Spring 2020

Bayesian Analysis of Binary Diagnostic Tests and Panel Count Data

Chunling Wang

Bayesian Analysis of Binary Diagnostic Tests and Panel Count Data

by

Chunling Wang

Bachelor of Arts
Inner Mongolia University, 2010

Master of Arts
Sun Yat-sen University, 2012

Master of Science
University of South Carolina, 2017

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Statistics

College of Arts and Sciences

University of South Carolina

2020

Accepted by:

Xiaoyan Lin, Major Professor

John Grego, Committee Member

Lianming Wang, Committee Member

Bo Cai, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

Dedication

This dissertation is dedicated to my husband, Tengxing Wang, parents, Fucai Wang and Chengping Zhu, and younger brother and sister-in-law, Xiaowei Wang and Min Mao. A happy family!

Acknowledgments

The PhD life gives me a great opportunity to discover myself. In this cultivating journey, I am so lucky to meet and know so many interesting, knowledgeable and helpful people who see my growth as important as theirs.

First of all, I want to thank my academic advisors, Dr. Timothy Hanson and Dr. Xiaoyan Lin. Dr. Hanson supervised me with the first project. It was he who brought me into the Bayesian world, and it set the tone of my academic research. He led and enlightened me to this field. Dr. Lin supervised me with another two projects. She consistently encouraged me to tackle problems from different perspectives and clearly present ideas. I learned that being a researcher needs to pay more attention to details. I am so grateful to have these two advisors, open-minded, inspiring and caring.

Also, I would like to thank Dr. John Grego, Dr. Lianming Wang and Dr. Bo Cai for being my committee members, and giving me many insightful suggestions and valuable comments, especially Dr. Grego for helping me edit my dissertation manuscript. I want to thank all professors that I have taken courses with, all classmates and friends. Spending time with them was enjoyable and entertaining.

I am deeply indebted to my husband, Dr. Tengxing Wang. Without his understanding and encouragement, my achievement would not have been possible. I know the furthest distance in the world is not necessary the physical distance, but the closest distance is alway the appreciation of two minds. Being with an interesting soul I never fear darkness. I also want to show my gratitude to my parents, Fucai Wang and Chengping Zhu, and my brother and sister-in-law, Xiaowei Wang and Min Mao for their unconditional love and support.

Abstract

This dissertation mainly explores several challenging topics that arise in diagnostic tests and panel count data in the Bayesian framework. Binary diagnostic tests, particularly multiple diagnostic tests with repeated measures and diagnostic procedures with a large number of raters, are studied. For panel count data, most traditional methods only handle panel count data for a single type of recurrent event. In this dissertation, we primarily focus on the case with multiple types of recurrent events.

In Chapter 1, an introduction to the binary diagnostic tests data and panel count data is presented and related literature works are briefly reviewed. To make the dissertation more coherent for the later chapters, some preliminary theories and algorithms, for instance the Metropolis Hastings algorithm, are presented. Finally, an outline of the dissertation organization is put forward.

In Chapter 2, a model for multiple diagnostic tests, applied repeatedly over time on each subject, is proposed; gold standard data are not required. The model is identifiable with as few as three tests; and correlation among tests at each time point in the diseased and non-diseased populations, as well as across time points is explicitly included. An efficient Markov chain Monte Carlo (MCMC) scheme allows for straightforward posterior inference. The proposed model is broadly illustrated via simulations and scaphoid fracture data from a prospective study (Duckworth et al., 2012) is analyzed. In addition, omnibus tests constructed from individual tests in parallel and serial are considered.

In Chapter 3, a Bayesian hierarchical conditional independence latent class model for estimating sensitivities and specificities for a large group of tests or raters is

proposed, which is applicable to both with-gold-standard and without-gold-standard situations. Through the hierarchical structure, not only are the sensitivities and specificities of individual tests estimated, but also the diagnostic performance of the whole group of tests. For a small group of tests or raters, the proposed model is further extended by introducing pairwise covariances between tests to improve the fitting and to allow for more modeling flexibility. Correlation residual analysis is applied to detect any significant covariance between multiple tests. Just Another Gibbs Sampler (JAGS) implementation is efficiently adopted for both models. Three real data sets from literature are analyzed to explicitly illustrate the proposed methods.

In Chapter 4, a Bayesian semiparameteric approach is proposed to analyze panel count data for multiple types of recurrent events. For each type of event, the proportional mean model is adopted to model the mean count of the event, where its baseline mean function is approximated by monotone I-splines (Ramsay et al., 1988). Correlation between multiple events is modeled by common frailty terms and scale parameters. Unlike many frequentist estimating equation methods, our approach is based on the observed likelihood and makes no assumption on the relationship between the recurrent processes and the observation process. Under the Poisson process assumption, an efficient Gibbs sampler based on a novel data augmentation is developed for the MCMC sampling. Simulation studies show good estimation performance of the baseline mean functions and the regression coefficients; meanwhile the importance of including the scale parameter to flexibly accommodate the correlation between events is also demonstrated. Finally, a skin cancer data example is fully analyzed to illustrate the proposed methods.

In Chapter 5, a brief summary of the studies we have completed in the previous chapters is delivered and at the same time we put forward some ideas for future work in each topic covered.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Binary Diagnostic Tests

Diagnostic tests are defined as an instrument or procedure performed to determine the presence or absence of some condition of interest such as disease. The primary medical interest is to evaluate the performance of the diagnostic tests and to estimate the disease prevalence of a population. Depending on the measure of the disease data, diagnostic tests can provide binary (positive/negative) outcomes, ordinal categorical responses or continuous biomarker levels. This dissertation chiefly focuses on binary diagnostic tests. The main parameters to be studied include sensitivity, specificity and disease prevalence. When the disease status is known, the estimation becomes very straightforward. Pepe (2003) and Broemeling (2007) proposed evaluation approaches for gold standard tests using frequentist and Bayesian perspectives. However, on most occasions, the available diagnostic tests are imperfect and cannot discriminate completely between the diseased and non-diseased individuals. As a result, more attention and efforts have been devoted to the methods for handling non-gold-standard tests.

The latent class model (LCM), which can naturally link the test responses with the underlying unknown disease status, is widely accepted and updated (Walter and Irwig, 1988; Vacek, 1985; Torrance-Rynard and Walter, 1997; Yang and Becker, 1997; Qu et al., 1996; Dendukuri and Joseph, 2001; van Smeden et al., 2013; Collins and Albert, 2016; Albert and Dodd, 2004). The fundamental assumption of LCM extends

from conditional independence to conditional dependence. The covariance structure has also attracted much attention. Under the LCM scheme, several models have been proposed. Vacek (1985), Torrance-Rynard and Walter (1997), Yang and Becker (1997), and Jones et al. (2010), among others, directly incorporated conditional pairwise covariances between tests. Qu et al. (1996) adopted latent Gaussian random effects (GRE) to model the conditional dependence between multiple tests. Albert et al. (2001) proposed a finite mixture model (FM) to account for the dependence between tests. Many researchers have realized that when the structure of the dependence is misspecified, the estimation tends to be biased. As a result, Albert and Dodd (2004) provide some guidance on the utilization of these models.

One core question facing LCM is the identifiability issue. Identifiability cannot happen when the number of parameters in the model is greater than the degrees of freedom. Even with sufficient degrees of freedom, identifiability cannot be guaranteed. Gustafson et al. (2005) and Gustafson (2009) showed that estimation of parameters can proceed based on the natural constraints imposed by the model itself. In a more general situation, Jones et al. (2010) gave some instructions on determining whether the model design of the multiple diagnostic tests is identifiable. For Bayesian approaches, estimation can borrow information from informative priors, so identifiability will not be a problem if researchers can construct some informative priors from historical data, expert's experience and other resources.

Data from a longitudinal study can give more information than that from a cross-sectional study. Evaluating each individual at multiple time points strengthens the test characteristics, and enables exploration of possible temporal covariance. An obvious benefit is that it offers more degrees of freedom. Repeated measures for multiple tests over time can enable consistent estimation of test accuracy and disease prevalence in the absence of a gold-standard test (Engel et al., 2010; Jones et al., 2012; Cook et al., 2000; Norris et al., 2009). In Chapter 2, the proposed model deploys

multiple binary tests with repeated measures. One of the important conclusions is that our model is identifiable for as few as three tests. Without repeated measures, this identifiability property cannot hold for three tests or four tests since the number of parameters exceeds the degrees of freedom.

The diagnostic accuracy of an individual test or rater has a crucial impact on clinical decision making. The assessment of diagnostic accuracy for multiple tests or raters also merits more attention. Zhang et al. (2012) proposed a latent class model with crossed random effects for the subjects and raters to estimate the diagnostic accuracy of a group of raters. Nelson and Edwards (2008) defined a model based kappa statistic for measuring the agreement among raters for binary classification. Lin et al. (2018) described a modeling approach to assess each rater's diagnostic skills by linking the rater's binary decisions with patient true disease status through patient latent disease severity. Chapter 3 proposes a hierarchical model to estimate the group rating agreement through a concentration parameter in a beta distribution with the tests' sensitivities and specificities estimated as well. Modified conditional independence models and conditional dependence models are introduced to handle large- and medium- size groups of tests or raters.

1.2 Panel Count Data

Panel count data come from recurrent events, where no exact occurrence times are recorded but the counts of event occurrences during adjacent observation times are observed. Unlike recurrent event data which have complete information of continuously observed event times, panel count data only have several discrete observation times and occurrence counts between adjacent time points. Panel count data are very common in fields such as social science, epidemiology, demographic studies and medical studies. Examples of panel count studies include hospitalizations or tumor

occurrences in medical studies, disease infection studies and system break downs in reliability studies.

The primary interest of panel count data includes estimation of the mean/rate function, treatment comparison and regression analysis (covariate effects). Researchers also explore topics on variable selection, analysis of mixed recurrent events and analysis of panel count data arising from multi-state models (Sun and Zhao, 2013). In panel count data analysis literature, there is a comprehensive system of theory and approaches from parametric models to nonparametric and semiparametric models. The motivation for using the parametric approach is, where the mean count is under a Poisson process assumption and the likelihood function can be directly derived, making the statistical inference immediate (Albert, 1991; Lawless, 1987; Thall, 1988). However, the appropriate parametric model may not exist or information or data is unavailable to validate the parametric model assumption. In this circumstance, the nonparametric model and semiparametric model, which don't require strong assumptions, can provide a more robust analysis. For the nonparametric procedure, the main goal is to estimate the mean function. There are likelihood-based estimators, such as nonparametric maximum likelihood estimators (NPMLE) and nonparametric maximum pseudolikelihood estimators (NPMPLE) (Wellner et al., 2000; Lu et al., 2007), and regression-based estimators, for instance isotonic regression estimators (IRE) (Sun and Kalbfleisch, 1995; Hu et al., 2009). Semiparametric models have recently gained more popularity, where the main focus is on the regression analysis, i.e. the covariate effects on the mean model. This is also the partial objective of the project in Chapter 4. For the semiparametric procedures, the baseline function is mostly modeled by a piecewise function (Lawless and Zhan, 1998) or monotone splines (Lu et al., 2009); the covariate effects are mainly estimated through the proportional mean model. Lin et al. (2001) have proposed a semiparametric transformation model to relax the assumption that two sets of covariate values are proportional over time.

In the frequentist framework, the estimation procedure largely depends on the estimation equation method, and discussion of the relationship between the counting process and the observation process cannot be avoided. This dissertation provides a Bayesian approach to estimate the covariate effects and baseline mean function, where the relationship between the recurrent event process and the observation process does not require any assumption.

Studies on univariate panel count data are considerably mature, and some research has begun to investigate multivariate panel count data. A notable difference for the latter is that the correlation among the multiple related recurrent events needs to be taken into account. Two approaches that can be employed to analyze multivariate panel count data are the marginal model approach, where the correlation is left unspecified, and the joint model approach where the correlation can be modeled through some latent or random variables. In the literature, most work is devoted to the marginal model approach. He et al. (2008) and Chen et al. (2005) studied regression analysis for multivariate panel count data under the assumption that the recurrent event process and observation process are independent. Li et al. (2011) and Zhang et al. (2013) also studied regression analysis but assume that there is some dependence between the recurrent event process and the observation process. All the above work has employed estimation equation methods in the frequentist framework. It is hard for the estimation equation based approach to derive the theoretical properties and conduct calculations when directly imposing a correlation structure to the model (Sun and Zhao, 2013). On the contrary, the joint model approach is not too complicated when using the Bayesian approach. In fact, the model proposed in Chapter 4 studies the correlation between any two recurrent events through common frailty terms and a scale parameter and provides a compact expression for the correlation.

1.3 Computing techniques

Performing Bayesian inferences require use of the joint posterior distribution over a set of parameters. In practice, there is no closed form of the posterior distribution, and the intractable integrals are hard to derive in most situations. Instead of solving the analytical equations, Markov chain Monte Carlo (MCMC) provides a solution to obtain statistical inferences by sampling. The most commonly used sampling techniques are the Metropolis-Hastings (MH) sampling and the Gibbs sampling, where the later is a special case of the former (Gelman, 1993). The MH algorithm was first developed by Metropolis et al. (1953) and subsequently generalized by Hastings (1970). Because of its importance and usefulness, its application (Müller, 1991; Chib and Greenberg, 1996) is steadily increasing in the literature. To make this algorithm more intuitive and understandable, Chib and Greenberg (1995) provide a tutorial exposition of the MH algorithm. Basically, the goal of the MH algorithm is to draw samples from the desired target distribution π. It starts with a function $f(x) = \pi K$ which is proportional to the target distribution and a proposal distribution (or candidate kernel) $q(x)$ from which new candidates can be drawn. The sampling process then proceeds iteratively as follows.

1. Initialize $x^{(0)} \sim q(x)$
2. Propose $x^{(cand)} \sim q(x^{(i)}|x^{(i-1)})$, for $i = 1, 2, ...$
3. Accept the proposal $x^{(i)} = x^{(cand)}$ with acceptance rate α, where $\alpha = min\{1, \frac{\pi(x^{(cand)})q(x^{(i-1)}|x^{(cand)})}{\pi(x^{(i-1)})q(x^{(cand)}|x^{(i-1)})}\}$. If the candidate is rejected, $x^{(cand)} \sim x^{(n-1)}$.
4. Repeat steps 2-3 until convergence.

If the proposal distribution is symmetric, for instance the random walk algorithm where the proposal distribution is the normal distribution, the acceptance rate can be simplified into the form $\alpha = min\{1, \frac{\pi(x^{(cand)})}{\pi(x^{(i-1)})}\}$. In this way, the calculation becomes

much simpler and faster. The tricky question for a random walk algorithm is to determine the updating step size, i.e. σ. If the step size is too large, the candidate may skip the more probable value and the Markov chain tends to stick to some value for a long time. If the step size is too small, the chain converges slowly. In order to find the proper step size, pre-runs and manual tuning are necessary. For high dimensional MCMC, the tuning task is laborious. With the development of the algorithm, some adaptive approaches appear (Haario et al., 2001; Roberts et al., 2001; Haario et al., 2005) to automatically select the step size and to elevate the computation speed. Chapter 2 mainly adopts Haario et al. (2005) 's componentwise adaptation to MCMC due to its flexibility in programming. In the adaptive algorithm, the variance of the proposal distribution depends on the historical values of the chain and hence is not a Markovian technique. The formula used to calculate the variance is presented below:

$$v_i^t = \begin{cases} v_0^t & t \leq t_0, \\ sVar(x_i^0, x_i^1, \ldots, x_i^{t-1}) + \epsilon & t > t_0, \end{cases}$$

where v_0^t is the initial variance whose choice proves not critical (Haario et al., 2005). t_0 is regarded as a burn-in period (the author used $t_0 = 10$), after which the variance of the proposal is related to the variance of the historical values in the Markov chain. s is a scale parameter, and the author suggests using $s = 2.4$ (in this disseratation $s = 0.5$ is used which is empirically more efficient). Finally ϵ is a constant, and it is used to avoid the situation in which the variance shrinks to 0. Given the proposal distributions, the multiple dimensional MCMCs are updated componentwise.

Chapter 3 and Chapter 4 particularly use the Gibbs sampler approach for Bayesian inference. Chapter 3 also introduces the syntax of Just Another Gibbs Sampler (JAGS) and presents some sample code. The Gibbs sampler is a special form of MH (Gelman, 1993), and it is mainly used to generate random variables from a (marginal) distribution to avoid the direct calculation of density. Geman and Geman (1987) used this method to study image processing. Gelfand and Smith (1990) reviewed and

compared several sampling-based density calculation methods. Casella and George (1992) exploited simple cases to demystify the working mechanism behind the Gibbs Sampler. Since then, the Gibbs sampler approach has been widely used for posterior sampling. The basic idea is to sequentially draw samples for each random variable from the conditional distribution with the remaining variables fixed to the current values. For $X \in R^K$, the algorithm proceeds as follows:

1. Initialize $x^{(0)} \sim q(x)$, where q() is usually a prior distribution

2. for iteration $i = 1, 2, ...$
   $x_1^{(i)} \sim p(X_1 = x_1 | X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)}, ..., X_K = x_K^{(i-1)})$
   $x_2^{(i)} \sim p(X_2 = x_2 | X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}, ..., X_K = x_K^{(i-1)})$
   ...
   $x_K^{(i)} \sim p(X_K = x_K | X_1 = x_1^{(i)}, X_2 = x_2^{(i)}, ..., X_{K-1} = x_{K-1}^{(i)})$

3. end

JAGS provides an easy way to conduct Gibbs sampling without the need to derive the full conditional distributions. The syntax for JAGS is simple and straightforward, since only models and priors need to be directly specified. All the derivations of full conditional distributions and sampling from them are carried out automatically. Plummer (2017) provides further details.

1.4 Outline

The rest of this dissertation is organized as follows. Chapter 2 and Chapter 3 present two models to analyze binary diagnostic tests results. Chapter 4 proposes a model to analyze multivariate panel count data. Chapter 5 is a summary.

In Chapter 2, a model for multiple tests with repeated measures is proposed in a situation without a gold standard. The primary goal is to estimate the sensitivity,

specificity, prevalence, test covariance and temporal covariance. An efficient adaptive MCMC algorithm is developed. Extensive simulation studies empirically prove the identifiability of the proposed model. Traceplots from the real data analysis section show that all the Markov chains attain convergence. Lastly, some omnibus tests are constructed which can provide some instructions for practitioners on how to collectively use the imperfect tests.

In Chapter 3, a Bayesian hierarchical model is proposed to study the tests' characteristics at both the individual and group levels. Depending on the group size, conditional independence and conditional dependence can be assumed accordingly. For guiding the addition of covariance terms into the model, correlation residual analysis is proposed. In addition, four algorithms (with and without multinomial imposition) are developed. Clear and compact JAGS sample codes are provided. Lastly, the proposed models and residual analysis are broadly illustrated by simulation studies and analysis of real data from the literature.

In Chapter 4, a Bayesian approach for studying panel count data of multiple types of events is proposed, where the primary interest lies in the estimation of covariate effects and the correlation between multiple events. For modeling the mean count of each type of event the proportional mean model is adopted where the baseline function is approximated by monotone I-splines. Furthermore, the correlation between any two events is modeled by the common frailty term and a scale parameter. Under the Poisson process assumption, an efficient Gibbs sampler is developed based on augmented data. Finally, comprehensive simulation studies and a skin cancer data analysis are conducted to show the performance of the proposed method.

In the last chapter, a brief wrap-up and ideas for future work are presented.

Chapter 2

Estimation of Sensitivity and Specificity of Multiple Repeated Binary Tests without a Gold Standard

2.1 Introduction

Sensitivity and specificity are the primary measures of diagnostic test accuracy. When true disease status is known via a gold-standard test, estimation of sensitivity and specificity is straightforward. However, when no gold standard is available, or its use prohibitive due to cost or ethical concerns, it is still possible to consistently estimate test accuracy in some sampling situations under certain modeling assumptions. The pioneering work of Hui and Walter (1980) shed light on estimating error rates for two tests in two populations with differing prevalence. An important result from their paper is that when two tests are simultaneously applied to the same individuals across two different populations, the assumption of conditional independence of tests given disease status renders the model identifiable and maximum likelihood estimates of sensitivity, specificity, and prevalence are consistent; this result is immediately generalizable to more than two tests and more than two populations. When only one population is considered, the two test model is clearly not identifiable since the three degrees of freedom in the resulting 2×2 table is less than the five unknown parameters. One remedy for such overparameterized models is to impose additional constraints (Walter and Irwig, 1988), e.g. treat some of the parameters as known,

yielding an identifiable model and consistent maximum likelihood estimators. Joseph et al. (1995) considered this situation but from a Bayesian approach where constraints are instead replaced with informative priors for a subset of the parameters. Similarly, Johnson et al. (2001) revisited the approach of Hui and Walter (1980) from a Bayesian perspective and compared the two tests, two population model with the two tests, one population model of Joseph et al. (1995). They showed that when the model is identifiable Bayesian posterior inference would ultimately converge to true parameter values regardless of the quality of prior distribution; however, if the model is unidentifiable, reliable inference depends heavily on the validity and preciseness of the prior information. However, in some circumstances, learning on model parameters can proceed even in the unidentifiable case based on natural constraints imposed by the model itself (Gustafson, 2009; Gustafson et al., 2005).

Generalizations to more than one or two populations and more than two possibly conditionally dependent tests, allowing for increased flexibility in data collection and modeling assumptions, have been accomplished through the addition of appropriate parameters (Walter and Irwig, 1988; Vacek, 1985; Dendukuri and Joseph, 2001; Hanson et al., 2003). Alternatively, conditional dependence among binary diagnostic tests at one time point can be induced via a subject-specific random effect (Qu et al., 1996; Qu and Hadgu, 1998; Shih and Albert, 1999; Engel et al., 2010). Other information (Branscum et al., 2008; Jones et al., 2012; Luo et al., 2013, 2014) correlated with disease status can be incorporated to help focus prevalence and accuracy estimation in the absence of a gold-standard. Repeated measures over time (Engel et al., 2010; Jones et al., 2012; Cook et al., 2000; Norris et al., 2009) also enable consistent estimation of test accuracy and prevalence in the absence of a gold-standard.

In this chapter, under the assumption that the latent disease status is temporally static, a repeated measures model for multiple, conditionally dependent binary tests is developed starting from Jones et al. (2010). An important finding of our work is

that, under certain assumptions, only three tests are necessary to consistently estimate sensitivity, specificity, and prevalence in only one population. This is in marked contrast to non-longitudinal data. For simple cross-sectional data, the tenuous assumption of conditional independence is necessary for three diagnostic tests (Walter and Irwig, 1988). Even with multiple time points, when only one fallible test is considered, the conditional independence assumption is still unavoidable (Yanagawa and Gladen, 1984; Espeland et al., 1989); see Hui and Zhou (Hui and Zhou, 1998) for a review. Furthermore, our model explicitly models pairwise between– and within–test covariance, which extends covariance from only test level to both test and temporal level. Such explicit parameterization makes calculation of common loci of inference such as conditional correlation (between pairs of tests at one time point or between repeated applications of one test) straightforward, as well as the incorporation of missing tests, time points, etc.

For the diagnosis of scaphoid fractures, both clinical and radiological assessment play an important role. The diagnosis of scaphoid fractures is a longitudinal process because some occult fractures cannot be identified on the initial assessment. For initial assessment of the patients with a scaphoid injury, several clinical tests are commonly used by emergency room doctors or nurse practitioners (Duckworth et al., 2012). It is recognized that no clinical test can perfectly distinguish people with a scaphoid fracture from those without; hence there is no gold standard. The sensitivity of clinical assessment has been historically overestimated and the specificity underestimated, which leads to over-treatment; false positives are subject to unnecessary restriction and medical resources are not properly allocated. We consider scaphoid fracture data from a prospective study on adult patients presenting suspected scaphoid fracture within 72 hours of injury. Among the $n = 223$ patients considered, 205 were examined twice (and so have repeated measures) between 10 and 18 days after the initial injury; 18 were only examined initially (i.e. once).

Conditional on the disease status, the repeated measures model developed here considers dependence at both test level and temporal level separately. In particular, the model does not require subject-specific random effects to induce the longitudinal correlation. Based on several simple but realistic assumptions, we construct formulae for likelihood construction and then use a flexible component-wise adaptive Metropolis-Hastings algorithm to obtain posterior inference. This chapter is organized as follows: Section 2 develops the model; Section 3 presents a simulation study; an analysis on repeated measures scaphoid fracture data is given in Section 4 and a comparison of parallel and serial multiple tests is given in section 5. The last section is the discussion and conclusion.

2.2 Model

Consider K binary diagnostic tests administered over J time points on individual i, yielding vectors of test results $\mathbf{T}_{i1}, \ldots, \mathbf{T}_{iJ}$ where $\mathbf{T}_{ij} = (T_{ij1}, \ldots, T_{ijK})'$. Denote $T_{ijk} = 1$ as testing positive and $T_{ijk} = 0$ negative. Assume that the operating characteristics of the tests, namely the sensitivity and specificity, do not change over time, and each individual's disease status ($D+$ is diseased and $D-$ is non-diseased) is also static. Denote the sensitivity and specificity of test k as $Se_k = P(T_{ijk} = 1|D_i+)$ and $Sp_k = P(T_{ijk} = 0|D_i-)$ respectively for $j = 1, 2 \ldots, J$. Let C_{ij}^+ and C_{ij}^- denote the pairwise covariance between any two tests conditional on subjects diseased and non-diseased respectively. Following Vacek (1985) and Dendukuri and Dendukuri and Joseph (2001), tests 1 and 2 have the following joint probabilities at time j in the diseased population.

$$\begin{aligned} P(T_{ij1} = 1, T_{ij2} = 1|D_i+) &= Se_1Se_2 + C_{12}^+ \\ P(T_{ij1} = 1, T_{ij2} = 0|D_i+) &= Se_1(1 - Se_2) - C_{12}^+ \\ P(T_{ij1} = 0, T_{ij2} = 1|D_i+) &= Se_2(1 - Se_1) - C_{12}^+ \end{aligned}$$

$$P(T_{ij1} = 0, T_{ij2} = 0|D_i+) \quad = \quad (1 - Se_1)(1 - Se_2) + C_{12}^+$$

In the non-diseased population C_{12}^+ is replaced with C_{12}^-, Se_k replaced with $(1 - Sp_k)$ and $(1 - Se_k)$ replaced with Sp_k. Similar expressions follow for all pairs of tests. Repeated applications of test k over time are assumed to also be dependent conditional on an individual's disease status; exchangeable covariance is assumed and denoted R_k^+ in the diseased population and R_k^- in the non-diseased leading to the following joint probabilities involving the repeated use of test 1 across any two time points j_1 and j_2.

$$\begin{aligned}
P(T_{ij_1 1} = 1, T_{ij_2 1} = 1|D_i+) &= Se_1^2 + R_1^+ \\
P(T_{ij_1 1} = 1, T_{ij_2 1} = 0|D_i+) &= Se_1(1 - Se_1) - R_1^+ \\
P(T_{ij_1 1} = 0, T_{ij_2 1} = 1|D_i+) &= Se_1(1 - Se_1) - R_1^+ \\
P(T_{ij_1 1} = 0, T_{ij_2 1} = 0|D_i+) &= (1 - Se_1)^2 + R_1^+
\end{aligned}$$

As before, the non-diseased versions replace R_1^+ with R_1^-, Se_1 with $(1 - Sp_1)$, and $(1 - Se_1)$ with Sp_1. Similar expressions are obtained for the remaining tests.

The resulting model is of the general type studied by Jones et al. (2010), with important differences being that test sensitivity and specificity remain constant over time and a temporal covariance is introduced. Using the methods of Jones et al. (2010), local identifiability of the resulting model can be shown for as few as $K = 3$ tests under the reasonable assumption of $Se_k + Sp_k > 1$. This is remarkable as in the non-repeated measures case, the typically untenable assumption of conditional independence *must* be assumed to estimate the $K = 3$ sets of sensitivities and specificities, as well as the prevalence (Walter and Irwig, 1988).

Subject i's likelihood contribution is obtained through the law of total probability

$$P(\mathbf{T}_i = \mathbf{t}_i) = \pi P(\mathbf{T}_i = \mathbf{t}_i|D_i+) + (1 - \pi)P(\mathbf{T}_i = \mathbf{t}_i|D_i-),$$

where π is the disease prevalence and $\mathbf{T}_i = (\mathbf{T}_{i1}, \ldots, \mathbf{T}_{iJ})$, etc. Note that $\mathbf{T}_i$ has a distribution over a contingency table with 2^{KJ} elements, a 2^K contingency table at

each of J time points. Under conditional independence across time and among tests, for individual i, we have

$$d_i^+ = P(\mathbf{T}_i = \mathbf{t}_i | D_i +) = \prod_{k=1}^{K} Se_k^{t_{i\bullet k}} (1 - Se_k)^{J - t_{i\bullet k}},$$

$$d_i^- = P(\mathbf{T}_i = \mathbf{t}_i | D_i -) = \prod_{k=1}^{K} Sp_k^{J - t_{i\bullet k}} (1 - Sp_k)^{t_{i\bullet k}}.$$

Here, $\bullet$ indicates summation over that index. For a fixed time point, pairwise covariance terms are introduced. Generalizing (5) and (6) in Jones et al. (2010), the "test covariance positive" summed over all time levels is a simple modification of the independence case above

$$\begin{aligned}
tc_i^+ &= \sum_{j=1}^{J} \sum_{u<v} (-1)^{t_{iju} + t_{ijv}} C_{uv}^+ Se_u^{t_{i\bullet u} - t_{iju}} (1 - Se_u)^{J-1-t_{i\bullet u} + t_{iju}} Se_v^{t_{i\bullet v} - t_{ijv}} \\
&\quad (1 - Se_v)^{J-1-t_{i\bullet v} + t_{ijv}} \prod_{k \neq u,v} Se_k^{t_{i\bullet k}} (1 - Se_k)^{J - t_{i\bullet k}} \\
&= \left[\prod_{s=1}^{K} Sp_s^{J - t_{i\bullet s}} (1 - Se_s)^{t_{i\bullet s}} \right] \sum_{j=1}^{J} \sum_{u<v} C_{uv}^+ \frac{[(Se_u - 1)/Se_u]^{t_{iju}} (Se_v - 1)/Se_v]^{t_{ijv}}}{(1 - Se_u)(1 - Se_v)} \\
&= d_i^+ \sum_{u=1}^{K-1} \sum_{v=u+1}^{K} C_{uv}^+ \sum_{j=1}^{J} \frac{[(Se_u - 1)/Se_u]^{t_{iju}} [Se_v - 1)/Se_v]^{t_{ijv}}}{(1 - Se_u)(1 - Se_v)} \\
&= d_i^+ \widetilde{tc}_i^+ .
\end{aligned}$$

Similarly, the "test covariance negative" is

$$\begin{aligned}
tc_i^- &= \sum_{j=1}^{J} \sum_{u<v} (-1)^{t_{iju} + t_{ijv}} C_{uv}^- (1 - Sp_u)^{t_{i\bullet u} - t_{iju}} (Sp_u)^{J-1-t_{i\bullet u} + t_{iju}} (1 - Sp_v)^{t_{i\bullet v} - t_{ijv}} \\
&\quad Sp_v^{J-1-t_{i\bullet v} + t_{ijv}} \prod_{k \neq u,v} (1 - Sp_k)^{t_{i\bullet k}} Sp_k^{J - t_{i\bullet k}} \\
&= \left[\prod_{s=1}^{K} (1 - Sp_s)^{J - t_{i\bullet s}} Se_s^{t_{i\bullet s}} \right] \sum_{j=1}^{J} \sum_{u<v} C_{uv}^- \frac{[Sp_u/(Sp_u - 1)]^{t_{iju}} [Sp_v/((Sp_v - 1))]^{t_{ijv}}}{Sp_u Sp_v} \\
&= d_i^- \sum_{u=1}^{K-1} \sum_{v=u+1}^{K} C_{uv}^- \sum_{j=1}^{J} \frac{[Sp_u/(Sp_u - 1)]^{t_{iju}} [Sp_v/(Sp_v - 1)]^{t_{ijv}}}{Sp_u Sp_v} \\
&= d_i^- \widetilde{tc}_i^- .
\end{aligned}$$

Now consider the longitudinal aspect; the "repeated measures covariance positive" is

$$\begin{aligned} rc_i^+ &= \sum_{k=1}^{K}\sum_{u<v}(-1)^{t_{iuk}+t_{ivk}}R_k^+ Se_k^{t_{i\bullet k}-t_{iuk}-t_{ivk}}(1-Se_k)^{J-t_{i\bullet k}-2+t_{iuk}+t_{ivk}} \\ &\quad \prod_{s\neq k} Se_s^{t_{i\bullet s}}(1-Se_s)^{J-t_{i\bullet s}} \\ &= \left[\prod_{s=1}^{K} Se_s^{t_{i\bullet s}}(1-Se_s)^{J-t_{i\bullet s}}\right]\sum_{k=1}^{K} R_k^+ \sum_{u<v}\frac{[(Se_k-1)/Se_k]^{t_{iuk}+t_{ivk}}}{(1-Se_k)^2} \\ &= d_i^+ \sum_{k=1}^{K} R_k^+ \sum_{u=1}^{J-1}\sum_{v=u+1}^{J}\frac{[(Se_k-1)/Se_k]^{t_{iuk}+t_{ivk}}}{(1-Se_k)^2} \\ &= d_i^+ \widetilde{rc}_i^+ \end{aligned}$$

Similarly, the "repeated measures covariance negative" is

$$\begin{aligned} rc_i^- &= \sum_{k=1}^{K}\sum_{u<v}(-1)^{t_{iuk}+t_{ivk}}R_k^- Sp_k^{J-t_{i\bullet k}-2+t_{iuk}+t_{ivk}}(1-Sp_k)^{t_{i\bullet k}-t_{iuk}-t_{ivk}} \\ &\quad \prod_{s\neq k} Sp_s^{J-t_{i\bullet s}}(1-Se_s)^{t_{i\bullet s}} \\ &= \left[\prod_{s=1}^{K} Sp_s^{J-t_{i\bullet s}}(1-Se_s)^{t_{i\bullet s}}\right]\sum_{k=1}^{K} R_k^- \sum_{u<v}\frac{[Sp_k/(Sp_k-1)]^{t_{iuk}+t_{ivk}}}{Sp_k^2} \\ &= d_i^- \sum_{k=1}^{K} R_k^- \sum_{u=1}^{J-1}\sum_{v=u+1}^{J}\frac{[Sp_k/(Sp_k-1)]^{t_{iuk}+t_{ivk}}}{Sp_k^2} \\ &= d_i^- \widetilde{rc}_i^- . \end{aligned}$$

Here it is assumed that the temporal covariance for a given test does not change with each pair of time points, implying exchangeability in the repeated measures. If differing amounts of correlation are expected between different pairs of time points the model is immediately generalized to accommodate this, although at the cost of increased complexity and numbers of parameters. For example, $J = 3$ time points requires $R^+_{k,12}$, $R^+_{k,13}$, and $R^+_{k,23}$ in the $D+$ population for test k and $R^-_{k,12}$, $R^-_{k,13}$, and $R^-_{k,23}$ in the $D-$ population unless additional assumptions are made. One such assumption might be the usual tapering of covariance to zero as the length of time between observations increases, e.g. the covariances between times t_1 and t_2 are $R^+_{k,t_1t_2} = (R^+_k)^{|t_1-t_2|}$ and $R^-_{k,t_1t_2} = (R^-_k)^{|t_1-t_2|}$ in the $D+$ and $D-$ populations respec-

tively for test k. This simplifying assumption allows correlation to increase for more proximal timepoints without adding additional parameters to the model; choice of appropriate temporal covariance can be made via DIC or LPML.

Combining all the three terms, the likelihood function can be explicitly expressed as

$$L = \prod_{i=1}^{n} \pi d_i^+ (1 + \widetilde{rc}_i^+ + \widetilde{tc}_i^+) + (1-\pi) d_i^- (1 + \widetilde{rc}_i^- + \widetilde{tc}_i^-).$$

For subjects that are only seen one time, the likelihood contribution is simply $\pi d_i^+ (1 + \widetilde{tc}_i^+) + (1-\pi) d_i^- (1 + \widetilde{tc}_i^-)$; this was used for 18 subjects in the scaphoid fracture data. Similarly, those subjects that are *known* to be infected have likelihood contribution $d_i^+ (1 + \widetilde{rc}_i^+ + \widetilde{tc}_i^+)$, and those known to not be infected have contribution $d_i^- (1 + \widetilde{rc}_i^- + \widetilde{tc}_i^-)$.

An important benefit of the proposed model is that missing data are readily incorporated: the likelihood contributions simply reflect those tests that have been actually observed at a particular time point. For example, if subject i did not receive test k' at any time point then $d_i^+ = \prod_{k \neq k'} Se_k^{t_{i\bullet k}} (1 - Se_k)^{J - t_{i\bullet k}}$ and $d_i^- = \prod_{k \neq k'} Sp_k^{J - t_{i\bullet k}} (1 - Sp_k)^{t_{i\bullet k}}$; similarly, the sums in tc_i^+ and tc_i^- do not include u or v indices that equal k', i.e. $C_{uv}^+ = C_{uv}^- = 0$ for u or v equal to k'. If more than one test is missing, simply remove the parameters related to the missing tests. Similarly, if a subject is not observed at time point t', i.e. all tests are missing, then the double sums in rc_i^+ and rc_i^- simply skip terms with u or v equal to t'. These ideas are generalized further to subjects that are missing a subset of tests at *every* time point. Thus very general data collection schemes can be handled and no subjects with missing data need be "thrown out" or imputed.

A useful aspect of the proposed model is the ready computation of conditional correlation. The covariance between test k_1 and test k_2 is the $D+$ population is $C_{k_1 k_2}^+$ and $C_{k_1 k_2}^-$ in the $D-$ population. Thus the pairwise conditional correlations are $\rho_{k_1 k_2}^+ = C_{k_1 k_2}^+ / \sqrt{Se_{k_1}(1 - Se_{k_1}) Se_{k_2}(1 - Se_{k_2})}$ and $\rho_{k_1 k_2}^- = C_{k_1 k_2}^- / \sqrt{Sp_{k_1}(1 - Sp_{k_1}) Sp_{k_2}(1 - Sp_{k_2})}$.

Similarly the conditional correlation of test k between any two time points is $R_k^+/[Se_k(1-Se_k)]$ in the $D+$ population and $R_k^-/[Sp_k(1-Sp_k)]$ in the $D-$ population.

Obtaining posterior inference in Bayesian latent class models commonly employs the Gibbs sampler (Joseph et al., 1995; Dendukuri and Joseph, 2001; Norris et al., 2009; Engel et al., 2010). The model developed here has numerous awkward constraints on the parameter space, however the fundamental requirement is simply that the elements of $P(\mathbf{T}_i = \mathbf{t}|D_i+)$ and $P(\mathbf{T}_i = \mathbf{t}|D_i-)$ are between zero and one for all 2^{KJ} values of $\mathbf{t}$. Posterior inference proceeds via an adaptive componentwise Metropolis-Hastings algorithm (Haario et al., 2005) that immediately discards any parameter proposal that does not satisfy the basic requirements on $P(\mathbf{T}_i = \mathbf{t}|D_i+)$ and $P(\mathbf{T}_i = \mathbf{t}|D_i-)$. Proposals are focused by considering easy-to-compute pairwise covariance requirements (Dendukuri and Joseph, 2001) (See Appendix A.1). This improves the mixing in the resulting Markov chain by an order of magnitude, allowing for reasonably small MCMC samples to be used and dramatically decreasing computation time.

2.3 Simulation

Here, data are simulated according to the model and the estimation procedure is validated. The case of $K = 3$ tests and $J = 2$ time points is considered yielding 19 parameters

$$Se_1, Se_2, Se_3, Sp_1, Sp_2, Sp_3, R_1^+, R_2^+, R_3^+, R_1^-, R_2^-, R_3^-, C_{12}^+, C_{13}^+, C_{23}^+, C_{12}^-, C_{13}^-, C_{23}^-, \pi.$$

Given the true parameter values, each cell probability, for instance $P(T_{11} = 0, T_{21} = 0, T_{31} = 0, T_{12} = 0, T_{22} = 0, T_{32} = 0)$, can be calculated based on the model in section 2.2. Then data is drawn from a multinomial distribution with these cell probabilities. Three samples of size $n = 200$, $n = 400$, and $n = 1000$ are replicated 200 times;

each data set is fitted using the adaptive Metropolis-Hastings algorithm run out 5000 iterations with no burn-in. All priors are uniform.

Table 2.1: $n = 200$; average of posterior means, average of posterior standard deviations, standard deviation of posterior means, and actual coverage level of 95% CIs.

Par.	Truth	Ave. Post. Mean	Ave. Post. SD	SD Post. Mean	95%CI
π	0.300	0.336	0.080	0.039	0.985
Se_1	0.800	0.733	0.065	0.058	0.920
Se_2	0.700	0.655	0.071	0.056	0.980
Se_3	0.600	0.566	0.078	0.054	0.980
Sp_1	0.600	0.592	0.047	0.041	0.975
Sp_2	0.700	0.692	0.043	0.039	0.970
Sp_3	0.800	0.792	0.040	0.033	1.000
R_1^+	0.020	0.023	0.027	0.018	0.995
R_2^+	0.020	0.023	0.030	0.021	0.980
R_3^+	0.020	0.013	0.034	0.023	0.995
R_1^-	0.020	0.021	0.025	0.021	0.985
R_2^-	0.020	0.018	0.021	0.018	0.965
R_3^-	0.020	0.021	0.018	0.015	0.975
C_{12}^+	0.010	0.017	0.019	0.014	0.980
C_{13}^+	0.015	0.019	0.020	0.015	0.975
C_{23}^+	0.020	0.019	0.021	0.016	0.990
C_{12}^-	0.020	0.021	0.015	0.013	0.975
C_{13}^-	0.015	0.014	0.013	0.012	0.960
C_{23}^-	0.010	0.011	0.013	0.010	0.990

Table 2.1 presents simulation results from 200 simulated data sets each of size $n = 200$; Tables 2.2 and 2.3 are for $n = 400$ and $n = 1000$. For each sample size the sample mean of the 200 posterior means, sample mean of the 200 posterior standard deviations, sample standard deviation of the posterior means and 95% coverage rate are calculated. Note that with increasing sample size, the mean of the posterior standard deviation of each parameter becomes narrower, empirically verifying identifiability. With sample size $n = 200$, close to the sample size of our scaphoid fracture data, the posterior estimates show more variability than larger sample sizes. The simulation also indicates that the mean of posterior standard deviation is ap-

Table 2.2: $n = 400$; average of posterior means, average of posterior standard deviations, standard deviation of posterior means, and actual coverage level of 95% CIs.

Par.	Truth	Ave. Post. Mean	Ave. Post. SD	SD Post. Mean	95%CI
π	0.300	0.335	0.068	0.041	0.975
Se_1	0.800	0.760	0.048	0.038	0.940
Se_2	0.700	0.669	0.056	0.047	0.945
Se_3	0.600	0.581	0.062	0.049	0.960
Sp_1	0.600	0.605	0.036	0.029	0.995
Sp_2	0.700	0.704	0.032	0.028	0.995
Sp_3	0.800	0.802	0.029	0.022	0.985
R_1^+	0.020	0.018	0.021	0.014	1.000
R_2^+	0.020	0.023	0.025	0.021	0.985
R_3^+	0.020	0.016	0.028	0.022	0.985
R_1^-	0.020	0.019	0.019	0.018	0.970
R_2^-	0.020	0.018	0.016	0.014	0.970
R_3^-	0.020	0.019	0.014	0.012	0.960
C_{12}^+	0.010	0.017	0.015	0.012	0.975
C_{13}^+	0.015	0.019	0.016	0.012	0.985
C_{23}^+	0.020	0.020	0.017	0.013	0.985
C_{12}^-	0.020	0.017	0.011	0.009	0.985
C_{13}^-	0.015	0.013	0.010	0.009	0.960
C_{23}^-	0.010	0.009	0.010	0.007	0.985

proaching the standard deviation of the posterior means with increasing sample size; thus the posterior standard deviation provides a reasonable estimate of variance of the posterior mean – as a point estimator – in repeated samples. The 95% coverage rate also approaches the nominal *frequentist* 95% level with more data. Note that estimation of parameters in the non-diseased group are more precise than that for diseased group, due to the fact that the disease prevalence is set to less than one-half, i.e. $\pi = 0.3$.

Table 2.3: $n = 1000$; average of posterior means, average of posterior standard deviations, standard deviation of posterior means, and actual coverage level of 95% CIs.

Par.	Truth	Ave. Post. Mean	Ave. Post. SD	SD Post. Mean	95%CI
π	0.300	0.307	0.054	0.044	0.975
Se_1	0.800	0.787	0.039	0.033	0.935
Se_2	0.700	0.692	0.042	0.037	0.955
Se_3	0.600	0.598	0.048	0.041	0.950
Sp_1	0.600	0.600	0.025	0.025	0.945
Sp_2	0.700	0.698	0.023	0.021	0.975
Sp_3	0.800	0.802	0.021	0.018	0.980
R_1^+	0.020	0.021	0.017	0.013	0.980
R_2^+	0.020	0.018	0.019	0.016	0.975
R_3^+	0.020	0.016	0.022	0.019	0.955
R_1^-	0.020	0.019	0.013	0.012	0.970
R_2^-	0.020	0.023	0.011	0.011	0.955
R_3^-	0.020	0.020	0.010	0.008	0.955
C_{12}^+	0.010	0.015	0.012	0.009	0.975
C_{13}^+	0.015	0.015	0.012	0.011	0.965
C_{23}^+	0.020	0.017	0.014	0.011	0.985
C_{12}^-	0.020	0.020	0.008	0.008	0.950
C_{13}^-	0.015	0.014	0.007	0.007	0.960
C_{23}^-	0.010	0.010	0.007	0.007	0.940

2.4 Data analysis

In Section 1, the scaphoid fracture data was introduced, based on $n = 223$ patients' clinical tests, of which 205 were repeated at a second time point. Duckworth et al. (2012) developed a set of clinical prediction rules by studying several clinical tests and other demographic factors predictive of scaphoid fracture. Based on their analysis we consider the $K = 3$ most predictive clinical tests: $(k = 1)$ pain on thumb-index finger pinch, $(k = 2)$ scaphoid tubercle tenderness, and $(k = 3)$ anatomical snuff box (ABS) pain on ulnar deviation. Duckworth et al. (2012) noted that the assumption of conditional independence among the tests is unrealistic because each examiner knew the result of each test, so our model is an appealing approach to analyzing these data.

Table 2.4: Posterior mean, median, standard deviation (SD) and 95% CI of model parameters.

	Mean	Median	SD	2.5%	97.5%
π	0.352	0.350	0.051	0.258	0.452
Se_1	0.706	0.704	0.053	0.599	0.813
Se_2	0.894	0.900	0.039	0.803	0.953
Se_3	0.907	0.912	0.032	0.835	0.958
Sp_1	0.640	0.638	0.032	0.581	0.702
Sp_2	0.615	0.616	0.036	0.545	0.683
Sp_3	0.589	0.588	0.035	0.523	0.663
R_1^+	0.077	0.078	0.026	0.030	0.125
R_2^+	0.028	0.023	0.023	0.000	0.088
R_3^+	0.007	0.005	0.010	-0.007	0.031
R_1^-	0.100	0.100	0.019	0.060	0.138
R_2^-	0.083	0.083	0.020	0.043	0.118
R_3^-	0.046	0.047	0.017	0.009	0.079
C_{12}^+	0.016	0.014	0.014	-0.006	0.050
C_{13}^+	0.013	0.012	0.008	0.001	0.031
C_{23}^+	0.006	0.005	0.007	-0.003	0.022
C_{12}^-	0.088	0.088	0.011	0.065	0.110
C_{13}^-	0.122	0.122	0.013	0.094	0.145
C_{23}^-	0.032	0.032	0.011	0.010	0.055

The component-wise adaptive Metropolis-Hastings algorithm was implemented for 50,000 iterations including a 2,000 iterate burn-in; this took roughly 1.5 hours using our R code. The traceplots for each of the 19 parameters are shown in the Appendix A.2; different runs with differing starting values and random number generator seeds lead to essentially identical inferences. Model parameter estimates and 95% credible intervals are in Table 2.4. Duckworth et al. (2012) estimated the prevalence in this population to be 0.37 using latent class analysis and 0.28 using traditional calculation method; the posterior mean and median from our model is close at 0.35 and the 95% CI includes both 0.37 and 0.28. Duckworth et al. (2012) (unrealistically) calculated the sensitivity and specificity at each time point separately; the model proposed here assumes that the testing characteristics do not change over time but encourages cor-

relation between pairs of tests and over time. Broadly, however, the Duckworth et al. (2012) results are consistent with ours. Note significant temporal and pairwise correlation among tests in the non-diseased population; significant pairwise correlation is not evident in the diseased subpopulation.

2.5 Tests in parallel and series

Multiple tests allow the construction of new "omnibus" tests, e.g. in series or parallel (Hanson et al., 2000). Parallel interpretation (positive on any test is positive, negative otherwise) increases the sensitivity and serial interpretation (positive on all tests is positive, negative otherwise) increases the specificity of diagnosis. We use the union symbol $\cup$ to denote parallel interpretation between tests and the intersection symbol $\cap$ to denote serial interpretation. With $K = 3$ tests, as in our application, there are eight new tests we can consider if only testing at one time point. Let the test results for a randomly selected person be $\mathbf{T} = (T_1, T_2, T_3)$ at a given time point. Additional tests are

$$T_1 \cup T_2,\ T_1 \cup T_3,\ T_2 \cup T_3,\ T_1 \cup T_2 \cup T_3,\ T_1 \cap T_2,\ T_1 \cap T_3,\ T_2 \cap T_3,\ T_1 \cap T_2 \cap T_3.$$

With conditional dependence, it may be that only one or two tests hold all the information about the disease state; e.g. if $T_2 \perp D|T_1$ then T_2 provides no more information on D given T_1. Thus additional cost can be avoided.

The model provides closed forms for $P(\mathbf{T} = \mathbf{t}|D+)$ and $P(\mathbf{T} = \mathbf{t}|D-)$ in terms of 19 model parameters. Let $p^+_{t_1t_2t_3} = p^+_{\mathbf{t}} = P(\mathbf{T} = \mathbf{t}|D+)$ and $p^-_{t_1t_2t_3} = p^-_{\mathbf{t}} = P(\mathbf{T} = \mathbf{t}|D-)$. Then the sensitivity of $T_1 \cup T_2$ is $p^+_{10\bullet} + p^+_{01\bullet} + p^+_{11\bullet} = 1 - p^+_{00\bullet}$; the specificity is $p^-_{00\bullet}$, etc. The sensitivity of $T_1 \cup T_2 \cup T_3$ is $1 - p^+_{000}$; the specificity is p^-_{000}. The sensitivity of $T_1 \cap T_2$ is $p^+_{11\bullet}$; the specificity is $1 - p^-_{11\bullet}$. These sums can be monitored over the MCMC run and posterior summaries and credible intervals computed in the

Table 2.5: Posterior mean, median and 95% CI of sensitivity and specificity of parallel and serial combinations of tests

Test	Sensitivity				Specificity			
	Mean	Median	2.5%	97.5%	Mean	Median	2.5%	97.5%
$T_1 \cup T_2$	0.959	0.965	0.894	0.989	0.475	0.475	0.412	0.536
$T_1 \cup T_3$	0.964	0.966	0.922	0.990	0.493	0.493	0.432	0.558
$T_2 \cup T_3$	0.985	0.987	0.961	0.998	0.384	0.385	0.320	0.453
$T_1 \cup T_2 \cup T_3$	0.993	0.995	0.980	0.999	0.370	0.372	0.306	0.440
$T_1 \cap T_2$	0.668	0.661	0.536	0.821	0.771	0.771	0.733	0.815
$T_1 \cap T_3$	0.675	0.666	0.553	0.821	0.728	0.723	0.681	0.785
$T_2 \cap T_3$	0.827	0.822	0.717	0.916	0.801	0.803	0.741	0.852
$T_1 \cap T_2 \cap T_3$	0.624	0.620	0.501	0.758	0.844	0.843	0.807	0.882

usual way. Note that for any $\mathbf{t}$, one simply need compute "likelihood contributions" $d^+(1+\widetilde{tc}^+)$ and/or $d^-(1+\widetilde{tc}^-)$ corresponding to $\mathbf{t}=(t_1,t_2,t_3)$ to obtain $p^+_{\mathbf{t}}$ and $p^-_{\mathbf{t}}$.

Table 2.5 shows the posterior mean, median and 95% CI for the sensitivities and specificities of the 8 new tests constructed in parallel and series. Evidently, sensitivity of the parallel tests and specificity of serial tests increase drastically. Furthermore, it is noted that the $T_1 \cup T_2 \cup T_3$ has the highest sensitivity and $T_1 \cap T_2 \cap T_3$ has the highest specificity which suggests each test contributes to the parallel sensitivity and serial specificity to some degree; that is to say, all these three tests bring some information in diagnosing the scaphoid fracture. Among these newly constructed tests, $T_2 \cap T_3$ is the only one with sensitivity and specificity greater than 80%, which provides clinical insight in combining these tests. This test is positive if both scaphoid tubercle tenderness and ABS pain on ulnar deviation are present, otherwise negative. Note that T_2 and T_3 have conditional pairwise covariances (Table 2.4) much smaller than the other two pairs, i.e. are much more closer to being conditionally independent, shedding light on why $T_2 \cap T_3$ appears to be the best omnibus test. Note that using $T_2 \cap T_3$ decreases the proportion of false positives by 20% compared to using either T_2 or T_3 by themselves.

Parallel and serial tests can also be considered over time for individual tests or groups of tests to improve either sensitivity (parallel) or specificity (serial) as well. For example, a parallel test for the kth diagnostic administered over two time points $T_{i1k} \cup T_{i2k}$ (kth test comes up positive at either time) increases sensitivity vs. just one time point T_{i1k}.

2.6 Discussion and Conclusions

A model for conditionally dependent binary tests repeated over time is developed and applied to a data set on clinical symptoms of scaphoid fracture. A notable feature is the accommodation of pairwise dependence with as few as three tests in one population. Without repeated measures one must make the untestable assumption of conditional independence among the three tests to move forward. For the scaphoid data, there is a clear indication of both correlation among pairs of tests and correlation over time, especially among those who do not have scaphoid fractures. Unlike previous approaches, the proposed model does not include subject-specific random effects and model parameters are immediately interpretable. In addition, correlation parameters can offer some useful information on how tests interact with each other and behave over time. Parallel and serial test construction further facilitate evaluation of multiple tests and provide insight on how to apply groups of tests simultaneously.

Although generalization to higher order (than pairwise) dependence is possible, Jones et al. (2010) suggest that in practice, pairwise dependence is likely to be more than enough. The model developed here is immediately generalizable to more than one population with differing prevalence, or prevalence modeled through binary regression. Populations could also be based on stratification over aspects such as gender, age group, etc. and model-fitting proceeds as usual. In fact, each strata only need introduce one new prevalence parameter, allowing highly efficient borrowing of information across sub-populations.

Chapter 3

Bayesian Hierarchical Latent Class Models for Estimating Diagnostic Accuracy

3.1 Introduction

In medical practice, multiple diagnostic tests (or raters) are often utilized to diagnose the disease status of a patient. Assessing the diagnostic accuracy of individual tests is important. The diagnostic accuracy based upon results from multiple raters also merits much attention. For a diagnostic test with a binary outcome, when the true disease status (with or without disease) is known or a gold standard reference standard exists, the estimation of sensitivity and specificity of the test is straightforward. However, on many occasions, the true disease status or the gold standard is prohibitive to obtain due to the cost or ethical concerns. For this situation, many latent class models, where the true disease status is unknown and therefore treated as latent, have been proposed to assess the diagnostic accuracy of tests.

The early development of latent class models lay in conditional independence models, where diagnostic results for the same patient across multiple tests are independent conditional on a patient's true disease status. Hui and Walter (1980) laid a solid foundation in studying the false positive and negative error rates of two diagnostic tests in two populations. Joseph et al. (1995) discussed a similar problem from a Bayesian perspective. However, this conditional independence between multiple tests does not always hold due to some common factors connecting tests (Dendukuri and Joseph, 2001). Vacek (1985) and Brenner (1996), among others, demonstrated that

ignoring the possible dependence between tests can lead to biased estimates of the prevalence of disease and accuracy of tests.

In the literature of latent class models, there are two general approaches for handling the conditional dependence between multiple tests. Vacek (1985), Torrance-Rynard and Walter (1997), Yang and Becker (1997), Jones et al. (2010) and Wang and Hanson (2019), among others, directly incorporated conditional pairwise covariances between tests. Qu et al. (1996) developed latent Gaussian random effects (GRE) to model the conditional dependence between multiple tests. Albert and Dodd (2004) provided a cautionary note and guidance in using latent class models with various dependence structures. In this chapter, we adopt the first type of the conditional dependence model due to the ease of interpretation. A systematic review of latent class models can be found in van Smeden et al. (2013) and Collins and Albert (2016).

Dendukuri and Joseph (2001) proposed Bayesian approaches for handling both of these two conditional dependence structures. However, they only considered positive correlation between tests and only dealt with two tests without providing a generalized method to handle more tests. Their full conditional distributions (for two tests) for the Gibbs Sampler implementation were complicated. When the number of tests increases, their likelihood based on the multinomial distribution of the test-result frequency grows exponentially. Our proposed model, an extension of Jones et al. (2010) , is amenable to accommodating more tests and allows for positive and negative correlations between tests. The multinomial property is just a special case used in our computational algorithms.

There has been much discussion about model identifiability for latent class models. For example, see Rothenberg et al. (1971), Dendukuri et al. (2004), and Gustafson et al. (2005). Jones et al. (2010) investigated the identifiability of the first type of the conditional dependence model. They concluded that a sufficient number of degrees of freedom does not guarantee unique estimates of prevalence and test performance.

They provided some symbolic algebra methodology to determine whether a proposed study design would lead to an identifiable model. When using Bayesian approaches, identifiability is not mandatory if good prior information is available (Jones et al., 2010). For example, Dendukuri and Joseph (2001) adopted informative priors to avoid the nonidentifiability. In this chapter, we propose to apply correlation residual analysis to reduce the number of parameters by only including significant covariance terms into the model to improve the identifiability.

In the literature, most methods can only accommodate a small number of tests or raters. For a large group of tests or raters, not only is individual diagnostic accuracy of interest to be estimated, but also the diagnostic performance of the whole group of tests or raters. Zhang et al. (2012) proposed a latent class model with crossed random effects for the subjects and raters to estimate the diagnostic accuracy of a group of raters. Lin et al. (2018) described a modeling approach to assess each rater's diagnostic skills by linking rater binary decisions with patient true disease status through patient latent disease severity. In this chapter, for the first time, the conditional independence model and the pairwise covariance model are further developed to analyze a large number of tests under the Bayesian framework. Unlike the approach of Dendukuri and Joseph (2001) where individual beta priors are assigned for sensitivities (specificities), our method assumes that all the sensitivities (specificities) follow a common beta prior with the two hyperparameters reflecting the group level sensitivity (specificity). Our proposed "Poisson zero trick" JAGS (Plummer, 2003) are easy to implement to flexibly incorporate a large number of tests. The algorithms are feasible for analyzing dozens of tests under the conditional dependence model and even hundreds of tests under the conditional independence model.

The rest of this chapter is organized as follows. Section 3.2 explicitly describes the conditional independence model and the pairwise covariance model with their

hierarchical priors. The correlation residual analysis is also introduced in this section. Section 3.3 shows details about the computation strategies with regard to their implementation in JAGS. Simulation studies to investigate the performance of the proposed methods are shown in Section 3.4. Section 3.5 illustrates the proposed methods with three real data examples. Lastly, Section 3.6 summarizes the main results with some discussions.

3.2 Models

Suppose that K tests (or raters) are used to evaluate n subjects, yielding vectors of binary test results $\mathbf{T}_1, \ldots, \mathbf{T}_n$, where $\mathbf{T}_i = (T_{i1}, \ldots, T_{iK})'$ for the ith subject ($i = 1, \ldots, n$). Denote $T_{ij} = 1$ if the test result is positive and $T_{ij} = 0$ if negative. Let D_i denote the latent true disease status for subject i with 1 being positive and 0 negative, respectively. Denote by π the underlying disease prevalence $P(D_i = 1)$. Then, the likelihood function for a latent class model for binary results is:

$$L = \prod_{i=1}^{n} \Big\{ \pi P(\mathbf{T}_i = \boldsymbol{t}_i \,|\, D_i = 1) + (1-\pi) P(\mathbf{T}_i = \boldsymbol{t}_i \,|\, D_i = 0) \Big\} \equiv \prod_{i=1}^{n} L_i, \tag{3.1}$$

where $L_i = \pi P(\mathbf{T}_i = \boldsymbol{t}_i \,|\, D_i = 1) + (1-\pi) P(\mathbf{T}_i = \boldsymbol{t}_i \,|\, D_i = 0)$.

3.2.1 Hierarchical conditional independence model (Model M1)

We first consider the conditional independence model, where the diagnostic results of a subject are independent across all tests conditional on the true disease status. The sensitivity and specificity of the jth test are denoted as Se_j and Sp_j, respectively. Then, according to the conditional independence structure,

$$\begin{aligned} P(\mathbf{T}_i = \boldsymbol{t}_i \,|\, D_i = 1) &= \prod_{j=1}^{K} P(T_{ij} = t_{ij} \,|\, D_i = 1) = \prod_{j=1}^{K} Se_j^{t_{ij}} (1 - Se_j)^{1-t_{ij}}; \\ P(\mathbf{T}_i = \boldsymbol{t}_i \,|\, D_i = 0) &= \prod_{j=1}^{K} P(T_{ij} = t_{ij} \,|\, D_i = 0) = \prod_{j=1}^{K} (1 - Sp_j)^{t_{ij}} Sp_j^{1-t_{ij}}. \end{aligned} \tag{3.2}$$

Individual sensitivities and specificities are assumed to be random effects and to independently follow common beta distributions, respectively:

$$\begin{aligned} \mathrm{Se}_j &\overset{iid}{\sim} Beta(\omega_1(\kappa_1 - 2) + 1, (1 - \omega_1)(\kappa_1 - 2) + 1), \\ \mathrm{Sp}_j &\overset{iid}{\sim} Beta(\omega_2(\kappa_2 - 2) + 1, (1 - \omega_2)(\kappa_2 - 2) + 1), \end{aligned}$$

where ω_1 is the mode of the beta distribution of test sensitivities and the concentrate parameter κ_1 reflects the spread of the distribution. The larger the value of κ_1, the more concentrated the distribution of test sensitivities is around the mode. To ensure the existence of the mode ω_1, κ_1 needs to be larger than 2. Similarly, ω_2 and κ_2 describe the mode and spread of the distribution of test specificities. An advantage of using this parameterization to denote a beta distribution is that it naturally provides group-level diagnostic accuracy and variation by the mode ω and concentrate term κ. See page 129 in Kruschke (2014) for more details of this parameterization. To allow data to inform these group-level parameters, we assign vague priors to them. We assign uniform(0.5,1) priors to ω_1 and ω_2, and diffused gamma priors, such as gamma(0.01,0.01), to $\kappa_1 - 2$ and $\kappa_2 - 2$. We assign a uniform(0,1) prior to the disease prevalence π.

Note that assigning uniform(0.5,1) priors to ω_1 and ω_2 is a natural prior choice as for most reasonable tests, we would expect their sensitivity and specificity are above 0.5. To avoid the "label-switching" problem in the fitting, we may follow Jones et al. (2010) to add the restriction of $\mathrm{Se}_j + \mathrm{Sp}_j > 1$. We can still assign Se_j the same beta prior, but let Sp_j follow the truncated beta prior with the support $(1 - \mathrm{Se}_j, 1)$. Similarly, to avoid the sampling of π stuck in 0 or 1 extremes, we may adjust the uniform(0,1) prior of the disease prevalence to a uniform$(\frac{1}{n}, 1 - \frac{1}{n})$.[20]

3.2.2 Hierarchical conditional dependence model (Model M2)

Following Dendukuri and Joseph (2001), our second model takes into account the pairwise dependence between multiple tests. Conditional covariances, such as between

test j and test k, are denoted as C_{jk}^{+} and C_{jk}^{-} given the subject being diseased or non-diseased, respectively. For instance, given subject i is diseased, the joint probabilities for test j and test k classifying subject i are as follows:

$$\begin{aligned}
P(T_{ij}=1, T_{ik}=1 \,|\, D_i=1) &= Se_j Se_k + C_{jk}^{+}, \\
P(T_{ij}=1, T_{ik}=0 \,|\, D_i=1) &= Se_j(1-Se_k) - C_{jk}^{+}, \\
P(T_{ij}=0, T_{ik}=1 \,|\, D_i=1) &= Se_k(1-Se_j) - C_{jk}^{+}, \\
P(T_{ij}=0, T_{ik}=0 \,|\, D_i=1) &= (1-Se_j)(1-Se_k) + C_{jk}^{+}.
\end{aligned}$$

Similarly, given subject i is non-diseased, the joint probabilities for test j and test k are

$$\begin{aligned}
P(T_{ij}=1, T_{ik}=1 \,|\, D_i=0) &= (1-Sp_j)(1-Sp_k) + C_{jk}^{-}, \\
P(T_{ij}=1, T_{ik}=0 \,|\, D_i=0) &= (1-Sp_j)Sp_k - C_{jk}^{-}, \\
P(T_{ij}=0, T_{ik}=1 \,|\, D_i=0) &= Sp_j(1-Sp_k) - C_{jk}^{-}, \\
P(T_{ij}=0, T_{ik}=0 \,|\, D_i=0) &= Sp_j Sp_k + C_{jk}^{-}.
\end{aligned}$$

It is clear that a positive (negative) value of C_{jk}^{+} reflects the positive (negative) dependence between test j and test k on diagnosing the same subjects when the true disease status is positive. A similar interpretation is applied to C_{jk}^{-} when the true disease status is negative. To be valid covariances, C_{jk}^{+} and C_{jk}^{-} need to ensure that the probability of each combination above is between 0 and 1. Therefore, necessary constraints are

$$\begin{aligned}
&(1-Se_j)(Se_k-1) < C_{jk}^{+} < min(Se_j, Se_k) - Se_j Se_k, \\
&(1-Sp_j)(Sp_k-1) < C_{jk}^{-} < min(Sp_j, Sp_k) - Sp_j Sp_k. \qquad (3.3)
\end{aligned}$$

Under this conditional pairwise dependence model, the joint probabilities of all k test results for subject i conditional on the true disease status can be expressed as follows using a generalized form in Jones et al. (2010):

$$P(\mathbf{T}_i = \mathbf{t}_i \,|\, D_i = 1) = \prod_{k=1}^{K} Se_k^{t_{ik}}(1-Se_k)^{1-t_{ik}} + \sum_{u=1}^{K-1}\sum_{v=u+1}^{K}(-1)^{t_{iu}+t_{iv}} cov_{uv}^{+} \prod_{k\neq u,v}^{K} Se_k^{t_{ik}}(1-Se_k)^{1-t_{ik}},$$

$$P(\mathbf{T}_i = \mathbf{t}_i \,|\, D_i = 0) = \prod_{k=1}^{K} Sp_k^{1-t_{ik}}(1-Sp_k)^{t_{ik}} + \sum_{u=1}^{K-1}\sum_{v=u+1}^{K}(-1)^{t_{iu}+t_{iv}} cov_{uv}^{-} \prod_{k\neq u,v}^{K} Sp_k^{1-t_{ik}}(1-Sp_k)^{t_{ik}}. \tag{3.4}$$

For the covariances, we assign uniform priors with the constraints (3.3) to them. For the other parameters, we adopt the same prior specification as in Subsection 3.2.1.

3.2.3 Correlation residual plot

It is worth pointing out that not every pair of covariances are needed in the conditional dependence model. Adding unnecessary covariances is not just introducing redundant parameters but also adding computational burden and uncertainty to the model. Therefore, we introduce a correlation residual analysis (Qu et al., 1996) to detect significant dependence between tests. The correlation between each pair of tests, such as test j and test k, is defined as

$$\mathrm{r}_{jk} = \frac{P(T_j = 1, T_k = 1) - P(T_j = 1)P(T_k = 1)}{\sqrt{P(T_j = 1)\Big(1 - P(T_j = 1)\Big)P(T_k = 1)\Big(1 - P(T_k = 1)\Big)}}.$$

The correlation residual is the difference between observed correlation and model-based correlation. For the observed correlation, $P(T_j = 1)$, $P(T_k = 1)$, and $P(T_j = 1, T_k = 1)$ are estimated by their sample proportions as $\frac{1}{n}\sum_{i=1}^{n} t_{ij}$, $\frac{1}{n}\sum_{i=1}^{n} t_{ik}$, and $\frac{1}{n}\sum_{i=1}^{n} t_{ik}t_{ij}$, respectively. For the model-based correlation, the joint probability is

$$P(T_j = 1, T_k = 1) = \pi Se_j Se_k + (1-\pi)(1-Sp_j)(1-Sp_k)$$

for model M1 in Section 2.1 and is

$$P(T_j = 1, T_k = 1) = \pi(Se_j Se_k + C_{jk}^{+}) + (1-\pi)\Big((1-Sp_j)(1-Sp_k) + C_{jk}^{-}\Big)$$

for model M2 in Section 2.2. The marginal probabilities for models M1 and M2 have the same form $P(T_j = 1) = \pi Se_j + (1 - \pi)(1 - Sp_j)$. After plugging in the estimates of the disease prevalence π, sensitivities and specificities, the model-based correlation is calculated.

For the correlation residual analysis, plotting all pairwise correlation residuals from the conditional independence model (M1) provides a simple visual way to check whether M1 sufficiently explains the data. If M1 provides a good fit, then all correlation residuals are expected to be close to zero. Otherwise, the conditional dependence model (M2) should be applied. The formal way to identify significant correlations should take into account the uncertainty of the model-based correlation. We can use the Markov Chain Monte Carlo (Gelman et al., 2013) samples of parameters to obtain the MCMC chain of the correlation residuals for each pair of tests. Based upon multiple simulation studies we conducted, the following criterion is proposed to identify significant covariance terms for M2: when a one-sided 95% credible interval (CI) of the correlation residual from M1, either (0, 95%) or (5%,100%) CI, includes only negative values or positive values, we define that correlation pair as significant and therefore the corresponding covariance terms are added in the pairwise covariance model (M2). For nonsignificant pairs, we set the corresponding covariance terms to 0 in M2. After fitting M2, we can check correlation residuals again and a good fitting in M2 model should now have all correlation residuals close to 0. Simulation results in Section 4 show that the one-sided credible interval criterion works well to identify significant correlations between tests.

3.3 Computational techniques

To configure the hierarchical structure of models for implementation, we apply MCMC computation in Just Another Gibbs Sampler (JAGS) (Plummer, 2003). The advantage of using JAGS is that the user does not need to derive the complicated full

conditional distributions or to write the sampling code by himself. Instead, the user only needs to prepare the JAGS model declaration syntax and then, in R, to pass data and model syntax to JAGS, for example, via the function *j*ags() in *R*2jags library (Su and Yajima, 2015). There are many built-in distributions in JAGS; however, none of them is directly proper for our proposed models. Consequently, we utilize the "Poisson zero trick" method to accommodate the likelihood in equation (3.1). Specifically, for $i = 1, ..., n$, $z_i = 0$ is introduced and z_i follows a Poisson distribution with mean parameter as $-log(L_i)$. In this way, the likelihood function ingenuously links the original data with the introduced z_i's, that is, the z_i's contribute the same likelihood as the original data.

3.3.1 Multinomial property

For K tests, there are in total 2^K possible test-result combinations for each patient. Given the parameters, we can calculate the probability for each test-result combination according to the models. For instance, in the case of two tests, there are 4 possible test-result combinations, i.e. 00, 01, 10 and 11, where 01 represents a non-diseased result from the first test and a diseased result from the second test. The probability of each test-result combination can be calculated by plugging the test-result combination values as $\mathbf{t}_i$ into (3.2) or (3.4) (depending on whether the M1 or M2 model is used) and then evaluating the likelihood formulae L_i. For example, the probability of the test results 00 is

$$P_{00} = P(T_{i1} = 0, T_{i2} = 0) = \pi P(T_{i1} = 0, T_{i2} = 0 \,|\, D_i = 1) + (1 - \pi) P(T_{i1} = 0, T_{i2} = 0 \,|\, D_i = 0).$$

Denote the frequency of subjects for the four combinations as N_{00}, N_{01}, N_{10}, and N_{11}, respectively. The vector (N_{00}, N_{01}, N_{10} N_{11}) follows a multinomial distribution (n, $\mathbf{p}$) with n (= $N_{00} + N_{01} + N_{10} + N_{11}$) the total number of subjects and $\mathbf{p} = (P_{00}, P_{01}, P_{10}, P_{11})$ denoting the probability vector. This multinomial property

holds for both models M1 and M2. Adopting this multinomial property for the fitting, especially for model M2, can significantly alleviate the computational burden. In Section 3.4, for the second simulation scenario, we have observed that it takes on average six or seven seconds to fit model M2 when using the multinomial property, while it takes more than one and a half hours to fit model M2 without using the multinomial property. However, for the situation when the sample size is relatively small in the sense that many test-result combinations are observed with zero frequency, imposing this multinomial distribution will distort the estimation instead. This phenomenon has been observed in the first simulations scenario in Section 3.4.

The R code of data preparation and JAGS model syntax for models M1 and M2 with and without using the multinomial property is included in the appendix B.1. Specifically, Algorithm 1 is for model M1 without the multinomial imposition; Algorithm 2 is for model M2 without the multinomial imposition; Algorithm 3 is for model M1 with the multinomial imposition; and Algorithm 4 is for model M2 with the multinomial imposition.

3.4 Simulation study

In this section, we conduct a simulation study to demonstrate the advantage of using the conditional dependence model when the dependence between tests does exist, to compare the performance of the four algorithms, and to investigate the operating characteristics of the correlation residual analysis. We consider two simulation scenarios, with the first scenario similar to the real data example in Section 5.2 with a relatively small number of patients and the second scenario similar to the real data example in Section 5.3 with a large number of patients. Specifically, in the first scenario, 400 (n) patients are rated by 4 (K) tests with test sensitivities $(0.96, 0.87, 0.81, 0.86)$ and test specificities $(0.97, 0.98, 0.99, 0.97)$. Among the 4 tests, $C_{23}^{+} = 0.05$, $C_{34}^{+} = 0.05$, $C_{23}^{-} = 0.001$, $C_{34}^{-} = 0.001$, and all other pairwise covariances set to 0. In the

second scenario, 4000 (n) patients are rated by 5 (K) tests with test sensitivities $(0.77, 0.65, 0.71, 0.68, 0.71)$ and test specificities $(0.89, 0.93, 0.88, 0.90, 0.86)$. Among the 5 tests, $C_{23}^{+} = 0.05$, $C_{14}^{+} = 0.05$, $C_{24}^{+} = 0.05$, $C_{25}^{+} = 0.05$, $C_{35}^{+} = 0.001$, $C_{23}^{-} = 0.001$, $C_{14}^{-} = 0.001$, $C_{24}^{-} = 0.001$, $C_{25}^{-} = 0.001$, $C_{35}^{-} = 0.05$, and all other pairwise covariances set to 0. For both scenarios, the underlying disease prevalence π is 0.45. Given these true parameter values, data are generated according to the conditional dependence model. Specifically, the test results for each patient are randomly simulated from a categorical distribution with 2^K possible test-result combinations and the probability of each possible test-result combination calculated using the individual likelihood defined in (3.1) with the two conditional probabilities calculated via equation (3.4). For each scenario, 200 data sets are simulated.

Table 3.1: Simulation results for scenario 1: $K = 4$, $n = 400$, and 200 data replicates.

	Truth	M1 (Algorithm 1) Bias (SD)	M1 (Algorithm 3) Bias (SD)	M2 (Algorithm 2) Bias (SD)	M2 (Algorithm 4) Bias (SD)
π	0.450	-0.011 (0.025)	-0.002 (0.025)	0.000 (0.026)	0.006 (0.026)
Se_1	0.960	-0.009 (0.017)	-0.027 (0.019)	-0.007 (0.017)	-0.021 (0.020)
Se_2	0.870	0.019 (0.024)	0.003 (0.025)	0.002 (0.026)	-0.009 (0.027)
Se_3	0.810	0.031 (0.029)	0.019 (0.029)	-0.002 (0.033)	-0.001 (0.029)
Se_4	0.860	0.016 (0.025)	0.001 (0.026)	0.001 (0.027)	-0.009 (0.028)
Sp_1	0.970	-0.017 (0.016)	-0.020 (0.016)	0.000 (0.013)	-0.006 (0.014)
Sp_2	0.980	-0.006 (0.011)	-0.011 (0.012)	-0.004 (0.011)	-0.011 (0.012)
Sp_3	0.990	-0.003 (0.007)	-0.008 (0.009)	-0.007 (0.009)	-0.017 (0.011)
Sp_4	0.970	-0.006 (0.013)	-0.011 (0.014)	0.003 (0.012)	-0.010 (0.013)
C_{23}^{+}	0.050	—	—	-0.001 (0.015)	-0.008 (0.013)
C_{34}^{+}	0.050	—	—	0.001 (0.015)	-0.006 (0.013)
C_{23}^{-}	0.001	—	—	0.004 (0.004)	0.005 (0.005)
C_{34}^{-}	0.001	—	—	0.003 (0.004)	0.005 (0.005)
	RMSE	0.01567	0.01501	0.00416	0.01195
	DIC	1224.128	238.467	1189.774	208.30
	Time	18.488 secs	1.237 secs	36.217 secs	1.893 secs

We fit each data set using the four algorithms and summarize the simulation results in Table 3.1 and Table 3.2 for the two simulation scenarios, respectively. In the tables, Bias stands for the difference between the average of 200 point estimates

Table 3.2: Simulation results for scenario 2: $K = 5$, $n = 4000$, and 200 data replicates.

	Truth	M1 (Algorithm 1) Bias (SD)	M1 (Algorithm 3) Bias (SD)	M2 (Algorithm 2) Bias (SD)	M2 (Algorithm 4) Bias (SD)
π	0.450	-0.020 (0.011)	-0.017 (0.011)	0.002 (0.013)	0.004 (0.013)
Se_1	0.770	-0.008 (0.013)	-0.011 (0.013)	-0.004 (0.015)	-0.006 (0.015)
Se_2	0.650	0.044 (0.015)	0.042 (0.015)	0.001 (0.016)	0.000 (0.016)
Se_3	0.710	0.040 (0.012)	0.038 (0.012)	0.003 (0.015)	0.001 (0.014)
Se_4	0.680	0.019 (0.014)	0.017 (0.014)	-0.002 (0.016)	-0.002 (0.016)
Se_5	0.710	0.048 (0.012)	0.045 (0.012)	0.002 (0.014)	0.000 (0.014)
Sp_1	0.890	-0.029 (0.009)	-0.029 (0.009)	-0.002 (0.010)	-0.002 (0.011)
Sp_2	0.930	0.080 (0.006)	0.079 (0.006)	0.069 (0.009)	0.067 (0.009)
Sp_3	0.880	0.010 (0.010)	0.010 (0.010)	0.003 (0.011)	0.002 (0.011)
Sp_4	0.900	-0.008 (0.008)	-0.009 (0.008)	-0.003 (0.009)	-0.003 (0.010)
Sp_5	0.860	-0.016 (0.010)	0.017 (0.010)	0.003 (0.011)	0.002 (0.011)
C_{23}^{+}	0.050	—	—	-0.002 (0.006)	-0.002 (0.006)
C_{14}^{+}	0.050	—	—	0.001 (0.007)	-0.001 (0.007)
C_{24}^{+}	0.050	—	—	0.000 (0.006)	0.000 (0.006)
C_{25}^{+}	0.050	—	—	-0.002 (0.006)	-0.002 (0.006)
C_{35}^{+}	0.001	—	—	-0.002 (0.006)	-0.002 (0.006)
C_{23}^{-}	0.001	—	—	0.000 (0.004)	0.000 (0.004)
C_{14}^{-}	0.001	—	—	0.002 (0.005)	0.001 (0.005)
C_{24}^{-}	0.001	—	—	-0.001 (0.002)	0.000 (0.002)
C_{25}^{-}	0.001	—	—	-0.001 (0.004)	0.001 (0.004)
C_{35}^{-}	0.050	—	—	-0.002 (0.006)	-0.002 (0.006)
	RMSE	0.01882	0.01914	0.00587	0.00586
	DIC	22017.830	937.885	21569.440	495.858
	Time	12.592 mins	2.693 secs	98.422 mins	6.261 secs

(posterior means) and the true values; SD stands for the average of 200 posterior standard deviations; RMSE is the square root of the average of 200 MSEs with each individual MSE calculated between each estimated diagnostic accuracy parameter (sensitivities and specificities) and their true parameter values; DIC stands for the average Deviance Information criterion (Speigelhalter, 2003), that is, the average of 200 DICs with each DIC produced by the JAGS. For both scenarios, Table 3.1 and Table 3.2 show that after incorporating the pairwise covariances, model M2 provides overall more accurate estimation of the disease prevalence, sensitivities and specificities than model M1. Specifically, the point estimates for model M2 are much

less biased with only slightly larger posterior standard deviations than model M1 and the RMSEs of model M2 are smaller than those of model M1. Furthermore, model M2 provides unbiased estimates of the covariances, which model M1 is unable to.

In terms of using the multinomial imposition, we can see that the algorithms with the multinomial imposition take much less time for the fitting. For example, Table 3.2 shows that for the second simulation scenario, it takes on average more than one and a half hours to fit model M2 without the multinomial imposition while it takes only 6 seconds to fit model M2 with the multinomial imposition. Table 3.1 shows that model M2 with the multinomial imposition (Algorithm 4) actually leads to worse estimation results with larger biases than model M2 without the multinomial imposition (Algorithm 2) for the first scenario, while Table 3.2 shows that M2 with the multinomial imposition (Algorithm 4) provides almost identical estimation results as those without the multinomial imposition (Algorithm 2) for the second scenario. This discovery implies that when the sample size n is large in the sense that the number of unique observed test result combinations equals to 2^K, using the multinomial property for the fitting can provide the estimation results as good as the original model and meanwhile it alleviates the computational burden. However, when the sample size is small, imposing the multinomial property will distort the estimation instead. Finally, comparing these two tables, we can see that as the sample size n increases, the biases and posterior standard deviations both clearly decrease for using each algorithm.

Table 3.3 and Table 3.4 summarize the correlation residual analysis for the 200 simulated data sets for each of the two simulation scenarios. Table 3.3 presents the selection frequency of significant correlation pairs for scenario 1 based on the one-sided credible interval criterion (described in Section 2.3) after fitting model M1. Table 3.3 shows that the number of correct inclusion of correlation pairs is 149. There are 24 times that only r_{34} is selected, 20 times only r_{23} is selected, and very few times that some other correlations are selected. Additional simulation studies (see Appendix

Table 3.3: Frequencies of correlation selection for scenario 1: $K = 4$, $n = 400$, and 200 data replicates. Note that "1" indicates "being selected"

Frequency	r_{12}	r_{13}	r_{23}	r_{14}	r_{24}	r_{34}
149	0	0	1	0	0	1
24	0	0	0	0	0	1
20	0	0	1	0	0	0
2	0	0	1	1	0	0
3	0	0	1	1	0	1
1	0	0	0	0	0	0
1	1	0	1	0	0	1

Table 3.4: Frequencies of correlation selection for scenario 2: $K = 5$, $n = 4000$, and 200 data replicates. Note that "1" indicates "being selected"

Frequency	r_{12}	r_{13}	r_{23}	r_{14}	r_{24}	r_{34}	r_{15}	r_{25}	r_{35}	r_{45}
124	0	0	1	1	1	1	0	1	1	1
23	0	0	1	1	1	1	1	1	1	1
2	0	1	1	1	1	1	1	1	1	1
30	1	0	1	1	1	1	0	1	1	1
17	0	1	1	1	1	1	0	1	1	1
4	1	0	1	1	1	1	1	1	1	1

B.2) show that partially missing importance covariances results in a small amount of estimation bias, but the bias is smaller than using the completely misspecified conditional independence model (M1). On average, the estimation results for scenario 1 when fitting with M2 with selected significant covariances by the one-sided credible interval criterion are very close to those when fitting with the true model. Table 3.4 presents the selection frequency of significant correlation pairs for scenario 2 where the true model has significant covariances C_{23}^{+}, C_{14}^{+}, C_{24}^{+}, C_{25}^{+} and C_{35}^{-}. Table 3.4 shows that besides the true significant pairs r_{23}, r_{14}, r_{24}, r_{25}, and r_{35} are selected, two extra pairs r_{34} and r_{45} are also always selected. The other three pairs r_{12}, r_{13}, and r_{15} are only selected 34, 19, and 29 times, respectively. Based on this observation, we fit the data with model M2 having these seven pairs of covariances added. The estimation results (see Appendix B.3) are comparable to those obtained from the true model.

Particularly, the estimates for the two extra covariance terms are approximately 0. A further simulation study shows that adding all ten pairs of covariances drastically worsens the estimation. In summary, when the covariance pattern is simple as in scenario 1, our correlation residual check method can frequently detect the pattern; when the covariance pattern is complicated as in scenario 2, a few more covariance terms may be selected without much influence on the estimation; when all pairwise covariance terms are included, estimation results can be poor.

3.5 Data analysis

In this section, we illustrate our proposed methods through three examples.

3.5.1 Example 1: Beam's mammogram data

Beam's mammogram data (Beam et al., 2003) was collected to study the variability in the interpretation of mammograms by a national sample of radiologists in the United States. It contains diagnostic results of 107 radiologists each evaluating 146 women from a breast cancer screening program. The true disease status of each patient (breast cancer or no breast cancer) was known from a biopsy examination or a minimum 2-year follow-up study. See more details of this data set in Beam et al. (2003). Beam's data was analyzed by Lin et al. (2018) to estimate radiologist diagnostic skills with a latent variable model, where the diagnostic results were dichotomized. Here we illustrate our Bayesian hierarchical conditional independence model with this dichotomous rating data. Pairwise covariances and the multinomial property are not convenient to use due to the large number (=107) of raters, and therefore we restrict our analysis to model M1 and Algorithm 1 in the Appendix B.1 is applied. The breast cancer disease prevalence π is estimated as 0.44 based upon the sample of 146 women. The mode of rater sensitivities, ω_1, is estimated as 0.94 and mode of rater specificities, ω_2, is estimated as 0.86. These modes measure the diagnostic accuracy

of the whole population of radiologists. The estimated concentrate term κ_1 is 16 and the estimated κ_2 is 10.4 indicating a wider spread of rater specificities. So, it is reasonable to conclude that the raters usually reach higher consensus on identifying the subjects as being diseased when the truth disease status is diseased. Figure 3.1 plots the estimated sensitivities and specificities obtained from model M1 versus the empirical sensitivities and specificities calculated from the data, demonstrating that the latent class model works well for the Beam's data.

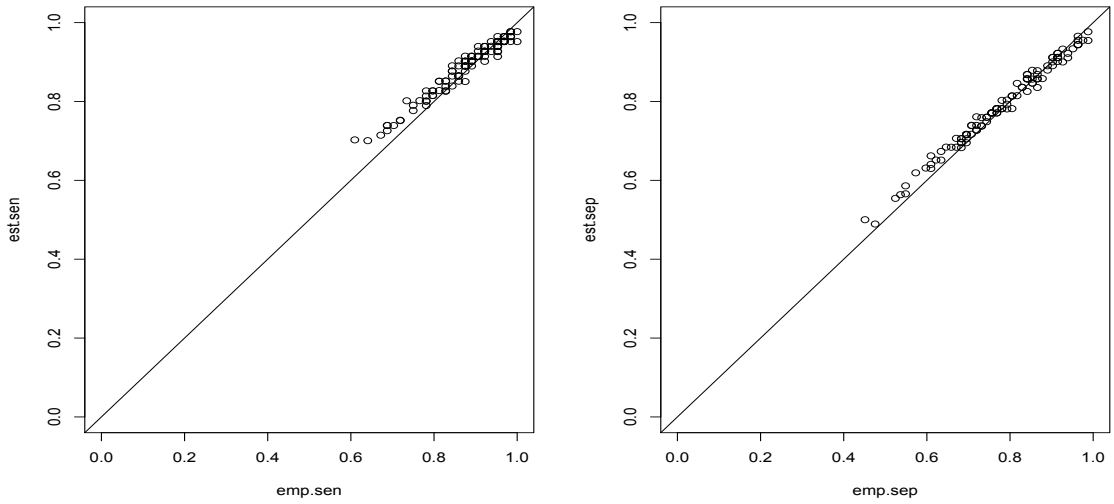

Figure 3.1: Plots of estimated sensitivities and specificities v.s. empirical sensitivities and specificities for the mammogram data in Example 1.

One advantage of using latent class models is that they do not require information of the true disease status. Theoretically, when the disease status is known, the nonidentifiability issue (if it exists) would be automatically resolved and the estimation is more accurate. Figure 3.2 shows that with and without true disease status, the estimated sensitivities and specificities are very close for this data set, which showing that there is no identifiability problem for the Beam's data and again the latent class model works well for the Beam's data. In addition, the estimated beta distributions for sensitivities and specificities remarkably catch the density curves of empirical sensitivities and specificities, which is presented in Figure 3.3.

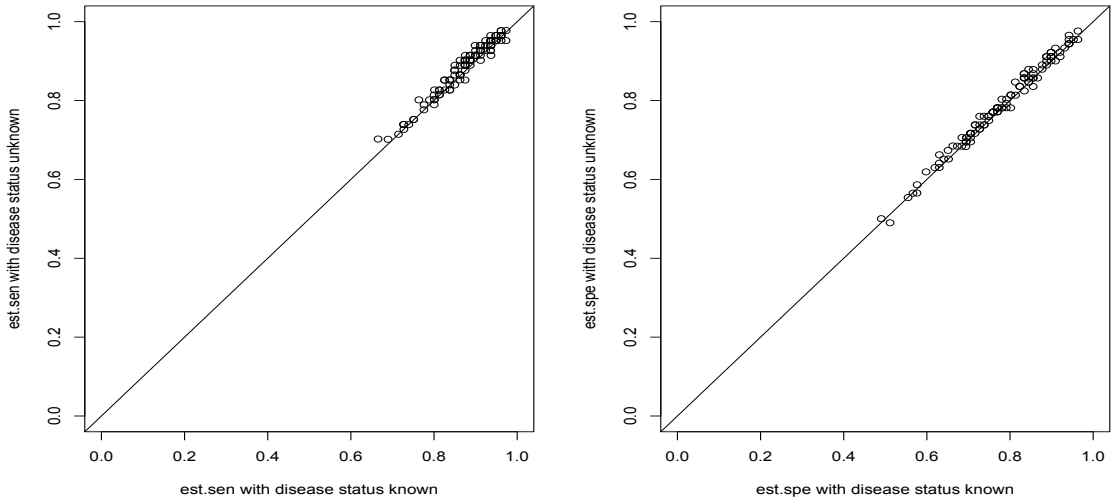

Figure 3.2: Plots of estimated sensitivities and specificities with true disease status known v.s. those without true disease status known for the mammogram data in Example 1.

3.5.2 Example 2: Alvord's HIV data

Alvord's HIV data (Alvord et al., 1988) was used to determine the diagnostic performance of HIV antibody assays. In the study, serum samples from each of the 428 subjects were tested by four conventional bioassays (tests). Alvord et al. (1988) showed that the traditional independent latent two-class model is inadequate to fit the data and a three-class model can properly resolve the issue. Qu et al. (1996) used correlation residual plot to check covariance pattern and applied Gaussian random effects to model the test dependence, which was shown to greatly improve the fitting greatly. One drawback of their model is that the parameters introduced are not easily interpretable and cannot directly explain the covariance. The pairwise covariance dependence model (M2) proposed in this chapter can naturally solve this problem and can easily be applied to this study with four bioassays. Because nine test result combinations have observed frequencies as 0, the multinomial property is not suitable to use.

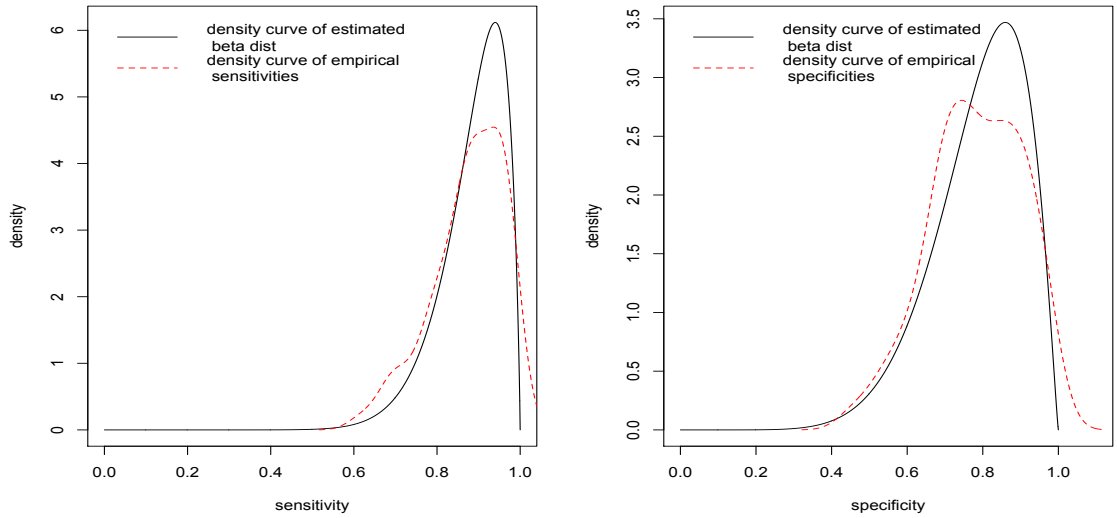

Figure 3.3: Density curves of the estimated and empirical distributions of sensitivities and specificities for the mammogram data in Example 1.

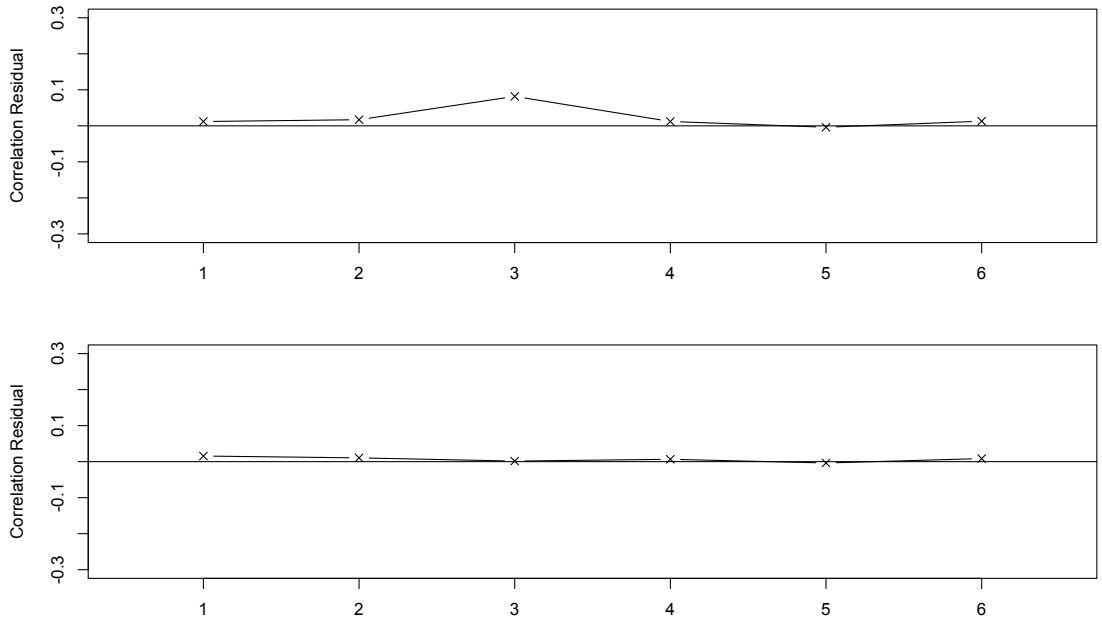

Figure 3.4: Pairwise correlation residual plots for models M1 and M2 for the HIV data in Example 2. The top panel is for model M1; the bottom panel is for model M2.

Table 3.5: Observed and expected frequencies for models M1 and M2 without the multinomial impose for the HIV data in Example 2

Test results	Observed frequency	Expected frequency			
		M1 (Algorithm 1)		M2 (Algorithm 2)	
		Mean	Median	Mean	Median
0000	170	169.12	168.72	168.62	168.31
1000	4	5.46	5.20	5.32	4.95
0100	6	6.40	6.18	6.48	6.27
1100	1	0.27	0.25	0.23	0.21
0001	15	13.87	13.58	13.97	13.68
1001	17	9.38	9.25	16.84	16.56
1101	4	11.93	11.70	4.77	4.47
1011	83	88.90	88.97	81.66	81.46
1111	128	118.67	118.24	126.19	125.88
Goodness of fit (Chi-square test p value)		0.005749	0.007699	0.9953	0.9979
DIC		1284.388		1273.979	

Table 3.6: Estimation results (posterior quantities) from models M1 and M2 without using the multinomial impose for the HIV data in Example 2

	M1 (Algorithm 1)			M2 (Algorithm 2)		
	Mean	Median	SD	Mean	Median	SD
π	0.540	0.540	0.025	0.542	0.542	0.024
Se_1	0.995	0.997	0.005	0.995	0.997	0.005
Se_2	0.572	0.572	0.032	0.572	0.571	0.032
Se_3	0.909	0.910	0.020	0.908	0.908	0.019
Se_4	0.995	0.996	0.005	0.995	0.996	0.005
Sp_1	0.969	0.971	0.013	0.970	0.972	0.013
Sp_2	0.964	0.965	0.013	0.960	0.961	0.013
Sp_3	0.993	0.995	0.006	0.993	0.995	0.006
Sp_4	0.924	0.926	0.020	0.924	0.925	0.018
C_{23}^{+}	—	—	—	0.032	0.032	0.010
C_{23}^{-}	—	—	—	0.003	0.002	0.004

The top panel of Figure 3.4 shows the pairwise correlation residuals in order of (T_1,T_2), (T_1,T_3), (T_2,T_3), (T_1,T_4), (T_2,T_4) and (T_3,T_4) after model M1 is fitted. Clearly, the correlation between T_2 and T_3 is not negligible. Based upon this information, model M2 is applied with C_{23}^{+} and C_{23}^{-} introduced and the other covariance terms set 0. The correlation residual plot of this model is on the bottom panel of Figure 3.4, showing all of the correlation residuals close to 0 and indicating an improved goodness of fit. Table 3.5 shows the observed frequencies and expected frequencies from M1 as well as M2. Clearly, M2 provides a better fit than M1. The chi-square goodness of fit test between the observed and expected frequencies indicates that M1 does not provide a good fit for the data since the p-values are only 0.0057 for posterior mean frequencies and 0.0077 for posterior median frequencies. The p-values of the chi-square test for M2 are above 0.9 for both posterior mean and median frequencies, suggesting M2 provides a more appropriate fit. Bayesian model comparison criterion DIC also confirms that M2 fits the data better with a smaller DIC value. Table 3.6 shows the estimated sensitivities and specificities from M1 and M2, which are very similar to those in Qu et al. (1996). The correlation residual plot shows the correlation pattern as a whole, while the estimated covariance terms nicely decompose this term into two parts based on the two latent classes. In this example, C_{23}^{-} is negligible compared to C_{23}^{+}, which tells the dependence between test 2 and test 3 mainly comes from diagnosing subjects who are diseased.

3.5.3 Example 3: Handelman's dentistry data

Handelman's dentistry data (Espeland and Handelman, 1989) focuses upon 5 dentists each evaluating 3869 dental x-rays according to a binary rating scale with 0 denoting sound and 1 denoting carious. See more details of this data set in Espeland and Handelman (1989). This data set contains a large number of subjects ($n = 3869$). It has the complete frequency information for all the possible test-result combinations, that

is, the number of unique observed test-result combinations of the data equals the total number 2^5 (= 32) of possible test-result combinations. Therefore, the multinomial property is imposed for fitting models M1 and M2.

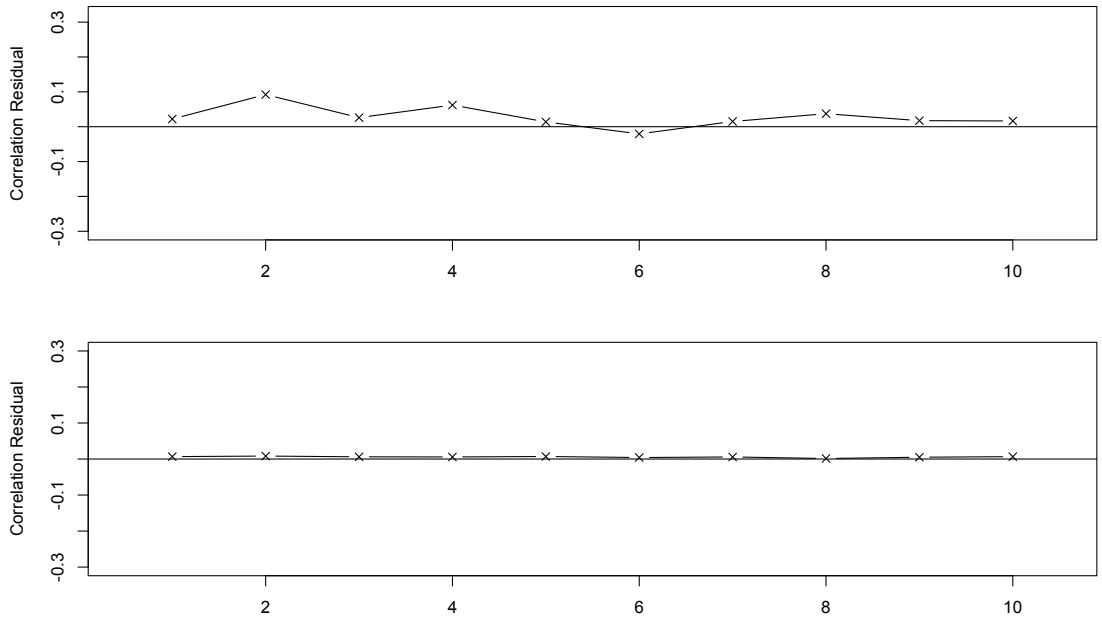

Figure 3.5: Pairwise correlation residual plots for models M1 and M2 for the dentistry data in Example 3. The top panel is for model M1; the bottom panel is for model M2.

From the top panel of Figure 3.5, we see the correlation pattern is more complicated. Both positive and negative correlations are displayed. The pairwise correlation from left to right are in the order of (T_1,T_2), (T_1,T_3), (T_2,T_3), (T_1,T_4), (T_2,T_4), (T_3,T_4), (T_1,T_5), (T_2,T_5), (T_3,T_5) and (T_4,T_5). Instead of handling each evident covariance separately, we recommend adding all pairwise covariance terms and letting the data decide the values of covariances. The correlation residual plot for M2 at the bottom panel of Figure 3.5 shows all the residuals are close to 0, which indicates M2 provides an improved and good fit for the data.

Table 3.7 shows the observed and expected frequencies estimated from models M1 and M2. The estimated frequencies from model M2 are much closer to the observed

Table 3.7: Observed and expected frequencies for models M1 and M2 with the multinomial impose for the dentistry data in Example 3

Test results	Observed frequency	Expected frequency			
		M1 (Algorithm 3)		M2 (Algorithm 4)	
		Mean	Median	Mean	Median
00000	1880	1822.27	1821.36	1860.31	1860.33
00001	789	821.36	820.14	780.61	780.52
00010	43	62.67	62.22	43.45	43.08
00011	75	50.72	50.70	76.91	76.55
00100	23	30.05	30.06	23.64	23.39
00101	63	48.65	48.71	63.57	63.21
00110	8	4.91	4.84	10.94	10.53
00111	22	36.26	36.09	22.17	21.80
01000	188	212.74	212.46	187.77	187.77
01001	191	150.99	150.84	194.21	193.36
01010	17	13.41	13.34	20.28	19.88
01011	67	61.42	61.15	64.10	63.70
01100	15	13.01	12.95	17.30	16.97
01101	85	90.50	90.50	83.62	83.35
01110	8	9.78	9.68	6.77	6.56
01111	56	85.87	85.73	59.41	59.11
10000	22	22.81	22.62	22.25	21.96
10001	26	26.60	26.49	27.41	26.97
10010	6	2.59	2.57	8.90	8.48
10011	14	17.10	17.04	13.91	13.67
10100	1	3.19	3.16	3.22	2.85
10101	20	25.77	25.67	19.99	19.82
10110	2	2.79	2.77	2.19	2.06
10111	17	24.69	24.53	19.88	19.64
11000	2	7.08	6.99	4.21	3.85
11001	20	42.79	42.60	20.01	19.79
11010	6	4.62	4.55	4.50	4.34
11011	27	40.20	40.09	30.90	30.66
11100	3	6.96	6.87	3.34	3.14
11101	72	61.40	61.19	75.26	75.17
11110	1	6.66	6.59	5.10	5.10
11111	100	59.15	58.93	92.89	92.63
Goodness of fit (Chi-square test p value)		$< 2.2e-16$	$< 2.2e-16$	0.4736	0.6402
DIC		619.1735		540.0874	

frequencies than those from model M1, especially for the frequencies of “00000" and “11111". The chi-square goodness of fit test also indicates that the M1 does not provide a good fit for the data with small p-values while M2 is shown to fit the data well with p-values larger than 0.4. A smaller value of DIC also indicates that M2 is the more appropriate model.

Table 3.8: Estimation results (Posterior quantities) from models M1 and M2 with the multinomial impose for the dentistry data in Example 3

	M1 (Algorithm 3)			M2 (Algorithm 4)		
	Mean	Median	SD	Mean	Median	SD
π	0.202	0.202	0.010	0.173	0.171	0.034
Se_1	0.407	0.407	0.021	0.442	0.439	0.073
Se_2	0.706	0.706	0.021	0.755	0.756	0.083
Se_3	0.597	0.596	0.023	0.604	0.601	0.090
Se_4	0.493	0.493	0.021	0.498	0.497	0.059
Se_5	0.899	0.900	0.014	0.937	0.949	0.038
Sp_1	0.989	0.989	0.002	0.979	0.979	0.005
Sp_2	0.898	0.898	0.007	0.884	0.881	0.018
Sp_3	0.986	0.986	0.003	0.963	0.963	0.009
Sp_4	0.968	0.968	0.004	0.951	0.951	0.012
Sp_5	0.695	0.695	0.009	0.683	0.680	0.029

Table 3.8 displays the estimated posterior means, medians and standard deviations for sensitivities and specificities from models M1 and M2. For our method, we observe that the posterior mean and median of sensitivities from model M2 tend to be larger than those from model M1 and the posterior mean and median of specificities tend to be smaller. In addition, the posterior standard deviations from M2 are overall larger than those from M1. Compared with Qu et al. (1996)’s results, our estimated sensitivities and specificities from model M1 (algorithm 3) correspond closely to those from their 2LC model, while the estimated sensitivities and specificities from model M2 (algorithm 4) are not as closely aligned with those from their 2LCR model. Table

3.9 displays the estimated pairwise covariances with C_{13}^{+} estimated apparently larger than other covariances.

Table 3.9: Estimated pairwise covariances conditional on the diseased class and the non-diseased class for the dentistry data in Example 3

		T_1	T_2	T_3	T_4	T_5
C^+	T_1	—	0.017	0.057	0.028	0.004
	T_2	—	—	0.019	0.010	0.003
	T_3	—	—	—	-0.007	0.007
	T_4	—	—	—	—	-0.006
	T_5	—	—	—	—	—
C^-	T_1	—	0.001	0.001	0.004	0.003
	T_2	—	—	0.006	0.005	0.014
	T_3	—	—	—	0.005	0.008
	T_4	—	—	—	—	0.010
	T_5	—	—	—	—	—

3.6 Summary

In this chapter, we propose two Bayesian hierarchical latent class models to allow the estimation of sensitivity and specificity of multiple diagnostic tests with or without gold standard data. Our proposed models build upon existing approaches by flexibly accommodating a large number of diagnostic tests. Further, our proposed pairwise covariance dependence model (M2), in contrast to Qu et al (1996)'s approach, provides easily interpretable estimates of parameters and direct interpretation of covariance parameters. Through the Bayesian hierarchical structure, individual sensitivities and specificities are modeled as random effects following two common overarching beta distributions, respectively. The mode parameters in the overarching beta distributions therefore reflect the group-level sensitivity and specificity. Further, the concentrate parameters implicitly reflect the rating agreement among the group with larger values indicating more consistency of group diagnostic performance.

Another contribution of this chapter is to provide easy-to-implement JAGS algorithms to apply these models. A guideline to using these algorithms is provided. Algorithm 1 for the independence model is always the first attempt to fit the data. If the correlation residual analysis detects significant dependence between at least two of the tests, then Algorithm 2 for the conditional dependence model is implemented to improve the fitting. Algorithm 3 and 4 are analogous to Algorithm 1 and 2 with the multinomial distribution imposition. When the number n of subjects is large relative to the number 2^K of possible test result combinations, using the multinomial distribution property can significantly reduce the computational burden. For example, for example 3, it takes 1.5 hours to run Algorithm 2 while it takes only 19 seconds to run Algorithm 4.

Although the pairwise covariance model is easy to apply and interpret, one limitation is the ability to handle data from other models, such as the GRE model (Qu et al., 1996) and the finite mixture (FM) model (Albert and Dodd, 2004). Our simulation study (see Appendix B.4) shows that for the data generated from these two models, our correlation residual analysis tends to include all pairs of covariances for the pairwise covariance model and our method leads to biased estimation results. One possible reason is that the pairwise covariance model explains covariances between tests directly while the GRE and FM model introduce the correlation between tests through common subject random effects. Finally, for real data analysis, we recommend using model selection criterion, such as DIC, to choose the best model among different models.

Chapter 4

Bayesian Semiparametric Regression Analysis of Multivariate Panel Count Data

4.1 Introduction

Panel count data often arise in epidemiological and medical studies, in which the events of interest have the property of recurrence and study subjects are monitored periodically. Since the subjects are not under continuous monitoring, the exact time of each recurrent event is not observed but the count of such events between adjacent observation times is known. For many studies, panel count data of several types of related recurrent events are collected. For example, the recurrent events can be different types of infections, tumors, and social behaviors such as drinking and drug use. The motivating example is a bivariate panel count data set on skin cancers from the literature (Sun and Zhao, 2013). The data arise from a skin cancer chemoprevention trial conducted by the University of Wisconsin Comprehensive Cancer Center. It is a double-blinded and placebo-controlled randomized III clinical trial. The main objective of the study is to evaluate the effectiveness of $0.5/m^2$/day PO difluoromethylornithine(DFMO) in reducing the recurrence rate of skin cancers in a population of patients with a history of non-melanoma skin cancers: basal cell carcinoma and squamous cell carcinoma. Two hundred and ninety one patients were randomized into either a placebo group or a DFMO group. During the study, the patients were scheduled to be examined every 6 months to check the development of both skin cancers. At each visit, the numbers of occurrences of both basal cell

carcinoma and squamous cell carcinoma since the previous visit were recorded. For this bivariate panel count data set, it's possible to analyze these two skin cancers separately to evaluate the effectiveness of DFMO. However, conducting a joint analysis is a better practice to investigate the correlation between the two cancers and to improve estimation efficiency.

Most methods established to analyze panel count data focus on a single type of recurrent event, for example, Sun and Kalbfleisch (1995), Wellner et al. (2000, 2007), Lu et al. (2009, 2007), Sun and Zhao (2013), Wang and Lin (2019), among others. There is only a limited number of papers proposed to analyze multivariate panel count data. He et al. (2008) considered regression analysis for multivariate panel count data and first proposed a class of marginal mean models which leave the dependence structures for related recurrent events completely unspecified. Zhang et al. (2013) then improved their model and provided a robust joint modeling approach for the regression analysis of multivariate panel count data with an informative observation process. Li et al. (2011) proposed semiparametric transformation models that allow for the dependence of the recurrent event processes on the observation process. Along the same line, Zhao et al. (2013) proposed a semiparametric additive model to analyze multivariate panel count data with dependent observation processes and a terminal event. All these models emphasize the dependence of the recurrent event processes on the observation process and require the model of the observation process to be explicitly specified. In this paper, the proposed approach is based on the observed likelihood, using only the observed counts and observation times, for which is not required to specify a model for the observation process. Instead of building up the correlation among multiple events through their dependence on the observation process, we tackle the correlations between different types of events directly by introducing common subject-specific gamma frailty terms and additional scale parameters. The resulted pairwise correlations can be calculated in a closed form and flexibly ac-

commodate different correlation structures including positive, negative, strong, and weak correlations. Sinha and Maiti (2004) applied the same strategy to model the dependence between panel count data and termination time. However, their application is only for a single type of recurrent event and requires that each subject have the same observation times with the termination time being one of those.

Frequentist approaches (He et al., 2008; Zhang et al., 2013; Li et al., 2010; Zhao et al., 2013) use estimating equation methods to estimate regression coefficients. These methods do not provide estimates for the baseline mean functions or the correlation between events. They usually require additional formulas or bootstrap procedures to estimate the standard errors of the estimated regression coefficients. In this paper, the Bayesian semiparametric approach can be used to estimate the regression coefficients and the baseline mean functions simultaneously. Unlike the frequentist methods, Markov chain Monte Carlo (MCMC) samples can also evaluate the posterior standard deviations of the regression coefficients directly (Gelman et al., 2013).

The remainder of this chapter is organized as follows. In Section 4.2, primary models and the correlation derivations and interpretations are introduced. In Section 4.3, a detailed description of monotone I-splines, augmented likelihood function construction, prior specification and posterior computation are presented. Section 4.4 evaluates our proposed methods via comprehensive simulation studies. The skin cancer example is used to demonstrate the performance of the proposed methods in Section 4.5. Lastly, Section 4.6 summarizes our findings and discusses some possible future research.

4.2 Model and Notation

4.2.1 Model Construction

Consider that n subjects participate in a long-term study involving K types of related recurrent events. Each subject is not under continuous monitoring and instead is

observed at discrete time points. Specifically, for each subject i, denoted by $J_i^{(k)}$ the total number of observation times for event k and the corresponding observation times are $0 = t_{i0}^{(k)} < t_{i1}^{(k)} < t_{i2}^{(k)} < ... < t_{iJ_i^{(k)}}^{(k)}$. Let $N_i^{(k)}(t)$ denote the cumulative count of the occurrence of event k prior to time t for subject i. Let $X_i^{(k)} = (x_{i1}^{(k)}, ..., x_{ip}^{(k)})'$ denote the $p \times 1$ covariate vector associated with subject i for event k. For simplicity, in this paper we assume the covariates for subject i are identical for all K events and denoted by X_i. The whole set of observed panel count data is denoted by $\mathcal{D} = \{t_{ij}^{(k)}, N_i^{(k)}(t_{ij}^{(k)}), X_i, \text{ for } k = 1, ..., K; i = 1, ..., n; j = 1, ..., J_i^{(k)}\}$. Finally, we assume there is a latent subject-specific positive frailty term w_i which affects the occurrence rates and connects the multiple events. The frailties $\{w_i\}$ are assumed to follow a gamma distribution $\mathcal{G}(\eta, \eta)$ with mean 1 and variance $1/\eta$. The model's identifiability is satisfied by setting the mean of the frailties equal to 1 (Sun et al., 2007; Sinha and Maiti, 2004).

Given the covariates X_i and frailty w_i, we assume the counting process $\{N_i^{(k)}(t),\ t > 0\}$ for event k has a proportional mean function in the following form:

$$E\left(N_i^{(k)}(t)|X_i, w_i\right) = w_i^{\alpha_k} U_0^{(k)}(t) \exp(X_i' \boldsymbol{\beta}^{(k)}), \tag{4.1}$$

where α_k is a scale parameter introduced to more flexibly accommodate the correlation between events, $U_0^{(k)}(t)$ is the baseline mean function for event k, and $\boldsymbol{\beta}^{(k)}$ is a $p \times 1$ vector of regression coefficients for event k. By default, the model considers different covariate effects for different recurrent events, but it is easy to extend to the situation where the covariate effects are identical across all K events. More details about α_k are discussed in Section 4.2.2.

Let $Z_{ij}^{(k)}$ denote the number of occurrences of event k in the jth time interval $(t_{i,j-1}^{(k)}, t_{ij}^{(k)}]$ for subject i, i.e. $Z_{ij}^{(k)} = N_i^{(k)}(t_{ij}^{(k)}) - N_i^{(k)}(t_{i,j-1}^{(k)})$, where we assume $N_i^{(k)}(t_{i0}) = 0$ for $i = 1, ..., n$. Through this transformation, the whole set of data can be expressed as $\mathcal{D} = \{t_{ij}^{(k)}, Z_{ij}^{(k)}, X_i, \text{for } k = 1, ..., K; i = 1, ..., n; j = 1, ..., J_i^{(k)}\}$. Under the Poisson process assumption, given the covariates X_i and latent frailty term w_i,

$Z_{ij}^{(k)}$'s independently follow Poisson distributions, which can be written as:

$$Z_{ij}^{(k)}|X_i, w_i \sim \mathcal{P}\left(w_i^{\alpha_k}\{U_0^{(k)}(t_{ij}^{(k)}) - U_0^{(k)}(t_{i,j-1}^{(k)})\}\exp(X_i'\boldsymbol{\beta}^{(k)})\right). \tag{4.2}$$

This form is particularly useful for constructing the likelihood function in Section 4.3.2. Note that when integrating out the frailty effect w_i, $Z_{ij}^{(k)}$'s marginally follow mixed Poisson distributions (Sun and Zhao, 2013).

4.2.2 Correlation expression

An advantage of our proposed model is its straightforwardness in deriving the correlation formula between two events. Among the K events, we predesignate one event that is of main interest, for instance event j, as a reference event, and let $\alpha_j = 1$. Then, given the covariates X_i and the unobservable frailty w_i, the cumulative counts of event j and event k ($k \neq j$) for subject i at any time point t are conditionally independent and have

$$N_i^{(j)}(t)|X_i, w_i \sim \mathcal{P}\left(w_i U_0^{(j)}(t)\exp(X_i'\boldsymbol{\beta}^{(j)})\right),$$

$$N_i^{(k)}(t)|X_i, w_i \sim \mathcal{P}\left(w_i^{\alpha_k} U_0^{(k)}(t)\exp(X_i'\boldsymbol{\beta}^{(k)})\right).$$

Using the law of total variance, we explicitly derive $Cov\Big(N_i^{(j)}(t), N_i^{(k)}(t)\Big)$, $Var\Big(N_i^{(j)}(t)\Big)$ and $Var\Big(N_i^{(k)}(t)\Big)$, and hence the correlation formula as below:

$$\begin{aligned} & Corr\Big(N_i^{(j)}(t), N_i^{(k)}(t)\Big) \\ = & \frac{Cov\Big(N_i^{(j)}(t), N_i^{(k)}(t)\Big)}{\sqrt{Var\Big(N_i^{(j)}(t)\Big)Var\Big(N_i^{(k)}(t)\Big)}} \\ = & \frac{\alpha_k}{\sqrt{\eta + \frac{\eta^2}{[U_0^{(j)}(t)\exp(X_i'\boldsymbol{\beta}^{(j)})]}}\sqrt{\frac{\Gamma(\eta+2\alpha_k)\Gamma(\eta)}{(\Gamma(\eta+\alpha_k))^2} - 1 + \frac{\Gamma(\eta)\eta^{\alpha_k}}{\Gamma(\eta+\alpha_k)[U_0^{(k)}(t)\exp(X_i'\boldsymbol{\beta}^{(k)})]}}} \end{aligned} \tag{4.3}$$

where $\eta > 0$ and $\alpha_k > -\eta/2$. For details on derivation of equation (4.3), see Appendix C.1. Clearly, $\alpha_k = 0$ implies that event j and event k are independent; $\alpha_k > 0$ implies that event j and event k are positively related; $\alpha_k < 0$ implies that event j and

event k are negatively related. When getting rid of the covariate effects, the baseline correlation between the two events is

$$Corr_0\Big(N^{(j)}(t), N^{(k)}(t)\Big) = \frac{\alpha_k}{\sqrt{\eta + \frac{\eta^2}{U_0^{(j)}(t)}}\sqrt{\frac{\Gamma(\eta+2\alpha_k)\Gamma(\eta)}{(\Gamma(\eta+\alpha_k))^2} - 1 + \frac{\Gamma(\eta)\eta^{\alpha_k}}{\Gamma(\eta+\alpha_k)U_0^{(k)}(t)}}}.$$

The benefit of this form is that it eliminates study at the individual level and provides a broader view of the correlation between two events. Let $U_0^{(j)}(t) = U_0^{(k)}(t) = 1$, we can explore the pure effect of α_k and η on the baseline correlation $Corr_0(N^{(j)}(t), N^{(k)}(t))$. Note that the correlation between any two events is controlled by the parameters η and α_k. Figure 4.1 shows that with the variance of frailties increasing (η decreasing), the correlation increases; with the magnitude of α_k increasing, the correlation also increases. When $\alpha_k > 0$, any $\eta > 0$ is a legitimate choice; while for $\alpha_k < 0$, the condition $\eta > -2\alpha_k$ must be satisfied to make $\Gamma(\cdot)$ valid.

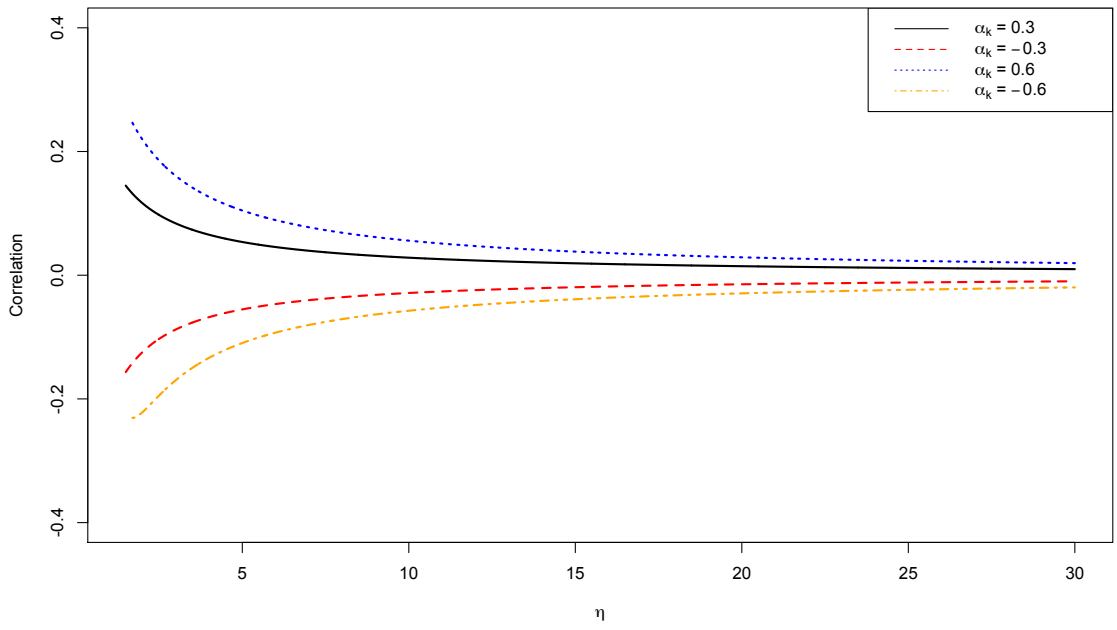

Figure 4.1: Baseline correlation between $N^{(j)}$ and $N^{(k)}$

In fact, interpretation of the pairwise correlation is not limited to the pairs of events involving the predesignated event. Conceptionally, any two powers α_p and α_q,

for $p \neq q$, can be transformed into 1 and $\frac{\alpha_q}{\alpha_p}$. A similar derivation can obtain the correlation between any two types of events. The general form of correlation between two events with α_p and α_q is shown in the Appendix C.2.

4.3 The Proposed Bayesian Semiparametric Approach

4.3.1 Modeling $U_0^{(k)}(t)$ with monotone I-splines

To accommodate the nondecreasing nature of the baseline mean functions in the proposed model, we choose monotone I-splines to model them. I-splines were first developed by Ramsay et al. (1988) and were then widely applied in many semiparametric models. To put it briefly, I-splines, a set of non-negative spline functions, are integrated M-splines (Ramsay et al., 1988). In our proposed model, each baseline function of event k is modeled as a linear combination of I-splines:

$$U_0^{(k)}(t) = \sum_{l=1}^{L} r_l^{(k)} I_l(t|d), \tag{4.4}$$

In the formula, $I_l(\cdot|d)$ is the I-spline basis function with degree d; L is the number of I-spline basis functions, which equals the number of interior knots plus the degree d; $\{r_l^{(k)}\}$ are the nonnegative spline coefficients. For more detailed information on the I-splines, refer to Ramsay et al. (1988) and Lin et al. (2015). The degree d and the placement of knots are two chief components that determine the basis functions. The former controls the smoothness and the latter controls the shape of the spline function. In general, 2 or 3 degrees is enough to provide adequate smoothness, and 10 to 30 knots can provide enough flexibility for a regression incorporating thousands of observations, according to Cai et al. (2011) and Wang and Dunson (2011). Equally spaced knots and quantile-based knots are two commonly used methods to select knots. Wang and Lin (2019) showed that the deviance information criterion (DIC) can be used to facilitate choosing the setup for the I-spline functions. In this paper, we

uniformly use degree 3 and 20 equally spaced knots (18 interior knots), which proves to provide sufficient flexibility for modeling the unknown baseline mean functions.

4.3.2 Likelihood augmentation with Poisson latent variables

Consider that n subjects participate in a long-term study involving K types of related recurrent events. The data structure and model are defined in Section 2.1. Under the Poisson process assumption, following equation (4.2), the observed likelihood function can be written in the following form:

$$L_{obs} = \prod_{k=1}^{K}\prod_{i=1}^{n}\prod_{j=1}^{J_i^{(k)}} \exp[-w_i^{\alpha_k}\{U_0^{(k)}(t_{i,j}^{(k)}) - U_0^{(k)}(t_{i,j-1}^{(k)})\}\exp(X_i'\boldsymbol{\beta}^{(k)})]$$

$$\times[w_i^{\alpha_k}\{U_0^{(k)}(t_{ij}^{(k)}) - U_0^{(k)}(t_{i,j-1}^{(k)})\}\exp(X_i'\boldsymbol{\beta}^{(k)})]^{Z_{ij}^{(k)}}/Z_{ij}^{(k)}!$$

Taking the baseline function $U_0^{(k)}(t)$ in the form of (4.4), the likelihood can further be written as

$$L(\boldsymbol{\theta}|\mathcal{D}) = \prod_{k=1}^{K}\prod_{i=1}^{n}\prod_{j=1}^{Ji^{(k)}} \exp[-w_i^{\alpha_k}\sum_{l=1}^{L} r_l^{(k)}\{I_l(t_{i,j}^{(k)}) - I_l(t_{i,j-1}^{(k)})\}\exp(X_i'\boldsymbol{\beta}^{(k)})]$$

$$\times[w_i^{\alpha_k}\sum_{l=1}^{L} r_l^{(k)}\{I_l(t_{i,j}^{(k)}) - I_l(t_{i,j-1}^{(k)})\}\exp(X_i'\boldsymbol{\beta}^{(k)})]^{Z_{ij}^{(k)}}/Z_{ij}^{(k)}!$$

where $\boldsymbol{\theta}$ represents the vector of all unknown parameters including $\boldsymbol{\beta}^{(k)} = (\beta_1^{(k)}, ..., \beta_p^{(k)})'$, $\boldsymbol{r^{(k)}} = (r_1^{(k)}, ..., r_L^{(k)})'$ for $k = 1, ..., K$, and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)'$. Note that for simplicity of notation, the d is omitted from the I-spline basis functions. With this likelihood format, the sampling for $r_l^{(k)}$ is especially difficult. To solve this problem, we further decompose $Z_{ij}^{(k)}$ as $\sum_{l=1}^{L} Z_{ijl}^{(k)}$, where the augmented Poisson latent variables $\{Z_{ijl}^{(k)}\}$ independently follow Poisson distributions as follows:

$$Z_{ijl}^{(k)}|X_i, w_i \sim \mathcal{P}\left(w_i^{\alpha_k}\{r_l^{(k)}[I_l(t_{i,j}^{(k)}) - I_l(t_{i,j-1}^{(k)})]\}\exp(X_i'\boldsymbol{\beta}^{(k)})\right).$$

The convolution property of Poisson distribution aids the transformation. Then with the augmented Poisson latent variables, the likelihood function can be unequivocally

expressed as

$$L_{aug}(\boldsymbol{\theta}|\mathcal{D}) = \prod_{k=1}^{K}\prod_{i=1}^{n}\prod_{j=1}^{J_i^{(k)}}\prod_{l=1}^{L}\exp[-w_i^{\alpha_k}r_l^{(k)}\{I_l(t_{i,j}^{(k)}) - I_l(t_{i,j-1}^{(k)})\}\exp(X_i'\boldsymbol{\beta}^{(k)})]$$

$$\times[w_i^{\alpha_k}r_l^{(k)}\{I_l(t_{i,j}^{(k)}) - I_l(t_{i,j-1}^{(k)})\}\exp(X_i'\boldsymbol{\beta}^{(k)})]^{Z_{ijl}^{(k)}}/Z_{ijl}^{(k)}!. \quad (4.5)$$

Based on this augmented likelihood 4.5, we develop the Bayesian computation algorithm in Section 4.3.3.

4.3.3 Prior specification and posterior computation

For Bayesian computation, we need to first specify prior distributions for unknown parameters. When we don't have much prior information about parameters, we usually assign vague priors for them. For $\boldsymbol{\beta}_m^{(k)}$, $m = 1, ..., p$, $k = 1, ..., K$, we assign $\mathcal{N}(0, \sigma^2)$ priors, where σ^2 takes a large value such as 100. For nonnegative $r_l^{(k)}$, $l = 1, ..., L$, $k = 1, ..., K$, we assign exponential priors with rate parameter λ_k, where λ_k itself follows a gamma prior $\mathcal{G}(a_\lambda, b_\lambda)$. This prior specification is appealing from the computational perspective because it leads to conjugate forms for each of the conditional posterior distributions of $\{r_l^{(k)}\}$ and $\{\lambda_k\}$. Theoretically, such a prior specification is closely related to Bayesian Lasso (Park and Casella, 2008). It is equivalent to the penalized likelihood approach with an L1 penalty on those spline coefficients, in which $\{\lambda_k\}$ serve as tuning parameters. Our simulation studies show that our approach is robust to the choice of hyperparameters, so we simply use $a_\lambda = 1$ and $b_\lambda = 1$. The frailty terms $\{w_i\}$ follow a gamma distribution $\mathcal{G}(\eta, \eta)$ with mean 1 and variance $1/\eta$. The parameter η, which controls the variance of frailties and hence the correlation of pairwise events, is very sensitive. We let η follow a vague gamma prior $\mathcal{G}(a_\eta, b_\eta)$, where a_η and b_η are assigned small values, such as $a_\eta = b_\eta = 0.01$, to let the data take the dominance in estimating η. At last, we assign vague uniform priors $Unif(-3, 3)$ to $\{\alpha_k\}$.

The Gibbs sampling algorithm is adopted for posterior computation. Basically, we derive the full conditional distribution of each parameter componentwise from the joint distribution of the likelihood function in (4.5) and the specified prior distributions. If the full conditional distribution of a parameter has a closed form, the sampling is straightforward. When the closed form is intractable, adaptive rejection sampling (ARS) (Gilks and Wild, 1992) is adopted if the full conditional posterior distribution preserves log-concavity. Even if log-concavity is not satisfied, we can still use Adaptive Rejection Metropolis Sampling (ARMS) (Gilks et al., 1995) to draw samples. The full conditional distributions of the Gibbs sampler are summarized as below.

1. Sample $(Z_{ij1}^{(k)}, ..., Z_{ijL}^{(k)})$ from a multinomial distribution $\mathcal{M}(Z_{ij}^{(k)}, \boldsymbol{P_{ij}^{(k)}})$, for $i = 1, ..., n$, $j = 1, ..., J_i^{(k)}$, $k = 1, ..., K$, where $\boldsymbol{P_{ij}^{(k)}} = (p_{ij1}^{(k)}, ..., p_{ijL}^{(k)})$ with $\sum_{l=1}^{L} p_{ijl}^{(k)} = 1$, and
$$p_{ijL}^{(k)} = \frac{r_l^{(k)}\{I_l(t_{ij}^{(k)}) - I_l(t_{i,(j-1)}^{(k)})\}}{\sum_{l=1}^{L} r_l^{(k)}\{I_l(t_{ij}^{(k)}) - I_l(t_{i,(j-1)}^{(k)})\}}.$$

2. Sample $r_l^{(k)}$ from a Gamma distribution $\mathcal{G}(A_l^{(k)}, B_l^{(k)})$, for $l = 1, ..., L$, $k = 1, ..., K$, with
$$A_l^{(k)} = \sum_{i=1}^{n} \sum_{j=1}^{J_i^{(k)}} Z_{ijl}^{(k)} + 1,$$
and
$$B_l^{(k)} = \sum_{i=1}^{n} w_i^{\alpha_k}\{I_l(t_{iJ_i^{(k)}}^{(k)}) - I_l(t_{i0}^{(k)})\} \exp(X_i'\boldsymbol{\beta}^{(k)}) + \lambda_k.$$

3. Sample λ_k from a Gamma distribution $\mathcal{G}(a_\lambda + L, b_\lambda + \sum_{l=1}^{L} r_l^{(k)})$, for $k = 1, ..., K$.

4. Sample $\beta_m^{(k)}$ by using adaptive rejection sampling (ARS) (Gilks and Wild, 1992) method, for $m = 1, ..., p$, $k = 1, ..., K$. The log full conditional distribution of

each β_m is proportional to

$$\exp[-\sum_{i=1}^{n}\sum_{l=1}^{L} w_i^{\alpha_k} r_l^{(k)} \{I_l(t_{iJ_i^{(k)}}^{(k)}) - I_l(t_{i0}^{(k)})\} \exp(X_i'\boldsymbol{\beta}^{(k)}) + \sum_{i=1}^{n}\sum_{j=1}^{J_i^{(k)}} X_i'\boldsymbol{\beta}^{(k)} Z_{ij}^{(k)} - (\beta_m^{(k)})^2/(2\sigma^2)].$$

5. Sample w_i for $i = 1, ..., n$, by using ARS. The log full conditional distribution of each w_i is proportional to

$$-\sum_{k=1}^{K} w_i^{\alpha_k} r_l^{(k)} \{I_l(t_{iJ_i^{(k)}}^{(k)}) - I_l(t_{i0}^{(k)})\} \exp(X_i'\boldsymbol{\beta}^{(k)}) + \sum_{k=1}^{K}\sum_{j=1}^{J_i^{(k)}}\sum_{l=1}^{L} Z_{ijl}^{(k)} \log(w_i^{\alpha_k}) - \eta w_i + (\eta - 1)\log(w_i).$$

6. Sample η by using ARMS, the log full conditional distribution of which is proportional to

$$(\eta - 1)\sum_{i=1}^{n} \log(w_i) - \eta(\sum_{i=i}^{n} w_i + b_\eta) + (n\eta + a_\eta - 1)\log(\eta) - n\log(\tau(\eta)).$$

7. Sample α_k, for $k = 1, ..., K$, by using ARMS, the log full conditional distribution of which is proportional to

$$-w_i^{\alpha_k} r_l^{(k)} \{I_l(t_{iJ_i^{(k)}}^{(k)}) - I_l(t_{i0}^{(k)})\} \exp(X_i'\boldsymbol{\beta}^{(k)}) + \sum_{j=1}^{J_i}\sum_{l=1}^{L} Z_{ijl}^{(k)} \log(w_i^{\alpha_k}).$$

4.4 Simulation studies

Simulation studies are conducted to evaluate the proposed methods. We only consider 2 (K) types of events for the purpose of demonstration. It is straightforward to extend to three or more types of events. By default, $\alpha_1 = 1$. For notational simplicity, we denote α_2 as α. We particularly assess the performance of estimating covariate coefficients $\{\boldsymbol{\beta}^{(k)}\}$ and baseline functions $\{U_0^{(k)}\}$ for different η and α values. We also compare the estimation results between models with and without the scale parameter α.

4.4.1 Data generation

Consider 2 types of events and 100 subjects. Data are simulated according to model (4.1) under the Poisson process assumption. Specifically, we consider four different values for α: -0.3, 0, 0.5, and 1, and two different values for η: 1 and 5. The baseline mean function for the first type of event is $U_0^{(1)}(t) = t + \log(1+t)$, which is approximately linear. For the second type of event, the baseline function is $U_0^{(2)}(t) = t^{0.5} + \log(1+t)$, which is curvilinear. Two covariates are involved for each subject, where X_1 is from a Bernoulli distribution with success probability $p = 0.5$ and X_2 is from a standard normal distribution. Each subject has the same observation times for the two types of events. The number of observation times for each subject is generated from a Poisson distribution with mean 7, and the time length between adjacent observation times follows an exponential distribution with mean 0.5. Given the covariates, the observation times and the generated frailty w_i, for each subject, the counts in each time interval for each type of event are generated from a Poisson distribution as in equation (4.2). Each set of simulations consists of 500 data replicates.

4.4.2 Simulation results

The first set of simulations aims at assessing the performance of estimating regression coefficients $\{\boldsymbol{\beta}^{(k)}\}$. For this set of simulations, data are generated with different true β values, but α is fixed at 1. Table 4.1 shows the estimation bias (Bias) defined as the average of the posterior means minus the true value, mean of the posterior standard deviations (SD), standard deviation of the posterior means (SE), and 95% coverage probability (CP95) of four combinations of β values. Clearly, the estimation of regression coefficients is good with small biases, SDs close to SEs, and 95% coverage probabilities around 95%. When η goes from 1 to 5 (i.e., the variance of the frailty goes from 1 to 0.2), the Bias, SD and SE uniformly become smaller. It is noted that

the estimation of $\{\beta_2\}$ is consistently better than that of $\{\beta_1\}$ in terms of giving smaller biases and SDs and SEs. This is because X_2 is a normal covariate, which carries more information than binary covariate X_1, and the variance of X_2 is four times larger than that of X_1. Our simulation study shows that when we add one more possible value to X_1 such that X_1 follows a uniform discrete distribution with $X_1 \in \{0, 1, -1\}$, the estimation of $\{\beta_1\}$ is improved and close to that of $\{\beta_2\}$.

Table 4.1: Estimation of regression coefficients from our proposed method when true $\alpha = 1$. Bias refers to the difference between the average of the 500 posterior means and the true value; SD refers to the mean of the 500 posterior standard deviations; SE refers to the standard deviation of the 500 posterior means, and CP95 refers to the 95% coverage probability

		$\eta = 1$				$\eta = 5$			
Parameter	Truth	Bias	SD	SE	CP95	Bias	SD	SE	CP95
$\beta_1^{(1)}$	0	0.0583	0.2238	0.1988	0.9520	0.0214	0.1287	0.1319	0.9400
$\beta_2^{(1)}$	1	0.0004	0.1215	0.1223	0.9520	0.0015	0.0696	0.0722	0.9300
$\beta_1^{(2)}$	1	0.0561	0.2225	0.1975	0.9700	0.0203	0.1291	0.1236	0.9600
$\beta_2^{(2)}$	1	-0.0002	0.1203	0.1220	0.9480	0.0004	0.0683	0.0689	0.9460
$\beta_1^{(1)}$	0	0.0392	0.2294	0.2179	0.9580	0.0178	0.1303	0.1311	0.9520
$\beta_2^{(1)}$	1	-0.0053	0.1245	0.1228	0.9520	-0.0008	0.0714	0.0709	0.9520
$\beta_1^{(2)}$	-1	0.0647	0.2523	0.2479	0.9560	0.0364	0.1651	0.1607	0.9500
$\beta_2^{(2)}$	1	-0.0109	0.1344	0.1297	0.9620	-0.0001	0.0858	0.0824	0.9620
$\beta_1^{(1)}$	1	0.0490	0.2198	0.2133	0.9500	0.0119	0.1203	0.1191	0.9540
$\beta_2^{(1)}$	1	-0.0078	0.1182	0.1202	0.9280	-0.0017	0.0646	0.0653	0.9380
$\beta_1^{(2)}$	-1	0.0771	0.2497	0.2373	0.9440	0.0400	0.1629	0.1593	0.9500
$\beta_2^{(2)}$	1	-0.0127	0.1323	0.1335	0.9440	-0.0026	0.0840	0.0825	0.9600
$\beta_1^{(1)}$	1	-0.0639	0.2152	0.2173	0.9320	-0.0178	0.1192	0.1194	0.9392
$\beta_2^{(1)}$	-1	0.0025	0.1138	0.1121	0.9500	0.0051	0.0633	0.0678	0.9392
$\beta_1^{(2)}$	1	-0.0663	0.2202	0.2247	0.9300	-0.0219	0.1280	0.1329	0.9196
$\beta_2^{(2)}$	1	0.0031	0.1177	0.1156	0.9440	0.0046	0.0681	0.0786	0.9412

The second set of simulations focuses on assessing the estimation performance of our proposed methods when data are generated with different α values. The true β values are fixed as $\beta_1^{(1)} = 1$, $\beta_2^{(1)} = -1$, $\beta_1^{(2)} = 1$, and $\beta_2^{(2)} = 1$. Table 4.2 summarizes the estimation results. The estimation results for regression coefficients are similar to those in Table 4.1. Overall, the estimation is good with small biases, SDs close

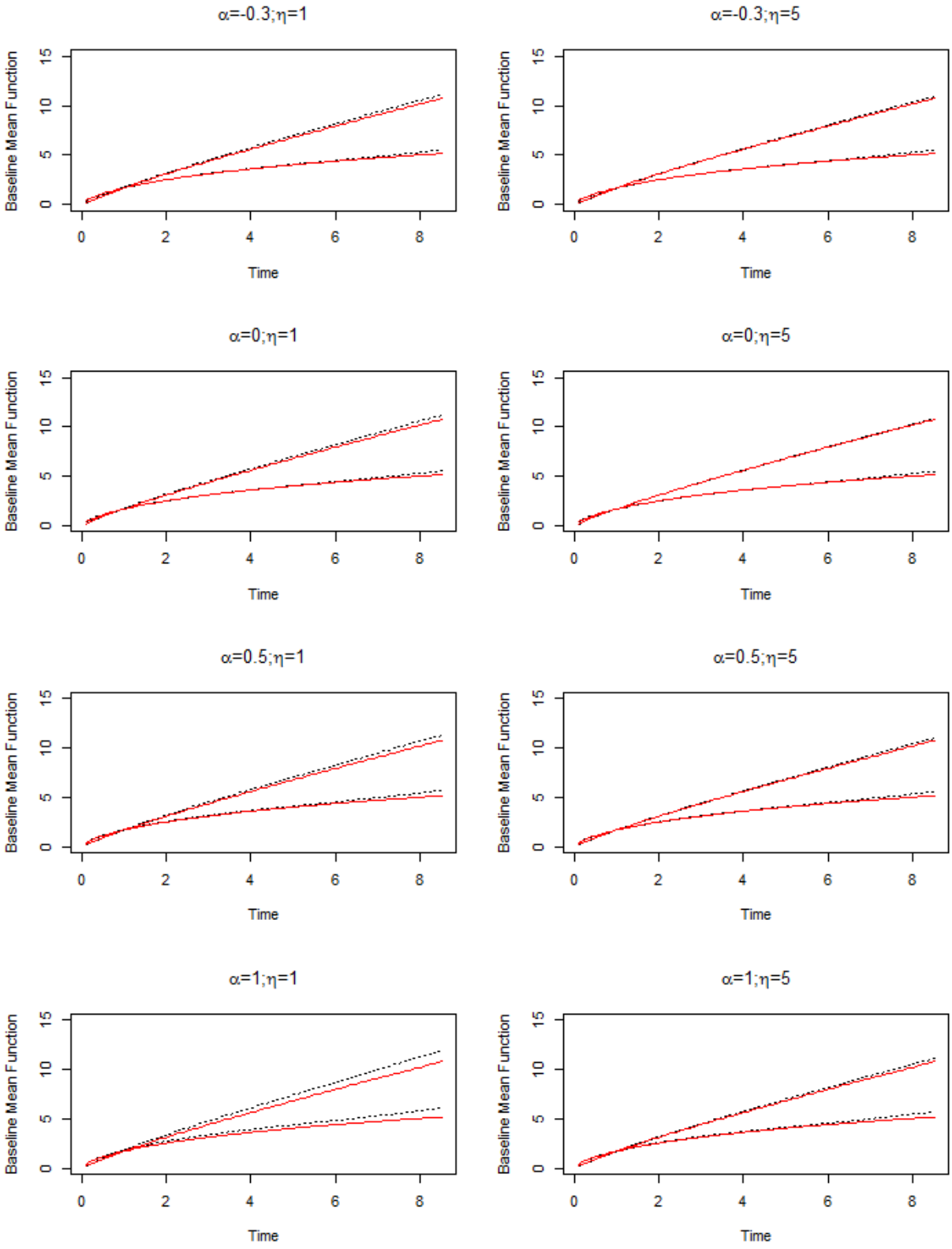

Figure 4.2: Baseline mean function for models having $\beta_1^{(1)} = 1$, $\beta_2^{(1)} = -1$, $\beta_1^{(2)} = 1$, $\beta_2^{(2)} = 1$, $\alpha = -0.3, 0, 0.5, 1$ and $\eta = 1, 5$

Table 4.2: Estimation of regression coefficients, α and η from our proposed method for data generated with different α values and true $\beta_1^{(1)} = 1$, $\beta_2^{(1)} = -1$, $\beta_1^{(2)} = 1$, and $\beta_2^{(2)} = 1$

		$\eta = 1$				$\eta = 5$			
Parameter		Bias	SD	SE	CP95	Bias	SD	SE	CP95
$\alpha = -0.3$	$\beta_1^{(1)}$	-0.0269	0.2190	0.2266	0.9300	-0.0124	0.1216	0.1252	0.9360
	$\beta_2^{(1)}$	-0.0054	0.1168	0.1202	0.9460	0.0006	0.0652	0.0685	0.9460
	$\beta_1^{(2)}$	-0.0086	0.1048	0.1049	0.9440	-0.0022	0.0829	0.0812	0.9540
	$\beta_2^{(2)}$	0.0021	0.0548	0.0580	0.9400	0.0006	0.0421	0.0421	0.9600
	α	-0.0083	0.0423	0.0462	0.9200	-0.0033	0.1010	0.1010	0.9400
	η	0.0412	0.1699	0.1754	0.9440	0.2494	1.0231	0.8814	0.9740
$\alpha = 0$	$\beta_1^{(1)}$	-0.0051	0.2229	0.2256	0.9400	-0.0017	0.1208	0.1175	0.9588
	$\beta_2^{(1)}$	-0.0043	0.1195	0.1195	0.9460	-0.0004	0.0654	0.0682	0.9294
	$\beta_1^{(2)}$	-0.0034	0.0758	0.0767	0.9460	-0.0048	0.0770	0.0749	0.9608
	$\beta_2^{(2)}$	-0.0010	0.0370	0.0378	0.9580	0.0007	0.0377	0.0365	0.9627
	α	0.0028	0.0344	0.0349	0.9420	0.0017	0.1152	0.1078	0.9784
	η	0.0328	0.1790	0.1781	0.9560	0.6264	1.3301	1.3349	0.9588
$\alpha = 0.5$	$\beta_1^{(1)}$	-0.0190	0.2194	0.2227	0.9360	-0.0076	0.1213	0.1179	0.9588
	$\beta_2^{(1)}$	-0.0136	0.1162	0.1190	0.9340	0.0001	0.0647	0.0652	0.9490
	$\beta_1^{(2)}$	-0.0263	0.1372	0.1338	0.9400	-0.0143	0.0958	0.0946	0.9549
	$\beta_2^{(2)}$	-0.0047	0.0719	0.0697	0.9600	0.0010	0.0497	0.0475	0.9588
	α	0.0066	0.0589	0.0578	0.9480	0.0163	0.1213	0.1246	0.9471
	η	0.0496	0.1729	0.1841	0.9500	0.4241	1.2655	1.2633	0.9569
$\alpha = 1$	$\beta_1^{(1)}$	-0.0639	0.2152	0.2173	0.9320	-0.0178	0.1192	0.1194	0.9392
	$\beta_2^{(1)}$	0.0025	0.1138	0.1121	0.9500	0.0051	0.0633	0.0678	0.9392
	$\beta_1^{(2)}$	-0.0663	0.2202	0.2247	0.9300	-0.0219	0.1280	0.1329	0.9196
	$\beta_2^{(2)}$	0.0031	0.1177	0.1156	0.9440	0.0046	0.0681	0.0786	0.9412
	α	0.0089	0.0804	0.0745	0.9760	0.0294	0.1529	0.1577	0.9392
	η	0.0469	0.1684	0.1734	0.9540	0.3745	1.2136	1.2777	0.9490

to SEs, and CP95s close to 95%. The estimation is more precise for the larger η value, and the estimation of $\{\beta_2\}$ is overall better than that of $\{\beta_1\}$. A new fact is observed when the magnitude of α approaches 0—the estimation of the regression coefficients for the second type of event becomes more precise with smaller biases and SDs and SEs. This observation is reasonable, as smaller values of α reflect the occurrence of the second type of events less affected by frailties, and thus more precise

estimation of the regression coefficients is expected. Furthermore, the estimated α and η have remarkably captured the truth. As opposed to the estimation of regression coefficients, a small value of η leads to better estimation of α. A glance at the table may find the bias of η is a little too high when the true η is 5. We need to point out that the difference between $\mathcal{G}(5,5)$ and $\mathcal{G}(6,6)$ is trivial, so the estimation can still accurately capture the shape of the distribution. Figure 4.2 shows the estimation of the baseline mean functions corresponding to different setups of α and η values. The solid lines represent the real baseline mean functions and the broken lines represent the estimated baseline mean functions. The two lines on the top are for the first type of event $U_0^{(1)}(t) = t + \log(1+t)$, and the curvilinear lines at the bottom are for the baseline mean function $U_0^{(2)}(t) = t^{0.5} + \log(1+t)$ of the second type of event. We can clearly see that all the estimated lines match the true lines well. The plots on the right side when $\eta = 5$ show an even better convergence to the truth than their counterparts on the left side when $\eta = 1$.

Finally, we evaluate the effect of the scale parameter α on estimation. Frailty models are well accepted for univariate panel count due to their flexibility and robustness. A naive extension to multivariate panel count data is to treat the frailty exactly the same for all types of events, i.e., each subject shares a common frailty for all the events. The following simulation shows that this practice leads to incorrect estimation results when the true α is not 1. Table 4.3 presents the estimation results for the same simulated data sets as in Table 4.2, but fitted with the naive model; i.e., the model (4.1) with α omitted. The last block of the table shows that the naive model does a comparable job in estimating the regression coefficients and η. This is because the naive model is the true model when the true value of α is 1. However, for data generated with α not equal to 1, the misspecified naive model provides poor estimation results, which is especially clear when the true η is 1. Compared with the results in Table 4.2, all bias's for the regression coefficients slightly increase. For

Table 4.3: Estimation of regression coefficients, α and η from the naive method with α omitted for data generated with different α values and true $\beta_1^{(1)} = 1$, $\beta_2^{(1)} = -1$, $\beta_1^{(2)} = 1$, and $\beta_2^{(2)} = 1$

		$\eta = 1$				$\eta = 5$			
Parameter		Bias	SD	SE	CP95	Bias	SD	SE	CP95
$\alpha = -0.3$	$\beta_1^{(1)}$	-0.0438	0.1344	0.2443	0.7220	-0.0130	0.1016	0.1288	0.8640
	$\beta_2^{(1)}$	0.0061	0.0710	0.1388	0.7080	0.0009	0.0536	0.0722	0.8500
	$\beta_1^{(2)}$	-0.0206	0.1353	0.1628	0.8880	-0.0103	0.1085	0.0938	0.9800
	$\beta_2^{(2)}$	0.0062	0.0715	0.0967	0.8540	0.0027	0.0574	0.0525	0.9720
	η	2.6937	0.6275	0.9058	0.0000	3.9252	0.8115	0.3937	0.0060
$\alpha = 0$	$\beta_1^{(1)}$	-0.0114	0.1459	0.2308	0.7820	-0.0046	0.1021	0.1205	0.9080
	$\beta_2^{(1)}$	-0.0028	0.0770	0.1294	0.7700	0.0010	0.0544	0.0690	0.8760
	$\beta_1^{(2)}$	-0.0302	0.1513	0.1166	0.9800	-0.0095	0.1106	0.0843	0.9900
	$\beta_2^{(2)}$	0.0080	0.0802	0.0709	0.9820	0.0031	0.0587	0.0474	0.9780
	η	1.9095	0.4903	0.6390	0.0000	3.6761	0.9184	0.5119	0.0320
$\alpha = 0.5$	$\beta_1^{(1)}$	-0.0282	0.1826	0.2277	0.8840	-0.0059	0.1073	0.1185	0.9160
	$\beta_2^{(1)}$	-0.0092	0.0962	0.1194	0.8780	-0.0002	0.0568	0.0646	0.9220
	$\beta_1^{(2)}$	-0.0536	0.1893	0.1383	0.9760	-0.0193	0.1162	0.0962	0.9780
	$\beta_2^{(2)}$	-0.0016	0.0997	0.0777	0.9880	0.0037	0.0617	0.0496	0.9920
	η	0.5948	0.2502	0.2712	0.2160	2.3688	1.1662	0.9781	0.4600
$\alpha = 1$	$\beta_1^{(1)}$	-0.0679	0.2161	0.2179	0.9320	-0.0181	0.1196	0.1194	0.9420
	$\beta_2^{(1)}$	0.0029	0.1136	0.1118	0.9500	0.0027	0.0631	0.0614	0.9500
	$\beta_1^{(2)}$	-0.0690	0.2202	0.2261	0.9380	-0.0208	0.1276	0.1315	0.9280
	$\beta_2^{(2)}$	0.0022	0.1170	0.1155	0.9520	0.0019	0.0673	0.0702	0.9440
	η	0.0401	0.1552	0.1614	0.9400	0.1853	0.9692	0.9736	0.9600

the regression coefficients for the first type of event, $\beta_1^{(1)}$ and $\beta_2^{(1)}$, the SEs are quite close to those in Table 4.2, but the SDs are smaller (the farther the true α is from 1, the smaller the SDs become), leading to CP95 much smaller than 95%. For the regression coefficients for the second type of event, $\beta_1^{(2)}$ and $\beta_2^{(2)}$, compared with those in Table 4.2, both SDs and SEs increase. For data generated with $\alpha_2 = 0$ or 0.5, the naive method inflates SDs more than SEs, leading to higher coverage probabilities than 95%. When $\alpha_2 = -.3$, the inflated SEs are actually larger than the inflated SDs, leading to lower coverage probabilities. The bias, SD and SE of η increase when the true α moves away from the value 1. The estimation results for $\eta = 5$ have a similar

pattern to those for $\eta = 1$, but are much improved. This is reasonable because a large η value implies a small variance for frailties and thus less correlation between two types of events. This set of simulations shows that including the scale parameter of α is important to correctly estimate the regression coefficients.

4.5 Real data analysis

In this section, we apply our proposed methods to analyze the motivating data set, the skin cancer data, and compare the results with those in the literature. Two hundred and ninety patients are analyzed with one deleted from the original group because there is no observation for that patient. The observation times were recorded in days. The covariate of major interest is denoted as X_1, which is equal to 1 if the patient was assigned to the DFMO group and 0 in the placebo group. The other three covariates of interest are X_2, the number of cancers prior to the trial; X_3, the patient's age; and X_4, gender with male 1 and female 0. Table 4.4 displays the results of the covariate coefficients of the two types of skin cancers from the proposed model. Clearly, gender (X_4) has no significant effect but the number of cancers prior to trial (X_2) and the patient's age (X_3) are significant for both skin cancers. The positive value of β_2 implies that the number of prior cancers has a positive relationship with the rate of new cancers. The negative value of β_3 suggests that older patients tend to have a lower rate of new cancers, which makes sense biologically since older people have a slow metabolism. However, DFMO has different effects on the recurrence rate of the two skin cancers. For basal cell carcinoma, the recurrence rate is decreased by a factor of 1.209; but for squamous cell carcinoma, the recurrence rate is increased by a factor of 1.111. However, neither of these two effects are significant due to the large posterior standard deviations.

Table 4.4: Estimation results (posterior mean, posterior standard deviation and 95% credible interval) of the covariate effects for basal cell carcinoma and squamous cell carcinoma

	β_1	β_2	β_3	β_4
basal	-0.1902	0.1028	-0.0376	0.0238
	0.1500	0.0145	0.0055	0.1506
	(-0.4894, 0.0987)	(0.0762, 0.1330)*	(-0.0487, -0.0273)*	(-0.2717, 0.3185)
squamous	0.1055	0.1451	-0.0196	0.3053
	0.2174	0.0213	0.0070	0.2142
	(-0.3153, 0.5251)	(0.1051, 0.0.1894)*	(-0.0333, -0.0060)*	(-0.1165, 0.7255)

Table 4.5: Estimation results of the covariate effects when common covariate effects are assumed for the two types of skin cancers from the proposed method, He et al. (2008)'s method, and Zhang et al. (2013)'s method

	β_1	β_2	β_3	β_4
Proposed	-0.1509	0.0810	-0.0264	0.0636
	0.1381	0.0122	0.0052	0.1409
	(-0.4203, 0.1197)	(0.0599, 0.1087)*	(-0.0367, -0.0169)*	(-0.2098, 0.3369)
He et al.	-0.0239	0.1440	-0.0116	0.3807
	0.1809	0.0212	0.0084	0.1778
	(-0.3785, 0.3307)	(0.1024, 0.1856)*	(-0.0281, 0.0049)	(0.0322, 0.7292)*
Zhang et al.	-0.2253	0.0784	0.0016	0.2534
	0.1831	0.0090	0.0087	0.1942
	(-0.5842, 0.1336)	(0.0608, 0.0960)*	(-0.0155, 0.0187)	(-0.1272, 0.06340)

The skin cancer data have also been analyzed by He et al. (2008) and Zhang et al. (2013), assuming the same covariate effects for both skin cancers. For comparison, we also modify step 4 in the Gibbs sampler in Section 4.3.3 and reanalyze the data with the same covariate coefficients for both cancers. The results are presented in Table 4.5.

Overall, the estimation results of the covariate coefficients for the three methods are quite similar. However, the Bayesian method we proposed produces smaller posterior standard deviations than those produced from the other two frequentist methods. The different effects of DFMO on the two skin cancers are partially hidden

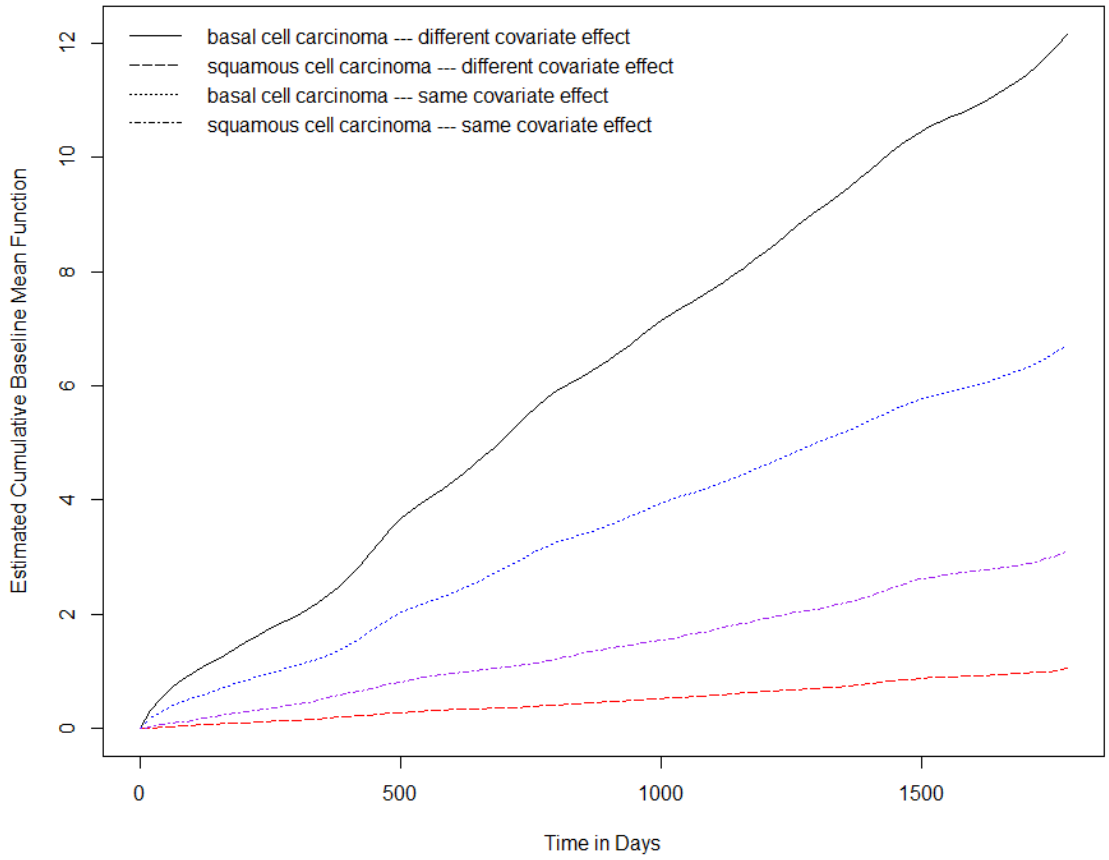

Figure 4.3: Estimated cumulative baseline mean function of models with and without the same covariate effects for basal cell carcinoma and squamous cell carcinoma

compared to the method when different covariate effects are assumed. The effect of the number of cancers prior to trial (X_2) is significant for all three methods. Our Bayesian method shows a significant effect on patient's age (X_3), while the other two frequentist methods do not. Our method and Zhang et al. (2013)'s method show no significant effect of gender (X_4), but He et al. (2008)'s method does.

In addition, our method can estimate the baseline mean functions simultaneously with the regression coefficients. Figure 4.3 presents the estimated baseline mean functions for basal cell carcinoma and squamous cell carcinoma, respectively. The solid line on the top and the broken line at the bottom represent the baseline mean func-

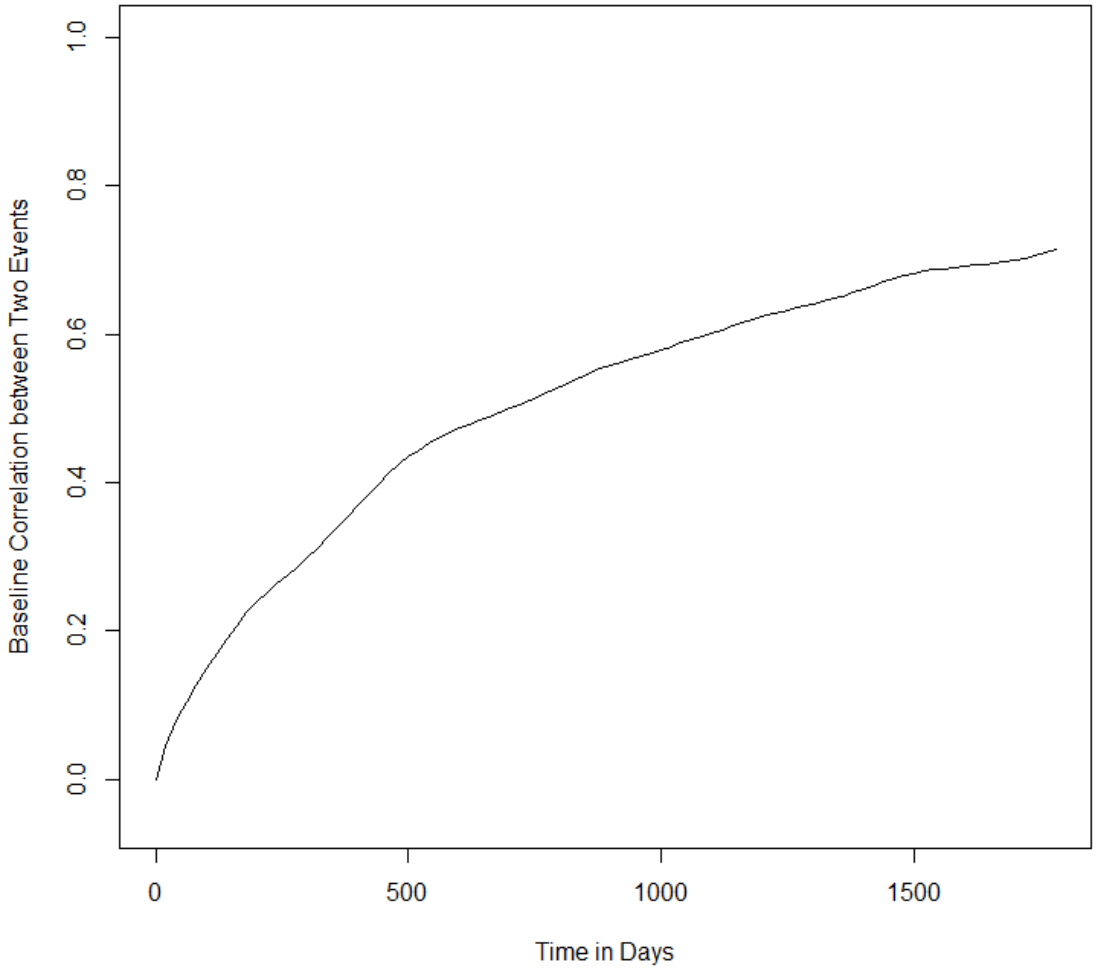

Figure 4.4: Baseline correlation between basal cell carcinoma and squamous cell carcinoma across time

tions of basal cell carcinoma and squamous cell carcinoma from the model assuming different covariate effects. The two broken lines in the middle represent the baseline mean functions of the two cancers when assuming the covariate effects are the same. Clearly, under the common covariate effects assumption, the baseline mean functions are both pulled towards the middle, which blurs the difference in the baseline mean functions between these two skin cancers.

Finally, Figure 4.4 displays the estimated baseline correlation function when co-

variate effects are assumed to be different for the two skin cancers. It can be seen that the correlation between the two recurrent skin cancers is positive and is strengthened across time.

4.6 Disussion

In this chapter, we have proposed a Bayesian estimation approach for the semiparametric regression analysis of multivariate panel count data. For each type of event, the proportional mean function is used to model the cumulative mean count of the event, where its baseline mean function is approximated by monotone I-splines. The correlation between two types of events is modeled by common frailty terms and a scale parameter. Based on a novel Poisson data augmentation, an efficient and easy-to-implement Gibbs sampler is developed for MCMC computation. Through the MCMC samples, the regression coefficients, the baseline mean functions, and the baseline correlation function between two events can be simultaneously estimated. Simulation studies have shown that the proposed approach provides accurate estimates of the regression coefficients and the baseline mean functions. Simulation studies have also demonstrated the importance of including the scale parameter in the model. The scale parameter provides substantive flexibility for modeling the correlation and meanwhile improves estimation performance.

Our approach can be slightly modified to accommodate common covariate effects for different types of events. However, real data analysis shows that doing so would hide significant covariate effects for individual types of events. Therefore, we suggest assuming different covariate effects for different events first for real data analysis, and then using the common covariate effects model later if similar effects are observed for different events.

Unlike the existing frequentist methods, our approach does not require model assumptions for the observation or censoring processes. Our approach is solely based

on the observed likelihood and only needs the observed counts and observation times for the analysis. This makes our proposed approach sufficiently generic to deal with panel count data arising from different observation schemes including dependent or independent censoring and/or observation processes. On the other hand, the proposed approach may lose a certain amount of efficiency due to not incorporating information on the observation or censoring processes when it is actually available.

Although each subject is assumed to have common covariates for different types of events and only time-independent covariates are considered in this chapter, our approach can accommodate different and time-dependent covariates for each subject by replacing X_i with $X_i^{(k)}(t)$ in model (4.1) and the Gibbs sampler.

Wang and Lin (2019) demonstrated the robustness of their approach to estimating the regression coefficients and the baseline function when the Poisson process assumption fails to hold for univariate panel count data. In this chapter, with the additional gamma frailty terms for the proportional mean models, we expect even greater robustness for our proposed approach. One way to further improve the performance of our approach is to nonparametrically model the frailties, such as using a Dirichlet process mixture of gammas to model the distribution of frailties.

Chapter 5

Conclusion

In this dissertation, two main topics, binary diagnostic tests and multivariate panel count data, are investigated in the Bayesian framework. Chapter 2 and Chapter 3 present ideas on two aspects of binary diagnostic tests, one for a longitudinal perspective with repeated measures and the other for a cross-sectional perspective with a medium or large group of tests (or raters). Chapter 4 extends the existing univariate panel count data literature to a multivariate cases. The commonality among the three chapters is that they all employ Bayesian sampling methods for implementation and all the algorithms are easy to implement, and work well in the situations involved.

For the binary diagnostic test models we developed in Chapter 2 and Chapter 3, the highest degree of correlation we considered was two, corresponding to pairwise correlation. Jones et al. (2010) suggested that in practice pairwise dependence is likely to be more than enough. The models we developed are easily generalizable to more than one population though here we only investigate one population. For further interest, we can study the problem at strata levels based on demographic information, such as age and gender, by introducing a prevalence for each subpopulation. In this dissertation we only consider diagnostic tests with binary results. For future investigation, ordinal data—data with several categories—also merit more exploration. Ordinal data convey more information than binary data, so potentially using ordinal data can further improve estimation results. Finally, when we deploy these developed latent class models, we need to be careful that the misspecified covariance structure may lead to biased estimation results (Albert and Dodd, 2004).

In Chapter 4, we demonstrate an application of the Bayesian method to the panel count data regime. We particularly explore the multivariate case. For each type of event, the proportional mean model is employed to model the cumulative mean count, where monotone I-splines are used to model the baseline mean function. Common frailties with scale parameters are utilized to accommodate the correlation between events. Unlike the frequentist estimating equation method, there is no need to specify the relationship between recurrent event processes and observation processes, making the model generally more applicable. It can also be applied to a situation with different observation times and different covariates for each event. For future study, we can incorporate different covariates for each event and/or time-dependent covariates in the model. In addition, we can model the frailties nonparametrically, for instance, using a Dirichlet process mixture of gammas to offer more flexibility. Another consideration would be to use the transformation of the proportional mean model when the proportional assumption does not hold.

Bibliography

Paul S Albert. A two-state markov mixture model for a time series of epileptic seizure counts. *Biometrics*, pages 1371–1381, 1991.

Paul S Albert and Lori E Dodd. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60 (2):427–435, 2004.

Paul S Albert, Lisa M McShane, Joanna H Shih, and US National Cancer Institute Bladder Tumor Marker Network. Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics*, 57(2):610–619, 2001.

W Gregory Alvord, JAMES E Drummond, LARRY O ARTHUR, ROBERT J BIGGAR, JAMES J GOEDERT, PAUL H LEVINE, EDWARD L MURPHY Jr, STANLEY H WEISS, and WILLIAM A BLATTNER. A method for predicting individual hiv infection status in the absence of clinical information. *AIDS research and human retroviruses*, 4(4):295–304, 1988.

Craig A Beam, Emily F Conant, and Edward A Sickles. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. *Journal of the National Cancer Institute*, 95(4):282–290, 2003.

AJ Branscum, WO Johnson, TE Hanson, and IA Gardner. Bayesian semiparametric roc curve estimation and disease diagnosis. *Statistics in Medicine*, 27(13):2474–2496, 2008.

Hermann Brenner. How independent are multiple ?independent?diagnostic classifications? *Statistics in Medicine*, 15(13):1377–1386, 1996.

Lyle D Broemeling. *Bayesian biostatistics and diagnostic medicine*. Chapman and Hall/CRC, 2007.

Bo Cai, Xiaoyan Lin, and Lianming Wang. Bayesian proportional hazards model for current status data with monotone splines. *Computational Statistics & Data Analysis*, 55(9):2644–2651, 2011.

George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

Bingshu E Chen, Richard J Cook, Jerald F Lawless, and Min Zhan. Statistical methods for multivariate interval-censored recurrent events. *Statistics in medicine*, 24 (5):671–691, 2005.

Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

Siddhartha Chib and Edward Greenberg. Markov chain monte carlo simulation methods in econometrics. *Econometric theory*, 12(3):409–431, 1996.

John Collins and Paul S Albert. Estimating diagnostic accuracy without a gold standard: a continued controversy. *Journal of biopharmaceutical statistics*, 26(6): 1078–1082, 2016.

RJ Cook, ETM Ng, and MO Meade. Estimation of operating characteristics for dependent diagnostic tests based on latent markov models. *Biometrics*, 56(4): 1109–1117, 2000.

N Dendukuri and L Joseph. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*, 57(1):158–167, 2001.

Nandini Dendukuri, Elham Rahme, Patrick Bélisle, and Lawrence Joseph. Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test. *Biometrics*, 60(2):388–397, 2004.

AD Duckworth, GA Buijze, M Moran, A Gray, CM Court-Brown, D Ring, and MM McQueen. Predictors of fracture following suspected injury to the scaphoid. *The Journal of Bone and Joint Surgery*, 94(7):961–968, 2012.

B Engel, J Backer, and W Buist. Evaluation of the accuracy of diagnostic tests from repeated measurements without a gold standard. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(1):83–100, 2010.

MA Espeland, OS Platt, and D Gallagher. Joint estimation of incidence and diagnostic error rates from irregular longitudinal data. *Journal of the American Statistical Association*, 84(408):972–979, 1989.

Mark A Espeland and Stanley L Handelman. Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics*, pages 587–599, 1989.

Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

Andrew Gelman. Iterative and non-iterative simulation algorithms. *Computing science and statistics*, pages 433–433, 1993.

Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. In *Readings in computer vision*, pages 564–584. Elsevier, 1987.

Wally R Gilks, NG Best, and KKC Tan. Adaptive rejection metropolis sampling within Gibbs sampling. *Applied Statistics*, pages 455–472, 1995.

Walter R Gilks and Pascal Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, pages 337–348, 1992.

P Gustafson. What are the limits of posterior distributions arising from nonidentified models, and why should we care? *Journal of the American Statistical Association*, 104(488):1682–1695, 2009.

P Gustafson, AE Gelfand, SK Sahu, WO Johnson, TE Hanson, L Joseph, and J Lee. On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Statistical Science*, 20 (2):111–140, 2005.

H Haario, E Saksman, and J Tamminen. Componentwise adaptation for high dimensional mcmc. *Computational Statistics*, 20(2):265–273, 2005.

Heikki Haario, Eero Saksman, Johanna Tamminen, et al. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

T Hanson, WO Johnson, and IA Gardner. Hierarchical models for estimating herd prevalence and test accuracy in the absence of a gold standard. *Journal of Agricultural, Biological, and Environmental Statistics*, 8(2):223, 2003.

TE Hanson, WO Johnson, and IA Gardner. Log-linear and logistic modeling of dependence among diagnostic tests. *Preventive Veterinary Medicine*, 45(1-2): 123–137, 2000.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

Xin He, Xingwei Tong, Jianguo Sun, and Richard J Cook. Regression analysis of multivariate panel count data. *Biostatistics*, 9(2):234–248, 2008.

X Joan Hu, Stephen W Lagakos, and Richard A Lockhart. Generalized least squares estimation of the mean function of a counting process based on panel counts. *Statistica Sinica*, 19:561, 2009.

SL Hui and SD Walter. Estimating the error rates of diagnostic tests. *Biometrics*, 36(1):167–171, 1980.

SL Hui and XH Zhou. Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research*, 7(4):354–370, 1998.

WO Johnson, JL Gastwirth, and LM Pearson. Screening without a "gold standard": the hui-walter paradigm revisited. *American Journal of Epidemiology*, 153(9): 921–924, 2001.

G Jones, WO Johnson, TE Hanson, and R Christensen. Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics*, 66(3): 855–863, 2010.

G Jones, WO Johnson, WD Vink, and N French. A framework for the joint modeling of longitudinal diagnostic outcome data and latent infection status: application to investigating the temporal relationship between infection and disease. *Biometrics*, 68(2):371–379, 2012.

L Joseph, TW Gyorkos, and L Coupal. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, 141(3):263–272, 1995.

John Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.

Jerald Franklin Lawless. Regression methods for poisson process data. *Journal of the American Statistical Association*, 82(399):808–815, 1987.

JF Lawless and M Zhan. Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *Canadian Journal of Statistics*, 26(4):549–565, 1998.

Ni Li, Do-Hwan Park, Jianguo Sun, and KyungMann Kim. Semiparametric transformation models for multivariate panel count data with dependent observation process. *Canadian Journal of Statistics*, 39(3):458–474, 2011.

DY Lin, LJ Wei, and Z Ying. Semiparametric transformation models for point processes. *Journal of the American Statistical Association*, 96(454):620–628, 2001.

Xiaoyan Lin, Bo Cai, Lianming Wang, and Zhigang Zhang. A Bayesian proportional hazards model for general interval-censored data. *Lifetime Data Analysis*, 21(3): 470–490, 2015.

Xiaoyan Lin, Hua Chen, Don Edwards, and Kerrie P Nelson. Modeling rater diagnostic skills in binary classification processes. *Statistics in medicine*, 37(4):557–571, 2018.

Minggen Lu, Ying Zhang, and Jian Huang. Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika*, 94(3):705–718, 2007.

Minggen Lu, Ying Zhang, and Jian Huang. Semiparametric estimation methods for panel count data using monotone b-splines. *Journal of the American Statistical Association*, 104(487):1060–1070, 2009.

S Luo, M Yi, X Huang, and KK Hunt. A bayesian model for misclassified binary outcomes and correlated survival data with applications to breast cancer. *Statistics in Medicine*, 32(13):2320–2334, 2013.

S Luo, X Su, SM DeSantis, X Huang, M Yi, and KK Hunt. Joint model for a diagnostic test without a gold standard in the presence of a dependent terminal event. *Statistics in Medicine*, 33(15):2554–2566, 2014.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Peter Müller. *A generic approach to posterior integration and Gibbs sampling*. Purdue University, Department of Statistics, 1991.

Kerrie P Nelson and Don Edwards. On population-based measures of agreement for binary classifications. *Canadian Journal of Statistics*, 36(3):411–426, 2008.

M Norris, WO Johnson, and I Gardner. Modeling bivariate longitudinal diagnostic outcome data in the absence of a gold standard. *Statistics and its Interface*, 2(2): 171–185, 2009.

Trevor Park and George Casella. The bayeisan lasso. *Journal of the American Statistical Association*, 103:681–686, 2008.

Margaret Sullivan Pepe. *The statistical evaluation of medical tests for classification and prediction*. Medicine, 2003.

M Plummer. Jags: A program for analysis of bayesian graphical models using gibbs sampling proceedings of the 3rd international workshop on distributed statistical computing, march 20–22, vienna, austria. *Vienna, Austria: Technische Universit at Wien*, 2003.

M Plummer. Jags version 4.3. 0 user manual [computer software manual]. *Retrieved from sourceforge. net/projects/mcmc-jags/files/Manuals/4. x*, 2, 2017.

Y Qu and A Hadgu. A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. *Journal of the American Statistical Association*, 93(443):920–928, 1998.

Y Qu, M Tan, and MH Kutner. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, 52(3):797–810, 1996.

James O Ramsay et al. Monotone regression splines in action. *Statistical science*, 3 (4):425–441, 1988.

Gareth O Roberts, Jeffrey S Rosenthal, et al. Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

Thomas J Rothenberg et al. Identification in parametric models. *Econometrica*, 39 (3):577–591, 1971.

JH Shih and PS Albert. Latent model for correlated binary data with diagnostic error. *Biometrics*, 55(4):1232–1235, 1999.

Debajyoti Sinha and Tapabrata Maiti. A Bayesian approach for the analysis of panel-count data with dependent termination. *Biometrics*, 60(1):34–40, 2004.

DJ Speigelhalter. Bayesian measures of model complexity and fit (with discussion). *J. Royal Statis. Soc. Ser. B*, 64:583–616, 2003.

Yu-Sung Su and Masanao Yajima. R2jags: Using r to run ?jags? *R package version 0.5-7*, 34, 2015.

J Sun and JD Kalbfleisch. Estimation of the mean function of point processes based on panel count data. *Statistica Sinica*, pages 279–289, 1995.

Jianguo Sun and Xingqiu Zhao. *Statistical analysis of panel count data*. Springer, 2013.

Jianguo Sun, Xingwei Tong, and Xin He. Regression analysis of panel count data with dependent observation times. *Biometrics*, 63(4):1053–1059, 2007.

Peter F Thall. Mixed poisson likelihood regression models for longitudinal interval count data. *Biometrics*, pages 197–209, 1988.

Vicki L Torrance-Rynard and Stephen D Walter. Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in medicine*, 16(19):2157–2175, 1997.

PM Vacek. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 41(4):959–968, 1985.

Maarten van Smeden, Christiana A Naaktgeboren, Johannes B Reitsma, Karel GM Moons, and Joris AH de Groot. Latent class models in diagnostic studies when there is no reference standard?a systematic review. *American journal of epidemiology*, 179(4):423–431, 2013.

SD Walter and LM Irwig. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology*, 41(9):923–937, 1988.

Chunling Wang and Timothy E Hanson. Estimation of sensitivity and specificity of multiple repeated binary tests without a gold standard. *Statistics in medicine*, 38(13):2381–2390, 2019.

Jianhong Wang and Xiaoyan Lin. A Bayesian approach for semiparametric regression analysis of panel count data. *Lifetime Data Analysis*, 2019. doi: 10.1007/s10985-019-09471-3.

Lianming Wang and David B Dunson. Semiparametric Bayes' proportional odds models for current status data with underreporting. *Biometrics*, 67(3):1111–1118, 2011.

Jon A Wellner, Ying Zhang, et al. Two estimators of the mean of a counting process with panel count data. *The Annals of statistics*, 28(3):779–814, 2000.

Jon A Wellner, Ying Zhang, et al. Two likelihood-based semiparametric estimation methods for panel count data with covariates. *The Annals of Statistics*, 35(5): 2106–2142, 2007.

T Yanagawa and BC Gladen. Estimating disease rates from a diagnostic test. *American Journal of Epidemiology*, 119(6):1015–1023, 1984.

Ilsoon Yang and Mark P Becker. Latent variable modeling of diagnostic accuracy. *Biometrics*, pages 948–958, 1997.

Bo Zhang, Zhen Chen, and Paul S Albert. Estimating diagnostic accuracy of raters without a gold standard by exploiting a group of experts. *Biometrics*, 68(4): 1294–1302, 2012.

Haixiang Zhang, Hui Zhao, Jianguo Sun, Dehui Wang, and KyungMann Kim. Regression analysis of multivariate panel count data with an informative observation process. *Journal of Multivariate Analysis*, 119:71–80, 2013.

Hui Zhao, Yang Li, and Jianguo Sun. Semiparametric analysis of multivariate panel count data with dependent observation processes and a terminal event. *Journal of Nonparametric Statistics*, 25(2):379–394, 2013.

Appendix A

Chapter 2 Appendix and Supplementary Materials

A.1 Focused Metropolis-Hastings proposals

Conditional on the diseased (C_{ij}^{+}), the pairwise correlation between any two tests i, j satisfy the following conditions:

$$0 < Se_i Se_j + C_{ij}^{+} < 1$$

$$0 < (1 - Se_i)(1 - Se_j) + C_{ij}^{+} < 1$$

$$0 < (1 - Se_i) Se_j - C_{ij}^{+} < 1$$

$$0 < (1 - Se_j) Se_i - C_{ij}^{+} < 1,$$

which boils down to

$$(1 - Se_i)(Se_j - 1) < C_{ij}^{+} < \min(Se_i, Se_j) - Se_i Se_j.$$

Similarly, we have

$$(1 - Sp_i)(Sp_j - 1) < C_{ij}^{-} < \min(Sp_i, Sp_j) - Sp_i Sp_j.$$

In the same spirit, conditional on the diseased (C_{ij}^{+}), the pairwise correlation between any two time points for test k should satisfy the following conditions:

$$0 < Se_k^2 + R_k^{+} < 1$$

$$0 < (1 - Se_k)^2 + R_k^{+} < 1$$

$$0 < (1 - Se_k)Se_k - R_k^+ < 1,$$

which boils down to

$$max(-Se_k^2, -(1-Se_k)^2, (1-Se_k)Se_k - 1) < R_k^+ < min(1-Se_k^2, 1-(1-Se_k)^2, (1-Se_k)Se_k)$$

$$max(-Sp_k^2, -(1-Sp_k)^2, (1-Sp_k)Sp_k - 1) < R_k^- < min(1-Sp_k^2, 1-(1-Sp_k)^2, (1-Sp_k)Sp_k)$$

Assuming sensitivities and specificities are greater than 0.5 implies the simplification

$$-(1 - Se_k)^2 < R_k^+ < (1 - Se_k)Se_k, \quad -(1 - Sp_k)^2 < R_k^- < (1 - Sp_k)Sp_k.$$

These inequalities focus proposals in the Metropolis-Hastings algorithm, allowing the quick removal of improper proposals without having to check that every joint probability is between zero and one, thus improving computational efficiency.

A.2 MCMC traceplots for schaphoid fracture data

A.2.1 MCMC traceplot for π

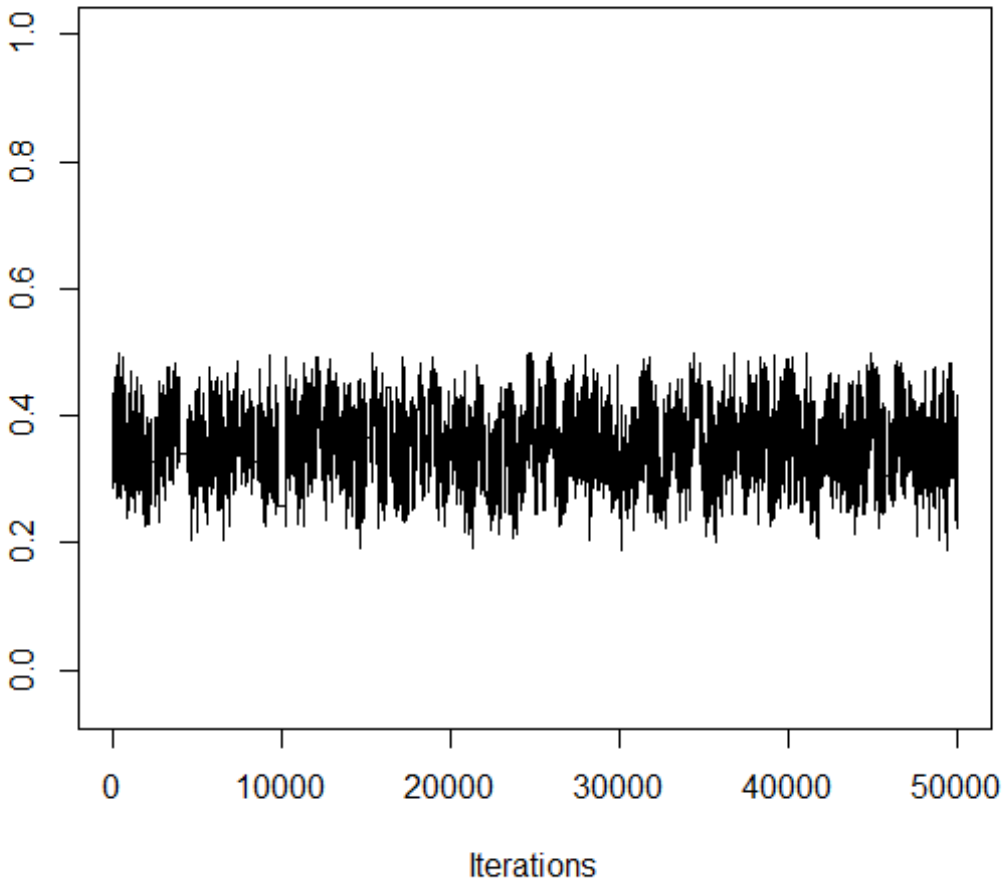

Figure A.1: Traceplot of π for the scaphoid fracture data. 50,000 iterations with first 2,000 iterates as burnin.

A.2.2 MCMC traceplots for Se_1, Se_2, Se_3, Sp_1, Sp_2 and Sp_3

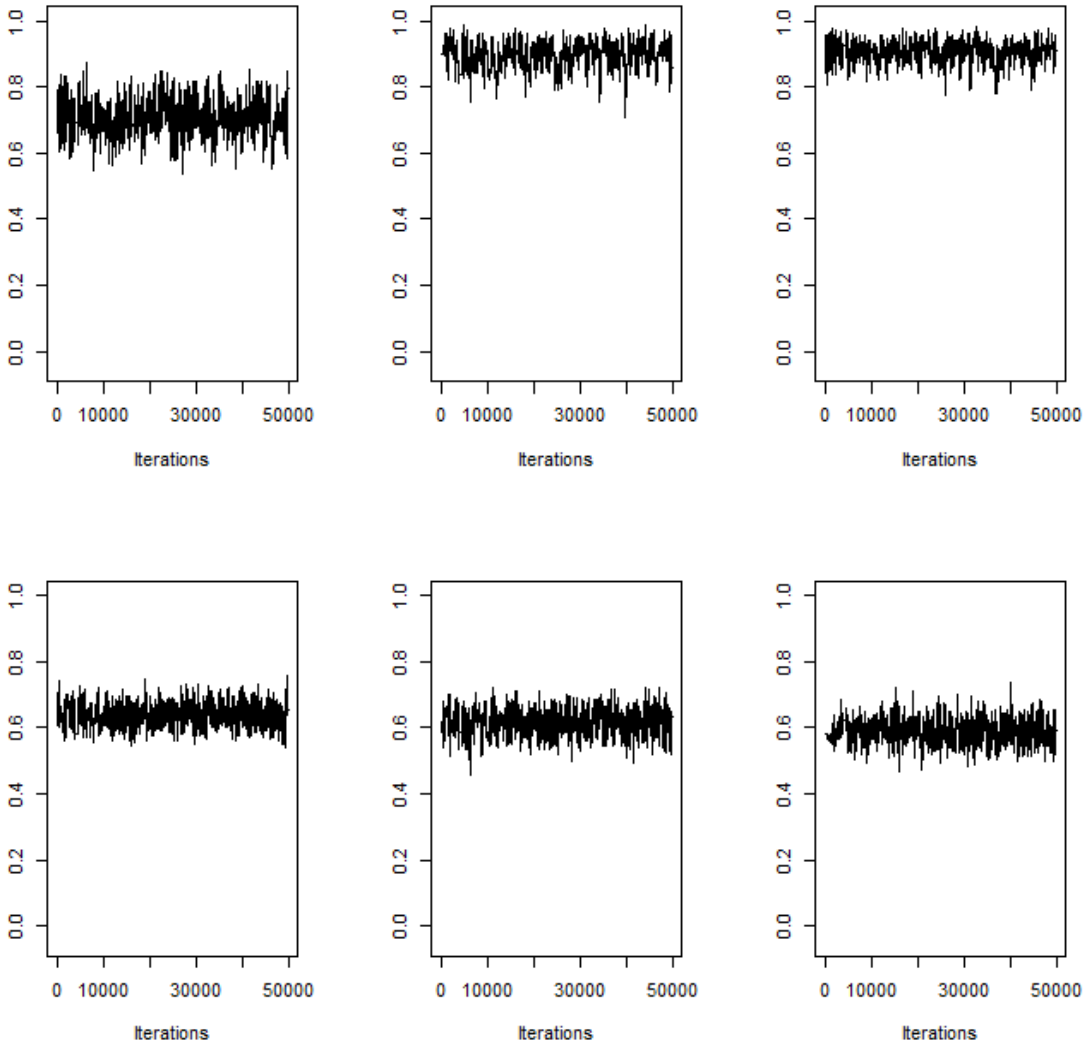

Figure A.2: Traceplots for Se_1, Se_2, Se_3 (top) and Sp_1, Sp_2, Sp_3 (bottom).

A.2.3 MCMC traceplots for R_1^+, R_2^+, R_3^+, R_1^-, R_2^- and R_3^-

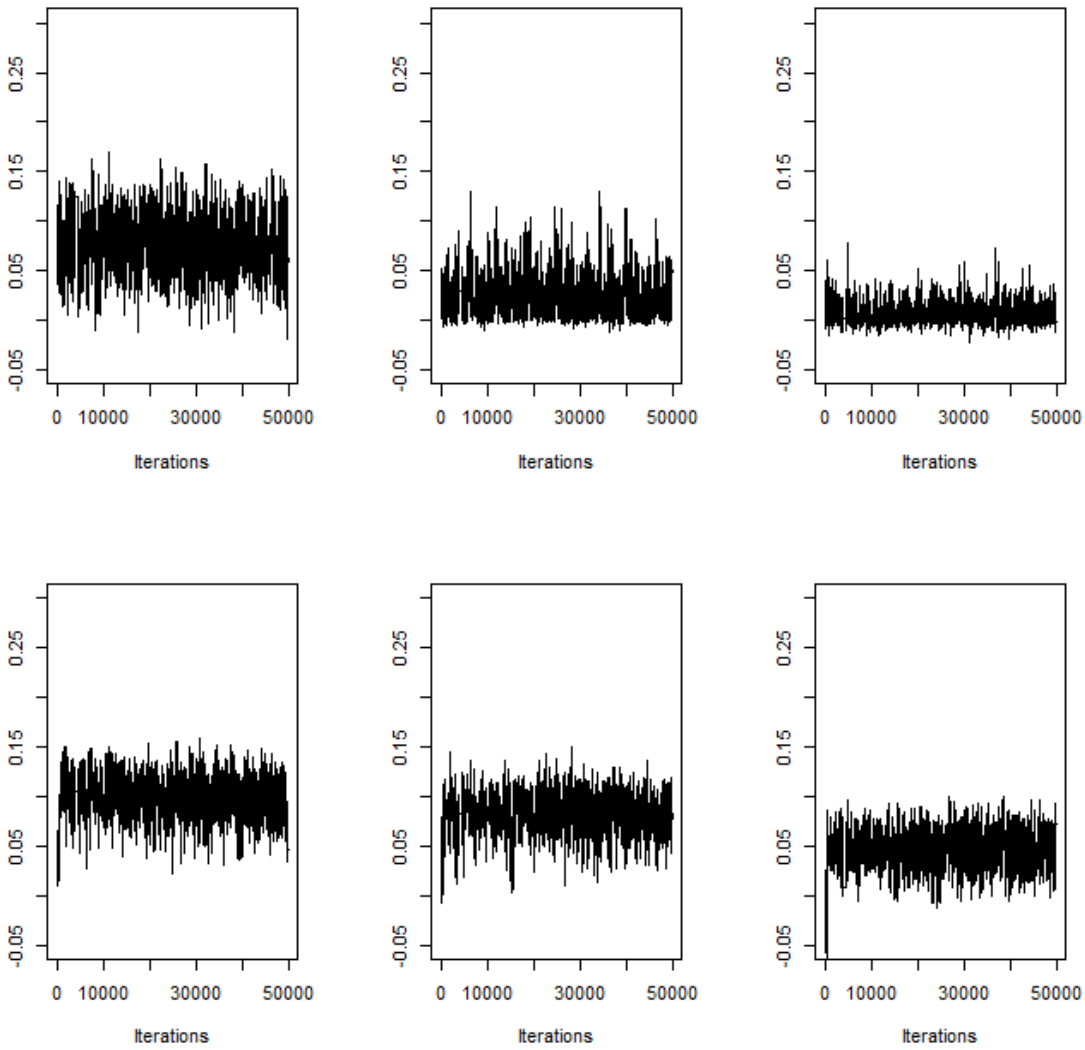

Figure A.3: Traceplots for R_1^+, R_2^+, R_3^+ (top) and R_1^-, R_2^-, R_3^- (bottom).

A.2.4 MCMC traceplots for C_{12}^{+}, C_{13}^{+}, C_{23}^{+}, C_{12}^{-}, C_{13}^{-} and C_{23}^{-}

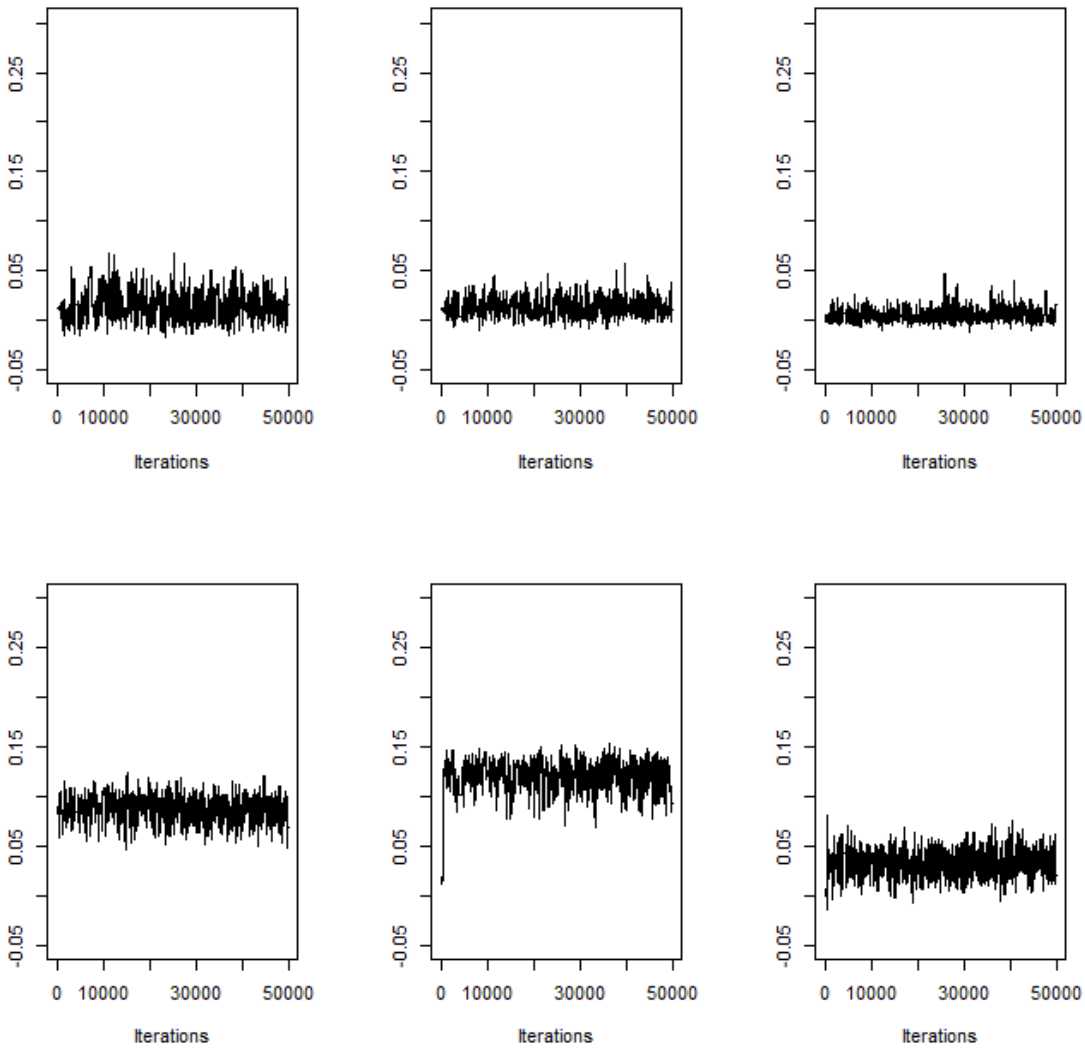

Figure A.4: Traceplots for C_{12}^{+}, C_{13}^{+}, C_{23}^{+} (top) and C_{12}^{-}, C_{13}^{-}, C_{23}^{-} (bottom).

Appendix B

Chapter 3 Appendix and Supplementary Materials

B.1 R and JAGS code

```
#########################################################################
#Algorithm 1: JAGS code for the Conditional Independent Model (M1)
#########################################################################
modelString1 = "
model{
for(i in 1:n){
   for (k in 1:K){
 s1[i,k]<- se[k]^x[i,k]*((1-se[k])^(1-x[i,k]))
 s2[i,k]<- sp[k]^(1-x[i,k])*((1-sp[k])^x[i,k])
 }
   prob[i]=pi*(prod(s1[i,1:K])) + (1-pi)*(prod(s2[i, 1:K]))
   z[i] ~ dpois( - log(prob[i]))
         }
for (k in 1:K){
   se[k] ~ dbeta( omega1*(kappa1 -2)+1, (1-omega1)*(kappa1-2) +1)
   sp[k] ~ dbeta( omega2*(kappa2 -2)+1, (1-omega2)*(kappa2-2) +1)
}
omega1 ~ dbeta(1,1)T(0.5,)
```

```
omega2 ~ dbeta(1,1)T(0.5,)
kappa1 = kappaMinusTwo1 +2
kappaMinusTwo1~ dgamma(0.01,0.01)
kappa2 = kappaMinusTwo2+2
kappaMinusTwo2 ~ dgamma(0.01,0.01)
pi ~ dbeta(1,1)
}"
writeLines( modelString1 , con="TEMPmodel1.bug" )

#############################################################################
#Algorithm 2: JAGS code for the Pairwise Covariance Model (M2)
#############################################################################
modelString2 = "
model{
for(i in 1:n){
   for (k in 1:K){
     s1[i,k]<- se[k]^x[i,k]*((1-se[k])^(1-x[i,k]))
     s2[i,k]<- sp[k]^(1-x[i,k])*((1-sp[k])^x[i,k])}
     for (j in 1:K){
     for (h in 1:K){
   cop[i,j,h]<- c1[j,h]*(-1)^(x[i,j] + x[i,h])/(s1[i,j]*s1[i,h])
   con[i,j,h]<- c2[j,h]*(-1)^(x[i,j] + x[i,h])/(s2[i,j]*s2[i,h])
                  }}
  eta[i] = (prod(s1[i,1:K]) *(1+ sum(cop[i,,])))
  theta[i] =(prod(s2[i, 1:K]) *(1+sum(con[i,,])))
  prob[i]=pi*eta[i] + (1-pi)*theta[i]
  z[i] ~ dpois( - log(prob[i]))
```

```
            }
for (k in 1:K){
   se[k] ~ dbeta( omega1*(kappa1 -2)+1, (1-omega1)*(kappa1-2) +1)
   sp[k] ~ dbeta( omega2*(kappa2 -2)+1, (1-omega2)*(kappa2-2) +1)
}
for (l in 1:(K-1)){
  for (h in (l+1):K){
   c1[l,h] ~ dunif((se[l]-1)*(1-se[h]), (min(se[l],se[h])-se[l]*se[h]))
   c2[l,h] ~ dunif((sp[l]-1)*(1-sp[h]), (min(sp[l],sp[h])-sp[l]*sp[h]))
 }}
for (h in 1:K){
  for (l in h:K){
c1[l,h] <-0
  c2[l,h] <-0
 }}
omega1 ~ dbeta(1,1)T(0.5, )
omega2 ~ dbeta(1,1)T(0.5,)
kappa1 = kappaMinusTwo1 +2
kappaMinusTwo1~ dgamma(0.01,0.01)
kappa2 = kappaMinusTwo2+2
kappaMinusTwo2 ~ dgamma(0.01,0.01)
pi ~ dbeta(1,1)
}"
writeLines( modelString2 , con="TEMPmodel2.bug" )
```

```
#########################################################################
# prepare data and use function jags() to run Algorithm 1 or Algorithm 2
#########################################################################
#n is the sample size,
#K is the number of raters (tests),
#m is the original dataset, an n by K matrix.
x<- structure(m, .Dim=c(n,K))
z<-rep(0,n)
dat=list(K=K,n=n,x=x, z=z) #prepare the data list
library(R2jags)
#run Algorithm 1 through function jags()
par1=c("se","sp","kappa2","kappa1","omega2","omega1", "pi","prob")
out1<- jags(model ="TEMPmodel1.bug", parameters.to.save=par1, data=dat,
           n.chains=1, n.iter=10000, n.burnin=1000)
#run Algorithm 2 through function jags()
par2=c("se","sp","kappa2","kappa1","omega2","omega1", "pi", "c1","c2","prob")
out2<- jags(model ="TEMPmodel2.bug", parameters.to.save=par2, data=dat,
           n.chains=1, n.iter=10000, n.burnin=1000)

#########################################################################
#Algorithm 3: JAGS code for the Conditional Independent Model (M1)
#             with the multinomial distribution imposition
#########################################################################
modelString3 = "
model{
for(i in 1:N){
   for (k in 1:K){
```

```
      s1[i,k]<- se[k]^x[i,k]*((1-se[k])^(1-x[i,k]))
s2[i,k]<- sp[k]^(1-x[i,k])*((1-sp[k])^x[i,k])
        }
   prob[i]=pi*(prod(s1[i,1:K])) + (1-pi)*(prod(s2[i, 1:K]))
   z[i] ~ dpois( - log(prob[i]))
            }
t[1:N] ~ dmulti(prob[1:N], n)
for (k in 1:K){
    se[k] ~ dbeta( omega1*(kappa1 -2)+1, (1-omega1)*(kappa1-2) +1)
    sp[k] ~ dbeta( omega2*(kappa2 -2)+1, (1-omega2)*(kappa2-2) +1)
        }
omega1 ~ dbeta(1,1)T(0.5,)
omega2 ~ dbeta(1,1)T(0.5,)
kappa1 = kappaMinusTwo1 +2
kappaMinusTwo1~ dgamma(0.01,0.01)
kappa2 = kappaMinusTwo2+2
kappaMinusTwo2 ~ dgamma(0.01,0.01)
pi ~ dbeta(1,1)
}"
writeLines( modelString3 , con="TEMPmodel3.bug" )

#########################################################################
#Algorithm 4: JAGS code for the Pairwise Covariance Model (M2)
#             with the multinomial distribution imposition
#########################################################################
modelStrinmodelString4 = "
model{
```

```
for(i in 1:N){
    for (k in 1:K){
     s1[i,k]<- se[k]^x[i,k]*((1-se[k])^(1-x[i,k]))
     s2[i,k]<- sp[k]^(1-x[i,k])*((1-sp[k])^x[i,k])
     }
     for (j in 1:K){
  for (h in 1:K){
   cop[i,j,h]<- c1[j,h]*(-1)^(x[i,j] + x[i,h])/(s1[i,j]*s1[i,h])
   con[i,j,h]<- c2[j,h]*(-1)^(x[i,j] + x[i,h])/(s2[i,j]*s2[i,h])
      }
      }
    eta[i] = (prod(s1[i,1:K]) *(1+ sum(cop[i,,])))
    theta[i] =(prod(s2[i, 1:K]) *(1+sum(con[i,,])))
    prob[i]=pi*eta[i] + (1-pi)*theta[i]
    z[i] ~ dpois( - log(prob[i]))
              }
t[1:N] ~ dmulti(prob[1:N], n)
for (k in 1:K){
    se[k] ~ dbeta( omega1*(kappa1 -2)+1, (1-omega1)*(kappa1-2) +1)
    sp[k] ~ dbeta( omega2*(kappa2 -2)+1, (1-omega2)*(kappa2-2) +1)
}
for (l in 1:(K-1)){
    for (h in (l+1):K){
c1[l,h] ~ dunif((se[l]-1)*(1-se[h]), (min(se[l],se[h])-se[l]*se[h]))
c2[l,h] ~ dunif((sp[l]-1)*(1-sp[h]), (min(sp[l],sp[h])-sp[l]*sp[h]))
 }}
for (h in 1:K){
```

```
  for (l in h:K){
     c1[l,h] <-0
     c2[l,h] <-0
 }}
omega1 ~ dbeta(1,1)T(0.5, )
omega2 ~ dbeta(1,1)T(0.5,)
kappa1 = kappaMinusTwo1 +2
kappaMinusTwo1~ dgamma(0.01,0.01)
kappa2 = kappaMinusTwo2+2
kappaMinusTwo2 ~ dgamma(0.01,0.01)
pi ~ dbeta(1,1)
}"
writeLines( modelString4 , con="TEMPmodel4.bug" )

#############################################################################
# prepare data and use function jags() to run Algorithm 3 or Algorithm 4
#############################################################################
#n is the sample size,
#K is the number of raters (tests),
#m is the original dataset, an n by K matrix.
#all 2^K test result combinations
N = 2^K
y = matrix(0, N, K)
for ( i in 1:K){
y[,i] = rep(c(0,1),each = 2^(K-i))}
#freq is the observed frequency vector of test result combinations
freq = numeric(N)
```

```
for ( i in 1:n){
for ( j in 1:N){
if (all (m[i,] == y[j,])){freq[j] = freq[j]+1}}}
t = freq
x<- structure(y, .Dim=c(N,K))
z<-rep(0,n)
dat=list(K=K,n=n, N=N, x=x, z=z, t=t)
#Output
par1=c("se","sp","kappa2","kappa1","omega2","omega1", "pi","prob")
out3<- jags(model ="TEMPmodel3.bug", parameters.to.save=par1, data=dat,
            n.chains=1, n.iter=10000, n.burnin=1000)
par2=c("se","sp","kappa2","kappa1","omega2","omega1", "pi", "c1","c2","prob")
out4<- jags(model ="TEMPmodel4.bug", parameters.to.save=par2, data=dat,
            n.chains=1, n.iter=10000, n.burnin=1000)
```

B.2 Estimation results for scenario 1 when fitting with model M2 with different covariance structures

We fit the same 200 simulated data sets of scenario 1 with M2 having one pair of important covariance parameters omitted intentionally and M2 with the selected significant covariances by the proposed one-sided credible interval criterion. The results are summarized in the middle three columns of Table B.1. In the table, Bias and SD stand for the average bias and average posterior standard deviation of each parameter for the 200 data sets, respectively. The first column of the table (also column 1 in Table 3.1 in Chapter 3) shows the estimation results when the completely misspecified conditional independence model (M1) is fitted. The fifth column of the table (also column 3 in Table 3.1 in Chapter 3) shows the estimation results when the true model (M2 with two pairs of covariance parameters) is fitted. It is clear that

missing a pair of important covariance parameters leads to larger biases than those using the true model, but smaller biases than those using the completely misspecified conditional independence model (M1). The estimation results when fitting with M2 with selected covariances by the one-sided credible interval criterion are very close to those when fitting with the true model. Finally, it's observed that the SD's are very close to each other when fitting with different covariance structures.

Table B.1: Simulation results for scenario one having different covariance structures.

		M1 no covariance	C_{23}^{+}/C_{23}^{-} omitted	C_{34}^{+}/C_{34}^{-} omitted	selected covariances	true model
	Truth	Bias (SD)	Bias (SD)	Bias (SD)	Bias (SD)	Bias (SD)
π	0.450	-0.011 (0.025)	-0.008 (0.026)	-0.006 (0.026)	-0.001 (0.026)	0.000 (0.026)
Se_1	0.960	-0.009 (0.017)	-0.008 (0.017)	-0.007 (0.017)	-0.007 (0.018)	-0.007 (0.017)
Se_2	0.870	0.019 (0.024)	0.018 (0.026)	0.008 (0.025)	0.004 (0.026)	0.002 (0.026)
Se_3	0.810	0.031 (0.029)	0.022 (0.030)	0.017 (0.030)	0.002 (0.033)	-0.002 (0.033)
Se_4	0.860	0.016 (0.025)	0.007 (0.026)	0.012 (0.026)	0.002 (0.027)	0.001 (0.027)
Sp_1	0.970	-0.017 (0.016)	-0.012 (0.015)	-0.009 (0.014)	-0.002 (0.013)	0.000 (0.013)
Sp_2	0.980	-0.006 (0.011)	-0.002 (0.010)	-0.009 (0.012)	-0.004 (0.011)	-0.004 (0.011)
Sp_3	0.990	-0.003 (0.007)	-0.004 (0.008)	-0.005 (0.008)	-0.006 (0.009)	-0.007 (0.009)
Sp_4	0.970	-0.006 (0.013)	-0.011 (0.014)	-0.001 (0.011)	-0.004 (0.012)	0.003 (0.012)
C_{23}^{+}	0.050	—	—	-0.009 (0.014)	—	-0.001 (0.015)
C_{34}^{+}	0.050	—	-0.007 (0.014)	—	—	0.001 (0.015)
C_{23}^{-}	0.001	—	—	0.004 (0.004)	—	0.004 (0.004)
C_{34}^{-}	0.001	—	0.003 (0.004)	—	—	0.003 (0.004)
	RMSE	0.01567	0.01243	0.00974	0.00446	0.00416

B.3 Estimation results for scenario 2 when 7 pairs of covariance terms are added in model M2

For the simulation scenario 2 where the true model has significant covariances C_{23}^{+}, C_{14}^{+}, C_{24}^{+}, C_{25}^{+}, and C_{35}^{-}, the correlation residual analysis always identifies 7 pairs of correlations as significant: r_{23}, r_{14}, r_{24}, r_{25}, r_{34}, r_{35}, and r_{45}. Table B.2 below

summarizes the estimation results when we fit the simulated 200 data sets with model M2 with these seven pairs of covariances added.

Table B.2: Simulation results for K=5, n=4000 with 7 pairs of covariance terms added

	Truth	Bias	Post.SD
π	0.450	-0.001	0.014
Se_1	0.770	0.002	0.016
Se_2	0.650	-0.001	0.016
Se_3	0.710	-0.000	0.016
Se_4	0.680	0.002	0.019
Se_5	0.710	0.001	0.015
Sp_1	0.890	0.002	0.011
Sp_2	0.930	0.004	0.009
Sp_3	0.880	0.001	0.011
Sp_4	0.900	0.003	0.012
Sp_5	0.860	0.002	0.011
C_{23}^{+}	0.050	0.001	0.007
C_{14}^{+}	0.050	0.001	0.008
C_{24}^{+}	0.050	0.001	0.007
C_{34}^{+}	0.000	-0.000	0.006
C_{25}^{+}	0.050	0.000	0.007
C_{35}^{+}	0.001	-0.000	0.006
C_{45}^{+}	0.000	-0.000	0.006
C_{23}^{-}	0.001	-0.001	0.004
C_{14}^{-}	0.001	-0.001	0.005
C_{24}^{-}	0.001	-0.001	0.003
C_{34}^{-}	0.000	-0.000	0.003
C_{25}^{-}	0.001	-0.001	0.004
C_{35}^{-}	0.050	0.001	0.006
C_{45}^{-}	0.000	-0.000	0.003

B.4 Robustness study for model M2

We investigate if our pairwise covariance model (M2) can handle the data generated from the Gaussian random effect (GRE) model and finite mixture (FM) model. Two

hundred data sets are simulated from the GRE and FM model, respectively, and then are fitted with our proposed methods. The correlation residual analysis after fitting M1 suggests that all the pairwise covariance terms should be selected for the GRE generated data and almost so for the FM generated data. When including all the covariance terms in model M2, the estimation results of disease prevalence, sensitivities and specificities, shown in Table B.3, are biased.

Table B.3: Simulation results with Model M2 for the data generated from GRE model and FM model

	Truth	GRE	FM
π	0.45	0.290 (0.012)	0.226 (0.019)
Se_1	0.77	0.708 (0.013)	0.707 (0.049)
Se_2	0.65	0.640 (0.013)	0.661 (0.044)
Se_3	0.71	0.692 (0.013)	0.717 (0.042)
Se_4	0.68	0.667 (0.013)	0.688 (0.043)
Se_5	0.71	0.702 (0.013)	0.724 (0.041)
Sp_1	0.89	0.712 (0.007)	0.679 (0.017)
Sp_2	0.93	0.788 (0.006)	0.763 (0.015)
Sp_3	0.88	0.735 (0.007)	0.710 (0.015)
Sp_4	0.90	0.758 (0.007)	0.732 (0.015)
Sp_5	0.86	0.724 (0.007)	0.698 (0.015)

Appendix C

Chapter 4 Appendix and Supplementary Materials

C.1 Derivation of $Cov(N_i^{(j)}, N_i^{(k)})$, $Var(N_i^{(j)})$, and $Var(N_i^{(k)})$ for $\alpha_j = 1$

$$\begin{aligned}
Cov(N_i^{(j)}, N_i^{(k)}) &= E(N_i^{(j)} N_i^{(k)}) - E(N_i^{(j)})E(N_i^{(k)}) \\
&= E[E(N_i^{(j)}|w_i)E(N_i^{(k)}|w_i)] - E[E(N_i^{(j)}|w_i)]E[E(N_i^{(k)}|w_i)] \\
&= U_0^{(j)} \exp(X'\boldsymbol{\beta}^{(j)}) U_0^{(k)} \exp(X'\boldsymbol{\beta}^{(k)}) [E(w_i^{1+\alpha_k}) - E(w_i)E(w_i^{\alpha_k})] \\
&= U_0^{(j)} \exp(X'\boldsymbol{\beta}^{(j)}) U_0^{(k)} \exp(X'\boldsymbol{\beta}^{(k)}) \times \frac{\alpha_k \eta^{-(\alpha_k+1)} \Gamma(\eta + \alpha_k)}{\Gamma(\eta)}
\end{aligned}$$

$$\begin{aligned}
Var(N_i^{(j)}) &= Var(E(N_i^{(j)}|w_i)) + E(Var(N_i^{(j)}|w_i)) \\
&= Var(w_i U_0^{(j)} \exp(X'\boldsymbol{\beta}^{(j)})) + E(w_i U_0^{(j)} \exp(X'\boldsymbol{\beta}^{(j)})) \\
&= \frac{(U_0^{(j)} \exp(X'\boldsymbol{\beta}^{(j)}))^2}{\eta} + U_0^{(j)} \exp(X'\boldsymbol{\beta}^{(j)})
\end{aligned}$$

$$\begin{aligned}
Var(N_i^{(k)}) &= Var(E(N_i^{(k)}|w_i)) + E(Var(N_i^{(k)}|w_i)) \\
&= Var(w_i^{\alpha_k} U_0^{(k)} \exp(X'\boldsymbol{\beta}^{(k)})) + E(w_i^{\alpha_k} U_0^{(k)} \exp(X'\boldsymbol{\beta}^{(k)})) \\
&= (U_0^{(k)} \exp(X'\boldsymbol{\beta}^{(k)}))^2 [\frac{\Gamma(\eta + 2\alpha_k)}{\Gamma(\eta)\eta^{2\alpha_k}} - (\frac{\Gamma(\alpha_k + \eta)}{\Gamma(\eta)})^2 \frac{1}{\eta^{2\alpha_k}}] + \\
&\quad U_0^{(k)} \exp(X'\boldsymbol{\beta}^{(k)}) \frac{\Gamma(\alpha_k + \eta)}{\Gamma(\eta)\eta^{\alpha_k}}
\end{aligned}$$

C.2 Derivation of $Cov(N_i^{(p)}, N_i^{(q)})$, $Var(N_i^{(p)})$, $Var(N_i^{(q)})$ and $Corr(N_i^{(p)}, N_i^{(q)})$ for any α_p and α_q

$$
\begin{aligned}
Cov(N_i^{(p)}, N_i^{(q)}) &= E(N_i^{(p)} N_i^{(q)}) - E(N_i^{(p)})E(N_i^{(q)}) \\
&= E[E(N_i^{(p)}|w_i)E(N_i^{(q)}|w_i)] - E[E(N_i^{(p)}|w_i)]E[E(N_i^{(q)}|w_i)] \\
&= U_0^{(p)} \exp(X'\boldsymbol{\beta}^{(p)}) U_0^{(q)} \exp(X'\boldsymbol{\beta}^{(q)}) [E(w_i^{\alpha_p+\alpha_q}) - E(w_i^{\alpha_p})E(w_i^{\alpha_q})] \\
&= U_0^{(p)} \exp(X'\boldsymbol{\beta}^{(p)}) U_0^{(q)} \exp(X'\boldsymbol{\beta}^{(q)}) \times \frac{\Gamma(\alpha_p+\alpha_q+\eta) - \frac{\Gamma(\alpha_p+\eta)\Gamma(\alpha_q+\eta)}{\Gamma(\eta)}}{\Gamma(\eta)\eta^{\alpha_p+\alpha_q}}
\end{aligned}
$$

$$
\begin{aligned}
Var(N_i^{(p)}) &= Var(E(N_i^{(p)}|w_i)) + E(Var(N_i^{(p)}|w_i)) \\
&= Var(w_i^{\alpha_p} U_0^{(p)} \exp(X'\boldsymbol{\beta}^{(p)})) + E(w_i^{\alpha_p} U_0^{(p)} \exp(X'\boldsymbol{\beta}^{(p)})) \\
&= (U_0^{(p)} \exp(X'\boldsymbol{\beta}^{(p)}))^2 [\frac{\Gamma(\eta+2\alpha_p)}{\Gamma(\eta)\eta^{2\alpha_p}} - (\frac{\Gamma(\alpha_p+\eta)}{\Gamma(\eta)})^2 \frac{1}{\eta^{2\alpha_p}}] + \\
&\quad U_0^{(p)} \exp(X'\boldsymbol{\beta}^{(p)}) \frac{\Gamma(\alpha_p+\eta)}{\Gamma(\eta)\eta^{\alpha_p}}
\end{aligned}
$$

$$
\begin{aligned}
Var(N_i^{(q)}) &= Var(E(N_i^{(q)}|w_i)) + E(Var(N_i^{(q)}|w_i)) \\
&= Var(w_i^{\alpha_q} U_0^{(q)} \exp(X'\boldsymbol{\beta}^{(q)})) + E(w_i^{\alpha_q} U_0^{(q)} \exp(X'\boldsymbol{\beta}^{(q)})) \\
&= (U_0^{(q)} \exp(X'\boldsymbol{\beta}^{(q)}))^2 [\frac{\Gamma(\eta+2\alpha_q)}{\Gamma(\eta)\eta^{2\alpha_q}} - (\frac{\Gamma(\alpha_q+\eta)}{\Gamma(\eta)})^2 \frac{1}{\eta^{2\alpha_q}}] + \\
&\quad U_0^{(q)} \exp(X'\boldsymbol{\beta}^{(q)}) \frac{\Gamma(\alpha_q+\eta)}{\Gamma(\eta)\eta^{\alpha_q}}
\end{aligned}
$$

$$
\begin{aligned}
Corr\Big(N_i^{(p)}, N_i^{(q)}\Big) &= \frac{\Gamma(\eta)\Gamma(\alpha_p+\alpha_q+\eta) - \Gamma(\alpha_p+\eta)\Gamma(\alpha_p+\eta)}{A \cdot B}, \text{where} \\
A &= \sqrt{\eta^{2\alpha_q}\Gamma(\eta)^2[\Gamma(\eta)\Gamma(2\alpha_p+\eta) - \Gamma(\alpha_p+\eta)^2] + \frac{\Gamma(\eta)^3\Gamma(\alpha_p+\eta)\eta^{\alpha_p+2\alpha_q}}{U_0^{(p)}(t)\exp(X_i'\boldsymbol{\beta}^{(p)})}} \\
B &= \sqrt{\eta^{2\alpha_p}\Gamma(\eta)^2[\Gamma(\eta)\Gamma(2\alpha_q+\eta) - \Gamma(\alpha_q+\eta)^2] + \frac{\Gamma(\eta)^3\Gamma(\alpha_q+\eta)\eta^{2\alpha_p+\alpha_q}}{U_0^{(q)}(t)\exp(X_i'\boldsymbol{\beta}^{(q)})}}
\end{aligned}
$$

$$
\begin{aligned}
Corr_0\Big(N_i^{(p)}, N_i^{(q)}\Big) &= \frac{\Gamma(\eta)\Gamma(\alpha_p+\alpha_q+\eta)-\Gamma(\alpha_p+\eta)\Gamma(\alpha_p+\eta)}{C\cdot D}, \text{where} \\
C &= \sqrt{\eta^{2\alpha_q}\Gamma(\eta)^2[\Gamma(\eta)\Gamma(2\alpha_p+\eta)-\Gamma(\alpha_p+\eta)^2]+\frac{\Gamma(\eta)^3\Gamma(\alpha_p+\eta)\eta^{\alpha_p+2\alpha_q}}{U_0^{(p)}(t)}} \\
D &= \sqrt{\eta^{2\alpha_p}\Gamma(\eta)^2[\Gamma(\eta)\Gamma(2\alpha_q+\eta)-\Gamma(\alpha_q+\eta)^2]+\frac{\Gamma(\eta)^3\Gamma(\alpha_q+\eta)\eta^{2\alpha_p+\alpha_q}}{U_0^{(q)}(t)}}
\end{aligned}
$$