_+^S$ denotes the Euclidean vector space of all p-tuples of positive real numbers and $TN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{A})$ denotes a p-variate truncated normal distribution for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ lying within the hyperplane $\mathbf{A}$. Defining $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top,_2^\top, \ldots,_n^\top)^\top$, $\mathbf{t} = (t_1, \ldots, t_n)^\top$, $\mathbf{u} = (u_1, \ldots, u_n)^\top$ and $\mathbb{I}_{\{A\}}(.)$ the indicator function of the set A, the corresponding likelihood function is given by

$$\begin{aligned} L(\Omega|\mathbf{y},\boldsymbol{\theta},\mathbf{u},\mathbf{t}) &\propto \prod_{i=1}^{N}[\phi_S(\mathbf{y}_i;\mathbf{X}_i\boldsymbol{\beta}+\boldsymbol{\theta}_i,u_i^{-1}\sigma^2\mathbf{I}_S)\phi_S(\boldsymbol{\theta}_i;-\sqrt{\frac{2}{\pi}}\kappa_1(\boldsymbol{\nu})\lambda\mathbf{1}_S+u_i^{-1/2}\lambda\mathbf{t}_i,u_i^{-1}\boldsymbol{\Sigma}) \\ &\quad\times\phi_S(\mathbf{t}_i;0,u_i^{-1}\mathbf{I}_q)\mathbb{I}_{\{\mathbb{R}_p^+\}}(\mathbf{t}_i)h(u_i|\boldsymbol{\nu})[\prod_{t=1}^{T}p_{it}^{\delta_{it}}(1-p_{it})^{\delta_{it}}] \qquad (3.10) \end{aligned}$$

where $p_{it}=\frac{\exp\{a_0+b_0\mathbf{Z}_t^\top\boldsymbol{\theta}_i\}}{1+\exp\{a_0+b_0\mathbf{Z}_t^\top\boldsymbol{\theta}_i\}}, i=1,\ldots,N, t=1,\ldots,T$. Now, to complete the Bayesian specification of the model, we need to put prior distribution on all the unknown parameters in $\Omega=(\boldsymbol{\beta},\sigma^2,\lambda,\tau^2,\rho,a_0,b_0)$. Since we have no prior information from historical data or from previous experiments, we assign conjugate but weakly informative priors to obtain well-defined and proper posteriors. A popular choice to ensure posterior propriety in a LMM is to consider proper (but diffuse) conditionally conjugate priors, such as non-informative Normal priors (with large variance) for the fixed-effects, and inverse gamma priors for a single variance component (Zhao et al., 2006). In general, we choose:

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathrm{N}_p(\boldsymbol{\beta}_0,\mathbf{S}_\beta), \\ a_0,\ b_0 &\sim \mathrm{N}(\mu_0,\sigma_0^2), \\ \sigma^2 &\sim \mathrm{IG}(\tau_\sigma,T_\sigma), \\ \lambda &\sim N(\lambda_0,S_\lambda), \\ \tau^2 &\sim \mathrm{IG}(a,b), \\ \rho &\sim \mathrm{Uniform}(c,d), \end{aligned}$$

where $N_p(.\ ,\ .)$ is the multivariate normal density, $\mathrm{IG}(\tau_\sigma,T_\sigma)$ is the inverse gamma (IG) density with parameters τ_σ and T_σ respectively. In particular, we used a Uniform$(0.95,1)$ prior for ρ [the spatial association parameter] to elucidate some measurable spatial association in our setup. Finally, the prior distribution for $\boldsymbol{\nu}$, with density $\pi(\boldsymbol{\nu})$, depends on the particular SNI distribution we use Cancho et al. (2011). These are as follows:

(a) Skew-t (ST) model: $\nu \sim \text{Gamma}(0.1, 0.01)\mathbb{I}_{\{(2,\infty)\}}$, i.e., the degrees of freedom parameter ν has a truncated Gamma prior distribution on the interval $(2, \infty)$. The truncation point was chosen to assure a finite variance.

(b) Skew-slash (SSL) model: $\nu \sim \text{Gamma}(a, b)$ density, with small positive values of a and b ($b \ll a$).

(c) Skew contaminated normal (SCN) model: $\nu_1 \sim \text{Uniform}(0, 1)$ and $\nu_2 \sim \text{Beta}(a, b)$.

Next, assuming elements of the full parameter vector Ω to be independent, the joint prior distribution is given by

$$\pi(\Omega) = \pi(\boldsymbol{\beta})\pi(a_0)\pi(b_0)\pi(\sigma^2)\pi(\lambda)\pi(\tau^2)\pi(\rho)\pi(\boldsymbol{\nu}). \tag{3.11}$$

Combining the likelihood function (3.10) and the prior distributions, the joint posterior distribution for Ω is now,

$$\pi(\Omega, \mathbf{u}, \mathbf{t}|\mathbf{y}) \propto L(\Omega|\mathbf{y}, \boldsymbol{\theta}, \mathbf{u}, \mathbf{t})\pi(\Omega). \tag{3.12}$$

Distribution (3.12) is analytically intractable, but MCMC methods such as the Gibbs sampler and Metropolis-Hastings algorithm can be used to draw samples, from which features of marginal posterior distribution of interest can be inferred. In this paper, we automate this MCMC sampling through a combination of `R` and `WinBUGS` software via the `R` package `R2WinBUGS`. Further details on the choice of hyper-parameters and assessments of convergence appear in the application section.

3.4.2 Bayesian Model Selection

To select our best model from various competing models such as the SN-CAR, ST-CAR, SSL-CAR, SCN-CAR and the basic N-CAR using Bayesian model selection tools, we consider both deviance-based criterion (Spiegelhalter et al., 2002) and measures based on posterior predictive performance (Gelman et al., 2014a). Due to the mixture framework and the presence of non-random missingness, we avoided using the popular Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). Instead, we used a variant of the DIC, called DIC$_3$ (Celeux et al., 2006). This is defined as DIC$_3 = \overline{D(\boldsymbol{\theta})} + \tau_D$, $\overline{D(\boldsymbol{\theta})} = -2\mathrm{E}\{\log[f(\mathbf{y}|\boldsymbol{\theta})]|\mathbf{y}\}$, $f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\theta})$ is the likelihood function, $\mathrm{E}\{\log[f(\mathbf{y}|\boldsymbol{\theta})]|\mathbf{y}\}$ is the posterior expectation of $\log[f(\mathbf{y}|\boldsymbol{\theta})]$ and τ_D is a measure of the effective number of parameters in the model, given by $\tau_D = \overline{D(\boldsymbol{\theta})} + 2\log(\mathrm{E}[f(\mathbf{y}|\boldsymbol{\theta})|\mathbf{y}])$. Thus, we have DIC$_3 = -4\mathrm{E}\{\log[f(\mathbf{y}|\boldsymbol{\theta})]|\mathbf{y}\} + 2\log(\mathrm{E}[f(\mathbf{y}|\boldsymbol{\theta})|\mathbf{y}])$. Let $\boldsymbol{\theta}^{(q)}$ be the MCMC posterior sample generated at the iteration q of the algorithm, $q = 1, \ldots, Q$. The first expectation in this expression can be approximated by $\overline{D} = \frac{1}{Q}\sum_{q=1}^{Q}\sum_{i=1}^{n} \log\left[f(\mathbf{y}_i|\boldsymbol{\theta}^{(q)})\right]$. Next, as recommended by Celeux et al. (2006), the second term in the expression can be approximated by $\sum_{i=1}^{n} 2\log \hat{f}(\mathbf{y}_i|\boldsymbol{\theta})$ with $\hat{f}(\mathbf{y}_i|\boldsymbol{\theta}) = \frac{1}{Q}\sum_{q=1}^{Q} f(\mathbf{y}_i|\boldsymbol{\theta}^{(q)})$. Model selection follows the 'lower is better' law, i.e., the model with the lowest value of DIC$_3$ is selected.

Besides DIC$_3$, we also used LPML (see details in section 2.4.2). Because the harmonic-mean identity can be unstable (Raftery et al., 2007), we consider a more pragmatic route and compute the CPO (and associated LPML) statistics using 500 non-overlapping blocks of the Markov chain, each of size 2000 post-convergence (i.e., after discarding the initial burn-in samples), and report the expected LPML computed over the 500 blocks.

3.5 Simulation Studies

To demonstrate the effects of fitting various sub-classes of the SNI-CAR formulation and non-random missingness on subject-level fixed effects, we conduct a simulation study. We use the full mouth MRF graph leading to S = 168, $\rho = 0.99$, and no spatial covariates (such as site in gap). Data are generated from the model

$$\begin{aligned} P(y_i(s) = \text{observed}) &= 1 - \Phi(a_0 + b_0\theta_i(s)), \\ y_i(s)|y_i(s) \text{ observed} &\sim N(a_1 + b_1\theta_i(s), \sigma_i^2) \end{aligned}$$

where $\boldsymbol{\theta}_i \sim ST(x^\top\beta\mathbf{1}_S, \tau_i^2\boldsymbol{Q}^{-1}(\rho), \mathbf{3}\mathbf{1}_S, 4)$, and ST is the skew-$t$ density with location (mean) vector $x^\top\beta\mathbf{1}_S$, covariance matrix $\tau_i^2\boldsymbol{Q}^{-1}(\rho)$, the skewness parameter 3, and the shape parameter (degrees of freedom) 4. Each simulated data set contains data generated from this model for $N = 50$ patients. The $p = 3$ subject-level covariates x_i are generated independently from the N$(0, 1)$ density, and the regression coefficients are $\boldsymbol{\beta} = (0, 1, 2)/3$. Finally, $\tau_i^2 = a_1 = b_1 = 1$ and $a_0 = -1$. Under this setup, M = 200 datasets are generated from each of the two designs that varies with the missing data mechanism b_0. They are:

- Design 1: $b_0 = 0$ and $\sigma_i^2 = 1$,
- Design 2: $b_0 = 1$ and $\sigma_i^2 = 1$,

For all designs, the observations within patients are spatially correlated. The subject-specific variances were all fixed to 1. We analyze each simulated data set using six models:

- Model 1: Normal (N) model without non-random missingness, that is, $b_0 = 0$,
- Model 2: Skew-normal (SN) and $b_0 = 0$,

Design	Model	b_0	β_0	β_1	β_2	RB$_1$	RB$_2$	MSE
1	1	-	0.380	0.630	0.900	-0.318	-0.329	0.059
	2	-	0.630	0.855	0.995	-0.008	-0.045	0.044
	3	-	0.190	0.995	1.000	-0.008	-0.006	0.006
	4	0.055	0.520	0.740	0.975	-0.089	-0.162	0.051
	5	0.040	0.555	0.875	1.000	0.009	-0.025	0.042
	6	0.050	0.160	0.995	1.000	-0.011	-0.005	0.005
2	1	-	0.440	0.705	0.905	-0.341	-0.334	0.048
	2	-	0.530	0.885	0.995	-0.148	-0.176	0.027
	3	-	0.215	0.865	0.985	-0.128	-0.079	0.015
	4	0.780	0.480	0.715	0.930	-0.310	-0.308	0.048
	5	1.000	0.615	0.910	1.000	-0.170	-0.215	0.036
	6	0.970	0.210	0.920	0.990	-0.027	-0.018	0.011

Table 3.1: Simulation study results. Column labels b_0 - β_2 give the proportion of 95% intervals that exclude zero. Columns RB$_1$ and RB$_2$ denote the Relative Bias for parameters β_1 and β_2, while the column MSE stands for the overall mean squared error for all parameters.

- Model 3: Skew-t (ST) model and $b_0 = 0$,
- Model 4: Normal model (N) with non-random missingness,
- Model 5: Skew-normal model (SN) with non-random missingness, and
- Model 6: Skew-t (ST) model with non-random missingness,

where all models account for the spatial association via the CAR structure. While Models 2 and 5 only accommodates asymmetry, Models 3 and 6 includes asymmetry and heavy tail behavior.

The results are presented in Table 3.1. For each model and each design, we calculate the proportion of the 95% posterior intervals for b_0 and the regression coefficients that exclude zero. We also compute the (overall) Mean Squared Error (MSE) and Relative Bias (RB) for the parameters, which are also used in the simulation studies in the occupational hygiene project (see 2.5.3). MSE $= \frac{1}{p\times M}\sum_{i=1}^{M}\sum_{j=1}^{p}(\hat{\beta}_j^{(i)} - \beta_j)^2$, and RelBias$_j = \frac{1}{M}\sum_{i=1}^{M}\frac{\hat{\beta}_j^{(i)}-\beta_j}{\beta_j}$, where $\hat{\beta}_j^{(i)}$ is the posterior mean of β_j from the ith simulated data set and β_j is the true value.

For Design 1 (that generates ignorable missing data), fitting non-ignorable missing N and SN models (Models 4 and 5) leads to enhanced power for β_1, compared to the respective Models 1 and 2. However, power remains the same for Models 3 and 6 (the ST cases). For estimating the null β_0, quite interestingly, the power increases in Model 4 compared to Model 1, but reduces for Models 5 and 6 compared to Models 2 and 3, respectively. The RB for both β_1 and β_2 reduces for Model 4, compared to Model 1 (the N models). However, for SN and ST models, the RBs of β_1 and β_2 are mostly comparable between their non-ignorable and ignorable missing counterparts, except for the SN cases (Models 2 and 5) in β_2 where it reduces for Model 5 compared to 2.

For Design 2 (which generates non-randomly missing data), there is a clear improvement in the performance for models that handle non-ignorable missing data (Models 4-6) over the ones that doesn't (Models 1-3), on the overall. Specifically, for β_1, there is improvement in power (see Column 5) in the non-ignorable models over their ignorable counterparts. However, for β_2, the power is comparable across both Designs and the 6 models. In addition, RB also reduces for the N and ST non-ignorable missingness models (Models 4 and 6) over their counterparts (Models 1 and 3) for both β_1 and β_2. However, this was reversed for the SN models, i.e., the non-ignorable SN model (Model 5) exhibited slightly increased bias over Model 2 for both β_1 and β_2. The estimated value of the overall MSE is lower for Model 6 compared to Model 4, same comparing Models 3 and 1, but strangely, higher for Model 5 versus Model 2.

Overall, we conclude that when the underlying dataset exhibit skewness, tail behavior and non-ignorable missingness (Design 2), the skew-t model turns out to be more flexible and efficient than the skew-normal and the usual normality based CAR models for parameter estimation. Quite interestingly, even when the data is generated under ignorable missingness pattern, some non-ignorable missingness model (such as the Normal) can present substantially improved parameter estimation compared to its ignorable counterpart. However, not much differences are noticed in the estimates from the SN and the ST models. Note that the introduction of various sources of random heterogeneity via skewness, thick-tails, spatial referencing and non-ignorable missingness indeed complicates our framework. Quite often, these sources are not individually identifiable, and that precludes us from understanding and estimating the individual influence of each one of these to the fixed effects estimation.

3.6 Application

In this section, we illustrate our method via application to the GAAD dataset. In particular, we posit 5 competing models with various choices of densities for $\theta_i(s)$ from the SNI-CAR class, and perform model comparison to choose the best fitting model. These models are the (i) N density [Model 1], (ii) SN density [Model 2], (iii) ST density [Model 3], (iv) SSL density [Model 4], and (v) SCN density [Model 5].

For specific prior choices, we assign the components of $\boldsymbol{\beta}$, a_0, and b_0 independent Normal(0, Precision = 0.01) priors. For the variance components σ^2 and τ^2, we assign a moderately diffuse IG(0.1, 0.01) [with mean 10], and for the asymmetry parameter λ, Normal(0, Precision = 0.01) prior to accommodate either positive or negative skewness, and allow the data to determine it. Finally, prior choices for ν are as follows. For the ST density, note that the choice of $\nu \sim \text{Gamma}(0.1,\ 0.01)\mathbb{I}_{\{(2,\infty)\}}$ to achieve a finite variance led to some issues with convergence of ν and some $\boldsymbol{\beta}$. Hence, we decided to use to $\nu \sim \text{Gamma}(0.1,\ 0.01)$. For the SSL density, the prior for ν is a Gamma(a, b) with small positive values of a and b ($a = 0.01, b = 0.001$), primarily to ensure conjugacy. For the SCN density, $\nu = (\nu_1,\ \nu_2)^T$, and once again for posterior conjugacy, both ν_1 and ν_2 are chosen Beta(1, 1)(= U(0, 1)). For each of these models, we ran 2 chains with widely

Criterion	N-CAR	SN-CAR	ST-CAR	SSL-CAR	SCN-CAR
DIC$_3$	47138.79	46430.73	38560.54	42560.37	46470.19
LPML	-42187.28	-41944.21	-36468.87	-41358.7	-41966.88

Table 3.2: Model comparison using DIC$_3$ and LPML

dispersed initial values. Posterior estimates were computed using 30000 iterations with an initial burn-in of 20000, and a thinning of 5. Posterior convergence was assessed using trace plots, autocorrelation plots, and the Gelman-Rubin scale-reduction factor $\hat{R}$ (Gelman and Rubin, 1992). `WinBUGS` code for fitting the skew-t model is available in B.2.

Table 3.2 presents the DIC$_3$ and LPML values after fitting the 5 competing models to the GAAD dataset. Note that all skewed versions produced better fit than the N-CAR model. In particular, the ST-CAR model produced the best fit among all models for both criteria. This model comparison result differs from what we got in section 2.4.2, where the models with skewed errors do not show a better fit compare to the model with Gaussian random error. Therefore, it's essential to model our skewed continuous CAL using a skewed-t distribution where spatial dependency and non-random missingness are both incorporated into the likelihood.

In Table 3.3, we summarize the posterior estimates of model parameters from the N-CAR and the (best-fitting) ST-CAR models. We observe that the parameter estimates for the covariates Age, Gender, Smoker and HbA1c have the same sign in both the models and are significant (with credible intervals excluding 0), implying that PD status is usually higher with increasing age, for males, for smokers, and for subjects with uncontrolled HbA1c. This overwhelmingly satisfies the adverse effect of uncontrolled diabetes on periodontal health, extensively explored earlier in oral epidemiology (Taylor and Borgnakke, 2008). However, the estimate of BMI which was significant in the Normal model turned out to be non-significant for the ST model. Next, while comparing parameter estimates corresponding to sites in various tooth-types (with Incisors are the baseline), we observe that the canines have lower degree of PD, with increasing degree of PD in the premolars, followed by molars, for both models. This is intuitive, given that there is a high proportion of diseased molars in this population, and our model assumes that the missingness is primarily due to previous onset of PD. In addition, tooth-site located in the 'gap' area, and a site in the maxilla (upper jaw) is indicative of a higher level of PD from both models. Furthermore, b_0 is positive and significant in both models, confirming our assumption that a higher degree of PD status may lead to a higher probability (proportion) of missing tooth.

Comparing posterior estimates of the variance components σ^2 (within-subject variance) and τ^2 (spatial variance), we observe that the posterior mean of σ^2 from the ST-CAR model is lower (with a tighter 95% CI) as compared to that from the N-CAR model. On the contrary, the estimated posterior mean of the CAR variance τ^2 is several fold higher in the ST model (also with a higher standard deviation), compared to the estimate from the N model. The posterior estimate of ρ, the spatial association parameter from the ST-CAR model is 0.995, reflecting a more realistic value, compared to the estimate of 0.951 from the Normal model. Posterior mean of δ, the skewness parameter from the ST model is 1.592, and is significant, conveying some degree of right-skewness in the non-transformed clustered CAL response. In addition, the estimate of ν (the t degrees of freedom), also from the ST model is 1.275, implying very thick tails, although the variance of the t density is undefined (since $\nu < 2$).

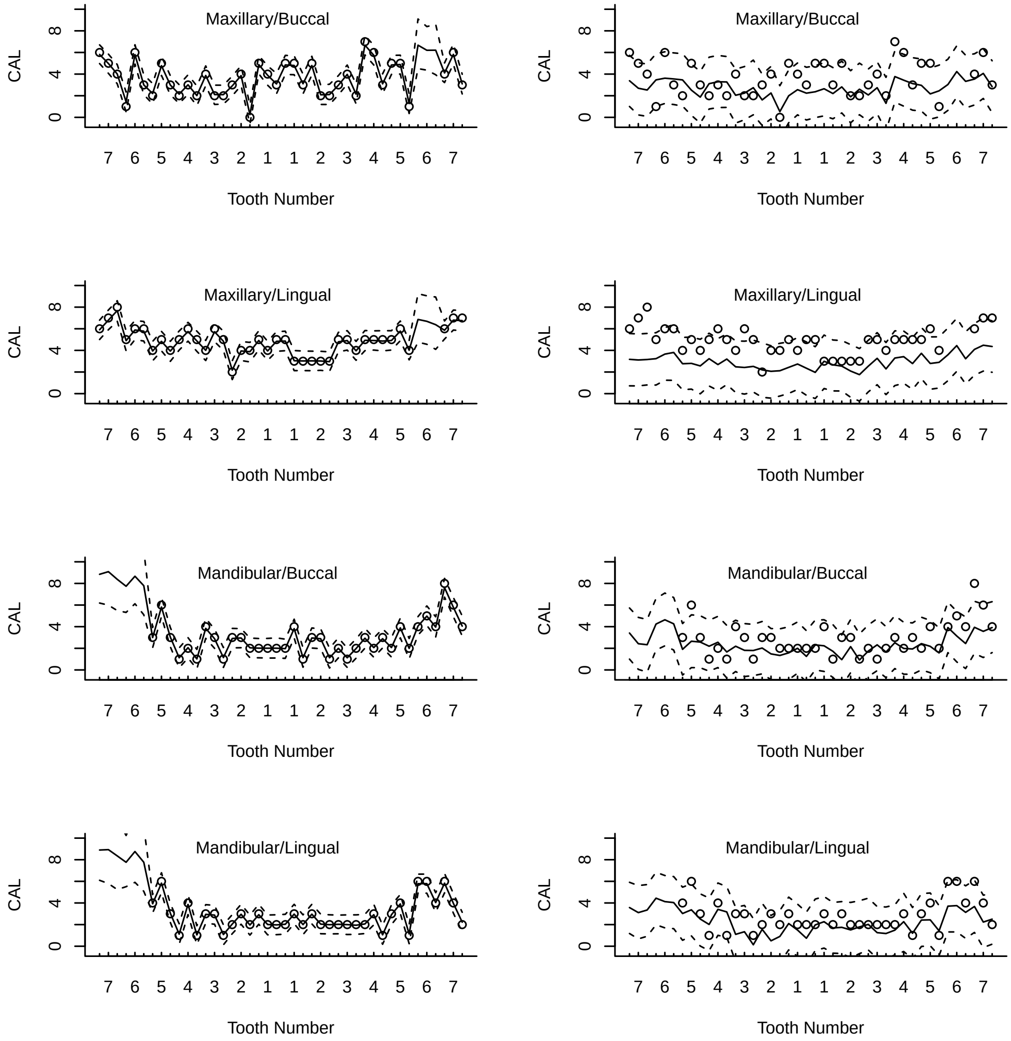
Figure 3.4 illustrates the difference between the ST-CAR and the N-CAR model in terms of prediction by comparing the fitted mean values and their 95% prediction intervals for a random subject (here, Subject # 52). From the plots, it is clear that the posterior means of the expected CAL values from the ST-CAR model closely resemble the true values (whenever they are non-missing), compared to the N-model which

	N-CAR				ST-CAR			
Parameter	Mean	SD	Lower	Upper	Mean	SD	Lower	Upper
Int.	-0.263	0.069	-0.402	-0.123	2.139	0.053	2.041	2.242
Age	0.025	0.0009	0.023	0.026	0.266	0.023	0.231	0.311
Gender	-0.163	0.022	-0.204	-0.117	-0.184	0.051	-0.262	-0.093
BMI	0.004	0.001	0.003	0.006	-0.002	0.030	-0.057	0.053
Smoker	0.406	0.027	0.356	0.461	0.234	0.042	0.159	0.302
HbA1c	0.157	0.020	0.119	0.197	0.215	0.029	0.153	0.261
a_0	-2.932	0.159	-3.225	-2.631	1.064	0.084	0.921	1.247
b_0	1.384	0.121	1.162	1.585	1.406	0.033	1.341	1.472
Canine	-0.087	0.033	-0.149	-0.018	-0.088	0.039	-0.164	-0.019
Premolar	0.553	0.237	0.506	0.599	0.442	0.027	0.395	0.494
Molar	1.505	0.036	1.436	1.578	1.629	0.042	1.556	1.708
Gap	0.437	0.026	0.389	0.486	0.655	0.018	0.616	0.684
Maxilla	0.242	0.018	0.205	0.277	0.273	0.054	0.191	0.357
ρ	0.951	0.0008	0.95	0.953	0.995	0.0003	0.995	0.996
τ^2	0.771	0.055	0.687	0.871	4.454	0.203	4.047	4.803
σ^2	1.445	0.061	1.347	1.557	0.111	0.011	0.093	0.136
δ	–	–	–	–	1.592	0.041	1.505	1.642
ν	–	–	–	–	1.275	0.009	1.258	1.288

Table 3.3: Parameter estimates derived from the Normal and the ST models, both under non-random missingness. ‘Lower’ and ‘Upper’ denotes the 2.5% and 97.5% level of the credible intervals, respectively.

produces various degrees of over- and under-fitting. In addition, the ST-CAR model leads to significantly precise estimates (as revealed from the much tighter intervals) compared to the N-CAR model. The wider intervals from the N-CAR model (which sometimes do not contain the true CAL response) reflect the increased uncertainty in prediction due to the choice of an inadequate random effects structure. Note that this subject (like many other subjects in this dataset) has missing molars. Here, the prediction uncertainty is considerable for these posteriorly-located (missing) molar sites in the mandibular/buccal and mandibular/lingual regions, for both models. The posterior prediction estimates for these missing sites from the N-CAR model appear to be substantially underestimated compared to the ST-CAR model, revealing once again that an over-simplified model may not estimate the true disease state.

For hierarchical generalized linear mixed-models, use of weakly-informative priors can lead to inference which are sensitive (Zhao et al., 2006; Natarajan and Kass, 2000) to the choice of priors on hyperparameters. To investigate this issue, we conducted sensitivity analysis on the routine use of inverse-gamma prior on variance components, and choice of the precision parameter for the Normal priors in the components of $\boldsymbol{\beta}$, a_0, and b_0. In all the results, we focused our attention on the estimation of the fixed effects parameters $\boldsymbol{\beta}$. In particular, we considered an array of weakly-informative to highly non-informative choice of priors. For example, we took $\sigma^2, \tau^2 \sim \text{IG}(10^{\kappa_1}, 10^{\kappa_1})$, where $\kappa_1 \in \{-4, -3, -2, -1, 0, 1, 2\}$, and the Normal precision to be $0.1, 0.01, 0.001$. Although we notice slight changes in the values of fixed effects estimates as well as model comparison measures, results were quite robust on the overall and did not change any conclusions regarding our best fitted model, and the posterior estimates.



(a) CAL fit from the ST−CAR Model

(b) CAL fit from the N−CAR Model

Figure 3.4: Fitted (prediction) curves and 95% prediction intervals obtained after fitting the (a) Normal-CAR model, and the (b)Skew-t-CAR model to the GAAD data. 'Dots' denote observed data, 'solid lines' denote the posterior mean estimates from the expected curves, and 'dashed lines' denote the corresponding 95% pointwise prediction intervals. Maxillary: upper jaw; Mandibular: lower jaw; Buccal: lip side, and Lingual: Tongue side

3.7 Conclusions

In this chapter, we extend the class of skew-normal/independent class proposed by Bandyopadhyay et al. (2010a) under a Bayesian framework to include spatial clustering, and non-random missingness, with an illustration on modeling PD data. Our analysis presents the necessity of considering skewness, thick-tails, and various other complexities observed in modeling PD data in terms of obtaining precise parameter estimates, and related prediction. The nice hierarchical representation given in (3.6 - 3.9) facilitates easy implementation using conventional free software, like `R2WinBUGS`, which seamlessly connects `R` with the popular Bayesian software `WinBUGS`. Furthermore, the methodology proposed can be easily extended to datasets of this kind observed in epidemiological studies.

A semi-parametric Bayesian treatment is certainly possible as an alternative (Müller and Quintana, 2004) ; however, we resorted to developing a parametric class of skew-densities primarily from the standpoint of easy implementation. Note that a semi-parametric or non-parametric proposition is often computationally challenging, and can lead to identifiability issues. Despite various skew-elliptical expositions (Genton, 2004) that are available, our development follows the skew-normal representation of Sahu et al. (2003) for elegant Bayesian implementation.

There exists a number of future directions to consider for further exploration. Note that our choice of using a robust density is primarily from a model-fitting and prediction standpoint, and precludes a thorough assessment of outliers and influential observations, the presence of which may have quite certainly affected the posterior estimates. In addition, we currently explore a spatial setting; whereas, a periodontal clinical trial may lead to a clustered-longitudinal setup where spatio-temporal modeling is of essence. These will be considered elsewhere. We can also try other methods to deal with the skewed

PD response as such adapting a link function that is suitable for spatial dependency and non-random missingness and we discuss this in depth in Chapter 4.

Chapter 4

A spatial joint factor model for extreme valued multivariate binary data with non-random missingness

4.1 Introduction

In this chapter we focus on the same GAAD data described in Chapter 3, but we model the CAL responses in a different manner. In Chapter 3, our response is the raw CAL which is treated as continuous. Here we follow the American Association of Periodontology 1999 classification (Armitage, 1999) and transform the continuous CAL into binary responses based on the periodontal disease status. The severity of PD recorded at each tooth site is classified in to the following categories:

(a) no PD (CAL = 0 mm),

(b) slight PD ($1 \leq \text{CAL} \leq 2$ mm),

(c) moderate PD ($3 \leq \text{CAL} \leq 4$ mm),

(d) severe PD ($\text{CAL} \geq 5$ mm),

(e) missing. (No measurement is taken)

We format the binary response as whether or not the CAL shows moderate to severe PD status, i.e. $\text{CAL} \geq 3$. As Figure 3.1 (a) shows the density of the continuous CAL is concentrated around smaller values of the raw CAL measurements, especially for CAL ≤ 2 mm. We are not surprised to observe that the summary statistic of the binary response reveals the asymmetry between the number of the observations with moderate to severe PD status and the number of observations without that condition. Based on this skewed binary response feature, we take a different modeling approach. Instead of using a skewed distribution directly to model the skewness as in Chapter 3, we use a link function that can work with the skewed binary response. The most popular method to model binary response is the logistic regression with a logit link, which is easy to explain and implement. Because of the skewness in the binary response, the normality assumption does not hold for our data. Under this condition, applying a symmetric link will result in link misspecification which leads to asymptotic bias and inefficient covariate estimates (Czado and Santner, 1992). One can also use non-Gaussian Markov random fields for the asymmetric binary response (Jin et al., 2016). However, this is not applicable to our extremely skewed situation neither. Therefore, we use a generalized extreme value (GEV) link with an unconstrained shape parameter to accommodate a variety of skewness situations (Wang and Dey, 2010) to model the extremely skewed binary response in the GAAD data, which we will show to have a better fit.

After figuring out the big picture, we want to focus on investigating and estimating the relationship between the patient-level and tooth-level covariates and the response by modeling the skewed binary response with the GEV link. However, this is complicated due to a few reasons. First, the GAAD data exhibit a large volume of missingness (around 27% of these data), typical of any PD dataset, given that PD is the major cause of tooth loss in adults. This missingness is non-ignorable, falls under the not-missing-at-random (NMAR) category for studying missing data patterns, and earlier was modeled using the shared parameter framework by Reich and Bandyopadhyay (2010) and Reich et al. (2013). Second, in addition to the traditional site-within-mouth clustering, PD progression is also considered to be spatially clustered, i.e., diseased status for a set of closely located tooth-sites are similar. Spatial modeling for PD data is not new; see Reich and Bandyopadhyay (2010); Reich et al. (2013); Boehm et al. (2013) for a variety of contexts in this vein. In addition to the skewed binary response, ignoring the aforementioned features of non-ignorable missingness and spatial clustering can bias parameter estimates and inference. Hence, we set forward to developing a spatial model using GEV link and jointly model the non-random missingness. This model insulates the estimation framework from all these shortcomings, and produces robust parameter estimates and precise prediction.

Given the complexities involved in this setup, the implementation of a classical inferential framework might appear to be daunting. Hence, we consider a hierarchical Bayesian formulation, with the ability to incorporate expert background (prior) information about the unknown parameters, and relying on the computation powers of the relevant Markov chain Monte Carlo (MCMC) steps for parameter estimation. The GEV link based Bayesian model provides a flexible modeling framework that simultaneously accommodates the complex features including skewness, spatial clustering, and non-random missingness of the GAAD data, but it also leads to a big computational challenge due to the large number of latent random variables which do not have closed-form full conditionals for direct posterior sampling. The convergence issue of MCMC arises when estimating the latent random variables from the corresponding posterior samples. When using the traditional random-walk algorithm, it makes the Markov chain move extremely slow and takes too many iterations for the chain to move to the target likelihood region, leading to slow convergence. We approach this issue by applying Hamiltonian Monte Carlo (HMC) algorithm within Gibbs sampling. The HMC algorithm is derived from the Hamiltonian dynamics, a derivation of the classical mechanics, which contribute to the formation of statistical mechanics. We use HMC algorithm to generate distant proposals by taking large jumps in the likelihood space to avoid the slow exploration that results from simple random-walk proposal. Thus, HMC allows the Markov chain to move faster across the likelihood regions and dramatically reduces the computation time when estimating the latent random variables. Posterior samples can also be used for model comparison. We use Watanabe-Akaike information criterion (WAIC), which is also called widely applicable information criterion (Watanabe, 2010), based on pointwise calculation of the posterior sample, to evaluate predictions for new data in a Bayesian context. We prefer WAIC over AIC or DIC because it averages over the posterior distribution, whereas AIC and DIC only estimate the performance a partial predictive density by conditioning on a point estimate.

The rest of the chapter is as follows. Section 4.2 introduces GEV distribution and unified model components. Section 4.3 develops the joint modeling under a Bayesian framework and some computation details using Hamiltonian Markov chain. Section 4.5 is the application of our proposed model to a PD data set followed by simulation studies (section 4.4). Section 4.6 is the discussion.

4.2 Model multivariate extreme value binary data with GEV link

To specify the notation, let $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)^\top$ denote an $n \times s$ binary response vector, where $y_i(s)$ denotes the response for subject i, $i = 1, 2, \ldots, N$ at location s, $s = 1, 2, \ldots, 168$. CAL measurements determine the value of the binary response. In order to model the moderate to severe PD status, we define the following:

- $\text{CAL}_i(s) < 3$, then $y_i(s) = 0$, i.e., no event.

- $\text{CAL}_i(s) \geq 3$, then $y_i(s) = 1$.

$\text{CAL}_i(s) \geq 3$ represents the "moderate to severe" of periodontal disease status. We find that about 17% of $y_i(s)$ is 1, which shows asymmetry behavior in the binary response. Hence, instead of using a skewed distribution directly to model the skewed continuous response as in Chapter 3, we use a link function that can work with the skewed binary response. Under non-normality condition, applying a symmetric link will result in link misspecification which leads to asymptotic bias and inefficient covariate estimates (Czado and Santner, 1992). Therefore, we use a generalized extreme value (GEV) link, which allows an unconstrained shape parameter to accommodate a variety of skewness situations (Wang and Dey, 2010) to model the extremely skewed binary response in the GAAD data.

The generalized extreme value (GEV) distribution is a family of continuous distributions which consist of Gumbel, Fréchet, and Weibull distributions (Wang and Dey, 2010). To build an appropriate and flexible model accommodating the skewed binary response, Wang and Dey (2010) proposed the GEV distribution as a link function. The proposed link function is different from the other generalized extreme value introduced by McFadden (1978). In McFadden (1978) definition, it is a family of multivariate distribution functions with marginal distribution being Type I extreme value distribution or Gumbel distribution (light-tailed) which is a special case of the GEV distribution proposed by Wang and Dey (2010). The advantage of the GEV link is the incorporation of an unconstrained shape parameter to fit a wide range of skewness, which is also identifiable and estimable based on the skewness of the response curve.

Because our response is not continuous, we introduce a latent variable $y_i^\star(s)$ to project the binary response onto the real line. In the GLM framework, let

$$p_i(s) = \mathrm{P}(y_i(s) = 1) = \mathrm{P}(y_i^\star(s) \geq 0). \tag{4.1}$$

We have the following linear mixed model:

$$\begin{aligned} y_i^\star(s) &= \mu_i(s) + \epsilon_i(s)\ , \\ \epsilon_i(s) &\sim GEV(0, \xi)\ , \end{aligned} \tag{4.2}$$

(4.1) can be written as:

$$p_i(s) = F(\mu_i(s) + \epsilon_i(s)), \tag{4.3}$$

where F is a cumulative distribution function and F^{-1} determines the link function.

According to Li et al. (2016), the GEV link is the inverse of H which is assumed as

$$\begin{aligned} p_i &= H(\boldsymbol{\mu}_i \,|\, \xi) = 1 - \mathrm{GEV}(-\boldsymbol{\mu}_i; \xi) \\ &= \begin{cases} 1 - \exp\{-(1 - \xi\boldsymbol{\mu}_i)_+^{-\frac{1}{\xi}}\}, & \xi \neq 0 \\ 1 - \exp\{-\exp(\boldsymbol{\mu}_i)\}, & \xi = 0 \end{cases} \end{aligned} \tag{4.4}$$

where $\mathrm{GEV}(\boldsymbol{\mu}_i; \xi)$ represents the cumulative probability at x for the GEV distribution with location parameter $\mu = \boldsymbol{\mu}_i$ and shape parameter ξ. Apply this model specification to our case, we can rewrite (4.3) with the generalized extreme value link, and our data likelihood becomes

$$p_i(s) = p(y_i(s) = 1) = H(\mu_i(s) \,|\, \xi) = 1 - \mathrm{GEV}(-\mu_i(s);\ \xi) \tag{4.5}$$

4.2.1 Incorporating spatial dependence via a multivariate latent variable

Note that the latent random variables, rather than being independent, are considered to be spatially clusters, which contribute to PD status and missingness. To account for the spatial correlation in the latent variable $\boldsymbol{\mu}_i$, we assume that for each i,

$$\boldsymbol{\mu}_i \sim \mathrm{MVN}(\mathbf{x}_i^\top \boldsymbol{\beta} + \boldsymbol{\omega}\boldsymbol{\alpha}, \Sigma) \ , \tag{4.6}$$

where MVN is the multivariate normal density, with $\mathbf{x}_i^\top \boldsymbol{\beta} + \boldsymbol{\omega}\boldsymbol{\alpha}$ denotes the mean and Σ denotes the positive definite $s \times s$ variance-covariance matrix (also see in section 3.3.2). $\mathbf{x}_i$ is the patient-level covariate matrix and $\boldsymbol{\beta}$ is the coefficient. $\boldsymbol{\omega}$ is the tooth-level covariate and $\boldsymbol{\alpha}$ is the coefficient. The multivariate latent variable $\boldsymbol{\mu}_i$ is modeled under a spatial framework by assigning a conditionally autoregressive covariance (CAR) prior to Σ (Besag, 1974). The CAR of $\mu_i(s)$ is $\sigma^2 Q(\rho)^{-1}$. In the spatial model, $\rho \in [0, 1]$ controls the degree of spatial association and $\sigma^2 > 0$ controls the magnitude of variation. $Q(\rho) = D - \rho W$, D is a $s \times s$ diagonal matrix where the ith diagonal entry represents the number of neighbors at location s_i, W is the adjacency matrix with $w_{ii'} = 1$ if s_i and $s_{i'}$ are considered adjacent, and otherwise $w_{ii'} = 0$. We consider horizontal neighboring teeth to be adjacent, whereas section 3.1 considered both horizontal and vertical neighbors to be adjacent.

4.2.2 Joint factor model with non-random missingness for the location of the missing teeth

As we described previously, a considerable amount of teeth are missing. The locations of the missing teeth are not random, and it is related to the periodontal health in that region of the mouth. Hence, we propose a joint model for the locations of missing teeth as a function of the latent random variable $\mu_i(s)$. The six observations on a tooth are either all observed or all unobserved in our data. Let $\delta_i(t)$, the observed data, be an indicator of whether tooth t =1, ..., T is missing for patient i and it can be modeled using the probit regression. Let $\delta_i(t) = 1$ if the tooth t is missing and $\delta_i(t) = 0$ if not. $\delta_i^\star(t)$ is a latent continuous variable, Reich and Bandyopadhyay (2010) proposed the following (also see section 3.3.2):

$$\begin{aligned} &\delta_i^\star(t) = a_0 + b_0 \mathbf{Z}_t^\top \boldsymbol{\mu}_i + \epsilon_i(t), \qquad (4.7) \\ &\epsilon_i(t) \sim \mathrm{N}(0,1) \ , \end{aligned}$$

where $\mathbf{Z}_t$ is a $s \times t$ transition matrix such that $\mathbf{Z}_t^\top \boldsymbol{\mu}_i$ represents the mean of $\boldsymbol{\mu}_i$ for all the six observations on tooth t. a_0 relates to the random missingness and b_0 relates the information from the latent variable $\boldsymbol{\mu}_i$ to the missingness tooth indicator $\delta_i^\star(t)$. If $\delta_i^\star(t) > 0$, i.e., it follows a truncated Normal distribution where it is bounded above zero and $\delta_i(t) = 1$. Otherwise when $\delta_i^\star(t) < 0$, it is a truncated Normal distribution bounded below zero. We can write the indicator function as $\delta_{i0}(t) = I(\delta_i^\star(t) > 0)$. Therefore, the probit regression to model the missingness through a shared latent variable with the data is :

$$\begin{aligned} Pr(\delta_i(t) = 1) &= Pr(\delta_i^\star(t) > 0) \\ &= Pr(a_0 + b_0 \mathbf{Z}_t^\top \boldsymbol{\mu}_i + \epsilon_i(t) > 0) \\ &= Pr(\epsilon_i(t) < a_0 + b_0 \mathbf{Z}_t^\top \boldsymbol{\mu}_i) \\ &= \Phi(a_0 + b_0 \mathbf{Z}_t^\top \boldsymbol{\mu}_i) \end{aligned} \tag{4.8}$$

4.3 Bayesian inference

We showed the prior information for the latent random variable $\boldsymbol{\mu}_i$ in section (4.2.1) which follows a multivariate Normal distribution with mean $\mathbf{x}_i^\top \boldsymbol{\beta} + \boldsymbol{\omega}\boldsymbol{\alpha}$ and the spatially dependence variance covariance matrix $\Sigma(\sigma^2;\ \rho)$. For the variance parameter σ^2, which we do not know much beyond the data, a non-informative prior distribution should be used (Gelman et al., 2006). The inverse-Gamma $(0.1, 0.1)$, can be used as a non-informative prior within the conditionally conjugate family. According to Wang and Dey (2011), the variances for $\boldsymbol{\beta}$ and α represent the prior belief of whether they will be near 0. A small value of the variance suggests strong prior belief that $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are centered at 0, while a large value suggests a less informative prior. In this sense, we choose the values to allow variances to be 1000, which gives weak prior belief in $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. We have $\boldsymbol{\beta} \sim$ MVN(0, Σ_β) and $\boldsymbol{\alpha} \sim$ MVN(0, Σ_α) where the diagonal elements of Σ_β and Σ_α equal 1000. This estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ relies on the information in the data which is more objective. The shape of the GEV distribution function is highly flexible with the tail behavior controlled by the shape parameter ξ. As for a less informative prior for the unconstrained shape parameter, we use Uniform $(-1, 1)$.

4.3.1 Posterior distributions for the GEV link model

With the data likelihood, spatial dependence, non-random missingness (section 4.2) and priors, we can write out the posterior distribution function:

$$
\begin{aligned}
&\prod_{i=1}^{n}\prod_{s=1}^{\text{site}}[(1-\text{GEV}(-\mu_i(s);\xi))^{y_i(s)}][(\text{GEV}(-\mu_i(s)))^{1-y_i(s)}]\\
&\times\prod_{i=1}^{n}\prod_{t=1}^{T}q_{i0}(t)^{\delta_i(t)}(1-q_{i0}(t))^{1-\delta_{i0}(t)}\times N(a_0\,|\,\mu_0,\sigma_0^2)\times N(b_0\,|\,\mu_0,\sigma_0^2)\\
&\times\prod_{i=1}^{n}\text{MVN}(\boldsymbol{\mu}_i\,|\,\mathbf{x}_i^\top\boldsymbol{\beta}+\boldsymbol{\omega}\boldsymbol{\alpha};\Sigma(\sigma^2;\rho))\times\text{Beta}(\rho\,|\,a_\rho,b_\rho)\times\text{IG}(\sigma^2\,|\,a_\sigma,b_\sigma)\\
&\times\text{MVN}(\boldsymbol{\alpha}\,|\,\mathbf{a}_\alpha,\mathbf{B}_\alpha)\times\text{MVN}(\boldsymbol{\beta}\,|\,\mathbf{a}_\beta,\mathbf{B}_\beta)\times\text{IG}(\xi\,|\,a_\xi,b_\xi)\\
&\text{Where } q_{i0}(t)=Pr\,(\delta_i(t)=1)=N(a_0+b_0\mathbf{Z}_t^\top\boldsymbol{\mu}_i\ ,1)=\Phi(a_0+b_0\mathbf{Z}_t^\top\boldsymbol{\mu}_i)
\end{aligned}
\tag{4.9}
$$

The full conditional distributions for each parameter follows:

$$
\begin{aligned}
&\sigma^2\,|\,\text{Rest}\propto\prod_{i=1}^{n}\text{MVN}(\boldsymbol{\mu}_i\,|\,\mathbf{x}_i^\top\boldsymbol{\beta}+\boldsymbol{\omega}\boldsymbol{\alpha};\Sigma(\sigma^2;\rho))\times\text{IG}(\sigma^2\,|\,a_\sigma,b_\sigma)\\
&\sim\text{IG}\left(\sigma^2\,|\,\frac{N*S}{2}+a_\sigma,b_\sigma+\sum_{i=1}^{n}\frac{1}{2}[\boldsymbol{\mu}_i-(\mathbf{x}_i\top\boldsymbol{\beta}+\boldsymbol{\omega}\boldsymbol{\alpha})]^\top(\mathbf{D}-\rho\mathbf{W})[\boldsymbol{\mu}_i-(\mathbf{x}_i\top\boldsymbol{\beta}+\boldsymbol{\omega}\boldsymbol{\alpha})]\right)
\end{aligned}
\tag{4.10}
$$

$$
\begin{aligned}
\boldsymbol{\beta}\,|\,\text{Rest}\ &\propto\text{MVN}(\boldsymbol{\beta}\,|\,\mathbf{a}_\beta,\mathbf{B}_\beta)\times\prod_{i=1}^{n}\text{MVN}(\boldsymbol{\mu}_i\,|\,\mathbf{x}_i^\top\boldsymbol{\beta}+\boldsymbol{\omega}\boldsymbol{\alpha};\Sigma(\sigma^2;\rho))\\
&\sim\text{MVN}\ (\mathbf{Gg},\ \mathbf{G}),\\
\mathbf{G}=(\mathbf{B}_\beta^{-1}+&\sum_{i=1}^{n}\mathbf{x}_i\Sigma^{-1}\mathbf{x}_i^\top)^{-1},\ \mathbf{g}=\left(\mathbf{a}_\beta^\top\mathbf{B}_\beta^{-1}+\sum_{i=1}^{n}\boldsymbol{\mu}_i\Sigma^{-1}\mathbf{x}_i^\top-\boldsymbol{\alpha}^\top\boldsymbol{\omega}^\top\Sigma^{-1}\sum_{i=1}^{n}\mathbf{x}_i^\top\right)^\top
\end{aligned}
\tag{4.11}
$$

$$\begin{aligned}
\boldsymbol{\alpha}\,|\,\text{Rest} \;&\propto \text{MVN}\;(\boldsymbol{\alpha}\,|\,\mathbf{a}_\alpha, \mathbf{B}_\alpha) \times \prod_{i=1}^{n} \text{MVN}\;(\boldsymbol{\mu}_i\,|\,\mathbf{x}_i^\top \boldsymbol{\beta} + \boldsymbol{\omega}\boldsymbol{\alpha}, \Sigma(\sigma^2;\rho)) \\
\boldsymbol{\alpha}\,|\,\text{Rest} \;&\sim \text{MVN}\;(\boldsymbol{\alpha}\,|\,\mathbf{H}\mathbf{h}, \mathbf{H}), \;\; \mathbf{H} = \left(\mathbf{B}_\alpha^{-1} + n \times \boldsymbol{\omega}^\top \Sigma^{-1} \boldsymbol{\omega}\right)^{-1}, \qquad (4.12) \\
\text{and } \mathbf{h}^\top &= \mathbf{a}_\alpha^\top \mathbf{B}_\alpha^{-1} + \sum_{i=1}^{n} \boldsymbol{\mu}_i^\top \Sigma^{-1} \boldsymbol{\omega} - \sum_{i=1}^{N} \boldsymbol{\beta}^\top \mathbf{x}_i \Sigma^{-1} \boldsymbol{\omega}.
\end{aligned}$$

$$\rho\,|\,\text{Rest} \;\propto \text{Beta}(\rho\,|\,a_\rho, b_\rho) \times \prod_{i=1}^{n} \text{MVN}(\boldsymbol{\mu}_i\,|\,\mathbf{x}_i^\top \boldsymbol{\beta} + \boldsymbol{\omega}\boldsymbol{\alpha}, \Sigma(\rho, \sigma^2)) \qquad (4.13)$$

$$\begin{aligned}
\xi\,|\,\text{Rest} \;&\propto \text{IG}(\xi\,|\,a_\xi, b_\xi) \times \prod_{i=1}^{n} \prod_{s=1}^{\text{site}} p_i(s)^{y_i(s)} [1 - p_i(s)]^{1 - y_i(s)} \\
&\propto \text{IG}(\xi\,|\,a_\xi, b_\xi) \times \prod_{i=1}^{n} \prod_{s=1}^{\text{site}} [1 - \text{GEV}\;(-\mu_i(s);\xi)]^{y_i(s)} [\text{GEV}\;(-\mu_i(s);\xi)]^{1 - y_i(s)}
\end{aligned}$$

$$(4.14)$$

$$\boldsymbol{\mu}_i(s)\,|\,\text{Rest} \propto \pi(\boldsymbol{\mu}_i) \times p(y_i\,|\,\boldsymbol{\mu}_i) \times p(\delta_i\,|\,\boldsymbol{\mu}_i) \qquad (4.15)$$

$$\begin{aligned}
\delta_i^\star(t)\,|\,a_0, b_0, \delta_i(t), \text{Rest} \propto \;&\text{Trun. N}\left(\begin{bmatrix} 1 & \mathbf{Z}_t^\top \boldsymbol{\mu}_i \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \end{bmatrix}, \; \epsilon_{i0}(t)\right)_{\{\delta_i^\star(t) > 0\}} \qquad \mathbf{I}(\delta_i(t) = 1) \\
&+ \text{Trun. N}\left(\begin{bmatrix} 1 & \mathbf{Z}_t^\top \boldsymbol{\mu}_i \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \end{bmatrix}, \; \epsilon_{i0}(t)\right)_{\{\delta_i^\star(t) < 0\}} \qquad (1 - \mathbf{I}(\delta_{i0}(t) = 1))
\end{aligned}$$

$$(4.16)$$

$$\begin{aligned}
\begin{pmatrix} a_0 \\ \\ b_0 \end{pmatrix} \,|\, \mathrm{Rest} &\propto \prod_{i=1}^{n} \prod_{t=1}^{T} N \left(\delta_i(t) \,|\, \mathbf{x}[i,t,] \begin{pmatrix} a_0 \\ \\ b_0 \end{pmatrix}, 1 \right) \\
&\propto \exp\left(-\frac{1}{2} \left(\left(\begin{smallmatrix} a_0 \\ b_0 \end{smallmatrix} \right) - \mathbf{M}m \right)^{\top} \mathbf{M}^{-1} \left(\left(\begin{smallmatrix} a_0 \\ b_0 \end{smallmatrix} \right) \right) - \mathbf{M}m \right) \\
\mathbf{M} &= \left(\sum_{i=1}^{n} \sum_{t=1}^{T} \mathbf{x}[i,t,]^{\top} \mathbf{x}[i,t,]\right)^{-1} \\
m &= \left(\sum_{i=1}^{n} \sum_{t=1}^{T} \delta_i(t) \mathbf{x}[i,t,]\right)^{\top}
\end{aligned} \tag{4.17}$$

We adopt the Gibbs sampler to draw posterior samples of the parameters from their full conditionals. For parameters σ, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$ with closed-form full conditionals, we can use direct sampling to get the posterior samples. For parameters ρ and ξ, we can use Metropolis-Hastings algorithm within Gibbs sampling. Whereas for the high-dimensional multivariate latent variables $\boldsymbol{\mu}$, we used within-Gibbs Hamiltonian algorithm to achieve faster mixing and convergence. More details about how to update $\boldsymbol{\mu}_i$ using an algorithm based on the Hamiltonian dynamics within a Metropolis Markov chain is in section C.1 and section C.2.

4.4 Simulation study

In the simulation study, we show the efficiency of using the GEV link and informative missingness on the analysis of model covariates as well as parameters. We use one quadrant for each patient leaving the total number of sites within each subject to be $= 42$. We consider 1 spatial covariate and 1 patient-level covariate. The complete data

$y_i(t)$ are generated from the following:

$$p(y_i(t) = 1) = 1 - GEV(-\mu_i(t),\ \xi)\ ,$$
$$\boldsymbol{\mu}_i \sim \text{MVN}\left(\mathbf{x}_i^\top \boldsymbol{\beta} + \boldsymbol{\omega}\boldsymbol{\alpha};\ \Sigma(\sigma^2;\ \rho)\right)\ .$$

After we have the complete data, we add missingness there. $\delta_i^\star(s)$ is the missingness indicator and we have $p(\delta_i^\star(s) = 1) = \Phi(a_0 + b_0 \mathbf{Z}^t \boldsymbol{\mu}_i)$. Each simulation data set has $n = 50$ patients. The one subject-level covariates β generated independently from $\log \text{N}(0, 1)$ density, and the regression coefficient is $\beta = -2$. The spatial covariate is generated from Binomial ($p = 0.5$) and the regression coefficient is $\alpha = -1$. And other simulation parameters are $\rho = 0.975$, $\xi = -0.5$, $a_0 = -1.5$, and $b_0 = 0.5$. Under this setup, we generated $M = 100$ datasets with non-random missingness. It's a 50 × 42 matrix for the binary response (about 20% being 1). The missing data are generated accordingly (about 25 percent). For the above design, we analyze each data set with the following 4 competing models:

1. Model 1: GEV link with informative missingness,
2. Model 2: GEV link with missing at random,
3. Model 3: probit link with informative missingness,
4. Model 4: probit link with missing at random,

and all the 4 models are under CAR structure for spatial association. Models 1 and 2 account for the extreme values in the binary response using a GEV link, and Models 3 and 4 are under the normality assumption for the random error and use a probit link. The simulation results are shown in Table 4.1. For each model, we summarize the posterior samples and get mean, 95% credible interval, and calculate the coverage probability for the regression coefficients and the parameters. We also compute the mean square error (MSE) and Relative Bias for the regression coefficients and the parameters. Table 4.1 shows GEV link with non-random missingness (Model 1) gives the largest coverage probability for the patient-level covariate coefficient β and spatial covariate coefficient α. As for the parameters, the coverage probabilities in Model 1 are the largest for σ^2 and ρ. For missingness parameters a_0 and b_0, Models 1 and 3 give similar coverage probabilities. The Relative Bias for β in Model 1 reduces compared to other models. When non-random missingness is included, the Relative Biases for α in Models 1 and 3 are comparable between the GEV link and probit link counterparts, and there's

	Param.	M1 (Truth)	M2	M3	M4
β	Coverage	88	83	76	80
	Relative bias	0.003	-0.004	0.025	0.030
	MSE	0.294	0.303	0.328	0.370
α	Coverage	69	45	17	13
	Relative bias	0.083	0.117	0.080	0.143
	MSE	3.236	3.528	7.647	10.100
ξ	Coverage	99	100	-	-
	Relative bias	-0.948	-0.961	-	-
	MSE	25.8	26.867	-	-
σ^2	Coverage	84	80	55	50
	Relative Bias	0.026	0.039	0.048	0.051
	MSE	0.442	0.429	0.716	0.916
ρ	Coverage	82	80	71	59
	Relative Bias	-0.007	-0.007	-0.002	-0.002
	MSE	0.011	0.014	0.015	0.018
a_0	Coverage	85	-	81	-
	Relative Bias	0.020	-	0.238	-
	MSE	271.796	-	331.371	-
b_0	Coverage	82	-	81	-
	Relative Bias	0.649	-	0.695	-
	MSE	69.105	-	75.374	-

Table 4.1: Simulation study results. Column labels M1 - M4 have give the coverage probability, relative bias, and mean square error (MSE) for the parameters from the 4 competing models with different link functions and missingness settings.

a slight increase in Models 2 and 4 with the missing-at-random scenario. There's a great decrease in MSE for α in Models 1 and 2 compared to Models 3 and 4. Overall, when the data exhibit extreme skewness, spatial correlation, and non-ignorable missingness, the GEV link with non-random missingness model (Model 1) is more flexible and efficient than the probit link model. Interestingly, the overall fit of the GEV link with random missingness model (Model 2) is better than that of the probit link with non-random missingness model. There's not so much difference between Models 3 and 4 in terms of Relative Bias and MSE. These simulation results show it is more important to choose the suitable link to accommodate the skewness feature in the data, while the information we gain by modeling non-random missingness is a plus.

4.5 Data analysis: the GAAD data

In this section, we analyze the GAAD data set using the model we developed in section 4.2. In the GAAD data, we find that about 27% of the responses are missing, and 17% of the non-missing responses have CAL ≥ 4 mm. In addition, the subject level covariates are Age (in years), Gender (1=Female, 0=Male), Body Mass Index or BMI (in kg/m^2), smoking status (1 = smoker, 0 = never smoker), glycemic status or Hba1c (1= High, 0 = controlled), etc. About 26% of the subjects are smokers. The mean age of the subjects is about 52 years with a range from 26-87 years. Female subjects seem to be predominant (about 73%) in our data, which is not uncommon in Gullah population (Johnson-Spruill et al. 2009). About 74% of subjects are obese (BMI >= 30) and 64% are with Hba1c = 1 (for subjects with blood sugar level higher than 7 percent), an indicator of high glycemic level. Furthermore, six tooth number indicators with the first tooth (molar) serving as the reference tooth, and upper jaw (1 = maxilla, 0 = otherwise). For this current analysis, we selected 100 patients with complete covariate information. To make computation convenient, instead of computing all the posterior samples for the 100×168 multivariate latent random variables, we average CAL measurement for each tooth and it is now 100×28 multivariate binary responses and this greatly reduces computation time. Whereas in section 3.6, we do not have this complexity and the data analysis is at site level. Including all the covariates mentioned here, we propose 4 competing models to fit the data, which are

(a) GEV link model with non-random missingness,

(b) GEV link model with random missingness ,

(c) probit link model with non-random missingness,

(d) probit link model with random missingness,

and they follow the same setting as in the simulation study (see section 4.4).

Because of the complexity of our model, model implementation using standard Bayesian software such as `WinBUGS` and `rjags` is impossible. We carried out our sampling using the free software `R` (http://www.r-project.org/). We adjust the MCMC sample size according the speed of convergence, which is monitored using trace plots as well as Gelman-rubin diagnostics. We ran 2 chains for sensitivity analysis with different initial values for all 4 models.

First, we want to show Bayesian model comparison results using Watanabe-Akaike information criterion (WAIC), also called widely applicable information criterion (Watanabe, 2010). WAIC is a fully Bayesian approach for estimating the out-of-sample-expectation. We choose WAIC over AIC or DIC because it averages over the posterior distribution, whereas AIC and DIC only estimate the performance of a partial predictive density by conditioning on a point estimate. Therefore, WAIC works better to find the model with precise prediction. We get WAIC based on pointwise calculation of the posterior sample to evaluate predictions for new data in a Bayesian context. It accounts for the computed log pointwise posterior predictive (lppd) density and a correction for effective number of parameters to adjust for over-fitting.

To calculate the lppd, let $\{\mu_i(t)^{s^\star}, \xi^{s^\star}\}$ denote the posteriors sample, and $S^\star$ is the number of posterior samples generated which we assume is big enough to capture the posterior distribution. The log pointwise predictive density (lppd) follows (Watanabe, 2010):

$$\begin{aligned}\text{Computed lppd} &= \text{computed log pointwise predictive density} \\ &= \sum_{i=1}^{n}\sum_{t=1}^{t}\log\left(\frac{1}{S^\star}\sum_{s^\star=1}^{S^\star} p\left(y_i(t)\,|\,\mu_i(t)^{s^\star}, \xi^{s^\star}\right)\right) .\end{aligned} \tag{4.18}$$

To calculate the correction for effective number of parameters, it follows:

$$\text{computed pWAIC} = \sum_{i=1}^{n}\sum_{t=1}^{t} V_{s^\star=1}S^\star\left(log\ p(y_i(t)\,|\,\mu_i(t)^{s^\star}, \xi^{s^\star}\right), \tag{4.19}$$

where $V_{s^\star=1}^{S^\star}\left(log\ p(y_i(t)\,|\,\mu_i(t)^{s^\star}, \xi^{s^\star}\right)$ represent the sample variance for data point $y_i(t)$. We get the effective number of parameters by summing over all the data points. With pWAIC as a bias correction, we have :

$$el\hat{p}pd_{\text{WAIC}} = \text{lppd - pWAIC.} \tag{4.20}$$

Gelman et al. (2014b) defined WAIC $= -2 \times el\hat{p}pd_{\text{WAIC}}$ so as to be on the deviance scale. Whereas in Watanabes original definition (Watanabe, 2010), WAIC is the negative of the average log pointwise predictive density. Here we scale WAIC as in Gelman et al. (2014b), so it is comparable with AIC, DIC, and other measures of deviance. Table 4.2 shows the WAIC for the GEV link model with informative missingness is

	GEV link		probit link	
Info missing	**Yes**	**No**	**Yes**	**No**
lppd	-467.982	-713.880	-6050.618	-5498.904
pWAIC	250.0503	456.810	4526.432	18261.59
WAIC	1436.064	2341.378	21154.1	47521

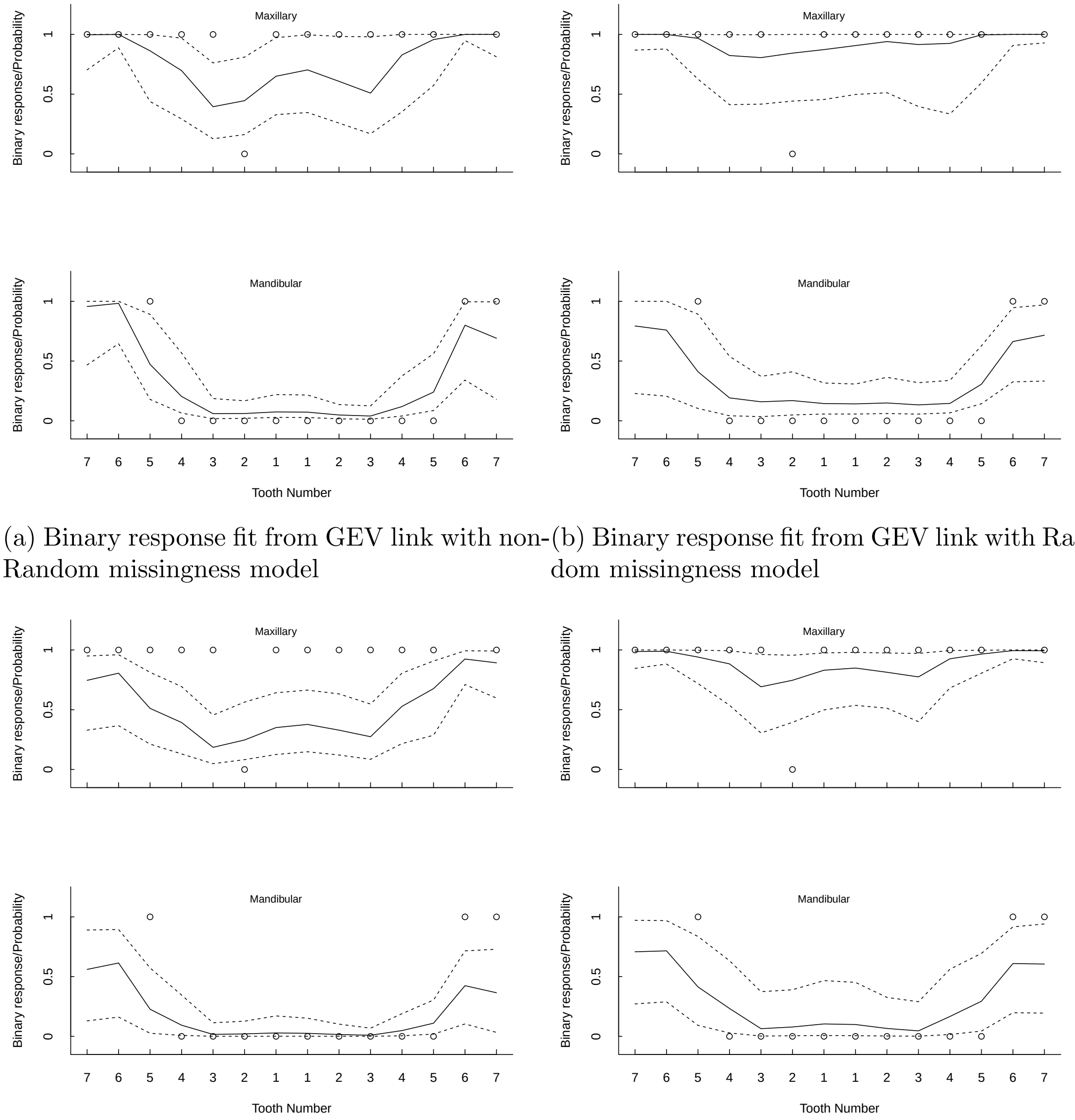
Table 4.2: Bayesian model comparison results using WAIC, log pointwise predictive density (lppd), and bias corrections for 4 competing models with different link functions and missingness setups. Smaller values of WAIC imply better predictive accuracy.

1436.064 and it is the smallest among all the 4 models. It suggests that the GEV link with informative missingness model has the best predictive performance among the 4 competing models. The results show the GEV link models have better fit than the probit link models and this shows the importance of choosing the correct link.

We further compare the model fitting using the fitted prediction interval. To do that, we first randomly pick a subject (subject 52, same as in Figure 3.4) and get the posterior samples for the multivariate latent random variable $\boldsymbol{\mu}(i, t),$ for $t = 1, 2, \ldots, 28$. Then we summarize the posterior to get mean and the 95% credible interval for each tooth for subject 52. Let $y_{52}(t)^{new}$ denote a new observation for subject 52 and tooth t. We can calculate $p(y_{52}(t)^{new} = 1) = 1 - \text{GEV}(\hat{\mu_{52}(t)}, \hat{\xi})$ as well as the 95% CI for $p(y_{52}(t)^{new} = 1)$. We then plot them against the observed data for $y_i(t)$. The probability should be close to 1 if $y_{52}(t) = 1$, and $p(y_{52}(t)^{new} = 1)$ should be small if $y_{52}(t) = 0$. Figure 4.1 illustrates the difference between the 4 competing models with different link functions and missingness assumptions. It shows the fitted probabilities from GEV link with informative missingness model closely represent the observed binary responses compared to those obtained from other models. It also reveals the GEV link with informative missingness model gives significantly precise estimates with much tighter intervals compared to the logit model where wider intervals show increased prediction uncertainty due to using misspecified link functions. In general, teeth at the lower jaw have better predictions compared to teeth at upper jaw. The CI bands for mandibular are tighter than those for maxillary.

Table 4.3 reports the posterior parameter estimates and the corresponding 95% credible intervals. Parameters whose 95% credible intervals exclude 0 are considered significant. The 95% credible interval for age does not include 0 and the posterior mean is positive. This suggests that periodontal health condition deteriorates with age. The posterior mean for Female is negative, revealing that males tend to have higher level of PD than females do. BMI and PD are positively associated for the GEV link and probit link models both with non-random missingness. Smoking has a positive influence on PD, with a much higher effect when missing-at-random is assumed in the model. Uncontrolled HbA1c is a positive indicator of PD which means a patient with uncontrolled HbA1c is more likely to have moderate to severe PD compared to a patient with controlled HbA1c after adjusting for all the other covariates. A tooth in the maxilla exhibits more advanced PD status than non-maxilla tooth, and the association for GEV link model with non-random missingness is strictly positive. T5, T6 and T7 are positively associated with PD status. The missingness coefficient a_0 accounts for missing at random, and b_0, which is the slope relating the latent spatial process, is significantly positive. This matches the intuition that patients with poor periodontal health generally have more missing teeth.

One obstacle of applying the GEV link is interpretations. Our model with the GEV link can not provide direct interpretations of the estimated covariate coefficients. Hence, we present Figure 4.2 to show the posterior predictive probabilities under various combination of covariates. Panel (a) presents the posterior predictive probabilities $Pr(y_i(s)^{\text{new}} = 1)$ from a random subject with Age = 52.03 and BMI = 37.41, under various combinations of gender, smoking status and HbA1c levels. For instance, the



(a) Binary response fit from GEV link with non-Random missingness model

(b) Binary response fit from GEV link with Random missingness model

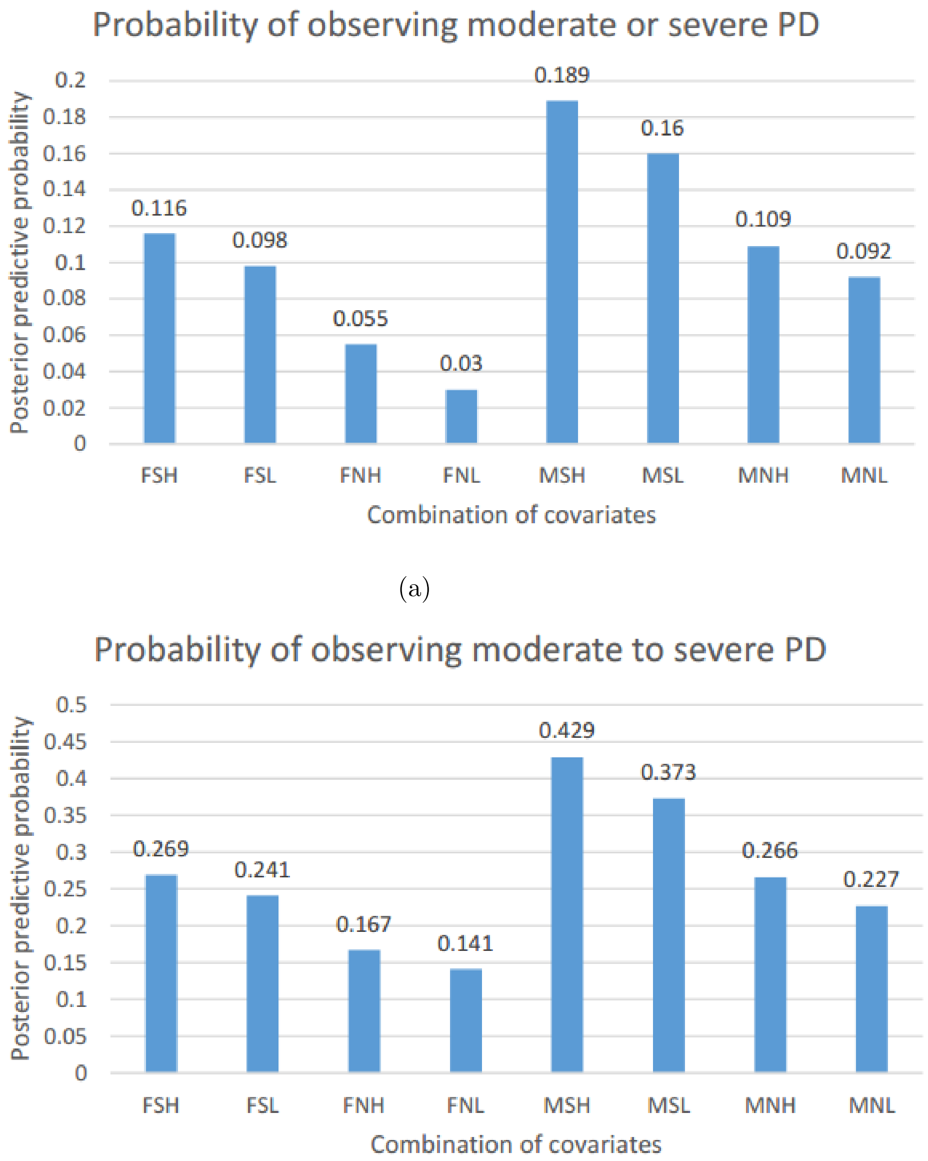
(c) Binary response fit from probit link with non-Random missingness model

(d) Binary response fit from probit link with Random missingness model

Figure 4.1: Fitted curves and 95% prediction intervals obtained after fitting (a) GEV link with informative missingness model, (b) GEV link with random missingness model, (c) Probit link with informative missingness model, and (d) Probit link with random missingness model to the GAAD data. Dots denote the observed binary response, 'solid lines' denotes the probability calculated with the posterior mean of the latent variable, and the 'dashed lines' are the corresponding 95% pointwise predictive intervals. Maxillary is the upper jaw and mandibular is the lower jaw.

	GEV link		Probit link	
Info missing	**Yes**	**No**	**Yes**	**No**
Age	0.092 $(0.059, 0.125)$	0.046 $(0.020, 0.072)$	0.046 $(0.028, 0.065)$	0.076 $(0.048, 0.109)$
Female	-0.524 $(-1.303, 0.252)$	-0.863 $(-1.480, -0.242)$	-0.421 $(-0.833, -0.008)$	-0.497 $(-1.120, 0.138)$
BMI	0.023 $(-0.011, 0.058)$	-0.001 $(-0.031, 0.027)$	0.011 $(-0.007, 0.030)$	-0.003 $(-0.033, 0.026)$
Smoker	0.593 $(-0.151, 1.330)$	1.199 $(0.609, 1.816)$	0.308 $(-0.094, 0.699)$	0.973 $(0.337, 1.594)$
HbA1c	0.182 $(-0.479, 0.848)$	0.133 $(-0.450, 0.694)$	0.129 $(-0.206, 0.481)$	0.393 $(-0.150, 0.953)$
Maxilla	0.648 $(0.011, 1.276)$	0.238 $(-0.340, 0.775)$	0.305 $(-0.020, 0.612)$	0.296 $(-0.219, 0.870)$
T2	-0.322 $(-0.592, -0.044)$	0.081 $(-0.171, 0.433)$	-0.205 $(-0.333, -0.063)$	-0.212 $(-0.546, -0.005)$
T3	-0.502 $(-0.774, -0.210)$	-0.106 $(-0.431, 0.111)$	-0.423 $(-0.639, -0.244)$	-0.402 $(-0.739, -0.067)$
T4	0.611 $(0.338, 0.931)$	-0.074 $(-0.292, 0.140)$	0.277 $(0.003, 0.511)$	0.292 $(-0.087, 0.605)$
T5	1.349 $(1.144, 1.595)$	0.659 $(0.471, 0.839)$	0.707 $(0.438, 0.975)$	0.673 $(0.330, 0.922)$
T6	3.051 $(2.695, 3.350)$	1.624 $(1.392, 1.821)$	1.681 $(1.468, 1.918)$	1.401 $(1.058, 1.684)$
T7	2.709 $(2.307, 3.217)$	1.724 $(1.489, 1.923)$	1.499 $(1.299, 1.799)$	1.339 $(0.937, 1.850)$
a_0:missing	-0.169 $(-0.249, -0.085)$	- -	-0.055 $(-0.139, 0.033)$	- -
b_0:missing	0.423 $(0.373, 0.507)$	- -	0.758 $(0.671, 0.853)$	- -
ρ	0.997 $(0.996, 0.998)$	0.998 $(0.996, 0.999)$	0.997 $(0.992, 0.999)$	0.996 $(0.962, 0.999)$
ξ	-0.004 $(-0.953, 0.952)$	-0.004 $(-0.952, 0.954)$	- -	- -
σ^2	0.251 $(0.173, 0.352)$	0.136 $(0.078, 0.221)$	0.066 $(0.043, 0.192)$	0.085 $(0.032, 0.145)$

Table 4.3: Posterior mean and 95% credible intervals for the covariate coefficients derived from GEV link and probit link models, under either non-random missingness or random missingness.



(a)

(b)

Figure 4.2: These figures show the posterior predictive probabilities $\Pr(y_i(s)^{\text{new}} = 1)$ under different combinations of covariates. F is female and M is male. S is smoking and N is non-smoking. H is uncontrolled HbA1c and L is controlled HbA1c. Panel (a) $y_i(s)$ from a random subject with Age =52.03 and BMI = 37.41, and Panel (b) represent a random subject with Age = 70 and BMI = 58.

posterior predictive probability of observing moderate to severe PD status for a random female smoker at age 52.03 and BMI= 37.41 with uncontrolled / high HbA1c is 0.116, and it is 0.098 for a random female smoker with controlled / low HbA1c. These probabilities for a male smoker are 0.189 and 0.16, respectively. Panel (b) presents the posterior predictive probabilities $Pr(y_i(s)^{\text{new}} = 1)$ from a random subject with Age = 70 and BMI = 58, under the various combinations which are the same as in panel (a). For a female smoker with uncontrolled (high) HbA1c, the posterior predictive probability of observing moderate to severe PD is about 3% higher than that of a female smoker with low HbA1c. To interpret the effect of smoking status, Figure 4.2a shows that the predictive probability of observing a moderate to severe PD tooth from a random male smoker with controlled HbA1c is 8% higher than that of a male non-smoker with uncontrolled HbA1c. Literatures on studying the relationship between smoking status and PD show there's strong evidence that smoking increase the probability of PD (Kinane and Chestnutt, 2000; Ah et al., 1994). The difference between female smokers and non-smokers is smaller than the difference between male smokers and non-smokers. With the same smoking and HbA1c situation, a random female subject is less likely to have moderate to severe PD compared to a random male subject.

4.6 Discussion

In this chapter, we explore the GEV link and use it to model asymmetric binary data under a Bayesian framework which simultaneously takes into account the features of spatial clustering and non-random missingness presented in the data. Our simulation studies and data analyses show the importance of choosing the right link to accommodate skewness and the non-random missingness. When we choose the GEV link over other asymmetric link functions such as logit link, we get less biased coefficient estimates and more precise prediction results. A challenge from model implementation is the convergence of MCMC computation due to the large number of latent random variables, which do not have closed-form full conditionals for posterior sampling via Gibbs algorithm. We facilitate this by using the within-Gibbs Hamiltonian algorithm which makes the MCMC moves much faster compared to any random walk algorithms, which makes the GEV link applicable to high-dimensional spatial-clustered data.

For further exploration, there are a couple directions to go. First, our model comparison is from a model-fitting and prediction perspective, and we can try to study the influential observations which will certainty have an impact on the posterior samples. Furthermore, besides GEV link, there are other links that can work with asymmetric binary response such as the Pareto link. We can also compare the fit between GEV link and Pareto link using cross validation. In addition, exploring the skewed ordinal response is possible. Wang and Dey (2011) shows that GEV distribution can be applied to skewed ordinal data with applications to a survey data set. We will investigate the feasibility of applying the GEV-link based model to the skewed raw CAL response.

Chapter 5

Discussion

There's no shortcut when it comes to analyze the complicated health science data. This dissertation contributes to the manner of developing parametric linear models under the Bayesian framework to study continuous and discrete asymmetric responses in two different disciplines in public health research. The motivation to develop these models comes from the unique features and characteristics of the data. Through simulation studies and data analysis results, our models fit better in terms of bias reduction and give more precise prediction results.

In Chapter 2, we introduce the dynamic linear model discretized from a differential equation under the Bayesian framework with the Gaussian assumption for the random errors. We show its efficiency by comparing between the skewed-t and skewed-normal models when handling the skewed continuous response using both simulated data and real data collected from field experiments in terms of coefficient estimates and prediction. Besides systematic well-mixed room model and near-field far-field model evaluation (Arnold et al., 2017), we now have viable new approach in occupational hygiene for contaminant exposure assessment. This will help occupational hygienists make efficient exposure assessments and accurate decisions to protect workers from harmful chemical agents. Giving informative priors allows us to combine experts' knowledge. This is useful to adapt variations from the model settings in occupational hygiene. The dynamic components account for measurement errors can reduce the credible interval length and lead to less bias. And we can keep the Gaussian assumption on measurement errors even though the data exhibits non-normality because the differential equation can capture the skewness.

When analyze the GAAD data, the primary goal is to investigate and estimate the covariate-response relationships. The continuous response CAL exhibits skewness and discretizing from the physical model is infeasible. Therefore, we cannot hold on to the normality assumption for the random errors to model the skewed response. Besides skewness, PD progression is also spatially-clustered. This dataset also exhibits a large volume of missing responses which falls under the not-missing-at-random (NMAR) category for studying missing data pattern. Based on these attributes, we develop 2 models.

In Chapter 3, we use a flexible parametric distribution family to model the multivariate continuous response under standard linear mixed model setup. We use skew-normal and skew-t distributions to compare with the normality assumption in the simulation study. Simulation studies and data analysis results show the skew-t model has the best fit. This is not surprising because skew-t controls skewness and kurtosis.

In Chapter 4, we make the continuous CAL into extremely skewed binary responses. With simulation studies and data analyses results, we show that the GEV distribution is applicable to multivariate skewed binary data with spatial dependency. Jointly modeling the response and non-random missingness using a shared latent multivariate random variable via a GEV link is a major break-through in handling extremely skewed binary response. By utilizing within-Gibbs Hamiltonian Monte Carlo sampling, we tackle the convergence issue of MCMC which makes the spatially correlated latent random variables estimable from the corresponding posterior samples. Model comparisons show a better model fit when use a GEV link compared to a logit link under normality assumption for the random errors.

This thesis adds to the existing literature for modeling asymmetric responses under the Bayesian parametric linear model proposition where we explore both continuous and discrete cases. It offers gains in parameters and regression coefficients estimation and more precision in prediction when we give appropriate assumptions for the random errors. This thesis shows this Bayesian framework is rich and flexible and there are areas for further explorations. Besides occupational hygiene and dental epidemiology, the methods we introduced here can certainly be applied to accommodate skew data in other scientific research topics.

References

Abbey, H. (1952) An examination of the reed-frost theory of epidemics. *Human Biology*, **24**, 201–234.

Agresti, A. and Franklin, C. (2007) *The Art and Science of Learning from Data*. Upper Saddle River, NJ: Prentice Hall, 2 edn.

Ah, B., Michele, K., Johnson, G. K., Kaldahl, W. B., Patil, K. D. and Kalkwart, K. L. (1994) The effect of smoking on the response to periodontal therapy. *Journal of Clinical Periodontology*, **21**, 91–97.

Arellano-Valle, R., Bolfarine, H. and Lachos, V. (2007) Bayesian inference for skew-normal linear mixed models. *Journal of Applied Statistics*, **34**, 663–682.

Arellano-Valle, R. B., Branco, M. D. and Genton, M. G. (2006) A unified view on skewed distributions arising from selections. *Canadian Journal of Statistics*, **34**, 581–601.

Armitage, G. C. (1999) Development of a classification system for periodontal diseases and conditions. *Annals of Periodontology*, **4**, 1–6.

Arnold, S. F., Shao, Y. and Ramachandran, G. (2017) Evaluating well-mixed room and near-field–far-field model performance under highly controlled conditions. *Journal of Occupational and Environmental Hygiene*, **14**, 427–437.

Azzalini, A. (2005) The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, **32**, 159–188.

Azzalini, A. and Capitanio, A. (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**, 367–389.

Azzalini, A. and Valle, A. D. (1996) The multivariate skew-normal distribution. *Biometrika*, **83**, 715–726.

Bandyopadhyay, D. and Canale, A. (2016) Non-parametric spatial models for clustered ordered periodontal data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **65**, 619–640.

Bandyopadhyay, D., Lachos, V. H., Abanto-Valle, C. A. and Ghosh, P. (2010a) Linear mixed models for skew-normal/independent bivariate responses with an application to periodontal disease. *Statistics in Medicine*, **29**, 2643–2655.

Bandyopadhyay, D., Lachos, V. H., Castro, L. M. and Dey, D. K. (2012) Skew-normal/independent linear mixed models for censored responses with applications to HIV viral loads. *Biometrical Journal*, **54**, 405–425.

Bandyopadhyay, D., Marlow, N. M., Fernandes, J. K. and Leite, R. S. (2010b) Periodontal disease progression and glycaemic control among Gullah African Americans with type-2 diabetes. *Journal of Clinical Periodontology*, **37**, 501–509.

Bandyopadhyay, D., Reich, B. J. and Slate, E. H. (2009) Bayesian modeling of multivariate spatial binary data with applications to dental caries. *Statistics in Medicine*, **28**, 3492–3508.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014) *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, 2 edn.

Beck, J., Garcia, R., Heiss, G., Vokonas, P. S. and Offenbacher, S. (1996) Periodontal disease and cardiovascular disease. *Journal of Periodontology*, **67**, 1123–1137.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, **36**, 192–236.

Boehm, L., Reich, B. J. and Bandyopadhyay, D. (2013) Bridging conditional and marginal inference for spatially referenced binary data. *Biometrics*, **69**, 545–554.

Box, G. E. and Cox, D. R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, **26**, 211–252.

Cancho, V. G., Dey, D. K., Lachos, V. H. and Andrade, M. G. (2011) Bayesian nonlinear regression models with scale mixtures of skew-normal distributions: Estimation and case influence diagnostics. *Computational Statistics & Data Analysis*, **55**, 588–602.

Carlin, B. P. and Louis, T. A. (2008) *Bayesian Methods for Data Analysis*. CRC Press, 2 edn.

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M. et al. (2006) Deviance information criteria for missing data models. *Bayesian Analysis*, **1**, 651–673.

Chen, M.-H., Dey, D. K. and Shao, Q.-M. (1999) A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, **94**, 1172–1186.

Chen, T., Fox, E. and Guestrin, C. (2014) Stochastic Gradient Hamiltonian Monte Carlo. In *Proceedings of the 31st International Conference on Machine Learning* (eds. E. P. Xing and T. Jebara), vol. 32 of *Proceedings of Machine Learning Research*, 1683–1691. Bejing, China: PMLR.

Coles, S., Bawa, J., Trenner, L. and Dorazio, P. (2001) *An introduction to Statistical Modeling of Extreme Values.* Springer, 1 edn.

Control, D., Group, C. T. R. et al. (1993) The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl j Med*, **1993**, 977–986.

Czado, C. and Santner, T. J. (1992) The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference*, **33**, 213–231.

Darby, M. L. and Walsh, M. (2014) *Dental Hygiene: Theory and Practice.* Elsevier Health Sciences, 4 edn.

Diebold, F. X., Schuermann, T. and Stroughair, J. D. (2000) Pitfalls and opportunities in the use of extreme value theory in risk management. *The Journal of Risk Finance*, **1**, 30–35.

Doucet, A. (2000) On sequential simulation-based methods for Bayesian filtering. *Statistics and Computing*, 197–208.

Elveback, L. and Varma, A. (1965) Simulation of mathematical models for public health problems. *Public Health Reports*, **80**, 1067–1076.

Fernandes, J. K., Wiegand, R. E., Salinas, C. F., Grossi, S. G., Sanders, J. J., Lopes-Virella, M. F. and Slate, E. H. (2009) Periodontal disease status in Gullah African Americans with type 2 diabetes living in South Carolina. *Journal of Periodontology*, **80**, 1062–1068.

Follmann, D. and Wu, M. (1995) An approximate generalized linear model with random effects for informative missing data. *Biometrics*, **51**, 151–168.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992) Model determination using predictive distributions with implementation via sampling-based methods. *Tech. rep.*, DTIC Document.

Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005) Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, **100**, 1021–1035.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2014a) *Bayesian Data Analysis.* Chapman & Hall/CRC Boca Raton, FL, USA, 2 edn.

Gelman, A., Hwang, J. and Vehtari, A. (2014b) Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, **24**, 997–1016.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472.

Gelman, A. et al. (2006) Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, **1**, 515–534.

Genton, M. G. (2004) *Skew-elliptical Distributions and Their Applications: A Journey Beyond Normality.* CRC Press, 1 edn.

Genton, M. G. and Zhang, H. (2012) Identifiability problems in some non-Gaussian spatial random fields. *Chilean Journal of Statistics*, **3**, 171–179.

Ghidey, W., Lesaffre, E. and Eilers, P. (2004) Smooth random effects distribution in a linear mixed model. *Biometrics*, **60**, 945–953.

Harshman, R. A., Green, P. E., Wind, Y. and Lundy, M. E. (1982) A model for the analysis of asymmetric data in marketing research. *Marketing Science*, **1**, 205–242.

Hornung, R. W., Greife, A. L., Stayner, L. T., Kyle Steenland, N., Herrick, R. F., Elliott, L. J., Ringenburg, V. L. and Morawetz, J. (1994) Statistical model for prediction of retrospective exposure to ethylene oxide in an occupational mortality study. *American Journal of Industrial Medicine*, **25**, 825–836.

Hosseini, F., Eidsvik, J. and Mohammadzadeh, M. (2011) Approximate Bayesian inference in spatial GLMM with skew-normal latent variables. *Computational Statistics & Data Analysis*, **55**, 1791–1806.

Irincheeva, I., Cantoni, E. and Genton, M. G. (2012) A non-Gaussian spatial generalized linear latent variable model. *Journal of Agricultural, Biological, and Environmental Statistics*, **17**, 332–353.

Jara, A., Quintana, F. and San Martín, E. (2008) Linear mixed models with skew-elliptical distributions: A bayesian approach. *Computational Statistics & Data Analysis*, **52**, 5033–5045.

Ji, C., Merl, D., Kepler, T. B. and West, M. (2009) Spatial mixture modelling for unobserved point processes: Examples in immunofluorescence histology. *Bayesian Analysis*, **4**, 297–316.

Jin, I. H., Yuan, Y. and Bandyopadhyay, D. (2016) A Bayesian hierarchical spatial model for dental caries assessment using non-Gaussian Markov random fields. *The Annals of Applied Statistics*, **10**, 884–905.

Johnson-Spruill, I., Hammond, P., Davis, B., McGee, Z. and Louden, D. (2009) Health of Gullah families in South Carolina with type 2 diabetes diabetes self-management analysis from project sugar. *The Diabetes Educator*, **35**, 117–123.

Kalman, R. E. (1963) Mathematical description of linear dynamical systems. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, **1**, 152–192.

Kauppinen, T. P. (1994) Assessment of exposure in occupational epidemiology. *Scandinavian Journal of Work, Environment & Health*, **20**, 19–29.

Keil, C. (2000) A tiered approach to deterministic models for indoor air exposures. *Applied Occupational and Environmental Hygiene*, **15**, 145–151.

Kerry, R. and Oliver, M. (2007) Determining the effect of asymmetric data on the variogram. i. underlying asymmetry. *Computers & Geosciences*, **33**, 1212–1232.

Kim, H.-M. and Mallick, B. K. (2004) A Bayesian prediction using the skew Gaussian distribution. *Journal of Statistical Planning and Inference*, **120**, 85–101.

Kinane, D. and Chestnutt, I. (2000) Smoking and periodontal disease. *Critical Reviews in Oral Biology & Medicine*, **11**, 356–365.

Lange, K. and Sinsheimer, J. S. (1993) Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, **2**, 175–198.

Li, D., Wang, X., Lin, L. and Dey, D. K. (2016) Flexible link functions in nonparametric binary regression with Gaussian process priors. *Biometrics*, **72**, 707–719.

Lin, T.-I. (2010) Robust mixture modeling using multivariate skew-t distributions. *Statistics and Computing*, **20**, 343–356.

McFadden, D. (1978) Modeling the choice of residential location. *Transportation Research Record*, 72–77.

Monteiro, J. V., Banerjee, S. and Ramachandran, G. (2014) Bayesian modeling for physical processes in industrial hygiene using misaligned workplace data. *Technometrics*, **56**, 238–247.

Müller, P. and Quintana, F. A. (2004) Nonparametric Bayesian data analysis. *Statistical Science*, **19**, 95–110.

Natarajan, R. and Kass, R. E. (2000) Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, **95**, 227–237.

Nathoo, F. S. and Ghosh, P. (2013) Skew-elliptical spatial random effect modeling for areal data with application to mapping health utilization rates. *Statistics in Medicine*, **32**, 290–306.

Neal, R. M. et al. (2011) MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, **2**, 113–162.

Neelon, B., Gelfand, A. E. and Miranda, M. L. (2014) A multivariate spatial mixture model for areal data: Examining regional differences in standardized test scores. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **63**, 737–761.

Neelon, B., Zhu, L. and Neelon, S. E. B. (2015) Bayesian two-part spatial models for semicontinuous data with application to emergency department expenditures. *Biostatistics*, **16**, 465–479.

Nelder, J. and Wedderburn, R. (1972) Generalized linear models. *Journal of the Royal Statistical Society*, **135**, 370–384.

Oliver, R. C., Brown, L. J. and Löe, H. (1998) Periodontal diseases in the United States population. *Journal of Periodontology*, **69**, 269–278.

Osborne, J. W. (2010) Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation*, **15**, 1–9.

Palacios, M. B. and Steel, M. F. J. (2006) Non-Gaussian Bayesian geostatistical modeling. *Journal of the American Statistical Association*, **101**, 604–618.

Pourabbas, R., Kashefimehr, A., Rahmanpour, N., Babaloo, Z., Kishen, A., Tenenbaum, H. C., Azarpazhooh, A., Mdala, I., Olsen, I., Haffajee, A. D. et al. (2005) Position paper: Epidemiology of periodontal diseases. *Journal of Periodontology*, **76**, 1406–1419.

Raftery, A. E., Newton, M. A., Satagopan, J. M. and Krivitsky, P. N. (2007) Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In *Bayesian Statistics 8* (eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West), vol. 8, 1–45. Oxford University Press.

Reich, B. J. and Bandyopadhyay, D. (2010) A latent factor model for spatial data with informative missingness. *The Annals of Applied Statistics*, **4**, 439–459.

Reich, B. J., Bandyopadhyay, D. and Bondell, H. D. (2013) A nonparametric spatial model for periodontal data with non-random missingness. *Journal of the American Statistical Association*, **108**, 820–831.

Reich, B. J. and Fuentes, M. (2007) A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *The Annals of Applied Statistics*, **1**, 249–264.

Rosa, G., Padovani, C. R. and Gianola, D. (2003) Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. *Biometrical Journal*, **45**, 573–590.

Sahu, S. K., Dey, D. K. and Branco, M. D. (2003) A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics*, **31**, 129–150.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639.

Taylor, G. W. and Borgnakke, W. (2008) Periodontal disease: Associations with diabetes, glycemic control and complications. *Oral Diseases*, **14**, 191–203.

Teeuw, W. J., Gerdes, V. E. and Loos, B. G. (2010) Effect of periodontal treatment on glycemic control of diabetic patients. *Diabetes Care*, **33**, 421–427.

Tsai, C., Hayes, C. and Taylor, G. W. (2002) Glycemic control of type 2 diabetes and severe periodontal disease in the US adult population. *Community Dentistry and Oral Epidemiology*, **30**, 182–192.

Tsonaka, R., Verbeke, G. and Lesaffre, E. (2009) A semi-parametric shared parameter model to handle nonmonotone non-ignorable missingness. *Biometrics*, **65**, 81–87.

Verbeke, G. and Lesaffre, E. (1996) A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, **91**, 217–221.

Wang, X. and Dey, D. K. (2010) Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption. *The Annals of Applied Statistics*, **4**, 2000–2023.

— (2011) Generalized extreme value regression for ordinal response data. *Environmental and Ecological Statistics*, **18**, 619–634.

Watanabe, S. (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**, 3571–3594.

Zeger, S. L. and Karim, M. R. (1991) Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.

Zhang, D. and Davidian, M. (2001) Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, **57**, 795–802.

Zhang, H. and El-Shaarawi, A. (2010) On spatial skew-Gaussian processes and applications. *Environmetrics*, **21**, 33–47.

Zhang, Y., Banerjee, S., Yang, R., Lungu, C. and Ramachandran, G. (2009) Bayesian modeling of exposure and airflow using two-zone models. *Annals of Ooccupational Hygiene*, **53**, 404–424.

Zhao, Y., Staudenmayer, J., Coull, B. A. and Wand, M. P. (2006) General design Bayesian generalized linear mixed models. *Statistical Science*, **21**, 35–51.

Zhou, X. and Schmidler, S. C. (2009) Bayesian parameter estimation in Ising and Potts models: A comparative study with applications to protein modeling. *Tech. rep.*, Technical report, Duke University.

Appendix A

Appendix for Chapter 2

A.1 Solution to 2.1

Given the differential equation that represents the well-mixed room model

$$\frac{dC}{dt} + (\frac{Q + K_L V}{V})C(t) = \frac{G + C_{IN}Q}{V},$$

we want to find the solution to $C(t)$. Let $a = \frac{Q+K_LV}{V}, b = \frac{G+C_{IN}Q}{V}$ where both a and b are constants. Multiplying both sides of (2.1) by exp(at), we get $\exp(\mathrm{at})C'(t) + a\exp(\mathrm{at})C(t) = b\exp(\mathrm{at})$ which is equivalent to

$$\frac{d}{d(t)}\exp(\mathrm{at})C(t) = b\exp(\mathrm{at}) \tag{A.1}$$

Take derivative of both sides of Equation (A.1), we can get $\int \frac{d}{d(t)} \exp(\text{at}) C(t) = \int b \exp(\text{at})$, which is equivalent to

$$
\begin{aligned}
&\exp(\text{at})C(t) = \frac{b}{a}\exp(\text{at}) + \text{constant}, \Rightarrow \\
&C(t) = \frac{b}{a} + \text{constant} \times \exp(-at), \Rightarrow \\
&C(0) = \frac{b}{a} + \text{constant}, \ \text{t=0. Therefore, constant} = C(0) - \frac{b}{a}, \Rightarrow \\
&C(t) = \frac{b}{a}\Big\{1 - \exp\Big(-\frac{Q+K_L V}{V}t\Big)\Big\}
\end{aligned}
$$

Therefore, $C(t) = \frac{G+C_{IN}Q}{Q+K_L V}\Big\{1 - \exp\Big(-\frac{Q+K_L V}{V}t\Big)\Big\} + C(0)\exp\Big(-\frac{Q+K_L V}{V}t\Big)$. In most cases including our chamber study, $C_{IN} = C(0)$. Therefore, the final solution can be written as: $C(t) = \frac{G+C(0)Q}{Q+K_L V}\Big\{1 - \exp\Big(-\frac{Q+K_L V}{V}t\Big)\Big\} + C(0)\exp\Big(-\frac{Q+K_L V}{V}t\Big)$.

A.2 Solution to (2.3)

Equation (2.3) can be simplified to produce the following unique solution (Zhang et al., 2009):

$$
\begin{aligned}
C_N(\boldsymbol{\theta};\mathbf{x},t) &= \tfrac{G}{Q} + \tfrac{G}{\beta} + G\left(\tfrac{\beta Q+\lambda_2 V_N(\beta+Q)}{\beta Q V_N(\lambda_1-\lambda_2)}\right)e^{\lambda_1 t} - G\left(\tfrac{\beta Q+\lambda_1 V_N(\beta+Q)}{\beta Q V_N(\lambda_1-\lambda_2)}\right)e^{\lambda_2 t}, \\
C_F(\boldsymbol{\theta};\mathbf{x},t) &= \tfrac{G}{Q} + G\left(\tfrac{\lambda_1 V_N+\beta}{\beta}\right)\left(\tfrac{\beta Q+\lambda_2 V_N(\beta+Q)}{\beta Q V_N(\lambda_1-\lambda_2)}\right)e^{\lambda_1 t} - G\left(\tfrac{\lambda_2 V_N+\beta}{\beta}\right)\left(\tfrac{\beta Q+\lambda_1 V_N(\beta+Q)}{\beta Q V_N(\lambda_1-\lambda_2)}\right)e^{\lambda_2 t}.
\end{aligned}
\tag{A.2}
$$

where λ_1 and λ_2 are the eigenvalues of $\mathbf{W}(\boldsymbol{\theta};\mathbf{x})$.

A.3 JAGS model and code

1. JAGS code to implement the Well-mixed model, where $\mu(0)$ is fixed at a nonzero value.

```
model {
a <- 1- (Q+ K_L * V)/V
b <- (G+ C_0*Q)/V
MU[1] ~ dnorm(a*C_0+b, inv_tau2)
for (i in 1:80){
Y[i] ~ dnorm(MU[i], inv_sigma2)
MU[i+1] ~ dnorm(a*MU[i]+b, inv_tau2)
}
K_L ~ dunif(0.00001, 0.06)
G ~ dunif(34, 100)
Q ~ dunif(0.0001, 0.5)
inv_sigma2 ~ dunif(20, 100)
inv_tau2 ~ dunif(20, 100)
sigma2 <- inverse(inv_sigma2)
tau2 <- inverse(inv_tau2)
}
```

2. `JAGS` code to implement the Well-mixed model, where $\mu(0)$ is fixed at 0, and the parameters have informative priors

```
model {
a <- 1- (Q+ K_L * V)/V
b <- G/V
MU[1] ~ dnorm(b, inv_tau2)
for (i in 1:80){
Y[i] ~ dnorm(MU[i], inv_sigma2)
MU[i+1] ~ dnorm(a*MU[i]+b, inv_tau2)
}
K_L ~ dunif(0.00001, 0.02)
G ~ dunif(41, 45)
Q ~ dunif(0.5, 0.3)
inv_sigma2 ~ dunif(20, 100)
inv_tau2 ~ dunif(20, 100)
sigma2 <- inverse(inv_sigma2)
tau2 <- inverse(inv_tau2)
}
```

3. `JAGS` code to implement the two-zone model, where $\boldsymbol{\mu}(0,0)$ has a multivariate rormal prior.

```
model
{
for (i in 1:70){
Y[i,1:2] ~ dmnorm(MU[i,1:2],V)
MU[i+1,1:2] ~ dmnorm(G1%*%MU[i,1:2] + g, W)
}
MU[1,1:2] ~ dmnorm(mu0, W)
beta~ dunif(0,14.5)
G~dunif(123, 136)
Q~dunif(0.2, 0.4)
V~dwish(r[,],3)
W~dwish(r[,],3)
mu0 ~ dmnorm(a, c0)
W.inv <- inverse(W)
V.inv <- inverse(V)
g[1]<-G/vn
g[2]<-0
G1<-I2[,]+W_theta[,]
W_theta[1,1]<-(-beta/vn)
W_theta[1,2]<-beta/vn
W_theta[2,1]<-beta/vf
W_theta[2,2]<-(-(beta+Q)/vf)
}
```

4. `JAGS` code to implement the two-zone model, where $\boldsymbol{\mu}(0,0)$ is fixed at a nonzero value.

```
model
{
for (i in 1:80){
Y[i,1:2] ~ dmnorm(MU[i,1:2],V)
MU[i+1,1:2] ~ dmnorm(G1%*%MU[i,1:2] + g, W)
}
MU[1,1:2] ~ dmnorm(mu0, inverse_c0)
beta~ dunif(0,14.5)
G~dunif(123, 127)
Q~dunif(0.28, 0.32)
V~dwish(r[,],3)
W~dwish(r[,],3)
W.inv <- inverse(W)
V.inv <- inverse(V)
g[1]<-G/vn
g[2]<-0
G1<-I2[,]+W_theta[,]
W_theta[1,1]<-(-beta/vn)
W_theta[1,2]<-beta/vn
W_theta[2,1]<-beta/vf
W_theta[2,2]<-(-(beta+Q)/vf)
}
```

A.4 Prior settings for the two-zone model

See Table A.1 for priors used in the data analysis for the two-zone model.

	Model		
Parameters	$\mathbf{C}(0) = (0,0)$	$\mathbf{C}(0) = (39,30)$	$\mathbf{C}(0) \sim$ MVN
β	$U(0,14.5)$	$U(0,14.5)$	$U(0,14.5)$
G	$U(123,127)$	$U(123,127)$	$U(123,127)$
Q	$U(0.28,0.32)$	$U(0.28,0.32)$	$U(0.28,0.32)$
$\mathbf{V}$	Wishart($\left(\begin{smallmatrix}10 & 0\\ 0 & 10\end{smallmatrix}\right)$, 4)	Wishart($\left(\begin{smallmatrix}10 & 0\\ 0 & 10\end{smallmatrix}\right)$, 4)	Wishart($\left(\begin{smallmatrix}10 & 0\\ 0 & 10\end{smallmatrix}\right)$, 4)
$\mathbf{W}$	Wishart($\left(\begin{smallmatrix}10 & 0\\ 0 & 10\end{smallmatrix}\right)$, 4)	Wishart($\left(\begin{smallmatrix}10 & 0\\ 0 & 10\end{smallmatrix}\right)$, 4)	Wishart($\left(\begin{smallmatrix}10 & 0\\ 0 & 10\end{smallmatrix}\right)$, 4)
$\mathbf{C}(0)$	NA	NA	MVN($\left(\begin{smallmatrix}0\\ 0\end{smallmatrix}\right)$, $\left(\begin{smallmatrix}10 & 0\\ 0 & 10\end{smallmatrix}\right)$)

Table A.1: Prior settings for two-zone models. It shows priors for 3 models with different assumptions for the initial contaminant concentrations. U is Uniform distribution and MVN stands multivariate Normal distribution

A.5 Analysis results of low and medium ventilation rate in Well-mixed model

Table A.2 shows analysis results from experimental data with informative priors when ventilation is low. At steady state, the concentration level is about 140 ppm, which is 600 mg/m^3. Therefore the initial state is set to be 45 mg/m^3. With informative priors, the value of $C(0)$ has little effect on the estimates or Monte Carlo standard error (MCSE). For the 2 scenarios with informative priors, the estimates are almost identical and the MCSE are also very close. The true values for both G (43.18) and Q (0.067) are included in the 95% CI. As for the 2 non-informative priors scenarios, Q is underestimated and the true value is included in the 95% CI. But G is not estimable with non-informative priors where the 95% CI doesn't include the true value.

	$C(0) = 0$, informative priors		$C(0) = 6$, informative priors	
	Estimate $(2.5\%, 97.5\%)$	MCSE	Estimate $(2.5\%, 97.5\%)$	MCSE
G	45.12 $(44.55, 45.33)$	1.53×10^{-3}	45.12 $(44.52, 45.33)$	1.61×10^{-3}
K_L	6.85×10^{-5} $(1.15 \times 10^{-5}, 0.000223)$	4.01×10^{-7}	6.99×10^{-5} $(1.16 \times 10^{-5}, 0.000227)$	4.27×10^{-7}
Q	6.44×10^{-2} $(6.38 \times 10^{-2}, 0.0662)$	4.88×10^{-6}	6.44×10^{-2} $(6.37 \times 10^{-2}, 0.0623)$	4.87×10^{-6}
σ^2	3.92×10^{-2} $(3.24 \times 10^{-2}, 0.0473)$	2.67×10^{-5}	3.93×10^{-2} $(3.24 \times 10^{-2}, 0.0472)$	2.68×10^{-5}
τ^2	5.12 $(4.19, 6.24)$	$3.70\times^{-3}$	5.13 $(4.19, 6.25)$	3.75×10^{-3}
	$C(0) = 0$, non-informative priors		$C(0) = 6$, non-informative priors	
	Estimate $(2.5\%, 97.5\%)$	MCSE	Estimate $(2.5\%, 97.5\%)$	MCSE
G	78.91 $(72.01, 85.77)$	2.50×10^{-2}	78.81 $(71.89, 85.81)$	2.46×10^{-2}
K_L	4.88×10^{-3} $(6.83 \times 10^{-4}, 0.00934)$	1.81×10^{-5}	4.41×10^{-3} $(6.87 \times 10^{-4}, 0.00904)$	1.77×10^{-5}
Q	5.25×10^{-2} $(3.32 \times 10^{-3}, 0.0979)$	2.08×10^{-4}	5.58×10^{-2} $(4.88 \times 10^{-3}, 0.0982)$	2.01×10^{-4}
σ^2	3.93×10^{-2} $(3.23 \times 10^{-2}, 0.0473\ )$	2.65×10^{-5}	3.93×10^{-2} $(3.24 \times 10^{-2}, 0.0472)$	2.69×10^{-5}
τ^2	2.90 $(2.39, 3.53)$	2.11×10^{-3}	2.91 $(2.39, 3.54)$	2.05×10^{-3}

Table A.2: Well-mixed model analysis results from experimental data with informative priors. Ventilation rate is Low. The true values are $G = 43.18$ and $Q = 0.067$. We fit the data to 4 competing models with different initial contaminant concentration and priors. The estimate and $(2.5\%, 97.5\%)$ are computed from the posterior sample. MCSE stands for Monte Carlo standard error for each model.

Analysis results for data from medium ventilation rate experimental setup are shown in Table A.3. The steady state concentration is around 178 mg/m^3 and the initials are set at 0 or 13.35 mg/m^3. With informative priors, the initial value of $C(t)$ doesn't affect the estimates or the Monte Carlo standard error (MCSE) where the estimates and MCSE are almost identical among these 2 scenarios. The true value for and $Q\ = 0.244$

is included in the 95% CI. The estimates of G and Q and are very close to the true values. As for the 2 non-informative priors scenarios, we still get similar estimates and MCSE. Q is underestimated and the true value is included in the 95% CI. But without enough prior information, G is not estimable.

	$C(0) = 0$, informative priors		$C(0) = 13.35$, informative priors	
	Estimate $(2.5\%, 95.5\%)$	MCSE	Estimate $(2.5\%, 95.5\%)$	MCSE
G	44.80 $(43.37, 45.33)$	2.67×10^{-3}	44.80 $(43.40, 45.32)$	3.71×10^{-3}
K_L	4.60×10^{-4} $(2.24 \times 10^{-5}, 0.00016)$	2.15×10^{-6}	4.46×10^{-4} $(4.46 \times 10^{-4}, 2.24 \times 10^{-5})$	3.07×10^{-6}
Q	0.24 $(0.23, 0.25)$	2.45×10^{-5}	0.24 $(0.23, 0.25)$	3.54×10^{-5}
σ^2	3.24×10^{-2} $(2.39 \times 10^{-2}, 0.044)$	2.53×10^{-5}	3.25×10^{-2} $(2.40 \times 10^{-2}, 0.044)$	3.55×10^{-5}
τ^2	3.44 $(2.52, 4.60)$	2.65×10^{-3}	3.45 $(2.54, 4.63)$	3.75×10^{-3}
	$C(0) = 0$, non-informative priors		$C(0) = 13.35$, non-informative priors	
	Estimate $(2.5\%, 97.5\%)$	MCSE	Estimate $(2.5\%, 97.5\%)$	MCSE
G	71.71 $(61.81, 82.11)$	3.78×10^{-2}	71.25 $(60.53, 81.50)$	3.79×10^{-2}
K_L	1.51×10^{-2} $(6.75 \times 10^{-4}, 0.0317)$	6.53×10^{-5}	1.48×10^{-2} $(7.20 \times 10^{-4}, 0.0316)$	6.31×10^{-5}
Q	0.189 $(9.47 \times 10^{-3}, 0.39)$	7.76×10^{-4}	0.190 $(9.36 \times 10^{-3}\ 0.38)$	7.59×10^{-4}
σ^2	0.032 $(2.39 \times 10^{-2}, 0.0438)$	3.56×10^{-5}	0.032 $(2.38 \times 10^{-2}, 0.0436)$	3.57×10^{-5}
τ^2	2.38 $(1.743.24)$	2.70×10^{-3}	2.40 $(1.763.26)$	2.71×10^{-3}

Table A.3: Well-mixed model analysis results from experimental data. Ventilation rate is Medium. The true values are $G = 43.18$ and $Q = 0.244$. We fit the data to 4 competing models with different initial contaminant concentration and priors. The estimate and $(2.5\%, 97.5\%)$ are computed from the posterior sample. MCSE stands for Monte Carlo standard error for each model.

A.6 Simulation study results for Well-mixed model

A.6.1 Prior Settings

In $C(0) = 0$ case, prior on G is Uniform(12, 18). Since when $C(0) = 0$, the term $K_L + G$ is identifiable but not the individuals. Both K_L and G can only be estimated from the priors. To get solid estimates, we need to use more information from the priors. So we put very informative priors on these two parameters where we have $K_L \sim U\ (0.04, 0.05)$ and $G \sim U\ (100, 110)$. We also put uniform priors on the precision. $\frac{1}{\sigma^2} \sim U\ (20, 80)$ and $\frac{1}{\tau^2} \sim U\ (20, 80)$. In the $C(0) = 0.03$ case, we have two sets of priors for K_L and G. We can put very informative priors on both K_L and G just as $C(0) = 0$ case. We also tried put $K_L \sim U(0.01, 0.09)$ and $G \sim U\ (73.5, 136.5)$ as in the non-informative priors scenario. The priors on precision are the same as the $C(0) = 0$ case.

A.6.2 Simulation Schemes and Results

The following tables in the section are simulation results from the well-mixed room model where 100 independent samples were drawn and analyzed in every simulation setting. The estimates, 95% CI and MCSE are summarized from 1 of the 100 samples. Relative bias, MSE and coverage probability are based on the 100 independent samples. Table A.4 shows results when we assume $C(0) = 0$ with informative priors. In general we have good coverage probability for all the parameters. As we can see from Table A.4, the true value for τ^2 is not available. That's because when generating C_t, it's not from a normal distribution therefore no variance is need to finish the job. The true parameter values for G, K_L, Q and σ^2 are included in the 95% posterior intervals.

Table A.5 shows the simulation results from the well-mixed room model where we assume $C(0) = 0$ with non-informative priors. The coverage probabilities are good for all the parameters. The true parameter values for G, K_L, Q and σ^2 are included in the

Param.	Estimate (2.5%, 97.5%)	MCSE	RB	MSE	Cover.
$G(105)$	104.54 $(100.21, 109.61)$	1.972×10^{-2}	-3.56×10^{-3}	1.456×10^{1}	100
$K_L(0.05)$	0.0495 $(0.0406, 0.0592)$	3.89×10^{-5}	-9.88×10^{-4}	4.96×10^{-6}	100
$Q(15)$	15.022 $(12.338, 17.667)$	1.032×10^{-2}	-3.29×10^{-3}	6.03×10^{-1}	100
$\sigma^2(0.030)$	0.30 $(0.025, 0.036)$	2.1×10^{-5}	-5.74×10^{-3}	8.72×10^{-4}	95
$\tau^2(-)$	0.0101 $(0.010, 0.0104)$	7.139×10^{-7}	-	-	-

Table A.4: Well-mixed model simulation results from 100 independent samples using $C(0) = 0$ with informative priors on the parameters. Param. is the parameter and the true values are shown inside the parenthesis. Posterior mean and (2.5%, 97.5%) credible interval are computed from the posterior samples. MCSE is the Monto Carlo standard error. RB stands for relative bias , MSE stands for mean square error, cover. is the coverage probability.

95% posterior intervals.

Param.	Estimate (2.5%, 97.5%)	MCSE	Relative bias	MSE	Cover.
$G(105)$	97.85 $(79.10, 119.626)$	7.41×10^{-2}	-0.05	2751	100
$K_L(0.05)$	4.24×10^{-2} $(1.71 \times 10^{-2}, 7.07 \times 10^{-2})$	9.67×10^{-5}	-1.09×10^{-1}	3.08×10^{-3}	100
$Q(15)$	14.98 $(12.17, 17.85)$	1.22×10^{-2}	-1.22×10^{-3}	0.72	100
$\sigma^2(0.030)$	2.89×10^{-2} $(2.38 \times 10^{-2}, 3.52 \times 10^{-2})$	2.08×10^{-5}	-4.95×10^{-3}	1.01×10^{-3}	94
$\tau^2(-)$	1.01×10^{-2} $(1.00 \times 10^{-2}, 1.011 \times 10^{-2})$	7.33×10^{-7}	-	-	-

Table A.5: Well-mixed model simulation results from 100 independent samples using $C(0) = 0$ with non-informative priors on the parameters. Param. is the parameter and the true values are shown inside the parenthesis. Posterior mean and (2.5%, 97.5%) credible interval are computed from the posterior samples. MCSE is the Monto Carlo standard error. RB stands for relative bias , MSE stands for mean square error, cover. is the coverage probability.

Table A.6 shows results when we assume $C(0) = 0.03$ with informative priors. The coverage probabilities are still goof for all the parameters. As in Table A.6, The true parameter values for G, K_L, Q and σ^2 are included in the 95% posterior intervals.

Param.	Estimate $(2.5\%, 97.5\%)$	MCSE	RB	MSE	Cover.
$G(105)$	104.686 $(100.249, 109.641)$	1.96×10^{-2}	-3.58×10^{-3}	14.56	100
$K_L(0.05)$	0.050 $(0.0407, 0.0594)$	3.95×10^{-5}	-6.98×10^{-5}	4.85×10^{-6}	100
$Q(15)$	14.914 $(12.306, 17.615)$	1.05×10^{-2}	-4.10×10^{-3}	7.63×10^{-1}	100
$\sigma^2(0.03)$	0.027 $(0.022, 0.032)$	1.89×10^{-5}	-6.74×10^{-3}	7.71×10^{-4}	95
$\tau^2(-)$	0.01010 $(0.01000, 0.0104)$	7.16×10^{-7}	-	-	-

Table A.6: Well-mixed model simulation results from 100 independent samples using $C(0) = 0.03$ with informative priors on the parameters. Param. is the parameter and the true values are shown inside the parenthesis. Posterior mean and $(2.5\%, 97.5\%)$ credible interval are computed from the posterior samples. MCSE is the Monto Carlo standard error. RB stands for relative bias , MSE stands for mean square error, cover. is the coverage probability.

Table A.7 shows results when we assume $C(0) = 0.03$ with informative priors. The coverage probabilities are good for all the parameters. As in Table A.7, The true parameter values for G, K_L, Q and σ^2 are included in the 95% posterior intervals. However, the estimates for G is 97.85 which is much smaller than the true value (105). This shows that when the priors don't provide enough information, we are unlikely to get a estimate that is close to the true value. This condition also holds when we assume $C(0) = 0.03$ with non-informative priors.

In general, the model with informative priors has a smaller MSE compare to that with non-informative priors. For example, in the $C(0) = 0$ case, the values for MSE in Table A.4 is much smaller than that in Table A.5.

Table A.8 shows model comparison results where DIC is used. In general, the model has a better fit when the initial value for $C(t)$ is not 0. Based on the DIC values shown in Table A.8, the model with non-zero initial $C(0)$ and non-informative priors has the best fit among all the 4 models.

Param.	Estimate (2.5%, 97.5%)	MCSE	RB	MSE	Cover.
$G(105)$	100.47 $(80.25, 120.32)$	7.305×10^{-2}	-0.05	3068	100
$K_L(0.05)$	0.0446 $(0.0172, 0.0702)$	9.431×10^{-5}	-1.16×10^{-1}	3.5×10^{-3}	100
$Q(15)$	15.15 $(12.16, 17.86)$	1.224×10^{-2}	-3.22×10^{-4}	7.47×10^{-2}	100
$\sigma^2(0.03)$	0.0262 $(0.0215, 0.0318)$	1.878×10^{-5}	-5.34×10^{-3}	1.08×10^{-3}	97
$\tau^2(-)$	0.0101 $(0.010, 0.0104)$	7.163×10^{-7}	-	-	-

Table A.7: Well-mixed model simulation results from 100 independent samples using $C(0) = 0.03$ with non-informative priors on the parameters. Param. is the parameter and the true values are shown inside the parenthesis. Posterior mean and $(2.5\%, 97.5\%)$ credible interval are computed from the posterior samples. MCSE is the Monto Carlo standard error. RB stands for relative bias , MSE stands for mean square error, cover. is the coverage probability.

Model	DIC	$\bar{D}$	p_D	$D(\bar{\boldsymbol{\theta}})$
$C(0) = 0$, informative priors	-685.1	-687.2	2.038	-689.238
$C(0) = 0$, non-informative priors	-689.3	-692.2	2.836	-695.036
$C(0) = 0.3$, informative priors	-701.7	-703.8	2.052	-705.852
$C(0) = 0.3$, non-informative priors	-709.9	-712.8	2.875	-715.675

Table A.8: Bayesian model comparison with DIC for the simulated WMR model data. We compare 4 competing models with different initial contaminant concentrations and priors. Model with the smallest DIC has the best fit.

A.7 Model comparison for well-mixed room model under different random errors assumptions

Model	Assump.	Low Vent. Rate			Medium Vent. Rate			High Vent. Rate		
		$\overline{D}(\theta)$	p_D	DIC	$\overline{D}(\theta)$	p_D	DIC	$\overline{D}(\theta)$	p_D	DIC
M1	N	821.7	2.037	823.7	285.9	2.099	288	141.7	2.296	144
	1SN	820	3.124	824	287.5	2.375	289.9	143.1	2.507	145.6
	2SN	822.8	3.156	825.9	289.4	2.45	291.9	144.8	5.326	144.6
	1ST	784.6	43.93	828.6	241.4	65.47	306.9	128.5	25.87	154.3
	2ST	552.4	406.6	959	172	144.3	316.3	114.3	34.32	148.6
M2	N	707.2	3.88	711.1	255.1	4.013	259.1	137.2	3.759	141
	1SN	706.4	5.063	711.5	256.7	4.142	260.9	138.7	3.946	142.6
	2SN	706.3	7.055	713.3	257.9	5.134	263.1	139.3	5.326	144.6
	1ST	669.1	47.07	716.2	211.3	47.83	259.1	121.2	18.48	139.7
	2ST	630.5	94.79	725.3	173.4	90.99	264.4	115.9	29.46	145.4
M3	N	821.7	2.037	823.7	286	2.096	288.1	141.7	2.274	143.9
	1SN	809.9	3.147	813	280.1	2.413	282.5	141.1	2.531	143.6
	2SN	803.8	3.201	807	281.8	2.448	284.2	142.5	2.746	145.3
	1ST	772.9	44.52	817.4	232.2	61.82	294.2	124.5	17.83	142.3
	2ST	864.9	79.54	944.5	174.9	140	314.9	115	31.38	146.4
M4	N	707.4	3.983	711.4	255.3	3.985	259.2	136.7	3.406	140.1
	1SN	706.4	5	711.4	256.8	4.231	261	138.6	4.002	142.6
	2SN	780.9	4.165	785.1	278.2	3.358	281.6	138.3	3.447	141.7
	1ST	669.3	47.6	716.9	210.7	47.19	257.9	122.6	19.36	142
	2ST	823.2	74.9	898.1	174.4	103.7	278.1	115.5	29.32	144.8

Table A.9: Values of posterior predictive model choice criterion for WMR model with low, medium, and high ventilation rates. M1-M4 are 4 model setups that specify the initial concentration level and priors. For each setup, DIC, is calculated from the posterior samples from 5 competing models with different assumptions for the measurement errors.

		High vent. rate			
Assumption	DF	M1	M2	M3	M4
1ST	ν_1	55.95	54.97	54.97	56.7
2ST	ν_1	57.45	56.26	57.42	56.23
	ν_2	52.17	54.87	53.14	58.19

Table A.10: Degree of freedom estimates for the skew-t models, ν_1 denotes the df of observed concentration level Y and ν_2 is the df of it's mean concentration level C.

Model	Assump.	LPML	Bayesian p-value	L-measure		
				G	P	D
M1	N	39.63	0.467	7.88	7.86	15.74
	1SN	39.62	0.469	7.80	7.87	15.67
	2SN	39.62	0.465	7.81	7.86	15.67
	1ST	41.41	0.486	5.64	7.78	13.42
	2ST	54.22	0.59	6.37	8.31	14.68
M2	N	39.63	0.470	7.88	7.86	15.74
	1SN	39.62	0.468	7.80	7.869	15.669
	2SN	39.62	0.468	7.79	7.866	15.656
	1ST	41.21	0.49	5.6	7.78	13.38
	2ST	41.4	0.48	5.6	7.77	13.37
M3	N	39.63	0.469	7.88	7.87	15.75
	1SN	39.63	0.455	7.89	8.10	15.99
	2SN	39.62	0.49755	7.80	7.87	15.67
	1ST	41.46	0.481	5.6832	7.7764	13.459
	2ST	54.49822	0.60075	6.3723	8.3175	14.689
M4	N	39.62554	0.464	7.8798	7.8564	15.736
	1SN	39.62584	0.4511	7.87	7.82	15.69
	2SN	39.62355	0.4704	7.7985	7.8722	15.671
	1ST	41.34306	0.48537	5.753	7.806	13.56
	2ST	54.43791	0.5959	6.335773	8.29972	14.63549

Table A.11: Values of posterior predictive model choice criterion for WMR model with low ventilation rate. M1-M4 are 4 model setups that specify the initial concentration level and priors. For each setup, LPML, Bayesian p-value, and L- measure are calculated from the posterior samples from 5 competing models with different assumptions for the measurement errors.

Model	Assumption	LPML	Bayesian p-value	L-measure		
				G	P	D
M1	N	23.50	0.454	2.604	2.596	5.3
	1SN	23.49	0.456	2.642	2.656	5.298
	2SN	23.50	0.452	2.649	2.650	5.299
	1ST	25.250	0.562	1.086	2.626	3.712
	2ST	31.874	0.485	1.001	2.267	3.268
M2	N	23.504	0.454	2.610	2.60	5.21
	1SN	23.499	0.452	2.657	2.655	5.312
	2SN	23.499	0.456	2.650	2.66	5.31
	1ST	30.065	0.488	1.010	2.823	3.833
	2ST	30.376	0.493	1.025	2.30	3.325
M3	N	23.504	0.455	2.604	2.598	5.202
	1SN	23.499	0.485	2.601	2.642	5.243
	2SN	23.498	0.452	2.651	2.651	5.302
	1ST	25.958	0.576	1.098	2.729	3.827
	2ST	31.5935	0.489	0.985	2.261	3.247
M4	N	23.504	0.451	2.607	2.593	5.201
	1SN	23.498	0.454	2.648	2.656	5.303
	2SN	23.499	0.449	2.651	2.654	5.305
	1ST	25.087	0.561	1.076	2.601	3.677
	2ST	31.523	0.490	0.980	2.259	3.239

Table A.12: Values of posterior predictive model choice criterion for WMR model with medium ventilation rate. M1-M4 are 4 model setups that specify the initial concentration level and priors. For each setup, LPML, Bayesian p-value, and L- measure are calculated from the posterior samples from 5 competing models with different assumptions for the measurement errors.

Appendix B

Appendix for Chapter 3

B.1 Parametric forms of the SNI distributions

We now introduce the members of the SNI class.

(i) *Multivariate skew-normal (SN) distribution.* This is the case when $H = 1$ (degenerate random variable) in (3.3).

(ii) *Multivariate skew-t (ST) distribution.* This is derived from (3.3) by taking $H =$ Gamma$(\nu/2, \nu/2)$, $\nu > 0$ and is denoted as $St_{p,p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \nu)$. The pdf of $\mathbf{Y}$ (Lin, 2010) is:

$$f(\mathbf{y}) = 2^p \, t_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Omega}, \nu) T_p\left(\sqrt{\frac{p+\nu}{d+\nu}}\mathbf{A}; \boldsymbol{\Delta}, \nu + p\right), \quad \mathbf{y} \in \mathbb{R}^p, \tag{A-1}$$

where $\mathbf{A} = \boldsymbol{\Lambda}^\top \boldsymbol{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ and $d = (\mathbf{Y} - \boldsymbol{\mu})^\top \boldsymbol{\Omega}^{-1}(\mathbf{Y} - \boldsymbol{\mu})$ is the Mahalanobis distance, $t_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ denotes the p-dimensional multivariate Student-t distribution with location $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$ and degrees of freedom (df) ν, and $T_p(\cdot; \boldsymbol{\Sigma}; \nu)$ is the cdf of $t_p(\cdot; \mathbf{0}, \boldsymbol{\Sigma}, \nu)$. A particular case of the skew–t distribution is the skew–Cauchy distribution, when $\nu = 1$. Also, when $\nu \uparrow \infty$, we have the SN distribution as the limiting case.

(iii) *Multivariate skew-slash (SSL) distribution.* It is derived from (3.3), choosing

$H = \text{Beta}(\nu, 1)$, $\nu > 0$. It is denoted by $SSL_{p,p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \nu)$ and the p.d.f is given by

$$f(\mathbf{y}) = 2^p \nu \int_0^1 u^{\nu-1} \phi_p(\mathbf{y}; \boldsymbol{\mu}, u^{-1}\boldsymbol{\Omega}) \Phi_p(u^{1/2}\mathbf{A}; \boldsymbol{\Delta}) du, \quad \mathbf{y} \in \mathbb{R}^p. \tag{A-2}$$

The SL distribution reduces to the SN distribution when $\nu \uparrow \infty$.

(iv) *Multivariate skew contaminated normal (SCN) distribution.* This arises when H takes one of two states, i.e. either ν_2 or 1, resepctive probabilities ν_1 and $1 - \nu_1$. with $\boldsymbol{\nu} = (\nu_1, \nu_2)^\top$. It is denoted by $SCN_{p,p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, \nu_1, \nu_2)$. The probability function of U is

$$h(u|\boldsymbol{\nu}) = \nu_1 \mathbb{I}_{\{\nu_2\}}(u) + (1 - \nu_1)\mathbb{I}_{\{1\}}(u), \;\; 0 < \nu_1 < 1, \, 0 < \nu_2 \leq 1. \tag{A-3}$$

It then follows that

$$f(\mathbf{y}) \;\;=\;\; 2^p \left\{ \nu_1 \phi_p(\mathbf{y}; \boldsymbol{\mu}, \nu_2^{-1}\boldsymbol{\Omega}) \Phi_p(\nu_2^{1/2}\mathbf{A}; \boldsymbol{\Delta}) + (1 - \nu_1)\phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Omega}) \Phi_p(\mathbf{A}; \boldsymbol{\Delta}) \right\}.$$

Parameter ν_1 can be interpreted as the proportion of outliers, while ν_2 may be interpreted as a scale factor. The SCN distribution reduces to the SN distribution when $\nu_2 = 1$.

B.2 `WinBUGS` code for implementing the best-fitting skew-t model

```
model{ # Skew-t CAL model:

pi <- 22/7

for(s in 1:nsubs){
  eta[s] ~ dgamma(v1,v1)
```

```
   for(i in 1:n){
   CAL[s,i]~dnorm(muCAL[s,i],taueC)
   CAL1[s,i]~dnorm(muCAL[s,i],taueC)
   muCAL[s,i]<- theta1[s,i] + aC
   # for symmetric models
   # theta1[s,i] <- mn[s,i] + sqrt(1/eta[s])*(theta[s,i])
   # for skew models
   theta1[s,i] <-  mn[s,i] + sqrt(1/eta[s])*(theta[s,i] +
           delta*abs(z[s,i]))- sqrt(2/pi)*delta*k1
   z[s,i] ~ dnorm(0,1)
   #Covariates
   mn[s,i]<- inprod(x[s,],beta[]) + gap*GAP[i] + can*canine[i]
     + premol*premolar[i] + mol*molar[i] + maxl*maxilla[i]
       }
      }

# Reparameterize Intercept, given that some of the covariates
were mean-subtracted to assist convergence

Int <-  aC - beta[1]*valA - beta[3]*valB

# t model
 k1 <- sqrt(v/2)*exp(loggam((v-1)/2))/exp(loggam(v/2))
 v1 <- v/2
 v ~ dgamma(0.1,0.01)
# Skew prior; note this is "lambda" in the paper
 delta ~ dnorm(0,0.01)

# Missing tooth model:
 for(s in 1:nsubs){for(t in 1:nteeth){
 MISS[s,t] ~ dbern(pmiss[s,t])
 pmiss[s,t]<-max(0.001,min(0.999,pmiss1[s,t]))
 logit(pmiss1[s,t])<-a0+b0*(theta1[s,TOOTH[t,1]]
 + theta1[s,TOOTH[t,2]] + theta1[s,TOOTH[t,3]]
 + theta1[s,TOOTH[t,4]] + theta1[s,TOOTH[t,5]]
   + theta1[s,TOOTH[t,6]])/6
   } }

# Implementing the inbuilt proper CAR prior in WinBUGS
   for(s in 1:nsubs){
 theta[s,1:n]~ car.proper(zero[s,], C[], adj[], num[], m[],
     tau2.car, rho)
  for(i in 1:n){ zero[s,i]<- 0 }}

# setting up the CAR adjacency matrix:
  for(i in 1:n){m[i] <- 1/num[i]}
  cumsum[1] <- 0
  for(i in 2:(n+1)) {cumsum[i] <- sum(num[1:(i-1)])}
  for(k in 1:N){
   for(i in 1:n){
pick[k,i] <- step(k - cumsum[i] - epsilon)*step(cumsum[i+1] - k)}
C[k] <-  1 / inprod(num[], pick[k,])}
epsilon <- 0.0001

# Priors and hyper-priors
for(j in 1:5){beta[j]~dnorm(0,0.01)}
aC ~ dnorm(0, 0.01)
gap ~ dnorm(0, 0.01)
can ~ dnorm(0, 0.01)
premol ~ dnorm(0, 0.01)
mol ~ dnorm(0, 0.01)
maxl ~ dnorm(0,0.01)

#Prior on spatial association
rho ~ dunif(0.95,1)
#Prior on CAR variance
sigmasq.car <- 1/tau2.car
tau2.car ~ dgamma(0.1,0.01)
#Prior on within-subject variance
sigmasq.eC<- 1/taueC
taueC ~ dgamma(0.1,0.01)
# Prior on the intercept and slope for the
non-random missingness regression
a0~ dnorm(0, 0.01)
b0~ dnorm(0, 0.01)
}
```

Appendix C

Appendix for Chapter 4

C.1 Update latent variable $\boldsymbol{\mu}_i$ using Hamiltonian Markov chain

Hamiltonian dynamics is a two dimensional systems which is consisted of potential energy $U(q)$ and kinetic energy $K(p)$, where p is a auxiliary variable represents momentum. The Hamiltonian function can be written as $H(p, q) = U(q) + K(p)$ (as in (C.2) and C.3). As for our high-dimensional multivariate latent random variable μ_i under a spatial framework, we want to sample $\boldsymbol{\mu}_i$ from it's posterior distribution:

$$p(\boldsymbol{\mu}_i \,|\, \text{Data}) \propto \pi(\boldsymbol{\mu}_i) \times p(y_i \,|\, \boldsymbol{\mu}_i) \times p(y_{i0} \,|\, \boldsymbol{\mu}_i) \tag{C.1}$$

Let $U(\boldsymbol{\mu}_i)$ denote the potential energy function for subject i, and it's given by:

$$\begin{aligned} U(\boldsymbol{\mu}_i) &= -\log\Big[\pi(\boldsymbol{\mu}_i) \times L(\boldsymbol{\mu}_i \,|\, D)\Big] \\ &= -\log\Big[\text{MVN}(\boldsymbol{\mu}_i \,|\, \mathbf{x}_i^\top \boldsymbol{\beta} + \boldsymbol{\omega}\boldsymbol{\alpha}, \Sigma(\rho, \sigma^2))\Big] \\ &- \log\Big[(1 - \text{GEV}(-\boldsymbol{\mu}_i; \xi))^{\mathbf{y}_i} (\text{GEV}(-\boldsymbol{\mu}_i; \xi))^{\mathbf{1}-\mathbf{y}_i}\Big] \\ &- \log\Big[\Phi(a_0 + b_0^\top \mathbf{Z}_t^\top \boldsymbol{\mu}_i)^{\mathbf{y}_{i0}} (\mathbf{1} - \Phi(a_0 + b_0 \mathbf{Z}_t^\top \boldsymbol{\mu}_i))^{\mathbf{1}-\mathbf{y}_{i0}}\Big] \\ &= -\log L_1 - \log L_2 - \log L_3 \end{aligned} \tag{C.2}$$

Derive gradient of $U(\boldsymbol{\mu}_i)$:

$$\begin{aligned} \frac{\partial U(\boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}_i} &= -\frac{\partial \log L_1}{\partial \boldsymbol{\mu}_i} - \frac{\partial \log L_2}{\partial \boldsymbol{\mu}_i} - \frac{\partial \log L_3}{\partial \boldsymbol{\mu}_i} \\ &= \Sigma^{-1}\boldsymbol{\mu}_i - \Sigma^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta} + \omega\boldsymbol{\alpha}) \\ &- \mathbf{y}_i \frac{\text{gev}(-\boldsymbol{\mu}_i; \xi)}{1 - \text{GEV}(-\boldsymbol{\mu}_i; \xi)} + \big(\mathbf{1} - \mathbf{y}_i\big) \frac{\text{gev}(-\boldsymbol{\mu}_i; \xi)}{\text{GEV}(-\boldsymbol{\mu}_i; \xi)} \\ &- \mathbf{y}_{i0} \frac{\phi(a_0 + b_0^\top \mathbf{Z}_t^\top \boldsymbol{\mu}_i) b_0^\top \mathbf{Z}_t^\top}{\Phi(a_0 + b_0^\top \mathbf{Z}_t^\top \boldsymbol{\mu}_i)} + (1 - \mathbf{y}_{i0}) \frac{\phi(a_0 + b_0^\top \mathbf{Z}_t^\top \boldsymbol{\mu}_i) b_0^\top \mathbf{Z}_t^\top}{1 - \Phi(a_0 + b_0^\top \mathbf{Z}_t^\top \boldsymbol{\mu}_i)} \end{aligned} \tag{C.3}$$

The momentum function is defined as $K(\mathbf{p}) = \frac{\mathbf{p}^\top \mathbf{p}}{2}$, where $\mathbf{p} \sim \text{MVN}(\mathbf{0}, \mathbf{I})$. Neal et al. (2011) and Chen et al. (2014) both gave example of how to implement HMC using a leapfrog method and the algorithm is as follows:

$$p_i(t + \epsilon/2) = p_i(t) - (\epsilon/2) \times \frac{\partial U(q(t))}{\partial q_i} \tag{C.4}$$

$$q_i(t + \epsilon) = q_i(t) + (\epsilon) \times p_i(t + \epsilon/2) \tag{C.5}$$

$$p_i(t + \epsilon) = p_i(t + \epsilon/2) - (\epsilon/2) \times \frac{\partial U(q(t + \epsilon))}{\partial q_i} \tag{C.6}$$

There are 2 steps to implement 1 iteration with the algorithm. First, we follow the leapfrog algorithm to get a proposed q and the auxiliary variable p as shown in (C.4), (C.5) and (C.6). Then use the Metropolis to determine if we want to accept or reject the proposed state as the next state of the Markov chain (see a sample R code in Appendix C.2). When getting a proposed position variable q for the potential energy, it's of great importance to select a suitable leapfrog step size ϵ. When the step size is large, the Metropolis gives a low acceptance rate. And if the step size is too small, the chain moves slowly over the exploration area of the likelihood which can be a waste of computation time (Neal et al., 2011). After running multiple testing samples, we decided to start with the step size of $\epsilon = 0.0005$ which gives an acceptance rate between 75% and 95% dependent on subject i. We calculate the acceptance rate for each subject i every 500 iterations and adjust the step size if the acceptance rate dropped below 75% or increase above 95%.

C.2 R code for implementing Hamiltonian algorithm in R

The following R code is adapted from Neal et al. (2011).

```
update_mu_i <- function(sub, mu.star,accept.mu, r){
current_q <- t(mu.star[r-1,sub,])
q <- t(as.matrix(mu.star[r-1, sub,]) )
# independent standard normal variates
p <- t(as.matrix(c(rnorm(n.teeth,0,1) )))
current_p <- p
#update with leapfrog method
for (k in 1:L){
# Make a half step for momentum at the beginning
##update momentum
p = p - (epsilon[sub]/2) * calculate_gradient(mu_value=t(q))
# Make a full step for the position
q = q + epsilon[sub] * p
# Make a half step for the momentum, except at end
of trajectory
p = p - (epsilon[sub]/2) * calculate_gradient(mu_value=t(q))
} # close: for (k in 1:L)
# Negate momentum at end of trajectory to make
the proposal symmetric
p = -p
# Evaluate potential and kinetic energies at start
# and end of trajectory
current_U = calculate_U(t(current_q))
current_K = (current_p%*%t(current_p)) / 2
proposed_U = calculate_U(t(q))
proposed_K = (p%*%t(p)) / 2

# Accept or reject the state at end of trajectory, returning
#either the position at the end of the trajectory or the
#initial position calculate accept.rate for each mu

if (log(runif(1)) < (current_U-proposed_U+current_K-proposed_K))
{
accept.mu[sub] <- accept.mu[sub] + 1
mu.star[r,sub,] <- q  # accept
}else
{
mu.star[r,sub,]<- current_q
#reject, mu.star[r,i,] = mu.star[r-1,i,]
}
return(c(mu.star[r,sub,], accept.mu[sub]))
}
```