BAYESIAN INFLUENCE DIAGNOSTIC METHODS FOR PARAMETRIC REGRESSION MODELS

Hyunsoon Cho

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, Gillings School of Global Public Health.

Chapel Hill
2009

Approved by:
Dr. Joseph G. Ibrahim, Advisor
Dr. Hongtu Zhu, Co-advisor
Dr. Gary G. Koch, Reader
Dr. Haitao Chu, Reader
Dr. Karen A. Graham, Reader

ABSTRACT

HYUNSOON CHO: BAYESIAN INFLUENCE DIAGNOSTIC METHODS FOR PARAMETRIC REGRESSION MODELS. (Under the direction of Dr. Joseph G. Ibrahim and Dr. Hongtu Zhu.)

The goals of assessing the influence of individual observations in statistical analysis are not only to identify influential observations such as outliers and high leverage points, but also to determine the importance of each observation in the analysis for a better model fit. Thus, assessing the influence of individual observations on a model, choosing an appropriate dimensionality of a model and selecting the best model for a given dataset are very important and highly relevant problems in any formal statistical analysis.

Recently, Bayesian methodologies have been getting enormous attention in biomedical research due to the potential advantages of fitting a vast array of complex models posed by modern data. As the demand for Bayesian data analysis and modeling increases, we need good diagnostic methods for model assessment and selection. In this dissertation, we develop Bayesian diagnostic measures based on case-deletion to assess the influence of each observation to model fit and model complexity. First, we propose Bayesian case influence diagnostics for complex survival models. In detail, we develop case deletion influence diagnostics for both the joint and marginal posterior distributions based on the Kullback-Leibler divergence. Second, we introduce three types of Bayesian case influence measures based on case deletion, namely the *ϕ-divergence*, *Cook's posterior mode distance* and *Cook's posterior mean distance* to evaluate the effects of deleting a set of observations in general Bayesian parametric models. We also examine the statistical properties of these three Bayesian case influence measures and

their applications to identification of influential sets and model complexity.

In any deletion diagnostic, "size matters" issue persists and it is a fundamental issue of influence analysis, because the size of the deletion diagnostic is associated with the size of the perturbation. For Cook's distance, that is Cook's distance is a monotonic function of the size of perturbation. Thus, we develop a scaled version of Cook's distance to address the size issue for deletion diagnostics in general parametric models.

ACKNOWLEDGMENTS

I have been very fortunate to have great mentors and friends during my journey of graduate studies. I would especially like to thank my advisor, Dr. Joseph Ibrahim, for his mentorship, advice and financial support during the completion of this dissertation. His genuine advice and leadership as a mentor guided me to a successful completion of my graduate studies. Also, I would like to thank Dr. Hongtu Zhu, who also served as my advisor, for his important contributions and guidance. He helped me to get through the difficult problems. I would like to express my deepest appreciation to Dr. Gary Koch for his generosity with time, advice and financial support. His advice and help for my consulting work at Biometric Consulting Laboratory (BCL) as well as for my life as a biostatistician were priceless. In addition, I would like to give special thanks to my other committee members, Dr. Haitao Chu and Dr. Karen Graham, for their time and comments. I am grateful for all of my friends in BCL for their friendship and encouragement. They have been wonderful to be around and are indispensable to my graduate studies in Chapel Hill.

Most importantly, this dissertation would not have been possible without the love and patience of my husband and family in Korea. My parents always believed me and provided me selfless devotion. My younger brother always has been my advocate. Finally, this dissertation is dedicated to my husband Jung-il for his endless love, support and encouragement. Although he is a Ph.D. in mechanical engineering, he has been my best discussion partner for my research. My journey would never have been completed without his devotion.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
ARMS	Adaptive Rejection Metropolis Sampling
BCIC	Bayesian Case-influence Information Criterion
BIC	Bayesian Information Criterion
BPIC	Bayesian Predictive Information Criterion
CD	Cook's Distance
CPO	Conditional Predictive Ordinate
DIC	Deviance Information Criterion
K-L	Kullback-Leibler
LD	Likelihood Displacement
MCMC	Markov Chain Monte Carlo

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

The importance of identification of influential observations in a statistical analysis is a well recognized methodological problem, and the development of diagnostic measures to detect influential observations is of interest to many researchers. Influential observations in a given dataset can have a strong impact on statistical inference and conclusions. In these situations, such influential observations are an important part of the data, and hence require the most careful examination. In statistical analysis, the goals of assessing the influence of individual observations (or generally, a set of observations) are not only to identify influential observations (or sets of observations) such as outliers and high leverage points, but also to determine the importance of each observation in the analysis for a better model fit (Stone, 1974, 1977; Cook, 1977; Cook and Weisberg, 1982; McCulloch, 1989; Geisser, 1975, 1993; Zhang, 1993). Thus, assessing the influence of individual observations on a model, choosing an appropriate dimensionality of a model and selecting the best model for a given dataset are very important and highly relevant problems in any formal statistical analysis.

Recently, Bayesian methodologies have been getting enormous attention in biomedical research due to the potential advantages of fitting a vast array of complex models posed by modern data. As the demand for Bayesian data analysis and modeling increases, we need good diagnostic methods for model assessment and selection. However, development of Bayesian influence diagnostic methods for parametric regression models pose both theoretical and computational challenges. Motivated by this, the first paper proposes Bayesian case influence diagnostics for complex survival models. We develop case deletion influence diagnostics for both the joint and marginal posterior distributions based on the Kullback-Leibler divergence (K-L divergence) (Kullback and Leibler, 1951). We present a simplified expression for computing the K-L divergence between the posterior with the full data and the posterior based on single case deletion. In addition, we investigate a theoretical connection between the proposed diagnostics based on the K-L divergence and Conditional Predictive Ordinate (CPO) (Gelfand et al., 1992; Geisser, 1993), as well as a connection between diagnostics based on Cox's partial likelihood (Cox, 1975). The second paper introduces three types of Bayesian case influence measures based on case deletion, namely the *ϕ-divergence*, *Cook's posterior mode distance* and *Cook's posterior mean distance* to evaluate the effects of deleting a set of observations in general Bayesian parametric models. We examine the statistical properties of these three Bayesian case influence measures and their applications to identification of influential sets and model complexity. This complexity measure is related to the complexity terms in other information criteria such as the Akaike Information Criterion (AIC) (Akaike, 1973) and the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002), and the leave-k-out cross validation method (Stone, 1974, 1977, 2002; Geisser and Eddy, 1979).

Cook's distance is one of the most important diagnostic tools for evaluating the effects of deleting a subset of observations on a parameter estimate or a fitted value in a large class of statistical models. However, for many complex data structures (e.g.,

longitudinal data) no rigorous approach has been developed to address a fundamental issue of Cook's distance: "size matters", that is Cook's distance is a monotonic function of the size of the perturbation. This issue has been largely neglected in the literature. The size matters issue persists in any deletion diagnostic, because the size of the deletion diagnostic is associated with the size of the perturbation. The third paper develops a scaled version of Cook's distance to address the size issue for deletion diagnostics in general parametric models. We use stochastic ordering to quantify the relationship between the size of perturbation and the amount of the perturbation on Cook's distance. Our scaled Cook's distance properly accounts for the size of a perturbation and the fitted model to the data.

The rest of this dissertation is organized as follows. The next section presents literature reviews of case influence measures and model assessment tools based on the criterion-based methods. Then we proceed to present each of the three papers: The first paper is presented in Chapter 2, and it develops Bayesian case influence diagnostics for survival models for both continuous survival time data and grouped survival data. The second paper is discussed in Chapter 3, and it proposes methods to evaluate the effects of deleting a set of observations in Bayesian regression models. These models include linear models, mixed models, generalized linear models and generalized linear mixed models. The third paper is discussed in Chapter 4, and it is dedicated to resolving the size issues for Cook's distance in general parametric model.

1.2 Literature Review

1.2.1 Case Influence Measures

In frequentist analysis, enormous research has been done for detecting outliers, influential points, and leverage points by assessing the influence of individual observations. The techniques used in such analysis are residuals, leverages, case-deletion measures,

and local influence measures (Cook, 1977; Belsley et al., 1980; Cook and Weisberg, 1982; Cook, 1986). A general approach to influence analysis is studying the changes in the outcome or other aspects of an analysis caused by a small perturbations in the model. The most popular perturbation scheme is based on case deletion. In fact, case deletion is also a special case of perturbations in local influence analysis (Cook, 1986), which utilize the concept of normal curvature in differential geometry in assessing the local behavior of the likelihood displacement.

Two widely used case deletion measures for assessing case influence are Cook's distance (Cook, 1977) and likelihood displacement (Cook and Weisberg, 1982; Cook, 1986). Likelihood displacement and Cook's distance have been used to detect influential observations in various parametric and semiparametric models from the frequentist point of view (Cook and Weisberg, 1982; Thomas and Cook, 1989; Pettitt and Daud, 1989; Weissfeld, 1990; Escobar and Meeker, 1992). The likelihood displacement measures the effect of deleting one observation on overall model fit using the log-likelihood. Let $\boldsymbol{\beta}$ be $p \times 1$ vector of the parameter of interest, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$ be the parameter estimates, usually maximum likelihood estimates (MLE), with full data and with the ith case deleted data, respectively, and $L(\boldsymbol{\beta})$ be the likelihood function for $\boldsymbol{\beta}$. The likelihood displacement is defined by

$$\mathrm{LD}(i) = 2\{\log L(\hat{\boldsymbol{\beta}}) - \log L(\hat{\boldsymbol{\beta}}_{(i)})\}. \tag{1.1}$$

For more about likelihood displacement, see Cook and Weisberg (1982), p182-183. Cook's distance measures the effect of deleting one observation on a parameter estimate or a fitted value. The generalized Cook distance for $\boldsymbol{\beta}$ is defined by

$$\mathrm{CD}(i) = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T M (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}), \tag{1.2}$$

where M is a positive definite weight matrix and M is often set as the Fisher information matrix. Since the seminal work of Cook (1977) on Cook's distance in linear regression, considerable research has been devoted to developing deletion diagnostics including Cook's distance for detecting influential observations (or clusters) in various statistical models including generalized linear models and the general linear model with correlated error (Cook, 1977; Cook and Weisberg, 1982; Chatterjee and Hadi, 1988; Andersen, 1992; Davison and Tsai, 1992; Wei, 1998; Haslett, 1999; Zhu et al., 2001; Fung et al., 2002). For instance, Preisser and Qaqish (1996) developed Cook's distance for generalized estimating equations. Christensen et al. (1992) and Banerjee and Frees (1997) considered case deletion and subject deletion diagnostics, respectively. Zhu et al. (2001) developed deletion diagnostics for models with missing data.

In Bayesian analysis, considerable research has been devoted to developing single case influence measures for various specific statistical models including generalized linear models, time series models, and survival models (Johnson and Geisser, 1983; Johnson, 1985; Pettit, 1986; Kass et al., 1989; Carlin and Polson, 1991; Gelfand et al., 1992; Weiss and Cook, 1992; Geisser, 1993; Blyth, 1994; Peng and Dey, 1995; Weiss, 1996; Christensen, 1997; Bradlow and Zaslavsky, 1997). There are two distinct approaches in assessing influence of individual observations. One is assessing the influence on the posterior distribution and other is assessing the influence with regard to the predictive distribution. For those two approaches, a common way of assessing the influence of an observation on model fit is through case deletion.

The two most popular Bayesian case influence measures are the Conditional Predictive Ordinate (CPO) (Gelfand et al., 1992; Geisser, 1993) and $\phi-$divergence (Weiss and Cook, 1992; Csiszár, 1967). The CPO is defined as the predictive density of the ith case given the data without the ith case. A large value of CPO for the ith case implies better concordance of the ith case with the rest of the data, and hence a better model fit. The $\phi-$divergence or $\phi-$influence based on case deletion is a measure of

discrepancy between the posterior distributions with and without a particular case. Various forms of $\phi(\cdot)$ have been considered in the literature (Weiss, 1996; Weiss and Cook, 1992; Kass et al., 1989; Blyth, 1994), which include L_1-distance, χ^2-divergence and Kullback-Leibler divergence (K-L divergence). A large value of the $\phi-$divergence for the ith case implies more influence of the ith case on estimation, hypothesis testing, and model fit. Many researchers have been interested in developing case influence diagnostics using the $\phi-$divergence, especially K-L divergence, under various parametric models (Johnson and Geisser, 1985; Pettit, 1986; Carlin and Polson, 1991; Weiss and Cook, 1992; Peng and Dey, 1995; Weiss, 1996; Christensen, 1997; Weiss and Cho, 1998). Pettit (1986) suggested the use of the K-L divergence in detecting influential observations in his review of Bayesian diagnostics. Carlin and Polson (1991) proposed an expected utility approach using the K-L divergence as a utility function to define the influence of a set of observations in a parametric modeling framework, considering the normal linear model and mixed models. Weiss and Cook (1992) introduced the K-L divergence to assess the divergence between posteriors in the context of case deletion in generalized linear models. Peng and Dey (1995) also developed a Bayesian diagnostic measure using general divergence measures including the K-L divergence on the posterior distribution and applied this measure to several regression models, such as a nonlinear model. Weiss (1996) and Weiss and Cho (1998) proposed assessing the influence of case deletion using model perturbations as well as establishing its relationship to the K-L divergence and CPO. Bayesian influence measures for assessing marginal posterior distributions have also been developed for the multivariate linear model and normal random effects models (Johnson and Geisser, 1985; Weiss and Cho, 1998).

1.2.2 Criterion-Based Model Assessment

In statistical analysis, we often interested in choosing an appropriate dimensionality of a model and selecting the best model for a given dataset. In the classical modeling

framework, information criteria (IC) are fundamental criteria for model comparisons, which incorporate measures of fit and complexity for model choice. Typically, deviance statistics are used for the measure of fit, and the number of parameters or degrees of freedom of estimators are used for the complexity of a model.

Considerable research has been devoted to model comparison and evaluation using the concepts of information criteria in both the frequentist and Bayesian point of views (Akaike, 1973; Takeuchi, 1976; Schwarz, 1978; Murata et al., 1994; Konishi and Kitagawa, 1996; Spiegelhalter et al., 2002; Ando, 2007), in which they incorporate different complexity terms for model choice. Akaike (1973) proposed the Akaike Information Criterion (AIC) which is defined by $\mathrm{AIC} = -2\log\{p(\boldsymbol{y}|\hat{\boldsymbol{\theta}})\} + 2p$, where p is the number of parameters and $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$. The Takeuchi Information Criterion (TIC) (Takeuchi, 1976) and Generalized Information Criterion (GIC) (Konishi and Kitagawa, 1996) are generalizations of AIC which relax the following assumptions: (i) a specified parametric family of distributions include the true model; and (ii) a model is "estimated" by its MLE. TIC relaxed assumption (i) and GIC relaxed both (i) and (ii). Murata *et al.* (1994) proposed the Network Information Criterion (NIC) as a generalization of AIC for determining the optimal number of parameters in neural networks. Schwarz (1978) adopted Bayesian argument in the development of the Bayesian Information Criterion (BIC), which is defined by $\mathrm{BIC} = -2\log\{p(\boldsymbol{y}|\hat{\boldsymbol{\theta}})\} + p\log(n)$, where n is the number of observations in the dataset. The Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) is defined by the posterior mean of the deviance as a Bayesian measure of fit and the effective number of parameters as the complexity component. DIC is given by $\mathrm{DIC} = -2E_{\boldsymbol{\theta}|\boldsymbol{Y}}[\log\{p(\boldsymbol{y}|\boldsymbol{\theta})\}] + p_D$, where $E_{\boldsymbol{\theta}|\boldsymbol{Y}}[\cdot]$ is the expectation with respect to the posterior distribution, $p(\boldsymbol{\theta}|\boldsymbol{y})$. The effective number of parameters, p_D is defined by $p_D = E_{\boldsymbol{\theta}|\boldsymbol{Y}}[-2\log\{p(\boldsymbol{y}|\boldsymbol{\theta})\}] + 2\log\{p(\boldsymbol{y}|\tilde{\boldsymbol{\theta}})\}$, where $\tilde{\boldsymbol{\theta}}$ is the posterior mean of $\boldsymbol{\theta}$. Thus, DIC can be used for comparing complex hierarchical Bayesian models in which the number of parameter is not clearly defined. The Bayesian

Predictive Information Criterion (BPIC) was proposed by Ando (2007) as an estimator of the posterior mean of the expected log-likelihood of the predictive distribution. BPIC is defined as BPIC $= -2E_{\boldsymbol{\theta}|\boldsymbol{Y}}[\log\{p(\boldsymbol{y}|\boldsymbol{\theta})\}] + 2n\hat{b}_{\boldsymbol{\theta}}$, where $\hat{b}_{\boldsymbol{\theta}}$ is the estimated asymptotic bias of the predictive discrepancy measures. For more details about $\hat{b}_{\boldsymbol{\theta}}$, see Ando (2007). L measure (Ibrahim and Laud, 1994; Gelfand and Ghosh, 1998; Ibrahim et al., 2001b; Chen et al., 2004) is another type of Bayesian criterion-based method, defined as the expected squared Euclidean distance between the observed data $\boldsymbol{y}$ with joint sampling density $p(\boldsymbol{y}|\boldsymbol{\theta})$ and the future response vector with the same sampling density as $\boldsymbol{y}|\boldsymbol{\theta}$.

CHAPTER 2

BAYESIAN CASE INFLUENCE DIAGNOSTICS FOR SURVIVAL MODELS

2.1 Introduction

In Bayesian analysis, considerable research has been done for developing case influence diagnostics using the K-L divergence under various parametric models (Johnson and Geisser, 1985; Pettit, 1986; Carlin and Polson, 1991; Weiss and Cook, 1992; Weiss, 1996; Weiss and Cho, 1998). Despite the extensive literature on Bayesian diagnostic methods for parametric models, very little has been developed for semiparametric models, including survival models. Due to the potential advantages of fitting a vast array of complex survival models posed by modern survival data, semiparametirc Bayesian methodologies in survival analysis have been getting enormous attention in biomedical research. Bayesian case influence diagnostics for survival models pose both theoretical and computational challenges, which are discussed here.

The objective of this paper is to propose Bayesian case deletion influence diagnostics for survival models. First, we develop diagnostic measures to assess the influence of

a case on both the joint and marginal posterior distributions based on the directed K-L divergence. In this development, we derive a novel and simplified expression for computing the K-L divergence, which facilitates efficient computation of the proposed diagnostic measures using Markov chain Monte Carlo (MCMC) samples from full data posterior distribution. This avoids the burden of sampling from each of the n posterior distributions, each based on deletion of the ith case, $i = 1, \cdots, n$. Second, we apply the proposed methodology to Bayesian survival models with continuous survival time data and grouped survival data. The survival model we consider are the Cox model with a gamma process prior on the cumulative baseline hazard (Sinha et al., 2003) and a proportional hazards frailty model in the presence of continuous survival time data accommodating correlated and clustered data. In the presence of grouped survival data, we considered the Cox model with a beta process prior. In addition, we investigate a theoretical connection between the proposed diagnostics based on the K-L divergence and CPO, as well as a connection between diagnostics based on Cox's partial likelihood.

To motivate the proposed methodology, we consider a well known dataset, the Stanford heart transplant data (Miller and Halpern, 1982). The dataset contains 184 transplant cases with the following variables: time measured from the date of the transplant in days; status code (dead or alive); patient age at first transplant in years; T5 mismatch score (missing for 27 of the cases). This dataset have been analyzed by many, illustrating frequentist diagnostic measures (Pettitt and Daud, 1989; Escobar and Meeker, 1992). Here, it is of interest to carry out Bayesian diagnostic methods not only to compare our results with the frequentist results, but also to possibly find other influential (or noninfluential) cases not identified by the previous methods. As shown in Figure 2.2, our proposed Bayesian diagnostic method identified some cases as influential in this dataset. More details regarding this example are given in Section 2.6.2. To further illustrate the methodology, we also apply the proposed methods to simulated data and a phase III melanoma clinical trial (E1690) discussed in Sections

2.6.1 and 2.6.3, respectively.

The rest of this paper is organized as follows. In Section 2.2, we introduce Bayesian case influence diagnostics based on the K-L divergence. In Sections 2.3 and 2.4, we derive case influence diagnostics for the Cox model and Cox frailty model with a gamma process prior. In Section 2.5, we present case influence diagnostics for the Cox model with a beta process prior. In Section 2.6, we examine the performance of the influence diagnostics using simulated data, the Stanford Heart Transplant data and the E1690 trial. We conclude the paper with some discussion in Section 2.7.

2.2 The Proposed Method

2.2.1 General Development

Let D be full data and D_{-i} be the data with the ith case deleted. Let $L(\boldsymbol{\beta}|D)$ denotes the likelihood based on the full data and $L(\boldsymbol{\beta}|D_{-i})$ denotes the likelihood based on the data without the ith case. The posterior distributions for the the full data and the ith case deleted can be defined as $p(\boldsymbol{\beta}|D) \propto L(\boldsymbol{\beta}|D)\pi(\boldsymbol{\beta})$ and $p(\boldsymbol{\beta}|D_{-i}) \propto L(\boldsymbol{\beta}|D_{-i})\pi(\boldsymbol{\beta})$, respectively, where $\pi(\boldsymbol{\beta})$ is the prior distribution of $\boldsymbol{\beta}$. A typical choice of $\pi(\boldsymbol{\beta})$ is a $N_p(\mu_{\mathbf{0}}, \boldsymbol{\Sigma_0})$ distribution or a uniform improper prior.

Let $K(P, P_{-i})$ denote the K-L divergence between P and P_{-i}, where P denotes the posterior distribution of $\boldsymbol{\beta}$ for the full data, and P_{-i} denotes the posterior distribution of $\boldsymbol{\beta}$ without the ith case. Specifically,

$$K(P, P_{-i}) = \int p(\boldsymbol{\beta}|D) \log \left\{ \frac{p(\boldsymbol{\beta}|D)}{p(\boldsymbol{\beta}|D_{-i})} \right\} d\boldsymbol{\beta}. \tag{2.1}$$

$K(P, P_{-i})$ thus measures the effect of deleting the ith case from the full data on the joint posterior distribution of $\boldsymbol{\beta}$. Note that $K(P, P_{-i}) \neq K(P_{-i}, P)$ in general. After some algebra, as shown in the Appendix A, we can derive a simplified expression for

$K(P, P_{-i})$ as follows:

$$K(P, P_{-i}) = \log E_{\boldsymbol{\beta}} \left[\frac{L(\boldsymbol{\beta}|D_{-i})}{L(\boldsymbol{\beta}|D)} \middle| D \right] + E_{\boldsymbol{\beta}} \left[\log \left\{ \frac{L(\boldsymbol{\beta}|D)}{L(\boldsymbol{\beta}|D_{-i})} \right\} \middle| D \right], \tag{2.2}$$

where $E_{\boldsymbol{\beta}}[\cdot|D]$ represents the expectation with respect to the joint posterior distribution of $\boldsymbol{\beta}$ given D. Equation (2.2) enables us to compute $K(P, P_{-i})$ for $i = 1, \cdots, n$, using only samples from the full data joint posterior distribution of $\boldsymbol{\beta}$. Therefore, (2.2) implies that we completely avoid sampling from $p(\boldsymbol{\beta}|D_{-i})$ for the computation of $K(P, P_{-i})$, and this saves us enormous computational time and effort.

Now suppose that interest lies in assessing the influence of the ith case on the subset $\boldsymbol{\beta_1}$ of the parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta_1}, \boldsymbol{\beta_2})$. Weiss and Cho (1998), Weiss (1996), and Weiss and Cook (1992) pointed out that if the goal of an analysis is to assess the influence of the ith case on the marginal posterior distribution of $\boldsymbol{\beta_1}$, then using the joint posterior of $(\boldsymbol{\beta_1}, \boldsymbol{\beta_2})$ to assess this influence may overstate the influence. Hence, in these settings, we need to consider the influence of a case using the marginal posterior distribution of $\boldsymbol{\beta_1}$.

We can express the marginal influence diagnostics of Weiss and Cho (1998) based on directed the K-L divergence as

$$K(P_1, P_{1,-i}) = \int p_1(\boldsymbol{\beta_1}|D) \log \left\{ \frac{p_1(\boldsymbol{\beta_1}|D)}{p_1(\boldsymbol{\beta_1}|D_{-i})} \right\} d\boldsymbol{\beta_1}, \tag{2.3}$$

where $p_1(\boldsymbol{\beta_1}|D) = \int p(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}|D) d\boldsymbol{\beta_2}$. The marginal K-L divergence, $K(P_1, P_{1,-i})$, in (2.3) measures the effect of deleting the ith case from the full data on the marginal posterior distribution of $\boldsymbol{\beta_1}$. Using similar derivations as in (2.2), we can obtain a simplified expression for $K(P_1, P_{1,-i})$ as follows:

$$K(P_1, P_{1,-i}) = \log E_{\boldsymbol{\beta}} \left[\frac{L(\boldsymbol{\beta}|D_{-i})}{L(\boldsymbol{\beta}|D)} \middle| D \right] - E_{\boldsymbol{\beta_1}} \left[\log \int \frac{L(\boldsymbol{\beta}|D_{-i})}{L(\boldsymbol{\beta}|D)} p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, D) d\boldsymbol{\beta_2} \middle| D \right], \tag{2.4}$$

where $p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, D)$=$p(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}|D)/\int p(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}|D)d\boldsymbol{\beta_2}$ and $\int \frac{L(\boldsymbol{\beta}|D_{-i})}{L(\boldsymbol{\beta}|D)} p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, D)d\boldsymbol{\beta_2}$ can be evaluated as $E_{\boldsymbol{\beta_2}}\left[\frac{L(\boldsymbol{\beta}|D_{-i})}{L(\boldsymbol{\beta}|D)}\middle|\boldsymbol{\beta_1}, D\right]$.

Following McCulloch (1989), calibration of $K(P, P_{-i})$ can be done by solving for p_i such that $K(P, P_{-i})$=$K(B(0.5), B(p_i))$=$-\log\{4p_i(1-p_i)\}/2$, where $B(p)$ denotes the Bernuolli distribution with success probability p. This implies that describing outcomes using $p(\boldsymbol{\beta}|D_{-i})$ instead of $p(\boldsymbol{\beta}|D)$ is compatible with describing an unobserved event as having probability p_i when the correct probability is 0.5. After calculating $K(P, P_{-i})$ from (2.2), we can compute p_i using $p_i = 0.5\left[1+\sqrt{1-\exp\{-2K(P, P_{-i})\}}\right]$. This equation implies that $0.5 \leq p_i \leq 1$. $p_i \gg 0.5$ implies that the ith case is influential, because deleting the ith case changes the posterior distribution as much as describing an observed event as having probability p_i when the correct probability is 0.5. In this paper, we use p_i as the calibration of $K(P, P_{-i})$ in all of the examples.

2.2.2 Independence Model

As an illustration, we consider the proposed diagnostic for the independence model. Suppose that given $\boldsymbol{\beta}$, y_i, $i = 1, 2, \cdots, n$ are independent response variables, not subject to censoring. Then the full data likelihood is $L(\boldsymbol{\beta}|D) = \prod_{k=1}^{n} f(y_k|\boldsymbol{\beta})$, where $f(y_k|\boldsymbol{\beta})$ is the density of y_k and the likelihood without the ith observation is $L(\boldsymbol{\beta}|D_{-i}) = \prod_{k=1, k\neq i}^{n} f(y_k|\boldsymbol{\beta})$. Therefore, $L(\boldsymbol{\beta}|D)/L(\boldsymbol{\beta}|D_{-i})$=$f(y_i|\boldsymbol{\beta})$ and the CPO is given by $CPO_i = [E_{\boldsymbol{\beta}}[\{f(y_i|\boldsymbol{\beta})\}^{-1}|D]]^{-1}$ (Gelfand et al., 1992).

Using (2.2) and the above results, we can therefore show that

$$\begin{aligned} K(P, P_{-i}) &= \log E_{\boldsymbol{\beta}}[\{f(y_i|\boldsymbol{\beta})\}^{-1}|D] + E_{\boldsymbol{\beta}}[\log\{f(y_i|\boldsymbol{\beta})\}|D] \\ &= -\log(CPO_i) + E_{\boldsymbol{\beta}}[\log\{f(y_i|\boldsymbol{\beta})\}|D]. \end{aligned} \tag{2.5}$$

Similarly, using equation (2.4) we can obtain $K(P_1, P_{1,-i})$ for the influence of the

ith case on the marginal posterior distribution of $\boldsymbol{\beta_1}$ and its connection with CPO as follows:

$$\begin{aligned} K(P_1, P_{1,-i}) &= \log E_{\boldsymbol{\beta}}[\{f(y_i|\boldsymbol{\beta})\}^{-1}|D] \\ &\quad - \int p(\boldsymbol{\beta_1}|D) \log[\int \{f(y_i|\boldsymbol{\beta})\}^{-1} p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, D) d\boldsymbol{\beta_2}] d\boldsymbol{\beta_1} \\ &= -\log(CPO_i) - E_{\boldsymbol{\beta_1}}[\log \int \{f(y_i|\boldsymbol{\beta})\}^{-1} p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, D) d\boldsymbol{\beta_2}|D], \quad (2.6) \end{aligned}$$

where $p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, D) = p(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}|D)/\int p(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}|D) d\boldsymbol{\beta_2}$ and $\int \{f(y_i|\boldsymbol{\beta})\}^{-1} p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, D) d\boldsymbol{\beta_2}$ can be evaluated as $E_{\boldsymbol{\beta_2}}[\{f(y_i|\boldsymbol{\beta})\}^{-1}|\boldsymbol{\beta_1}, D]$. Since (2.2), (2.4), (2.5) and (2.6) are expressed as posterior expectations with respect to the full data posterior distribution, they can be easily calculated using only MCMC samples from the full data posterior distribution of $\boldsymbol{\beta}$.

2.3 Cox Model with Gamma Process Prior

2.3.1 Model

In the Cox proportional hazards model (Cox, 1972), the gamma process is a very commonly used nonparametric prior process for the cumulative baseline hazard (Kalbfleisch, 1978). The full data is denoted as $D = \{\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{X}\}$, where $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$ denotes the observed survival times, where y_i may be right censored. We assume that the survival times are all distinct and ordered, i.e., $0 < y_1 < y_2 < \cdots < y_n < \infty$. $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_n)'$ is an indicator vector with $\delta_i = 1$ if the ith subject failed, and $\delta_i = 0$ if the ith subject was right censored. Also, $\boldsymbol{X}$ is an $n \times p$ matrix of covariates with ith row $\boldsymbol{x}_i'$, and $D_{-i} = \{\boldsymbol{y}_{-i}, \boldsymbol{\delta}_{-i}, \boldsymbol{X}_{-i}\}$ denotes the data with the ith subject, (i.e., $(y_i, \delta_i, \boldsymbol{x}_i')$) deleted from D. The hazard function is given by $h(y_i|\boldsymbol{x}_i) = h_0(y_i) \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown regression coefficients, and $h_0(.)$ is an unknown

baseline hazard function.

Under the Cox model, the joint probability of the survival of n subjects is given by

$$P(\boldsymbol{Y} > \boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}, H_0) = \exp\left\{-\sum_{k=1}^{n} H_0(y_k)\exp(\boldsymbol{x}_k'\boldsymbol{\beta})\right\}, \tag{2.7}$$

where $H_0(y)$ is the cumulative baseline hazard (Ibrahim et al., 2001a). We take $H_0 \sim GP(cH^*(\cdot), c)$, where GP denotes gamma process, $H^*(y)$ is a known differentiable parametric function which represents a parametric guess for the cumulative baseline hazard $H_0(y)$, and $c \geq 0$ is a confidence parameter. $H^*(y)$ is thus the mean of the process. Letting $h_k = H_0(y_k) - H_0(y_{k-1})$, we take $h_k \sim Gamma(ch_{0k}, c)$, the h_k's are independent, where $h_{0k} = H^*(y_k) - H^*(y_{k-1})$ and $Gamma(\alpha, \lambda)$ denotes the gamma distribution with mean α/λ ($\alpha > 0$ and $\lambda > 0$).

The marginal likelihood function of $\boldsymbol{\beta}$ can now be written as follows (Ibrahim et al., 2001a; Sinha et al., 2003) :

$$\begin{aligned} L(\boldsymbol{\beta}|D) &= \prod_{k=1}^{n} L_k(\boldsymbol{\beta}|D) \qquad (2.8) \\ &= \prod_{k=1}^{n} \exp\left[cH^*(y_k)\log\left\{1 - \frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_k}\right\}\right]\left[-ch^*(y_k)\log\left\{1 - \frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_k}\right\}\right]^{\delta_k}, \end{aligned}$$

where $h^*(y) = \frac{d}{dy}H^*(y)$, $A_k = \sum_{l \in \mathcal{R}(y_k)} \exp(\boldsymbol{x}_l'\boldsymbol{\beta})$, and $\mathcal{R}(y_k) = \{l : y_l \geq y_k\}$ is the set of subjects at risk at time y_k.

We now derive the likelihood function without the ith subject. If $y_k < y_i$ then the risk set at time y_k involves the ith subject, otherwise, the risk set at y_k does not involve the ith subject. Therefore, after deleting the ith subject, the risk set changes to $\mathcal{R}(y_k) = \{l : y_l \geq y_k, l \neq i\}$ for $k < i$. As the risk set changes, the corresponding A_k in the denominators of (2.8) changes to $A_k - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ for $k < i$, whereas for $k > i$, the risk set and A_k remain the same (see Appendix A for details). Hence, the likelihood

function without the ith subject is given by

$$L(\boldsymbol{\beta}|D_{-i}) = \prod_{k=1}^{i-1} L_{k,-i}(\boldsymbol{\beta}|D) \prod_{k=i+1}^{n} L_k(\boldsymbol{\beta}|D), \tag{2.9}$$

where

$$\begin{aligned} L_{k,-i}(\boldsymbol{\beta}|D) &= \exp\left[cH^*(y_k)\log\left\{1 - \frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_k - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}\right] \\ &\quad \times \left[-ch^*(y_k)\log\left\{1 - \frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_k - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}\right]^{\delta_k}, \\ L_k(\boldsymbol{\beta}|D) &= \exp\left[cH^*(y_k)\log\left\{1 - \frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_k}\right\}\right]\left[-ch^*(y_k)\log\left\{1 - \frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_k}\right\}\right]^{\delta_k}. \end{aligned}$$

The posterior distributions based on the full data and the data without the ith subject are thus given by $p(\boldsymbol{\beta}|D) \propto L(\boldsymbol{\beta}|D)\pi(\boldsymbol{\beta})$ and $p(\boldsymbol{\beta}|D_{-i}) \propto L(\boldsymbol{\beta}|D_{-i})\pi(\boldsymbol{\beta})$, respectively.

2.3.2 Diagnostic Measures

For the Cox model in general, the likelihood function cannot be written as a product of n independent terms because the risk set for the kth subject involves observations other than the kth subject. Because of this dependency, we use (2.8) for the likelihood function. Another advantage of (2.8) is its computational feasibility. Since the hazard, h_k, has been integrated out from (2.8), (2.8) is only a function of $\boldsymbol{\beta}$. Therefore, sampling the h_k's is not necessary for Bayesian inference and diagnostics, and thus only samples from the posterior distribution of $\boldsymbol{\beta}$ are needed.

After some algebra, the ratio of likelihoods for the full data and the data without the ith subject can be written as $L(\boldsymbol{\beta}|D)/L(\boldsymbol{\beta}|D_{-i}) = g_i(\boldsymbol{\beta})L_i(\boldsymbol{\beta}|D)$. Thus, we can get a simplified expression for computing the influence of the ith subject on the joint posterior distribution of $\boldsymbol{\beta}$ as follows:

$$\begin{aligned} K(P, P_{-i}) &= \log E_{\boldsymbol{\beta}}[\{g_i(\boldsymbol{\beta})L_i(\boldsymbol{\beta}|D)\}^{-1}|D] + E_{\boldsymbol{\beta}}[\log\{g_i(\boldsymbol{\beta})L_i(\boldsymbol{\beta}|D)\}|D] \qquad (2.10) \\ &= -\log(CPO_i) + E_{\boldsymbol{\beta}}[\log L_i(\boldsymbol{\beta}|D)|D] + \log[E_{\boldsymbol{\beta}}[\{g_i(\boldsymbol{\beta})\}^{-1}|D]] + E_{\boldsymbol{\beta}}[\log g_i(\boldsymbol{\beta})|D], \end{aligned}$$

where

$$L_i(\boldsymbol{\beta}|D) = \exp\left[cH^*(y_i)\log\left\{1-\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{c+A_i}\right\}\right]\left[-ch^*(y_i)\log\left\{1-\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{c+A_i}\right\}\right]^{\delta_i}, \tag{2.11}$$

$g_i(\boldsymbol{\beta}) = \prod_{k=1}^{i-1} L_k(\boldsymbol{\beta}|D)/\prod_{k=1}^{i-1} L_{k,-i}(\boldsymbol{\beta}|D)$, which can be simplified as

$$g_i(\boldsymbol{\beta}) = \frac{\prod_{k=1}^{i-1}\left[1-\frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c+A_k}\right]^{cH^*(y_k)}\left[-\log\left\{1-\frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c+A_k}\right\}\right]^{\delta_k}}{\prod_{k=1}^{i-1}\left[1-\frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c+A_k-\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right]^{cH^*(y_k)}\left[-\log\left\{1-\frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c+A_k-\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}\right]^{\delta_k}}. \tag{2.12}$$

In addition, CPO_i can be written as,

$$CPO_i = \frac{E_{\boldsymbol{\beta}}[\{g_i(\boldsymbol{\beta})\}^{-1}|D]}{E_{\boldsymbol{\beta}}[\{g_i(\boldsymbol{\beta})L_i(\boldsymbol{\beta}|D)\}^{-1}|D]}\ . \tag{2.13}$$

Since (2.10) is expressed as a posterior expectation with respect to the full data, computation of (2.10) can be done using MCMC samples from the full data posterior $p(\boldsymbol{\beta}|D)$. The samples from $p(\boldsymbol{\beta}|D)$ can be easily obtained using Adaptive Rejection Metropolis Sampling (ARMS, Gilks et al. (1995)) within Gibbs. Specifically, we have

$$K(P, P_{-i}) = \log\left[\frac{1}{J}\sum_{j=1}^{J}\{g_i(\boldsymbol{\beta}^{(j)})L_i(\boldsymbol{\beta}^{(j)}|D)\}^{-1}\right] + \frac{1}{J}\sum_{j=1}^{J}\log\{g_i(\boldsymbol{\beta}^{(j)})L_i(\boldsymbol{\beta}^{(j)}|D)\}\ , \tag{2.14}$$

and

$$CPO_i = \frac{\frac{1}{J}\sum_{j=1}^{J}\{g_i(\boldsymbol{\beta}^{(j)})\}^{-1}}{\frac{1}{J}\sum_{j=1}^{J}\{g_i(\boldsymbol{\beta}^{(j)})L_i(\boldsymbol{\beta}^{(j)}|D)\}^{-1}}\ , \tag{2.15}$$

where J is the number of Gibbs samples after burn-in and $\boldsymbol{\beta}^{(j)} = (\beta_1^{(j)}, \cdots, \beta_p^{(j)})'$ is the

jth Gibbs sample, $j = 1, \cdots, J$.

Similarly, we obtain

$$
\begin{aligned}
&K(P_1, P_{1,-i}) \\
&= \log E_{\boldsymbol{\beta}}[\{g_i(\boldsymbol{\beta})L_i(\boldsymbol{\beta}|D)\}^{-1}|D] - E_{\boldsymbol{\beta_1}}[\log \int \{g_i(\boldsymbol{\beta})L_i(\boldsymbol{\beta}|D)\}^{-1} p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, D) d\boldsymbol{\beta_2}|D] \quad (2.16) \\
&= -\log(CPO_i) + \log E_{\boldsymbol{\beta}}[\{g_i(\boldsymbol{\beta})\}^{-1}|D] - E_{\boldsymbol{\beta_1}}[\log \int \{g_i(\boldsymbol{\beta})L_i(\boldsymbol{\beta}|D)\}^{-1} p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, D) d\boldsymbol{\beta_2}|D].
\end{aligned}
$$

Monte Carlo evaluation of $E_{\boldsymbol{\beta_1}}[\log \int \{g_i(\boldsymbol{\beta})L_i(\boldsymbol{\beta}|D)\}^{-1} p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, D) d\boldsymbol{\beta_2}|D]$ in (2.16) can be obtained using the following steps:

Step 1. We use Gibbs sampling to obtain the samples $\boldsymbol{\beta}^{(j)} = (\boldsymbol{\beta_1}^{(j)}, \boldsymbol{\beta_2}^{(j)})$ for $j = 1, \cdots, J$ from $p(\boldsymbol{\beta}|D)$ and record $(\boldsymbol{\beta_1}^{(1)}, \cdots, \boldsymbol{\beta_1}^{(J)})$ as J Gibbs samples from the marginal posterior of $\boldsymbol{\beta_1}$, $p(\boldsymbol{\beta_1}|D)$.

Step 2. We use Gibbs sampling to obtain the samples $\boldsymbol{\beta}^{(r)} = (\boldsymbol{\beta_1}^{(r)}, \boldsymbol{\beta_2}^{(r)})$ for $r = 1, \cdots, R$ from $p(\boldsymbol{\beta}|D)$ and record $(\boldsymbol{\beta_2}^{(1)}, \cdots, \boldsymbol{\beta_2}^{(R)})$ as R Gibbs samples from the marginal posterior of $\boldsymbol{\beta_2}$ given $\boldsymbol{\beta_1}$, $p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, D)$.

Step 3. For each $\boldsymbol{\beta_1}^{(j)}$, use $\boldsymbol{\beta_2}^{(r)}$ as nested Gibbs samples from $p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}^{(j)}, D)$ to get the Monte Carlo approximation of $E_{\boldsymbol{\beta_1}}[\log \int \{g_i(\boldsymbol{\beta})L_i(\boldsymbol{\beta}|D)\}^{-1} p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, D) d\boldsymbol{\beta_2}|D]$ as $\frac{1}{J}\sum_{j=1}^{J} \log[\frac{1}{R}\sum_{r=1}^{R}\{g_i(\boldsymbol{\beta_1}^{(j)}, \boldsymbol{\beta_2}^{(r)}), L_i(\boldsymbol{\beta_1}^{(j)}, \boldsymbol{\beta_2}^{(r)}|D)\}^{-1}]$.

Note that the Gibbs samples in the first and second steps need to be sampled independently. Now, we can get the MCMC approximation of (2.16) as

$$
\begin{aligned}
K(P_1, P_{1,-i}) \;=\; & \log\left[\frac{1}{J}\sum_{j=1}^{J}\left\{g_i(\boldsymbol{\beta_1}^{(j)}, \boldsymbol{\beta_2}^{(j)})L_i(\boldsymbol{\beta_1}^{(j)}, \boldsymbol{\beta_2}^{(j)}|D)\right\}^{-1}\right] \qquad (2.17) \\
& -\frac{1}{J}\sum_{j=1}^{J}\log\left[\frac{1}{R}\sum_{r=1}^{R}\left\{g_i(\boldsymbol{\beta_1}^{(j)}, \boldsymbol{\beta_2}^{(r)})L_i(\boldsymbol{\beta_1}^{(j)}, \boldsymbol{\beta_2}^{(r)}|D)\right\}^{-1}\right].
\end{aligned}
$$

After computing $K(P, P_{-i})$ or $K(P_1, P_{1,-i})$ for all subjects, we can plot $K(P, P_{-i})$

or $K(P_1, P_{1,-i})$ across subjects to identify influential cases.

Since $K(P, P_{-i})$ measures the effect of deleting the ith case on the joint posterior distribution of $\boldsymbol{\beta}$, it can be viewed as a Bayesian analogue of the likelihood displacement (LD), as discussed in Cook (1986). Specifically, for the Cox model, $K(P, P_{-i})$ is comparable to the likelihood displacement based on partial likelihood, which is available in Statistical Analysis Systems (SAS) version 9.1.3. For more on likelihood displacement for the Cox model, see Pettitt and Daud (1989). In addition, a limiting expression for $K(P, P_{-i})$ based on model (2.8) provides a method for computing $K(P, P_{-i})$ under Cox's partial likelihood.

2.3.3 Relationship to Partial Likelihood

In this subsection, we derive a limiting expression for $K(P, P_{-i})$ based on model (2.8) in Section 2.3. This result provides a method for computing $K(P, P_{-i})$ under Cox's partial likelihood. Kalbfleisch (1978) and Sinha et al. (2003) showed that the partial likelihood defined by Cox (1975) can be obtained as a limiting case of the marginal posterior for $\boldsymbol{\beta}$ in the Cox model with continuous time survival data under a gamma process prior for the cumulative baseline hazard. The partial likelihood can be written as (Sinha et al., 2003)

$$\lim_{c\to 0} \frac{L(\boldsymbol{\beta}|D)}{c^{\sum_{k=1}^{n}\delta_k}(-\log c)^{\delta_n}\prod_{k=1}^{n}\{h^*(y_k)\}^{\delta_k}} = \prod_{k=1}^{n}\left\{\frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{A_k}\right\}^{\delta_k}. \tag{2.18}$$

Therefore,

$$\lim_{c\to 0} \frac{L_i(\boldsymbol{\beta}|D)}{c^{\delta_i}\{h^*(y_i)\}^{\delta_i}} = \left\{\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{A_i}\right\}^{\delta_i}, \qquad \text{for } i = 1, \cdots, n-1, \tag{2.19}$$

and

$$\lim_{c\to 0} \frac{L_n(\boldsymbol{\beta}|D)}{c^{\delta_n}(-\log c)^{\delta_n}\{h^*(y_n)\}^{\delta_n}} = \left\{\frac{\exp(\boldsymbol{x}_n'\boldsymbol{\beta})}{A_n}\right\}^{\delta_n}, \qquad \text{for } i = n. \tag{2.20}$$

Using similar ideas and extensions of the proofs of (2.18), we can derive the limiting expression of $g_i(\boldsymbol{\beta})$ as $c \to 0$. Since $A_{n-1} - \exp(\boldsymbol{x}_n'\boldsymbol{\beta}) = \exp(\boldsymbol{x}_{n-1}'\boldsymbol{\beta})$, we have

$$\lim_{c\to 0} g_i(\boldsymbol{\beta}) = \frac{\prod_{k=1}^{i-1}\left\{\frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{A_k}\right\}^{\delta_k}}{\prod_{k=1}^{i-1}\left\{\frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{A_k - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}^{\delta_k}}, \qquad \text{for } i = 1, \cdots, n-1, \tag{2.21}$$

and

$$\lim_{c\to 0}(-\log c)^{\delta_{n-1}} g_i(\boldsymbol{\beta}) = \frac{\prod_{k=1}^{i-1}\left\{\frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{A_k}\right\}^{\delta_k}}{\prod_{k=1}^{i-1}\left\{\frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{A_k - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}^{\delta_k}}, \qquad \text{for } i = n. \tag{2.22}$$

Using the above results, it follows that

$$\begin{aligned}\lim_{c\to 0} K(P, P_{-i}) &= \lim_{c\to 0}\left[\log E_{\boldsymbol{\beta}}[\{g_i(\boldsymbol{\beta})L_i(\boldsymbol{\beta}|D)\}^{-1}|D] + E_{\boldsymbol{\beta}}[\log\{g_i(\boldsymbol{\beta})L_i(\boldsymbol{\beta}|D)\}|D]\right] \quad (2.23)\\ &= \log E_{\boldsymbol{\beta}}[\lim_{c\to 0}\{\alpha_i g_i(\boldsymbol{\beta})L_i(\boldsymbol{\beta}|D)\}^{-1}|D] + E_{\boldsymbol{\beta}}[\log\{\lim_{c\to 0}\alpha_i g_i(\boldsymbol{\beta})L_i(\boldsymbol{\beta}|D)\}|D],\end{aligned}$$

where

$$\alpha_i = \begin{cases} \frac{1}{c^{\delta_i}\{h^*(y_i)\}^{\delta_i}} & \text{for } i = 1, \cdots, n-1, \\ \frac{(-\log c)^{\delta_{n-1}}}{c^{\delta_n}\{h^*(y_n)\}^{\delta_n}(-\log c)^{\delta_n}} & \text{for } i = n. \end{cases} \tag{2.24}$$

Hence, we can obtain

$$K^*(P, P_{-i}) \equiv \lim_{c\to 0} K(P, P_{-i}) = \log E_{\boldsymbol{\beta}}[\{M_i(\boldsymbol{\beta})\}^{-1}|D] + E_{\boldsymbol{\beta}}[\log\{M_i(\boldsymbol{\beta})\}|D]\ , \tag{2.25}$$

where

$$M_i(\boldsymbol{\beta}) = \frac{\prod_{k=1}^{i-1}\left\{\frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{A_k}\right\}^{\delta_k}\left\{\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{A_i}\right\}^{\delta_i}}{\prod_{k=1}^{i-1}\left\{\frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{A_k - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}^{\delta_k}}\ . \tag{2.26}$$

Thus, we can compute $K^*(P, P_{-i})$ using MCMC samples from $p(\boldsymbol{\beta}|D)$ implied by model

(2.8) in Section 2.3.

2.4 Frailty Model with Gamma Process Prior

2.4.1 Model

In survival analysis, the hazard function for each individual may depend on a set of frailties representing unobservable risk factors. In this section, we extend the results of Sinha et al. (2003) to the proportional hazards frailty model and develop Bayesian case deletion diagnostic measures.

Let y_{ij} denote the survival times and $\boldsymbol{x}_{ij}$ denotes the $p \times 1$ covariate vector for the jth subject in the ith cluster for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m_i$. The total number of subjects is $N = \sum_{i=1}^{n} m_i$ and δ_{ij} is an indicator with $\delta_{ij} = 1$ if the jth subject in the ith cluster failed and $\delta_{ij} = 0$ otherwise. The hazard function is given by $h(y_{ij}|w_i, \boldsymbol{x}_{ij}) = h_0(y_{ij}) w_i \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown regression coefficients, $h_0(.)$ is an unknown baseline hazard function, and w_i is the frailty term for the ith cluster.

To extend the results of Sinha et al. (2003), we first rearrange the data as follows. Let $D = \{\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{w}\}$ denote the complete data. We assume that the survival times, $y = (y_1, y_2, \cdots, y_N)'$, are all distinct and ordered as $0 < y_1 < y_2 < \cdots < y_N < \infty$, $\boldsymbol{X}$ is an $N \times p$ matrix of covariates with kth row $\boldsymbol{x}'_k$, $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_N)'$ is the right censoring indicator vector, and $\boldsymbol{w} = (w_{(1)}, w_{(2)}, \cdots, w_{(N)})'$ is the frailty vector according to $(y_1, y_2, \cdots, y_N)$. Thus, $h(y_k|w_{(k)}, \boldsymbol{x}_k) = h_0(y_k) w_{(k)} \exp(\boldsymbol{x}'_k\boldsymbol{\beta})$, $k = 1, 2, \cdots, N$, and assuming a gamma process prior for $H_0(y)$ as in Section 2.3.1, the likelihood function can be obtained as

$$L(\boldsymbol{\beta}|D) = \prod_{k=1}^{N} L_k(\boldsymbol{\beta}|D) \tag{2.27}$$

$$= \prod_{k=1}^{N} \exp\left[cH^*(y_k) \log\left\{1 - \frac{w_{(k)} \exp(\boldsymbol{x}'_k\boldsymbol{\beta})}{c + A_{wk}}\right\}\right] \left[-ch^*(y_k) \log\left\{1 - \frac{w_{(k)} \exp(\boldsymbol{x}'_k\boldsymbol{\beta})}{c + A_{wk}}\right\}\right]^{\delta_k},$$

where $h^*(y) = \frac{d}{dy}H^*(y)$, $A_{wk} = \sum_{l \in \mathcal{R}(y_k)} w_{(l)} \exp(\boldsymbol{x}_l'\boldsymbol{\beta})$ and $\mathcal{R}(y_k) = \{l : y_l \geq y_k\}$ is the set of subjects at risk at time y_k.

Now, we consider the data with the ith subject deleted. If the ith subject is the only observation in a cluster ($m_i = 1$), the frailty term for that cluster is deleted along with the deletion of the ith subject. Otherwise, the frailty term for the cluster remains. Therefore, we denote the data with the ith subject deleted as $D_{-i} = (\boldsymbol{y}_{-i}, \boldsymbol{\delta}_{-i}, \boldsymbol{X}_{-i}, \boldsymbol{w})$ for $m_i \geq 2$ and $D_{-i} = (\boldsymbol{y}_{-i}, \boldsymbol{\delta}_{-i}, \boldsymbol{X}_{-i}, \boldsymbol{w}_{(-i)})$ for $m_i = 1$. Furthermore, let $D_{obs} = \{\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\delta}\}$ denote the observed data and $D_{obs,-i} = \{\boldsymbol{X}_{-i}, \boldsymbol{y}_{-i}, \boldsymbol{\delta}_{-i}\}$ denote the observed data with the ith subject deleted. Because of the change in the risk set with the deletion of the ith subject, A_{wk} in the denominators of (2.27) becomes $A_{wk} - w_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ for $k < i$ and remains as A_{wk} for $k > i$. Thus,

$$
\begin{aligned}
L(\boldsymbol{\beta}|D_{-i}) &= \prod_{k=1}^{i-1} L_{k,-i}(\boldsymbol{\beta}|D) \prod_{k=i+1}^{N} L_k(\boldsymbol{\beta}|D) \\
&= \prod_{k=1}^{i-1} \exp\left[cH^*(y_k) \log\left\{1 - \frac{w_{(k)} \exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_{wk} - w_{(i)} \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}\right] \\
&\qquad \times \left[-ch^*(y_k) \log\left\{1 - \frac{w_{(k)} \exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_{wk} - w_{(i)} \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}\right]^{\delta_k} \\
&\times \prod_{k=i+1}^{N} \exp\left[cH^*(y_k) \log\left\{1 - \frac{w_{(k)} \exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_{wk}}\right\}\right] \left[-ch^*(y_k) \log\left\{1 - \frac{w_{(k)} \exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_{wk}}\right\}\right]^{\delta_k}.
\end{aligned} \tag{2.28}
$$

We assume that $\boldsymbol{\beta}$ and $\boldsymbol{w}$ are independent a priori and $\pi(\boldsymbol{w}) = \prod_{j=1}^{n} \pi(w_j)$, where the w_j's are i.i.d. gamma random variables with mean 1. The posterior distribution for the observed data is $p(\boldsymbol{\beta}, \boldsymbol{w}|D_{obs}) \propto L(\boldsymbol{\beta}|D)\pi(\boldsymbol{w})\pi(\boldsymbol{\beta})$. The posterior distribution for the observed data with the ith subject deleted is given by

$$
p(\boldsymbol{\beta}, \boldsymbol{w}|D_{obs,-i}) \propto \begin{cases} L(\boldsymbol{\beta}|D_{-i})\pi(\boldsymbol{w})\pi(\boldsymbol{\beta}) & \text{for } m_i \geq 2, \\ L(\boldsymbol{\beta}|D_{-i}) \prod_{j=1, j\neq i}^{n} \pi(w_j)\pi(\boldsymbol{\beta}) & \text{for } m_i = 1. \end{cases} \tag{2.29}
$$

2.4.2 Diagnostic Measures

For the computation of $K(P, P_{-i})$, we assume that there are at least two subjects in each cluster ($m_i \geq 2$). The influence of the ith subject on the joint posterior distribution of $\boldsymbol{\beta}$ is given by

$$K(P, P_{-i}) = \int\int p(\boldsymbol{\beta}, \boldsymbol{w}|D_{obs}) \log \left\{ \frac{p(\boldsymbol{\beta}, \boldsymbol{w}|D_{obs})}{p(\boldsymbol{\beta}, \boldsymbol{w}|D_{obs,-i})} \right\} d\boldsymbol{\beta}\, d\boldsymbol{w}. \tag{2.30}$$

If we denote the ratio of likelihoods with full data and data without the ith subject as $L(\boldsymbol{\beta}|D)/L(\boldsymbol{\beta}|D_{-i}) = g_i(\boldsymbol{\beta}, \boldsymbol{w})L_i(\boldsymbol{\beta}|D)$, $K(P, P_{-i})$ can be computed as follows:

$$\begin{aligned} K(P, P_{-i}) &= \log[E_{\boldsymbol{\beta},\boldsymbol{w}}[\{g_i(\boldsymbol{\beta}, \boldsymbol{w})L_i(\boldsymbol{\beta}|D)\}^{-1}|D_{obs}]] + E_{\boldsymbol{\beta},\boldsymbol{w}}[\log\{g_i(\boldsymbol{\beta}, \boldsymbol{w})L_i(\boldsymbol{\beta}|D)\}|D_{obs}] \\ &= -\log(CPO_i) + E_{\boldsymbol{\beta},\boldsymbol{w}}[\log L_i(\boldsymbol{\beta}|D)|D_{obs}] \\ &\quad + \log[E_{\boldsymbol{\beta},\boldsymbol{w}}[\{g_i(\boldsymbol{\beta}, \boldsymbol{w})\}^{-1}|D_{obs}]] + E_{\boldsymbol{\beta},\boldsymbol{w}}[\log g_i(\boldsymbol{\beta}, \boldsymbol{w})|D_{obs}], \end{aligned} \tag{2.31}$$

where $E_{\boldsymbol{\beta},\boldsymbol{w}}[\,\cdot\,|D_{obs}]$ is the expectation with respect to the joint posterior of $(\boldsymbol{\beta}, \boldsymbol{w})$ given D_{obs}. The corresponding $L_i(\boldsymbol{\beta}|D)$, $g_i(\boldsymbol{\beta}, \boldsymbol{w})$ and CPO_i can be written as

$$\begin{aligned} L_i(\boldsymbol{\beta}|D) = & \\ & \exp\left[cH^*(y_k)\log\left\{1 - \frac{w_{(k)}\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_{wk}}\right\}\right]\left[-ch^*(y_k)\log\left\{1 - \frac{w_{(k)}\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_{wk}}\right\}\right]^{\delta_k}, \end{aligned} \tag{2.32}$$

$$\begin{aligned} g_i(\boldsymbol{\beta}, \boldsymbol{w}) = & \\ & \frac{\displaystyle\prod_{k=1}^{i-1}\left[1 - \frac{w_{(k)}\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_{wk}}\right]^{cH^*(y_k)}\left[-\log\left\{1 - \frac{w_{(k)}\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_{wk}}\right\}\right]^{\delta_k}}{\displaystyle\prod_{k=1}^{i-1}\left[1 - \frac{w_{(k)}\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_{wk} - w_{(i)}\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right]^{cH^*(y_k)}\left[-\log\left\{1 - \frac{w_{(k)}\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_{wk} - w_{(i)}\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}\right]^{\delta_k}}, \end{aligned} \tag{2.33}$$

and

$$CPO_i = \frac{E_{\boldsymbol{\beta},\boldsymbol{w}}[\{g_i(\boldsymbol{\beta},\boldsymbol{w})\}^{-1}|D_{obs}]}{E_{\boldsymbol{\beta},\boldsymbol{w}}[\{g_i(\boldsymbol{\beta},\boldsymbol{w})L_i(\boldsymbol{\beta}|D)\}^{-1}|D_{obs}]} \ . \tag{2.34}$$

The computation of (2.31) can be accomplished using MCMC samples from $p(\boldsymbol{\beta},\boldsymbol{w}|D_{obs})$. To obtain samples from $p(\boldsymbol{\beta},\boldsymbol{w}|D_{obs})$, we perform ARMS within Gibbs using the full conditional distributions (i) $p(\boldsymbol{\beta}|\boldsymbol{w},D_{obs})$ and (ii) $p(\boldsymbol{w}|\boldsymbol{\beta},D_{obs})$.

In assessing the influence of case deletion on $\boldsymbol{\beta_1}$ of $\boldsymbol{\beta}=(\boldsymbol{\beta_1},\boldsymbol{\beta_2})$, we define

$$K(P_1,P_{1,-i}) = \int\int p_1(\boldsymbol{\beta_1},\boldsymbol{w}|D_{obs}) \log\left\{\frac{p_1(\boldsymbol{\beta_1},\boldsymbol{w}|D_{obs})}{p_1(\boldsymbol{\beta_1},\boldsymbol{w}|D_{obs,-i})}\right\} d\boldsymbol{\beta_1}\, d\boldsymbol{w}, \tag{2.35}$$

where

$$\begin{aligned} p_1(\boldsymbol{\beta_1},\boldsymbol{w}|D_{obs}) &= \int p(\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{w}|D_{obs})d\boldsymbol{\beta_2} \\ p_1(\boldsymbol{\beta_1},\boldsymbol{w}|D_{obs,-i}) &= \int p(\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{w}|D_{obs,-i})d\boldsymbol{\beta_2}. \end{aligned}$$

A computational formula for $K(P_1,P_{1,-i})$ is therefore given by

$$\begin{aligned} K(P_1,P_{1,-i}) &= \log[E_{\boldsymbol{\beta},\boldsymbol{w}}[\{g_i(\boldsymbol{\beta},\boldsymbol{w})L_i(\boldsymbol{\beta}|D)\}^{-1}|D_{obs}]] \\ &\quad -E_{\boldsymbol{\beta_1},\boldsymbol{w}}[\log\int\{g_i(\boldsymbol{\beta},\boldsymbol{w})L_i(\boldsymbol{\beta}|D)\}^{-1}p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1},\boldsymbol{w},D_{obs})d\boldsymbol{\beta_2}|D_{obs}] \\ &= -\log(CPO_i)+\log[E_{\boldsymbol{\beta},\boldsymbol{w}}[\{g_i(\boldsymbol{\beta},\boldsymbol{w})\}^{-1}|D_{obs}]] \\ &\quad -E_{\boldsymbol{\beta_1},\boldsymbol{w}}[\log\int\{g_i(\boldsymbol{\beta},\boldsymbol{w})L_i(\boldsymbol{\beta}|D)\}^{-1}p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1},\boldsymbol{w},D_{obs})d\boldsymbol{\beta_2}|D_{obs}], \end{aligned} \tag{2.36}$$

where $p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1},\boldsymbol{w},D_{obs}) = p(\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{w}|D_{obs})/\int p(\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{w}|D_{obs})d\boldsymbol{\beta_2}$. The computation of (2.36) can also be carried using MCMC samples from $p(\boldsymbol{\beta},\boldsymbol{w}|D_{obs})$.

2.4.3 Relationship to Partial Likelihood

Using similar justifications as in Sinha et al. (2003), we obtain the frailty model based on Cox's partial likelihood (Sargent, 1998) as a limiting case of the marginal posterior

distribution of $\boldsymbol{\beta}$ based on the frailty model discussed in Section 2.4.1. For the likelihood given by (2.27), we can show that

$$\lim_{c\to 0} \frac{L(\boldsymbol{\beta}, \boldsymbol{w}|D)}{c^{\sum_{k=1}^{N}\delta_k}(-\log c)^{\delta_N}\prod_{k=1}^{N}\{h^*(y_k)\}^{\delta_k}} \simeq \prod_{k=1}^{N}\left\{\frac{w_{(k)}\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{A_{wk}}\right\}^{\delta_k}. \tag{2.37}$$

We see that the right-hand side of (2.37) is equal to the frailty model based on Cox's partial likelihood (Sargent, 1998).

Using a similar method as for proving (2.25), it can be shown that

$$\lim_{c\to 0} K(P, P_{-i}) = \log E_{\boldsymbol{\beta},\boldsymbol{w}}[\{M_i(\boldsymbol{\beta}, \boldsymbol{w})\}^{-1}|D_{obs}] + E_{\boldsymbol{\beta},\boldsymbol{w}}[\log M_i(\boldsymbol{\beta}, \boldsymbol{w})|D_{obs}], \tag{2.38}$$

where

$$M_i(\boldsymbol{\beta}, \boldsymbol{w}) = \frac{\prod_{k=1}^{i-1}\left\{\frac{w_{(k)}\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{A_{wk}}\right\}^{\delta_k}\left\{\frac{w_{(i)}\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{A_{wi}}\right\}^{\delta_i}}{\prod_{k=1}^{i-1}\left\{\frac{w_{(k)}\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{A_{wk}-w_{(i)}\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}^{\delta_k}}. \tag{2.39}$$

Therefore, we can compute $K(P, P_{-i})$ for the frailty model based on partial likelihood using MCMC samples from the $p(\boldsymbol{\beta}, \boldsymbol{w}|D_{obs})$ implied by (2.27).

2.5 Cox Model with Beta Process Prior

2.5.1 Model

The actual survival time is often unknown in medical studies. However, we can obtain the information whether the subject is failed or censored in a given interval. In this case, the data is available as grouped within the intervals and called grouped survival data. We construct a finite partition of the time axis, $0 < s_1 < s_2 < \cdots < s_J$, with $s_J > y_i$ for $i = 1, 2, \cdots, n$. Thus, we have J disjoint intervals and let $I_j = (s_{j-1}, s_j]$. The observed data D is available as grouped within theses intervals. We denote $D =$

$(\boldsymbol{X}, \mathcal{R}_j, \mathcal{D}_j : j = 1, 2, \cdots, J)$ as full data, where $\mathcal{R}_j$ is the risk set and $\mathcal{D}_j$ is the failure set of the jth interval I_j. To define the data with the ith subject deleted from the full data, we assume that the ith subject is in the ath interval, $I_a = (s_{a-1}, s_a]$. And we denote $D_{-i} = (\boldsymbol{X}, \mathcal{R}_a^{-i}, \mathcal{D}_a^{-i}, \mathcal{R}_j, \mathcal{D}_j : j = 1, \cdots, a-1, a+1, \cdots, J)$ as the data without the ith subject, where $\mathcal{R}_a^{-i}$ is the risk set and $\mathcal{D}_a^{-i}$ is the failure set of the ath interval without the ith subject. We also assume that the censoring indicator for the deleted subject is known.

To model the grouped survival data, we consider the discretized beta process (Hjort, 1990; Sinha, 1997) with a grouped data likelihood (Ibrahim et al., 2001a). Let h_j be the discretized baseline hazard rate in the interval $I_j = (s_{j-1}, s_j]$, $j = 1, 2, \cdots, J$ and we specify independent beta priors for the h_j's. Specifically, we take $h_j \sim Beta(c_{0k}\alpha_{0k}, c_{0k}(1-\alpha_{0k}))$, and h_j are independent for $j = 1, 2, \cdots, J$. The likelihood is given by (Ibrahim et al., 2001a)

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{h}|D) &= \prod_{j=1}^{J} L_j(\boldsymbol{\beta}, \boldsymbol{h}|D) \\ &= \prod_{j=1}^{J} \left[\prod_{k \in \mathcal{R}_j - \mathcal{D}_j} (1-h_j)^{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})} \prod_{l \in \mathcal{D}_j} \left\{ 1 - (1-h_j)^{\exp(\boldsymbol{x}_l'\boldsymbol{\beta})} \right\} \right], \quad (2.40) \end{aligned}$$

where $\boldsymbol{h} = (h_1, h_2, \cdots, h_J)'$.

After deleting the ith subject, the risk set and failure set of the ath interval change. Since the risk set of the ath interval always contains the ith subject, it becomes $\mathcal{R}_a - \{i\text{th subject}\}$ after deleting the ith subject. On the other hand, the failure set of the ath interval contains the ith subject only if the survival time is observed from the failure (event). Therefore, the failure set becomes $\mathcal{D}_a - \{i\text{th subject}\}$, if the ith subject is failed, and it remains same as $\mathcal{D}_a$, if the ith subject is censored. Thus, the likelihood

without the ith subject (ith subject $\in I_a$) is given by

$$L(\boldsymbol{\beta}, \boldsymbol{h}|D_{-i}) = \prod_{j=1, j\neq a}^{J} L_j(\boldsymbol{\beta}, \boldsymbol{h}|D) L_a(\boldsymbol{\beta}, \boldsymbol{h}|D_{-i}) \tag{2.41}$$
$$= \prod_{j=1}^{J} L_j(\boldsymbol{\beta}, \boldsymbol{h}|D) \left[(1-\delta_i)(1-h_a)^{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})} + \delta_i \left\{1-(1-h_a)^{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}\right]^{-1},$$

where δ_i is the indicator for the ith subject having 1 for failure and 0 for censoring. $L_a(\boldsymbol{\beta}, \boldsymbol{h}|D_{-i})$ is the likelihood for the ath interval without the ith subject and given by

$$L_a(\boldsymbol{\beta}, \boldsymbol{h}|D_{-i}) = L_a(\boldsymbol{\beta}, \boldsymbol{h}|D) \left[(1-\delta_i)(1-h_a)^{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})} + \delta_i \left\{1-(1-h_a)^{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}\right]^{-1}. \tag{2.42}$$

A typical prior distribution for $\boldsymbol{\beta}$ is a $N_p(\mu_0, \Sigma_0)$, which is independent of $\boldsymbol{h}$. The posterior distributions for full data and data without the ith subject are given by

$$p(\boldsymbol{\beta}, \boldsymbol{h}|D) = L(\boldsymbol{\beta}, \boldsymbol{h}|D)\pi(\boldsymbol{h})\pi(\boldsymbol{\beta}) / \int\int L(\boldsymbol{\beta}, \boldsymbol{h}|D)\pi(\boldsymbol{h})\pi(\boldsymbol{\beta})\, d\boldsymbol{h}\, d\boldsymbol{\beta}\ ,$$

and

$$p(\boldsymbol{\beta}, \boldsymbol{h}|D_{-i}) = L(\boldsymbol{\beta}, \boldsymbol{h}|D_{-i})\pi(\boldsymbol{h})\pi(\boldsymbol{\beta}) / \int\int L(\boldsymbol{\beta}, \boldsymbol{h}|D_{-i})\pi(\boldsymbol{h})\pi(\boldsymbol{\beta})\, d\boldsymbol{h}\, d\boldsymbol{\beta}. \tag{2.43}$$

Sampling from the joint posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{h})$, $p(\boldsymbol{\beta}, \boldsymbol{h}|D)$, can be done by using transformation $q_j = -\log(1-h_j)$, $j = 1, 2, \cdots, J$ and exponential auxiliary variables (Ibrahim et al., 2001a).

2.5.2 Diagnostic Measures

We define CPO statistics for the ith subject in the ath interval with the grouped data likelihood as

$$CPO_i = p(z_i|D_{-i})|_{z_i \in I_a}\ , \quad i = 1, 2, \cdots, n, \tag{2.44}$$

where $p(z_i|D_{-i})$ is the predictive density of the ith subject given D_{-i}. We denote the ratio of likelihoods with full data and data with the ith subject deleted as $L(\boldsymbol{\beta}, \boldsymbol{h}|D)/L(\boldsymbol{\beta}, \boldsymbol{h}|D_{-i}) = g_i(\boldsymbol{\beta}, h_a)$. Specifically, $g_i(\boldsymbol{\beta}, h_a) = (1-\delta_i)(1-h_a)^{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})} + \delta_i \left\{1-(1-h_a)^{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}$. We can show that the CPO statistics for the beta process model can be computed by

$$CPO_i = \left[E_{\boldsymbol{\beta},\boldsymbol{h}}\{g_i(\boldsymbol{\beta}, h_a)^{-1}|D\}\right]^{-1}, \quad \text{for } i = 1, 2, \cdots, n. \tag{2.45}$$

For the beta process model, K-L divergence is defined by

$$K(P, P_{-i}) = \int\int p(\boldsymbol{\beta}, \boldsymbol{h}|D) \log\left\{\frac{p(\boldsymbol{\beta}, \boldsymbol{h}|D)}{p(\boldsymbol{\beta}, \boldsymbol{h}|D_{-i})}\right\} d\boldsymbol{h}\, d\boldsymbol{\beta}. \tag{2.46}$$

We can obtain computational formula for $K(P, P_{-i})$ assessing the influence of the ith subject on the joint posterior distribution and establish its connection to CPO as follows:

$$\begin{aligned} K(P, P_{-i}) &= \log E_{\boldsymbol{\beta},\boldsymbol{h}}\left[g_i(\boldsymbol{\beta}, h_a)^{-1}|D\right] + E_{\boldsymbol{\beta},\boldsymbol{h}}\left[\log g_i(\boldsymbol{\beta}, h_a)|D\right] \\ &= -\log(CPO_i) + E_{\boldsymbol{\beta},\boldsymbol{h}}\left[\log g_i(\boldsymbol{\beta}, h_a)|D\right]. \end{aligned} \tag{2.47}$$

We can also obtain computational formula for K-L divergence assessing the influence of the ith subject on the marginal posterior distribution of $\boldsymbol{\beta_1}$ and establish its

connection to CPO as follows:

$$\begin{aligned} K(P_1, P_{1,-i}) &= \log E_{\boldsymbol{\beta},\boldsymbol{h}}\left[g_i(\boldsymbol{\beta}, h_a)^{-1}|D\right] \\ &- E_{\boldsymbol{\beta_1},\boldsymbol{h}}\left[\log\left\{\int g_i(\boldsymbol{\beta}, h_a)^{-1} p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, \boldsymbol{h}, D) d\boldsymbol{\beta_2}\right\}\middle| D\right] \\ &= -\log(CPO_i) - E_{\boldsymbol{\beta_1},\boldsymbol{h}}\left[\log\left\{\int g_i(\boldsymbol{\beta}, h_a)^{-1} p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, \boldsymbol{h}, D) d\boldsymbol{\beta_2}\right\}\middle| D\right], \end{aligned} \tag{2.48}$$

where $p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, \boldsymbol{h}, D) = p(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{h}|D)/\int p(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{h}|D) d\boldsymbol{\beta_2}$.
The actual computation of equations (2.45), (2.47) and (2.48) can be done using MCMC samples from $p(\boldsymbol{\beta}, \boldsymbol{h}|D)$.

2.6 Illustrative Examples

In this section, we illustrate our methodology with simulated data and two real data sets.

2.6.1 Simulated Data

To examine the performance of the proposed diagnostics measures, we considered simulated datasets with one or more of the generated cases perturbed. The covariate x_{i1}, $i = 1, \cdots, n$, was generated from a $N(30, 4)$ distribution and standardized for numerical stability. An additional covariate, x_{i2}, was independently generated from a $Bernoulli(0.5)$ distribution. The failure time T_i was generated from an exponential distribution with hazard rate λ_i, where $\lambda_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})$ with $\beta_0 = 1, \beta_1 = -0.5$ and $\beta_2 = 2$, and the censoring time C_i was generated from an exponential distribution with $\lambda_c = 2.56$, where T_i and C_i were assumed independent. The survival times y_i, $i = 1, \cdots, 150$, were taken as $y_i = \min(T_i, C_i)$, δ_i was the censoring indicator equal to 1, if $T_i \leq C_i$, and 0, if $T_i > C_i$. In the simulated data, y_i ranged from 0.000008 to 0.8269 with median=0.0553, mean=0.1045 and standard deviation= 0.1321, whereas λ_i ranged

from 1.11 to 58.79 with median=5.97, mean=12.89 and standard deviation=13.28. The observed censoring rate was 32%.

We selected cases 10, 59 and 62 for perturbation. To create influential observations in the dataset, we choose one or two of those selected cases and perturbed the survival time (y_i), the covariate (x_{i1}), or both the survival time and the covariate of the chosen case(s). For the above perturbations, the survival time for the ith case was perturbed as $\tilde{y}_i = y_i + 5\hat{\sigma}_y$, where $\hat{\sigma}_y$ is the standard deviation of the y_i's. And the covariate x_1 for the ith case was perturbed as $\tilde{x}_{i1} = x_{i1} - 5\hat{\sigma}_{x_1}$, where $\hat{\sigma}_{x_1}$ is the standard deviation of the x_{i1}'s. After perturbing the survival time, the survival time of cases 10, 59, and 62 were changed from 0.01820 to 0.67849, 0.06854 to 0.72883, and 0.06765 to 0.72795, respectively. After perturbing x_1, the value of this covariate for cases 10, 59 and 62 was then changed from -1.39632 to -6.39632, -0.60912 to -5.60912, and -1.31305 to -6.31305, respectively. Specifically, we considered 6 different types of perturbation schemes: (I) perturbation of a survival time for a case; (II) perturbation of a covariate for a case; (III) perturbation of both a survival time and a covariate for a case; (IV) perturbation of a survival time for two cases; (V) perturbation of a covariate for two cases; (VI) perturbation of a survival time for one case and a covariate for another case. Detailed descriptions regarding the perturbations are given in Table 2.1. In Table 2.1, dataset (a) denotes the original simulated dataset with no perturbation and datasets (b)-(o) denote datasets with perturbed case(s) added by the perturbation schemes (I)-(VI).

We fit the gamma process model of Section 2.3 with an exponential $H^*(y) = 2.7y$. We chose a noninformative prior distribution for $\boldsymbol{\beta}$ as $N_2(\mathbf{0}, 10^6\mathbf{I})$. We used ARMS within Gibbs to obtain posterior samples. After burn-in, 40,000 MCMC posterior samples were used in the analysis. The proposed joint and marginal K-L divergences, $K(P, P_{-i})$ in (2.10), $K(P_1, P_{1,-i})$, $K(P_2, P_{2,-i})$ in (2.16), and calibrations of those divergences were computed for the simulated data with and without perturbation of the cases. We used p_i in Section 2.2.1 to compute the calibrations of $K(P, P_{-i})$, $K(P_1, P_{1,-i})$

TABLE 2.1: Posterior means and standard deviations for the simulated data with c=0.01

Number of perturbed case	Perturbation scheme	Dataset names	Description of perturbation	Perturbed case number	Maximum likelihood estimates β_1 MLE	β_1 SE	β_2 MLE	β_2 SE	Posterior estimates β_1 Mean	β_1 SD	β_2 Mean	β_2 SD
none	no perturbation	a	original data	none	-0.5292	0.1082	1.9945	0.2545	-0.5411	0.1077	2.0348	0.2567
1 case	I	b	survival time	10	-0.3355	0.1019	1.5003	0.2191	-0.3424	0.1020	1.5256	0.2197
		c	survival time	59	-0.4412	0.1091	1.5765	0.2219	-0.4483	0.1090	1.6038	0.2229
		d	survival time	62	-0.3563	0.1036	1.5126	0.2191	-0.3619	0.1037	1.5380	0.2194
	II	e	covariate	10	-0.3480	0.0703	1.8916	0.2490	-0.3482	0.0703	1.9275	0.2508
		f	covariate	59	-0.2630	0.0654	1.8329	0.2472	-0.2621	0.0655	1.8663	0.2483
		g	covariate	62	-0.2413	0.0610	1.8280	0.2468	-0.2390	0.0613	1.8620	0.2475
	III	h	survival time and covariate	10	-0.0086	0.0566	1.4968	0.2259	-0.0061	0.0564	1.5199	0.2274
		i	survival time and covariate	59	-0.0279	0.0609	1.4819	0.2247	-0.0251	0.0607	1.5055	0.2263
		j	survival time and covariate	62	-0.0099	0.0574	1.4943	0.2255	-0.0071	0.0571	1.5198	0.2264
2 cases	IV	k	survival time	10	-0.3196	0.1046	1.3616	0.2155	-0.3257	0.1051	1.3870	0.2162
			survival time	59								
		l	survival time	10	-0.2677	0.1021	1.3416	0.2154	-0.2727	0.1026	1.3646	0.2149
			survival time	62								
	V	m	covariate	10	-0.2401	0.0601	1.8187	0.2471	-0.2385	0.0606	1.8529	0.2479
			covariate	59								
		n	covariate	59	-0.1911	0.0568	1.7958	0.2463	-0.1899	0.0572	1.8290	0.2479
			covariate	62								
	VI	o	survival time	10	-0.1896	0.0676	1.4792	0.2209	-0.1858	0.0678	1.5034	0.2206
			covariate	62								

and $K(P_2, P_{2,-i})$. We monitored convergence of the Gibbs chain using the method proposed by Geweke (1992), as well as trace plots. We conducted sensitivity analyses using c=0.01, 0.1, 1, 10 and 100. For brevity, we present results for only the low confidence value of c=0.01. For the computation of $K(P_1, P_{1,-i})$ and $K(P_2, P_{2,-i})$, we used every 5th sample from the 40,000 MCMC posterior samples to reduce the autocorrelations and yield better convergence results.

Table 2.1 shows that the posterior inferences are sensitive to the perturbation of the selected case(s). Overall, the inferences are most sensitive to the perturbation of both the survival time and the covariate. Since we used noninformative priors on $\boldsymbol{\beta}$ and c=0.01, the posterior estimates were similar to the maximum likelihood estimates based on partial likelihood. The results regarding the diagnostics showed that $K(P, P_{-i})$, as well as $K(P_1, P_{1,-i})$ and $K(P_2, P_{2,-i})$, changed very little for the non-perturbed cases, while they changed a lot for the perturbed case(s).

The results in Table 2.2 show that before perturbation (dataset (a)), all of the selected cases are not influential, each providing a small $K(P, P_{-i})$ with its calibration close to 0.5. However, after perturbation (datasets (b) through (o)), $K(P, P_{-i})$ for those perturbed cases increased a lot and the corresponding calibrations become much larger than 0.5, indicating those cases are influential. Specifically, perturbing both the survival time and the covariate of a case increases $K(P, P_{-i})$ a lot. For example, $K(P, P_{-i})$ (and its calibration) for case 10 in dataset (h) is increased from 0.0006 (0.5168) to 5.8040 (1). We also note that the perturbed cases are similarly identified as influential using the likelihood displacement (LD) based on partial likelihood. Moreover, Figure 2.1 clearly shows that $K(P, P_{-i})$ performed well for identifying influential case(s) in each dataset providing larger $K(P, P_{-i})$ for the perturbed case(s) compared to the other cases.

Moreover, in Table 2.1, we observe that perturbing the survival time of a case had influence on the posterior estimates of both β_1 and β_2, while perturbing the covariate (x_1) of a case alone had more influence on the estimates of β, corresponding to

TABLE 2.2: Case influence diagnostics for the simulated data

Perturbation	Dataset	Case		Joint Influence			Marginal Influence		
scheme	names	number	LD	$K(P, P_{-i})$	Cal.	$K(P_1, P_{1,-i})$	Cal.	$K(P_2, P_{2,-i})$	Cal.
no perturbation	a	10	0.0009	0.0006	0.5168	0.0014	0.5266	0.0008	0.5205
		59	0.0001	0.0001	0.5067	0.0002	0.5090	0.0003	0.5118
		62	0.0204	0.0109	0.5736	0.0107	0.5727	0.0036	0.5421
I	b	10	1.9946	3.0065	0.9994	1.8175	0.9934	2.1358	0.9965
	c	59	1.1502	1.6045	0.9898	0.4879	0.8947	1.4537	0.9862
	d	62	1.9791	3.0257	0.9994	1.3214	0.9819	1.7394	0.9922
II	e	10	0.9141	1.3144	0.9816	1.4099	0.9849	0.0160	0.5901
	f	59	1.9342	2.5355	0.9984	2.5159	0.9984	0.0771	0.7042
	g	62	2.0644	2.5896	0.9986	2.4073	0.9980	0.0618	0.6814
III	h	10	3.6459	5.8040	1.0000	5.8891	1.0000	0.5759	0.9135
	i	59	3.8906	6.1829	1.0000	5.1433	1.0000	0.6485	0.9262
	j	62	3.8218	6.2595	1.0000	5.4944	1.0000	0.3322	0.8484
IV	k	10	1.1233	0.9574	0.9617	0.5498	0.9084	0.4243	0.8782
		59	0.3366	0.2051	0.7900	0.0066	0.5572	0.1890	0.7805
	l	10	0.8173	0.5998	0.9179	0.2924	0.8327	0.2824	0.8284
		62	0.7663	0.5439	0.9071	0.2448	0.8111	0.2681	0.8221
V	m	10	0.0744	0.0627	0.6717	0.0630	0.6720	0.0041	0.5452
		59	1.1271	1.0910	0.9710	1.1114	0.9722	0.0249	0.6102
	n	59	0.4191	0.2971	0.8347	0.2833	0.8288	0.0010	0.5219
		62	0.7247	0.6499	0.9264	0.6118	0.9201	0.0309	0.6262
VI	o	10	1.0668	1.2878	0.9806	0.3136	0.8413	0.9959	0.9646
		62	0.8762	1.1853	0.9761	1.1941	0.9765	0.0685	0.6789

Note that Cal. represents calibration.

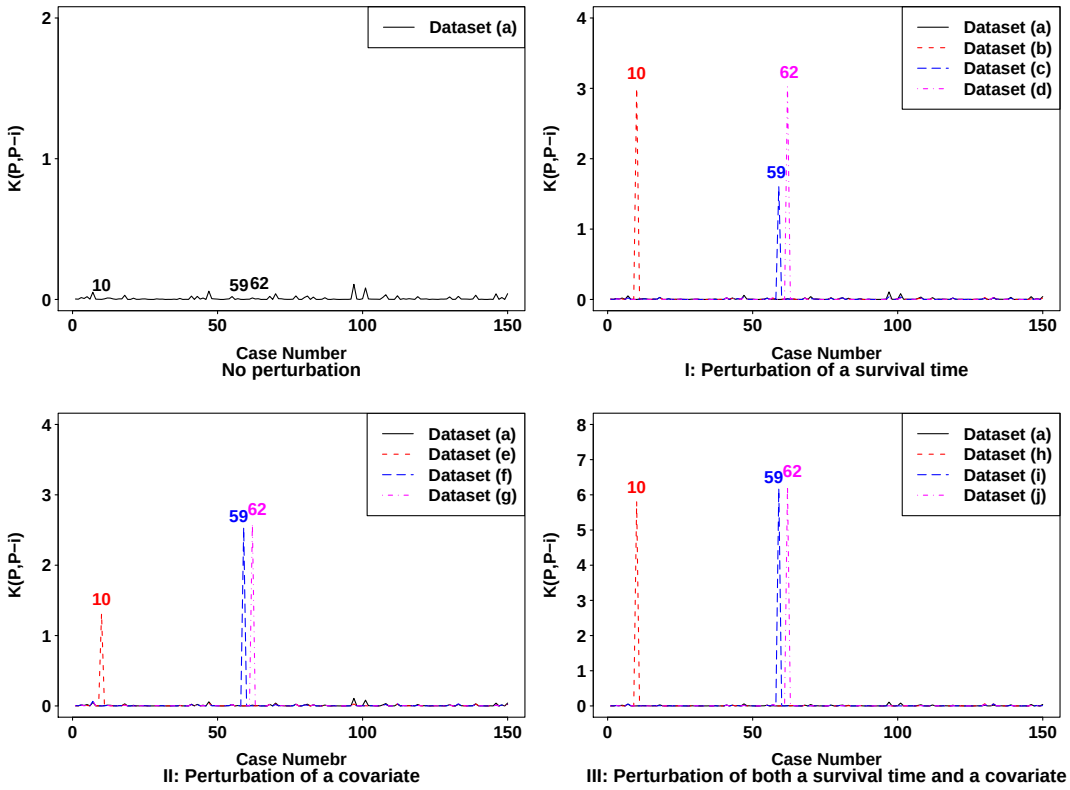

FIGURE 2.1: $K(P, P_{-i})$ for the simulated data with c=0.01.

the perturbed covariate. We see that $K(P_1, P_{1,-i})$ and $K(P_2, P_{2,-i})$ in Table 2.2 describe these marginal influences well. Specifically, both $K(P_1, P_{1,-i})$ and $K(P_2, P_{2,-i})$ increase for the perturbation of the survival time, while $K(P_1, P_{1,-i})$ increases relative to $K(P_2, P_{2,-i})$ for the perturbation of the covariate (x_1). For example, perturbing the survival time of case 62 in dataset (d) increases $K(P_1, P_{1,-i})$ and $K(P_2, P_{2,-i})$ from 0.0107 to 1.3214, and 0.0036 to 1.7394, respectively, while perturbing the covariate (x_1) of case 62 in dataset (g) increases $K(P_1, P_{1,-i})$ and $K(P_2, P_{2,-i})$ from 0.0107 to 2.4073, and 0.0036 to 0.0618, respectively.

Although there may be masking effects when there is more than one perturbed case (datasets (k) through (o)), $K(P, P_{-i})$ identifies the influential cases by providing a larger $K(P, P_{-i})$ and its calibration compared to the other cases. In addition, $K(P_1, P_{1,-i})$ and $K(P_2, P_{2,-i})$ also describe the influence of the cases on posterior inference regarding

β_1 and β_2, respectively. However, the magnitude of the measures become much smaller and the existence of an extremely influential case may mask the influence of other cases. This is not surprising since the proposed diagnostics are based on single case deletion.

Overall, the proposed joint and marginal influence diagnostic measures, $K(P, P_{-i})$, $K(P_1, P_{1,-i})$ and $K(P_2, P_{2,-i})$ performed well for identifying influential cases as well as describing the influence of a case on posterior inference.

2.6.2 Stanford Heart Transplant Data

To further illustrate the proposed methodology, we revisit the Stanford heart transplant data discussed in Section 2.1. Escobar and Meeker (1992) used 184 transplant cases to identify influential cases using an accelerated failure time lognormal regression model. We used the same dataset here with some minor modifications and identified influential cases using the proposed methodology. Of the 184 cases, 71 cases were right censored. The covariate included in this analysis was Age (x_1) (mean=41.09, and standard deviation=11.036) as well as a quadratic term of Age (x_2). Similar to Miller and Halpern (1982) and Escobar and Meeker (1992), the T5 mismatch score covariate was not used in this analysis due to its nonsignificance. For numerical stability in MCMC sampling, we standardized Age and divided survival time by 365 to make time in years instead of days.

We fit the gamma process model of Section 2.3 with $H^*(y) = 0.35y$, c=0.01 and c=100. We chose a noninformative prior distribution for $\boldsymbol{\beta} = (\beta_1, \beta_2)$ as $N_2(\mathbf{0}, 10^6\mathbf{I})$. MCMC computations were done similarly as described in Section 2.6.1, and 14,000 MCMC posterior samples were used in this analysis after burn-in. The posterior means (standard deviations) and 95% Highest Posterior Density (HPD) intervals for $\boldsymbol{\beta}$ were: For c=0.01, 0.4588 (0.1134) and (0.2404, 0.6830) for β_1, and 0.2323 (0.0841) and (0.0650, 0.3949) for β_2 ; For c=100, they were 0.3793 (0.1068) and (0.1746, 0.5916) for β_1, and 0.1117 (0.0766) and (-0.0385, 0.2606) for β_2.

TABLE 2.3: Case influence diagnostics for the heart transplant data

Case identification				c=0.01		c=100	
Patient no.	Time(days)	Status	Age	$K(P, P_{-i})$	Cal.	$K(P, P_{-i})$	Cal.
74	2006	Alive	15	0.1539	0.7573	0.1818	0.7761
159	10	Dead	13	0.0865	0.6993	0.0973	0.7102
119	1116	Alive	14	0.0743	0.6858	0.0628	0.6718
139	86	Dead	12	0.0530	0.6585	0.0871	0.6999
160	5	Dead	20	0.0307	0.6219	0.0337	0.6277
108	42	Dead	19	0.0303	0.6211	0.0359	0.6316
133	1	Dead	21	0.0270	0.6145	0.0289	0.6185

Note that Cal. represents calibration

Table 2.3 shows subjects having large $K(P, P_{-i})$ and calibration values compared to the other subjects in the dataset. For both small and large c, case 74 was identified as the most influential, having $K(P, P_{-i})$ (calibration)=0.1539 (0.7573) for c=0.01 and $K(P, P_{-i})$ (calibration)=0.1818 (0.7761) for c=100. Cases 159, 119 and 139 were also identified as influential. In addition, we identified cases 160, 108 and 133 as somewhat influential compared to other cases in the dataset for both small and large c. Figure 2.2 shows a plot of $K(P, P_{-i})$ for all the cases using c=0.01. Upon examination of these cases, it appears that these cases are influential due to low values of the covariate age, and because there were not many low age cases. Specifically, cases 159, 139, 160, 108 and 133 had small failure times in spite of their low age values. An analysis using the likelihood displacement based on partial likelihood showed that cases 74, 159, 119 and 139 were also identified as influential. In addition, our analysis identified similar cases as being influential as in Escobar and Meeker (1992), in which they identified influential cases using either case weight perturbations (patient number: 21, 74, 119, 133, 159) or response perturbations (patient number: 18, 21, 133, 139, 159) based on an accelerated failure time lognormal regression model. Although a different model than ours was

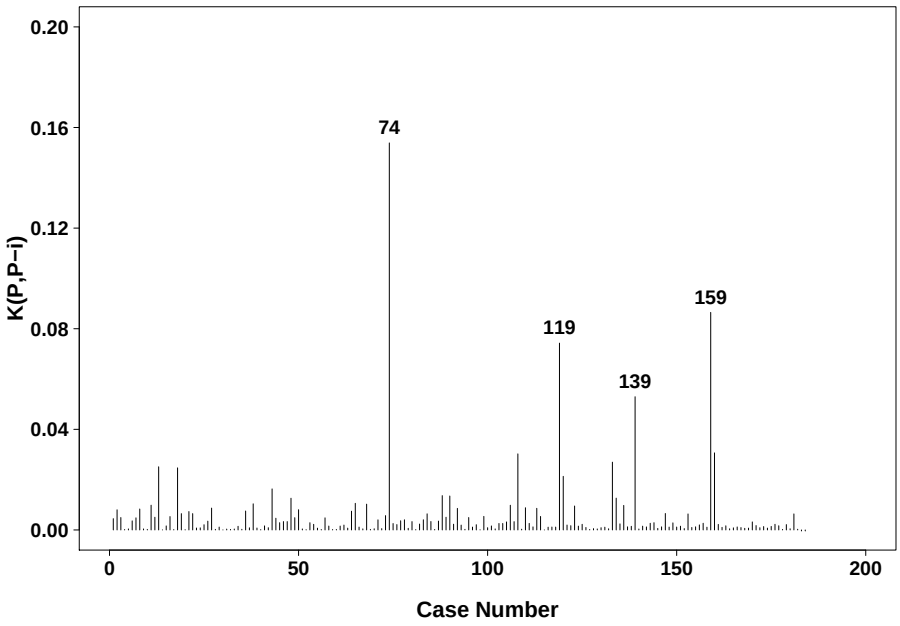

FIGURE 2.2: $K(P, P_{-i})$ for the heart transplant data with c=0.01

being fit, we used the results in Escobar and Meeker (1992) as a benchmark for the proposed Bayesian methodology to examine whether the proposed Bayesian methodology was at least consistent and yielding results in the same direction as commonly used frequentist methodology. We note that we used patient number as case number while Escobar and Meeker (1992) used case number sorted by Age.

2.6.3 Melanoma Data

As a further demonstration of the proposed methodology, we considered a phase III clinical trial conducted by the Eastern Cooperative Oncology Group (ECOG), labeled E1690 (Kirkwood et al., 2000). The trial evaluated the efficacy of interferon alfa-2b therapy on melanoma patients. The dataset used in this analysis consisted of 427 patients on the high-does interferon arm and observation arm combined. The response variable was relapse-free survival (RFS) time in years (a continuous variable, ranging from 0 to 6.9760 with mean=2.2596, and standard deviation=1.9487). Of the 427 patients, 241

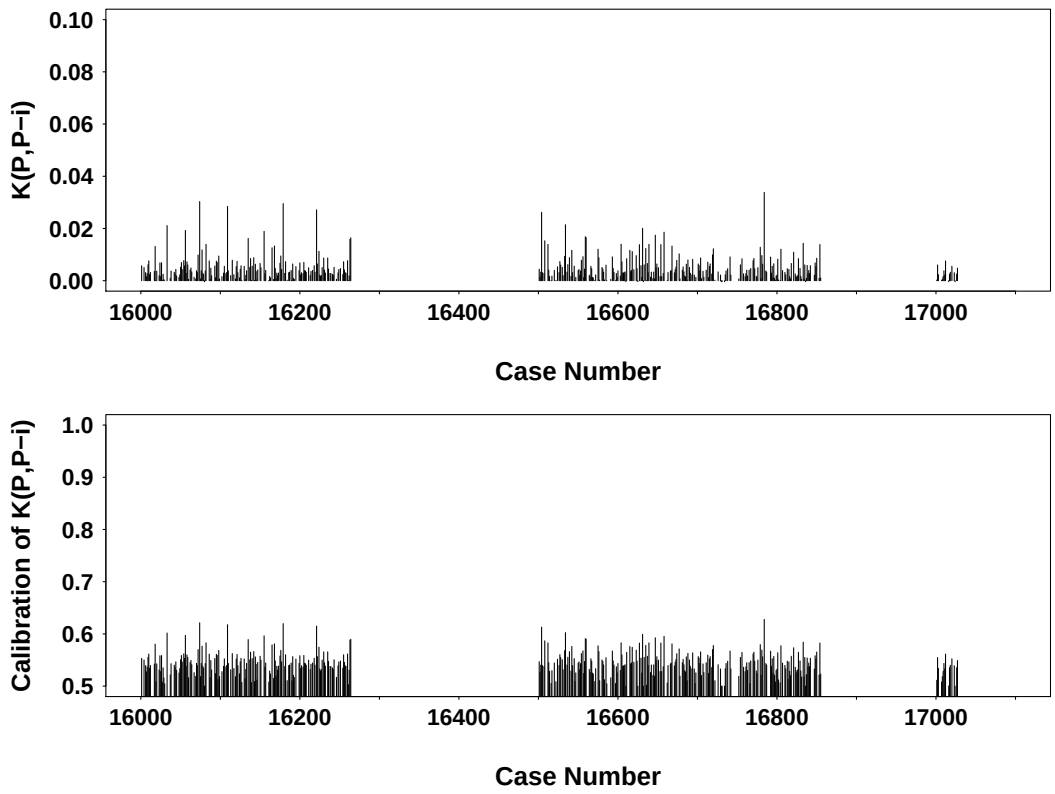

FIGURE 2.3: $K(P, P_{-i})$ and calibration for the E1690 data with c=0.01

patients experienced cancer relapse (event) and 186 patients were right censored. The covariates included in this analysis were age (a continuous variable ranging from 19.13 to 78.05 with mean=47.93, and standard deviation=13.15), treatment (215 patients on high-dose interferon arm (IFN), 212 patients on the observation arm (OBS)), sex (159 females, 268 males), and performance status (54 patients with moderate performance status, 373 patients with good performance status). For numerical stability in MCMC sampling, age was standardized. We fit the gamma process model of Section 2.3 with $H^*(y) = 0.26y$ and c=0.01. We chose a noninformative prior distribution for $\boldsymbol{\beta}$ as $N_4(\mathbf{0}, 10^6\mathbf{I})$. MCMC computations were done similarly as described in Section 2.6.1

For the E1690 data, we did not find any highly influential cases. The $K(P, P_{-i})$ was smaller than 0.034 for all cases and the corresponding calibrations were not much larger than 0.5 (Figure 2.3). However, cases 16784, 16074, 16179, 16109, 16221 and

TABLE 2.4: Case influence diagnostics for the E1690 data with c=0.01

Case				Data			Joint Influence	
number	RFS time	Status	Age	Treatment	Sex	Performance	$K(P, P_{-i})$	Cal.
16784	5.40999	Censored	64.0767	IFN	Male	Moderate	0.0338	0.6279
16074	5.10609	Censored	50.1848	OBS	Male	Moderate	0.0303	0.6213
16179	4.07666	Censored	64.0301	IFN	Male	Moderate	0.0295	0.6198
16109	5.15811	Censored	37.8617	OBS	Male	Moderate	0.0284	0.6176
16221	3.53730	Censored	57.5633	OBS	Male	Moderate	0.0271	0.6149
16504	6.58179	Censored	54.9049	OBS	Female	Moderate	0.0262	0.6130
Case				Marginal Influence				
number	$K(P_1, P_{1,-i})$	Cal.	$K(P_2, P_{2,-i})$	Cal.	$K(P_3, P_{3,-i})$	Cal.	$K(P_4, P_{4,-i})$	Cal.
16784	0.0030	0.5384	0.0047	0.5484	0.0020	0.5313	0.0235	0.6072
16074	0.0010	0.5223	0.0003	0.5131	0.0047	0.5484	0.0256	0.6118
16179	0.0024	0.5348	0.0041	0.5454	0.0019	0.5309	0.0207	0.6006
16109	0.0051	0.5502	0.0034	0.5409	0.0061	0.5551	0.0229	0.6057
16221	0.0012	0.5246	0.0017	0.5287	0.0028	0.5371	0.0217	0.6031
16504	0.0011	0.5238	0.0006	0.5175	0.0038	0.5436	0.0112	0.5744

Note that Cal. represents calibration

16504 had larger $K(P, P_{-i})$ compared to the other cases (Table 2.4). Specifically, case 16784 ($K(P, P_{-i})$=0.0338, calibration=0.6279) and case 16074 ($K(P, P_{-i})$=0.0303, calibration=0.6213) were identified as the most and the second most influential cases compared to the other cases. After an investigation as to the reason why these identified cases were more influential than others, we found that the identified cases had longer relapse free survival time (although they were censored) in spite of their large ages compared to other cases having moderate performance status. The marginal influence for the individual β_j's showed that the identified observations were more influential on posterior inference of β_4, which corresponds to the performance status covariate, compared to the other covariates (Table 2.4).

2.7 Discussion

We have proposed Bayesian case influence diagnostics using the Kullback-Leibler divergence for survival models. We have provided simple computational formulas for computing case influence on both the joint and marginal posterior distributions using MCMC techniques. We have only considered diagnostics based on single case deletion. This can be easily expanded to deletion of more than a single case or subsets of cases. In principle, this methodology can also be applied to any regression model by specifying the ratio of likelihoods with full data and data with a single case (or subset of cases) deleted. We have presented the full development for survival models here for focus and clarity of exposition.

CHAPTER 3

BAYESIAN CASE INFLUENCE MEASURES AND THEIR APPLICATIONS

3.1 Introduction

In Bayesian analysis, considerable research has been devoted to developing single case influence measures for various specific statistical models including generalized linear models, time series models, and survival models (Johnson and Geisser, 1983; Johnson, 1985; Pettit, 1986; Kass et al., 1989; Carlin and Polson, 1991; Gelfand et al., 1992; Weiss and Cook, 1992; Geisser, 1993; Blyth, 1994; Peng and Dey, 1995; Weiss, 1996; Christensen, 1997; Bradlow and Zaslavsky, 1997). Despite the extensive literature on Bayesian diagnostic measures in specific models, very little has been done on systematically examining Bayesian case influence measures in general parametric models when a small or large number of observations are deleted at a time.

The aims of this paper are to introduce three types of Bayesian case influence measures based on case deletion, namely the $\phi-$*divergence*, *Cook's posterior mode distance* and *Cook's posterior mean distance*, and to evaluate the effects of deleting a set of

observations in general parametric models including random effects models. When the number of observations in each set, denoted as N_S, is small, we will systematically derive their asymptotic approximations, which facilitate their computation and establish their asymptotic equivalence. We also propose a calibration method for evaluating their relative sizes. Moreover, we will extend these results to the case when N_S increases with sample size. In particular, we show that Cook's posterior mode and posterior mean distance have nice asymptotic properties even when $N_S \to \infty$. We will establish connections between Bayesian case influence measures, measures of Bayesian model complexity, as well as leave-k-out cross validation methods for model selection. Specifically, we will show that the sums of these proposed Bayesian case-deletion statistics are measures of model complexity, and show their asymptotic equivalence to the effective number of parameters in the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). Finally, based on the proposed measures, we construct Bayesian information criterion which can be used for model selection.

The rest of this paper is organized as follows. In Section 3.2, we introduce the three Bayesian case influence measures and propose computational formulas for computing these measures. We also examine their asymptotic properties and propose a calibration method to assess the relative sizes of these measures. Finally, we examine deleting sets of observations, whose numbers increase with sample size. In Section 3.3, we investigate applications of the proposed diagnostic measures to Bayesian model assessment. In Section 3.4, we illustrate the proposed methodologies with some Bayesian regression models. In Section 3.5, we illustrate the proposed methodology on generalized linear models and generalized linear mixed models using two real data examples involving a Los Angeles Heart Study and a clinical trial for epileptic patients. We conclude the paper with some discussion in Section 3.6.

3.2 Bayesian Case Influence Measures

3.2.1 Preliminaries

Let $p(\boldsymbol{Y}|\boldsymbol{\theta})$ be the probability function for a random vector $\boldsymbol{Y}^T = (\boldsymbol{Y}_1^T, \cdots, \boldsymbol{Y}_n^T)$, parameterized by an unknown parameter vector $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^T$ in an open subset Θ of R^p. Moreover, the dimension of $\boldsymbol{Y}_i = (y_{i1}, \cdots, y_{im_i})^T$, denoted by m_i, can vary across all i. For example, in longitudinal studies and mixed models, m_i is the number of observations in each cluster and this may vary significantly across the clusters. Let $p(\boldsymbol{\theta})$ be the prior distribution of $\boldsymbol{\theta}$. The posterior distribution for the full data $\boldsymbol{Y}$ is given by $p(\boldsymbol{\theta}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.

We are interested in assessing the influence of deleting a set of observations, denoted by S, on posterior inferences regarding $\boldsymbol{\theta}$. Let $N = \sum_{i=1}^{n} m_i$ and N_S be, respectively, the total number of observations and the number of observations in the set S. A subscript '[S]' denotes the relevant quantity with all observations in S deleted. For instance, if $S = \{i\}$, then $\boldsymbol{Y}_{[S]}$ is the corresponding observed data with all of $\boldsymbol{Y}_i$ deleted, whereas if $S = \{i_1, i_2\}$, then $\boldsymbol{Y}_{[S]}$ is the corresponding observed data with Y_{i_1} and Y_{i_2} deleted. Furthermore, we may set $S = \{i_1, \cdots, i_k\}$ and $S = \{(i_1, j_1), \cdots, (i_k, j_k)\}$ to allow more complicated case deletions. Let $\boldsymbol{Y}_S$ denote a subsample of $\boldsymbol{Y}$ consisting of all the observations in S and let $\boldsymbol{Y}_{[S]}$ denote a subsample of $\boldsymbol{Y}$ with all observations in S ($\boldsymbol{Y}_S$) deleted. The posterior distribution for a subsample of the data $\boldsymbol{Y}$ is given by $p(\boldsymbol{\theta}|\boldsymbol{Y}_{[S]}) \propto p(\boldsymbol{Y}_{[S]}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, where $p(\boldsymbol{Y}_{[S]}|\boldsymbol{\theta})$ is given by $p(\boldsymbol{Y}|\boldsymbol{\theta})/p(\boldsymbol{Y}_S|\boldsymbol{\theta})$.

Now, we introduce three types of Bayesian case influence measures based on case deletion. The first type is the $\phi-$influence of $\boldsymbol{Y}_{[S]}$, defined by

$$\mathrm{D}_\phi(S) = \int \phi(R_{[S]}(\boldsymbol{\theta}))p(\boldsymbol{\theta}|\boldsymbol{Y})d\boldsymbol{\theta}, \tag{3.1}$$

where $R_{[S]}(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{Y}_{[S]})/p(\boldsymbol{\theta}|\boldsymbol{Y})$ and $\phi(\cdot)$ is a convex function with $\phi(1) = 0$ (Weiss

and Cook, 1992; Weiss, 1996). $\mathrm{D}_\phi(S)$ directly measures the distance (discrepancy) between two posterior distributions $p(\boldsymbol{\theta}|\boldsymbol{Y}_{[S]})$ and $p(\boldsymbol{\theta}|\boldsymbol{Y})$ (Csiszár, 1967; Weiss and Cook, 1992) and a large value of $\mathrm{D}_\phi(S)$ corresponds to an influential set of observations. Various forms of $\phi(\cdot)$ have been widely considered in the literature (Kass et al., 1989; Weiss and Cook, 1992; Blyth, 1994; Peng and Dey, 1995; Weiss, 1996). For instance, $\phi(\cdot)$ can be chosen to be $\phi_\alpha(u)$, which is defined by $4\{1-u^{(1+\alpha)/2}\}/(1-\alpha^2)$ for $\alpha \neq \pm 1$, $u\log(u)$ for $\alpha = 1$, and $-\log(u)$ for $\alpha = -1$. In particular, $\phi_1(\cdot)$ and $\phi_{-1}(\cdot)$ lead to the Kullback-Leibler divergence (K-L divergence); moreover, $\phi(u) = \phi_1(u) + \phi_{-1}(u)$ leads to the symmetric K-L divergence. The L_1-distance and the χ^2-divergence correspond to $\phi(u) = 0.5|u-1|$ and $\phi(u) = (u-1)^2$, respectively (Kass et al., 1989).

The second Bayesian influence measure assesses the discrepancy between the posterior mode of $\boldsymbol{\theta}$ with and without the ith case (Cook and Weisberg, 1982). We call this measure *Cook's posterior mode distance*. Specifically, we define the posterior modes of $\boldsymbol{\theta}$ for the full sample $\boldsymbol{Y}$ and a subsample $\boldsymbol{Y}_{[S]}$ as $\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{Y})$ and $\hat{\boldsymbol{\theta}}_{[S]} = \mathrm{argmax}_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{Y}_{[S]})$, respectively. Then, Cook's posterior mode distance for comparing $\boldsymbol{Y}$ and $\boldsymbol{Y}_{[S]}$, denoted by $\mathrm{CP}(S)$, can be defined as follows:

$$\mathrm{CP}(S) = (\hat{\boldsymbol{\theta}}_{[S]} - \hat{\boldsymbol{\theta}})^T G_{\boldsymbol{\theta}} (\hat{\boldsymbol{\theta}}_{[S]} - \hat{\boldsymbol{\theta}}), \tag{3.2}$$

where $G_{\boldsymbol{\theta}}$ is chosen to be a positive definite matrix. For instance, $G_{\boldsymbol{\theta}}$ can be $-\partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{Y}) = -\partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}|\boldsymbol{\theta}) - \partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}}$, where $\partial^2_{\boldsymbol{\theta}}$ represents the second-order derivative with respect to $\boldsymbol{\theta}$. If we consider a uniform improper prior for $\boldsymbol{\theta}$, then $\mathrm{CP}(S)$ reduces to the well-known Cook's distance for deleting a set of observations (Cook and Weisberg, 1982). A large value of $\mathrm{CP}(S)$ implies more influence of the set S on the posterior mode.

The third type of Bayesian influence measure assesses the distance between the posterior mean of $\boldsymbol{\theta}$ with and without the observations in S. We define the posterior

mean of $\boldsymbol{\theta}$ for the full sample $\boldsymbol{Y}$ and a subsample $\boldsymbol{Y}_{[S]}$ as $\tilde{\boldsymbol{\theta}} = \int \boldsymbol{\theta} \cdot p(\boldsymbol{\theta}|\boldsymbol{Y})d\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}_{[S]} = \int \boldsymbol{\theta} \cdot p(\boldsymbol{\theta}|\boldsymbol{Y}_{[S]})d\boldsymbol{\theta}$, respectively. Cook's posterior mean distance for deleting the observations in the set S, denoted by $\mathrm{CM}(S)$, can then be defined as follows:

$$\mathrm{CM}(S) = (\tilde{\boldsymbol{\theta}}_{[S]} - \tilde{\boldsymbol{\theta}})^T W_{\boldsymbol{\theta}} (\tilde{\boldsymbol{\theta}}_{[S]} - \tilde{\boldsymbol{\theta}}), \tag{3.3}$$

where W_θ is chosen to be a positive definite matrix. A large value of $\mathrm{CM}(S)$ corresponds to an influential set S regarding the posterior mean.

Although all three Bayesian case influence measures assess the influence of a set of observations, there is a conceptual difference among those measures. The $\mathrm{D}_\phi(S)$ quantifies the effects of deleting a set of observations on the overall posterior distribution, whereas $\mathrm{CP}(S)$ and $\mathrm{CM}(S)$ quantify the effects of deleting a set of observations on the posterior estimates; the posterior mode and the posterior mean of $\boldsymbol{\theta}$, respectively. Since $\mathrm{D}_\phi(S)$ measures the overall differences between $p(\boldsymbol{\theta}|\boldsymbol{Y})$ and $p(\boldsymbol{\theta}|\boldsymbol{Y}_{[S]})$, and such differences may include shape, mode, mean etc., $\mathrm{D}_\phi(S)$ can be more sensitive to the changes of the posterior distributions due to the deletion of the observations in S compared to $\mathrm{CP}(S)$ and $\mathrm{CM}(S)$.

3.2.2 Computation, Approximation and Calibration

Ideally, the proposed case influence measures can all be computed using only MCMC samples from the full posterior distribution, $p(\boldsymbol{\theta}|\boldsymbol{Y})$. We define $p_S(\boldsymbol{\theta})$, the ratio of likelihoods with and without the observations in S as

$$p_S(\boldsymbol{\theta}) = \frac{p(\boldsymbol{Y}|\boldsymbol{\theta})}{p(\boldsymbol{Y}_{[S]}|\boldsymbol{\theta})} = p(\boldsymbol{Y}_S|\boldsymbol{Y}_{[S]}, \boldsymbol{\theta}), \tag{3.4}$$

which is the conditional distribution of $\boldsymbol{Y}_S$, which contains all observations in S, given $\boldsymbol{Y}_{[S]}$. Then, we have $p(\boldsymbol{\theta}|Y_{[S]}) = [p_S(\boldsymbol{\theta})]^{-1} p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) / \int [p_S(\boldsymbol{\theta})]^{-1} p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$. Thus,

the computational formula for $\mathrm{D}_\phi(S)$ can be obtained as

$$\mathrm{D}_\phi(S) = E_{\boldsymbol{\theta}|\boldsymbol{Y}}\left[\phi\left(\frac{[p_S(\boldsymbol{\theta})]^{-1}}{E_{\boldsymbol{\theta}|\boldsymbol{Y}}\{[p_S(\boldsymbol{\theta})]^{-1}\}}\right)\right], \tag{3.5}$$

where $E_{\boldsymbol{\theta}|\boldsymbol{Y}}$ denotes the expectation taken with respect to the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{Y})$. Specifically, for the K-L divergence ($\phi(u) = -\log(u)$), the computational formula is given by $\mathrm{D}_\phi(S) = \log E_{\boldsymbol{\theta}|\boldsymbol{Y}}\{[p_S(\boldsymbol{\theta})]^{-1}\} + E_{\boldsymbol{\theta}|\boldsymbol{Y}}\{\log[p_S(\boldsymbol{\theta})]\}$.

To compute $\mathrm{CP}(S)$, we need to evaluate $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_S$. In general, the posterior mode of $\boldsymbol{\theta}$ does not have a closed analytic form, thus we have to rely on iterative methods such as Newton-Raphson to obtain $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{[S]}$. However, this can be computationally intensive for most models, such as state space models. $G_{\boldsymbol{\theta}}$ in $\mathrm{CP}(S)$ can be analytically obtained by evaluating $J_N(\boldsymbol{\theta}) = -\partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{Y}) = -\partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}|\boldsymbol{\theta}) - \partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta})$ at $\hat{\boldsymbol{\theta}}$.

Since we can write $\tilde{\boldsymbol{\theta}} = E_{\boldsymbol{\theta}|\boldsymbol{Y}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\theta}}_{[S]} = E_{\boldsymbol{\theta}|\boldsymbol{Y}}\{\boldsymbol{\theta} \cdot [p_S(\boldsymbol{\theta})]^{-1}\}/E_{\boldsymbol{\theta}|\boldsymbol{Y}}\{[p_S(\boldsymbol{\theta})]^{-1}\}$, we can easily compute $\mathrm{CM}(S)$ using MCMC samples from the full posterior distribution, $p(\boldsymbol{\theta}|\boldsymbol{Y})$. Specifically, the posterior mean of $\boldsymbol{\theta}$, denoted $\tilde{\boldsymbol{\theta}}$, can be obtained directly by averaging the MCMC samples and $W_{\boldsymbol{\theta}}$ can be analytically obtained by evaluating $J_N(\boldsymbol{\theta})$ at $\tilde{\boldsymbol{\theta}}$. Furthermore, $W_{\boldsymbol{\theta}}$ as well as $G_{\boldsymbol{\theta}}$ can be approximated by the inverse of the posterior covariance matrix, obtained from the MCMC samples.

For diagnostic purposes, it is desirable to derive computationally feasible approximations to these case influence measures. We obtain the following theorems, whose detailed proofs can be found in the Appendix.

Theorem 3.1. *If Assumptions C1-C4 in the Appendix hold and N_S is bounded by a fixed constant, then we have the following results:*
(a) $D_\phi(S)$ can be approximated by

$$D_\phi(S) = 0.5\ddot{\phi}(1)[\partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})]^T [J_N(\hat{\boldsymbol{\theta}})]^{-1} [\partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})][1 + O_p(N^{-1})],$$

where $\ddot{\phi}(1) = \partial_u^2 \phi(u)|_{u=1}$.

(b) The one-step approximation for $\hat{\boldsymbol{\theta}}_{[S]}$ *is given by*

$$\hat{\boldsymbol{\theta}}_{[S]} = \hat{\boldsymbol{\theta}} + O_p(N^{-1}) = \hat{\boldsymbol{\theta}} - [J_N(\hat{\boldsymbol{\theta}})]^{-1} \partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})[1 + O_p(N^{-1})].$$

(c) The one-step approximation for $\tilde{\boldsymbol{\theta}}_{[S]}$ *is given by*

$$\tilde{\boldsymbol{\theta}}_{[S]} = \tilde{\boldsymbol{\theta}} - [J_N(\hat{\boldsymbol{\theta}})]^{-1} \partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})[1 + O_p(N^{-1})].$$

(d) $2D_\phi(S)/\ddot{\phi}(1)$*, CP*$(S)$*, and CM*$(S)$ *are asymptotically equivalent, that is,*

$$D_\phi(S) = 0.5\ddot{\phi}(1) \times CP(S) + O_p(N^{-2}) = 0.5\ddot{\phi}(1) \times CM(S) + O_p(N^{-2}).$$

Theorem 3.1 has several important implications. Theorem 3.1 (a) provides theoretical and computational approximations of $\mathrm{D}_\phi(S)$ as a quadratic form in $\partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})$. Theorem 3.1 (b) and (c) provide the one-step approximation of $\hat{\boldsymbol{\theta}}_{[S]}$ and $\tilde{\boldsymbol{\theta}}_{[S]}$, which reduces the burden of computing $\hat{\boldsymbol{\theta}}_{[S]}$ and $\tilde{\boldsymbol{\theta}}_{[S]}$ for each S. Moreover, to the best of our knowledge, Theorem 3.1 (d) is the first result that establishes a direct connection between $\mathrm{D}_\phi(S)$, $\mathrm{CP}(S)$ and $\mathrm{CM}(S)$ for any $\phi(\cdot)$ within the Bayesian framework. In particular, for $\phi_\alpha(u) = -\log(u)$, it can be shown that $\partial_u^2 \phi_\alpha(u)|_{u=1} = 1$, which leads to $\mathrm{D}_{\phi_\alpha}(S) = 0.5\mathrm{CP}(S) + O_p(N^{-2}) = 0.5\mathrm{CM}(S) + O_p(N^{-2})$ for all α. Furthermore, for the χ^2-divergence and the symmetric K-L divergence, we have $\partial_u^2 \phi(u)|_{u=1} = 2$, which gives $\mathrm{D}_\phi(S) = \mathrm{CP}(S) + O_p(N^{-2}) = \mathrm{CM}(S) + O_p(N^{-2})$. However, no simple connection exists between the three diagnostic measures based on the L_1-distance ($\phi(u) = 0.5|u-1|$), because $\ddot{\phi}(1) = 0$. Thus, $\mathrm{D}_\phi(S) = O_p(N^{-2})$ for the L_1-distance.

According to Theorem 3.1, to approximate these case influence measures, we only need to compute the posterior mean $\tilde{\boldsymbol{\theta}}$, the observed-data information matrix $J_N(\tilde{\boldsymbol{\theta}})$,

$\partial_{\boldsymbol{\theta}} \log p_S(\boldsymbol{\theta})$ evaluated at $\tilde{\boldsymbol{\theta}}$, and

$$\mathrm{AP}(S; \tilde{\boldsymbol{\theta}}) = [\partial_{\boldsymbol{\theta}} \log p_S(\tilde{\boldsymbol{\theta}})]^T [J_N(\tilde{\boldsymbol{\theta}})]^{-1} [\partial_{\boldsymbol{\theta}} \log p_S(\tilde{\boldsymbol{\theta}})]. \tag{3.6}$$

In particular, $\tilde{\boldsymbol{\theta}}$ and $J_N(\tilde{\boldsymbol{\theta}})$ can be easily computed from the MCMC samples. For most statistical models, the computation of $\partial_{\boldsymbol{\theta}} \log p_S(\boldsymbol{\theta}) = \partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}|\boldsymbol{\theta}) - \partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_{[S]}|\boldsymbol{\theta})$ is relatively straightforward. Here, we use the fact that the posterior mean and posterior mode are asymptotically equivalent under suitable regularity conditions which are satisfied for the regression models considered here.

To calibrate these case influence measures, we use posterior predictive p-value (Gelman et al., 1996, 2003), which is defined as the probability that replicate data could be more extreme than the observed data, as measured by the case influence measures. We present the five key steps to compute the posterior predictive p-value for $\mathrm{AP}(S; \tilde{\boldsymbol{\theta}})$ in (3.6).

Step 1. Using the observed data $\boldsymbol{Y}$, we obtain the MCMC sample $\boldsymbol{\theta}^{(j)}$ for $j = 1, \cdots, J$ from $p(\boldsymbol{Y}|\boldsymbol{\theta})$.

Step 2. For given $\boldsymbol{\theta}^{(j)}$, we compute $\mathrm{AP}(S; \boldsymbol{\theta}^{(j)})$ based on the observed data, and denote it by $\mathrm{AP}^{obs}(S; \boldsymbol{\theta}^{(j)})$ for $j = 1, \cdots, J$.

Step 3. We draw one replicate, denoted by $\boldsymbol{Y}^{rep,(j)}$, from the distribution $p(\boldsymbol{Y}^{rep}|\boldsymbol{\theta}^{(j)})$ for each $\boldsymbol{\theta}^{(j)}$ and compute $\mathrm{AP}(S; \boldsymbol{\theta}^{(j)})$ based on the replicate data, denote by $\mathrm{AP}^{rep}(S; \boldsymbol{\theta}^{(j)})$ for $j = 1, \cdots, J$.

Step 4. We repeat Step 3 M times, and therefore we have $\mathrm{AP}^{rep,(m)}(S; \boldsymbol{\theta}^{(j)})$ for $m = 1, \cdots, M$ and $j = 1, \cdots, J$.

Step 5. The Monte Carlo approximation of the posterior predictive p-value can be

obtained as

$$\hat{p}_{AP}(S) = \frac{1}{J}\frac{1}{M}\sum_{j=1}^{J}\sum_{m=1}^{M} I(\mathrm{AP}^{rep,(m)}(S;\boldsymbol{\theta}^{(j)}) \geq \mathrm{AP}^{obs}(S;\boldsymbol{\theta}^{(j)})), \tag{3.7}$$

where $I(\cdot)$ is an indicator function.

Note that $\hat{p}_{AP}(S)$ is not a regular p-value. A small value of $\hat{p}_{AP}(S)$ corresponds to an influential set S in the data for the given model.

3.2.3 Deleting Large Numbers of Observations

Although we have systematically examined the deletion of a bounded number of observations, there are some applications that require deleting relatively large numbers of observations. For instance, for clustered data, we may be interested in deleting all observations in some clusters, whose numbers may be comparable with the total number of clusters n. That is, $N_S \to \infty$. Moreover, the idea of multifold cross validation (Geisser, 1975; Zhang, 1993) requires deleting a large number of observations at a time. We obtain the following theorem.

Theorem 3.2. *If Assumptions C1, C2, C3', and C4 in the Appendix hold and* $N_S \to \infty$ *and* $N_S/N \to \gamma \in [0,1)$*, then we have the following results:*
(a) The one-step approximation for $\hat{\boldsymbol{\theta}}_{[S]}$ *is given by*

$$\hat{\boldsymbol{\theta}}_{[S]} = \hat{\boldsymbol{\theta}} + O_p(N^{-1/2}) = \hat{\boldsymbol{\theta}} - [J_{N,[S]}(\hat{\boldsymbol{\theta}})]^{-1}\partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})[1 + O_p(N^{-1/2})],$$

where $J_{N,[S]}(\boldsymbol{\theta}) = -\partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{Y}_{[S]})$*. If* $\gamma = 0$*, then*

$$\hat{\boldsymbol{\theta}}_{[S]} = \hat{\boldsymbol{\theta}} - [J_N(\hat{\boldsymbol{\theta}})]^{-1}\partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})[1 + O_p(N^{-1/2}) + O_p(N_S/N)]. \tag{3.8}$$

(b) The one-step approximation for $\tilde{\boldsymbol{\theta}}_{[S]}$ is given by

$$\tilde{\boldsymbol{\theta}}_{[S]} - \tilde{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_{[S]} - \hat{\boldsymbol{\theta}})[1 + o_p(1)].$$

(c) CP(S) and CM(S) can be asymptotically approximated by $[\partial_{\boldsymbol{\theta}} \log p_S(\tilde{\boldsymbol{\theta}})]^T \cdot [J_{N,[S]}(\tilde{\boldsymbol{\theta}})]^{-1}[J_N(\tilde{\boldsymbol{\theta}})][J_{N,[S]}(\tilde{\boldsymbol{\theta}})]^{-1}[\partial_{\boldsymbol{\theta}} \log p_S(\tilde{\boldsymbol{\theta}})]$. *If $\gamma = 0$, then*

$$CP(S) = CM(S)[1 + o_p(1)] = AP(S; \tilde{\boldsymbol{\theta}})[1 + o_p(1)]. \tag{3.9}$$

(d) $D_\phi(S)$ can be approximated by

$$D_\phi(S) = \phi(A_S) + O_p(N^{-1}),$$

where $A_S = \sigma \times p(\boldsymbol{Y}_{[S]}|\hat{\boldsymbol{\theta}})p(\hat{\boldsymbol{\theta}})/[\sigma_{[S]} \times p(\boldsymbol{Y}_{[S]}|\hat{\boldsymbol{\theta}}_S)p(\hat{\boldsymbol{\theta}}_S)]$, $\sigma^2 = [J_N(\hat{\boldsymbol{\theta}})/N]^{-1}$ *and* $\sigma^2_{[S]} = [J_{N,[S]}(\hat{\boldsymbol{\theta}}_S)/N]^{-1}$.

Theorem 3.2 has several important implications. Theorem 3.2 (a) and (b) provide the one-step approximations of $\hat{\boldsymbol{\theta}}_{[S]}$ and $\tilde{\boldsymbol{\theta}}_{[S]}$, which reduce the burden of computing $\hat{\boldsymbol{\theta}}_{[S]}$ and $\tilde{\boldsymbol{\theta}}_{[S]}$ for each S. Theorem 3.2 (c) provides theoretical and computational approximations of CP(S) and CM(S). If $N_S/N \to 0$, such as $N_S = \sqrt{N}$, then CP(S) and CM(S) can be well approximated by AP($S; \tilde{\boldsymbol{\theta}}$). Theorem 3.2 (d) shows that when $N_S \to \infty$, $\mathrm{D}_\phi(S)$, which can be approximated by $\phi(A_S)$, is not asymptotically equivalent to AP($S; \tilde{\boldsymbol{\theta}}$) in any case. Therefore, we cannot use AP($S; \tilde{\boldsymbol{\theta}}$) to characterize the asymptotic behavior of $\mathrm{D}_\phi(S)$. Since calculating $D_\phi(S)$ and $p(\boldsymbol{Y}_{[S]}|\boldsymbol{\theta})$ in A_S can be computationally tedious compared with CM(S) and AP($S; \tilde{\boldsymbol{\theta}}$) in many models, such as random effects models with or without missing data, we generally suggest using the Bayesian case influence measures CP(S) and CM(S) for diagnostic purposes.

3.3 Applications to Model Assessment

3.3.1 Model Complexity and Cross Validation

The three proposed Bayesian case influence measures are also associated with Bayesian measures of model complexity and fit. Specifically, we consider a Bayesian measure of model complexity (BMMC) based on the sum of $\mathrm{AP}(S;\tilde{\boldsymbol{\theta}})$ as follows:

$$\mathrm{MC}(I_S) = \sum_{S \in I_S} \mathrm{AP}(S;\tilde{\boldsymbol{\theta}}), \tag{3.10}$$

where I_S denotes all possible sets sharing the same pattern as the set S. For instance, if we consider single cluster deletion, that is $S = \{i\}$, then $I_S = \{\{1\},\cdots,\{n\}\}$. If we consider single observation deletion, that is $S = \{(i,j)\}$, then $I_S = \{\{(1,1)\},\cdots,\{(1,m_1)\},\cdots,\{(n,m_n)\}\}$. Similarly, we can define other BMMCs based on $\mathrm{D}_\phi(S)$, $\mathrm{CP}(S)$, and $\mathrm{CM}(S)$, which are asymptotically equivalent to $\mathrm{MC}(I_S)$.

An interesting question is how $\mathrm{MC}(I_S)$ is related to both classical and Bayesian measures of model complexity. The following theorem ensures that $\mathrm{MC}(I_S)$ is asymptotically equivalent to the effective number of parameters in other information criteria such as the Akaike information criterion (AIC) (Akaike, 1973), Takeuchi's information criterion (TIC) (Takeuchi, 1976) and DIC. Using the one-step approximation for $\tilde{\boldsymbol{\theta}}_{[S]}$ in Theorem 3.1, we are led to the following Theorem on the consistency of $\mathrm{MC}(I_S)$.

Theorem 3.3. *Suppose that Assumptions C1, C2, C5 and C6 in the Appendix hold and N_S is bounded by a fixed constant. Then, we have the following results:*
(a) (Consistency) Let N_{I_S} be the number of sets in I_S. Then

$$\frac{N}{N_{I_S}} MC(I_S) = \frac{N}{N_{I_S}} tr\{[J_N(\tilde{\boldsymbol{\theta}})]^{-1} K_N(I_S|\tilde{\boldsymbol{\theta}})\} = tr[J_*^{-1} K_*(I_S)] + o_p(1), \tag{3.11}$$

where $J_N(\tilde{\boldsymbol{\theta}}) = -\partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{Y})|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$ *and* $K_N(I_S|\tilde{\boldsymbol{\theta}}) = \sum_{S\in I_S}[\partial_{\boldsymbol{\theta}} \log p_S(\boldsymbol{\theta})]^{\otimes 2}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$. *Moreover,*

$$J_* = \lim_{N\to\infty} \frac{E[-\partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}_*|\boldsymbol{Y})]}{N} \quad \textit{and} \quad K_*(I_S) = \lim_{N_{I_S}\to\infty} \frac{E\{\sum_{S\in I_S}[\partial_{\boldsymbol{\theta}} \log p_S(\boldsymbol{\theta}_*)]^{\otimes 2}\}}{N_{I_S}}, \tag{3.12}$$

where $\boldsymbol{\theta}_*$ *denotes the pseudo-true parameter (Bunke and Milhaud, 1998) and* $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^T$ *for any vector* $\boldsymbol{a}$.

(b) (Asymptotic Normality) $\sqrt{N} \times SMC(I_S)$ *converges to a* $N(0, \sigma_S^2)$ *distribution, where* $SMC(I_S) = NN_{I_S}^{-1}\Big(MC(I_S) - tr\{[J_N(\tilde{\boldsymbol{\theta}})]^{-1}E[K_N(I_S|\tilde{\boldsymbol{\theta}})]\}\Big)$.

Theorem 3.3 has several important implications. First, we consider single cluster deletion $I_S = \{\{1\}, \cdots, \{n\}\}$ and examine the relationship of $\mathrm{MC}(I_S)$ with other model complexity measures based on clustered data, in which the $\boldsymbol{Y}_i$ are independent for different i, but the components in each $\boldsymbol{Y}_i$ may be correlated. In this case, we have $N_{I_S} = n$, $K_N(I_S|\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^n \{\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_i|\boldsymbol{\theta})\}^{\otimes 2}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$, $K_*(I_S) = \lim_{n\to\infty} n^{-1}E[\sum_{i=1}^n \{\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_i|\boldsymbol{\theta}_*)\}^{\otimes 2}]$, $J_N(\tilde{\boldsymbol{\theta}}) = -[\sum_{i=1}^n \partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_i|\boldsymbol{\theta}) + \partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta})]|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$ and $J_* = \lim_{N\to\infty} N^{-1}E[-\{\sum_{i=1}^n \partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_i|\boldsymbol{\theta}_*) + \partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}_*)\}]$. Let $p_* = \mathrm{tr}[J_*^{-1}K_*(I_S)]$. Using a uniform improper prior for $\boldsymbol{\theta}$, p_* is the measure of model complexity in TIC. Furthermore, if the model $p(\boldsymbol{Y}|\boldsymbol{\theta})$ is correctly specified, then p_* reduces to p, the number of parameters, and $\mathrm{MC}(I_S) = p + o_p(1)$. In this case, p is the measure of model complexity in AIC. For general priors, p_* is the effective number of parameters in the network information criterion (NIC) (Murata et al., 1994; Ripley, 1996). Moreover, $\mathrm{MC}(I_S)$ is also associated with the effective number of parameters, denoted by p_D, in DIC, where $p_D = E_{\boldsymbol{\theta}|\boldsymbol{Y}}[-2\log p(\boldsymbol{Y}|\boldsymbol{\theta})] + 2\log[p(\boldsymbol{Y}|\tilde{\boldsymbol{\theta}})]$. Under the two conditions of approximately normal likelihoods and a uniform improper prior for $\boldsymbol{\theta}$, it can be shown that $p_D = \mathrm{tr}\{J_N(\tilde{\boldsymbol{\theta}})E[(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\otimes 2}]\} + o_p(1)$ (Spiegelhalter et al., 2002). Moreover, using the fact that $E[(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\otimes 2}] = J_N(\boldsymbol{\theta}_*)^{-1}K_N(I_S|\boldsymbol{\theta}_*)J_N(\boldsymbol{\theta}_*)^{-1}[1 + o_p(1)]$ (Bunke and Milhaud, 1998), we can obtain the following connections between p_D and $\mathrm{MC}(I_S)$:

$p_D = \mathrm{MC}(I_S) + o_p(1)$. Thus, $\mathrm{MC}(I_S)$ has many of the same properties as p_D (Spiegelhalter et al., 2002). We also note that $\mathrm{MC}(I_S)$ is always nonnegative, whereas p_D is not. Finally, we can apply the Lindeberg-Fellner central limit theorem to establish the asymptotic normality of $\mathrm{SMC}(I_S)$ by noting that

$$\mathrm{SMC}(I_S) = \frac{1}{\sqrt{\sum_{i=1}^n m_i}} \mathrm{tr}\left([J_N(\tilde{\boldsymbol{\theta}})]^{-1} \{ [\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_i|\boldsymbol{\theta})]^{\otimes 2} - E[\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_i|\boldsymbol{\theta})]^{\otimes 2} \} \right).$$

Second, we consider single observation deletion $I_S = \{\{(1,1)\}, \cdots, \{(n, m_n)\}\}$ and examine $\mathrm{MC}(I_S)$ for clustered data. We have $N_{I_S} = N = \sum_{i=1}^n m_i$ and

$$\partial_{\boldsymbol{\theta}} \log p_{[(i,j)]}(\boldsymbol{\theta}) = \partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_i|\boldsymbol{\theta}) - \partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_{i,[(i,j)]}|\boldsymbol{\theta}), \tag{3.13}$$

where $\boldsymbol{Y}_{i,[(i,j)]}$ denotes $\boldsymbol{Y}_i$ with $y_{i,j}$ deleted. It can be shown that

$$\begin{aligned} K_N(I_S|\tilde{\boldsymbol{\theta}}) &= \sum_{i=1}^n m_i \{\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_i|\tilde{\boldsymbol{\theta}})\}^{\otimes 2} \\ &- \sum_{i=1}^n \partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_i|\tilde{\boldsymbol{\theta}}) \{\sum_{j=1}^{m_i} \partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_{i,[i,j]}|\tilde{\boldsymbol{\theta}})\}^T \\ &- \sum_{i=1}^n \{\sum_{j=1}^{m_i} \partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_{i,[i,j]}|\tilde{\boldsymbol{\theta}})\} [\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_i|\tilde{\boldsymbol{\theta}})]^T \\ &+ \sum_{i=1}^n \sum_{j=1}^{m_i} \{\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_{i,[i,j]}|\tilde{\boldsymbol{\theta}})\}^{\otimes 2}. \end{aligned} \tag{3.14}$$

Moreover, $p_* = \mathrm{tr}[J_*^{-1} K_*(I_S)]$ can be regarded as the measure of model complexity for clustered data. Even if the model $p(\boldsymbol{Y}|\boldsymbol{\theta})$ is correctly specified, p_* does not reduce to p, the number of parameters, and $\mathrm{MC}(I_S) \neq p + o_p(1)$. Compared with p as the measure of model complexity in AIC, $p_* = \mathrm{tr}[J_*^{-1} K_*(I_S)]$ accounts for the correlation structure in the clustered data. Although one may consider other case deletion mechanisms, we omit them here for the brevity and the sake of space.

The posterior predictive p-value of $\mathrm{MC}(I_S)$ can be computed in a similar fashion as

described in Section 3.2.2. Thus, we obtain

$$\hat{p}_{MC}(S) = \frac{1}{J}\frac{1}{M}\sum_{j=1}^{J}\sum_{m=1}^{M} I(\mathrm{MC}^{rep,(m)}(I_S)^{(j)} \geq \mathrm{MC}^{obs}(I_S)^{(j)}),$$

where $\mathrm{MC}^{obs}(I_S)^{(j)}=\sum_{S\in I_S}\mathrm{AP}^{obs}(S;\boldsymbol{\theta}^{(j)})$ and $\mathrm{MC}^{rep,(m)}(I_S)^{(j)}=\sum_{S\in I_S}\mathrm{AP}^{rep,(m)}(S;\boldsymbol{\theta}^{(j)})$. A small value of $\hat{p}_{MC}(S)$ corresponds to a potentially misspecified model $p(\boldsymbol{Y}|\boldsymbol{\theta})$.

$\mathrm{MC}(I_S)$ is also associated with the leave-k-out cross validation method (Stone, 1974, 1977, 2002; Geisser and Eddy, 1979). The cross-validation method usually divides the data into two subsamples: a training sample and a validation sample. The training sample is used for model fitting and the validation sample is used to assess model fit. For a given S, $\tilde{\boldsymbol{\theta}}_{[S]}$ is estimated from the training sample $\boldsymbol{Y}_{[S]}$, whereas we use the predictive distribution $p(\tilde{\boldsymbol{Y}}_S|\boldsymbol{Y}_{[S]})$ for model validation, where $\tilde{\boldsymbol{Y}}_S$ is an independent copies of $\boldsymbol{Y}_S$. One choice of the predictive distribution is to use $p(\tilde{\boldsymbol{Y}}_S|\boldsymbol{Y}_{[S]},\boldsymbol{\theta})$. By substituting $\boldsymbol{Y}_S$ and $\tilde{\boldsymbol{\theta}}_{[S]}$ into $p(\tilde{\boldsymbol{Y}}_S|\boldsymbol{Y}_{[S]},\boldsymbol{\theta})$, we can define the deleting-k multifold cross validation criterion as

$$\mathrm{MCV}(I_S) = \sum_{S\in I_S}\log p(\boldsymbol{Y}_S|\boldsymbol{Y}_{[S]},\tilde{\boldsymbol{\theta}}_{[S]}) = \sum_{S\in I_S}\log p_S(\tilde{\boldsymbol{\theta}}_{[S]}). \tag{3.15}$$

We now obtain the following result.

Theorem 3.4. *Suppose that Assumptions C1-C4 in the Appendix hold and N_S is bounded by a fixed constant. Then we have $MCV(I_S) = \sum_{S\in I_S}\log p_S(\tilde{\boldsymbol{\theta}}) - MC(I_S)[1 + o_p(1)]$.*

Theorem 3.4 establishes the connection between $\mathrm{MCV}(I_S)$ and $\mathrm{MC}(I_S)$. If we consider single cluster deletion for clustered data, then we have $\sum_{S\in I_S}\log p_S(\boldsymbol{\theta}) = \sum_{i=1}^{n}\log p(\boldsymbol{Y}_i|\boldsymbol{\theta}) = \log p(\boldsymbol{Y}|\boldsymbol{\theta})$. Thus, based on the previous discussion about BMMC, $\mathrm{MCV}(I_S)$ is also closely related to AIC, TIC and NIC.

3.3.2 Model Comparison Criterion

Since $\mathrm{MC}(I_S)$ measures the complexity of a fitted model, we can construct a Bayesian information model selection criterion based on $\mathrm{MC}(I_S)$. Let $\boldsymbol{Z} = (Z_1, \cdots, Z_n)$ be an independent copy of $\boldsymbol{Y}$. We consider the following quantity:

$$\eta = n^{-1} E_{\boldsymbol{Z}} E_{\boldsymbol{\theta}|\boldsymbol{Y}} \{\log p(\boldsymbol{Z}|\boldsymbol{\theta})\} = n^{-1} \int \left\{ \int \log p(\boldsymbol{Z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{Y}) d\boldsymbol{\theta} \right\} g(\boldsymbol{Z}) d\boldsymbol{Z},$$

where $g(\boldsymbol{Z})$ is the true distribution of $\boldsymbol{Z}$. Here, η is an extension of the predictive discrepancy measure in Ando (2007) for dependent data. When $Z_1, \cdots, Z_n$ are independent, then η is an average of the predictive discrepancy measure considered by (Ando, 2007).

Following the development of BPIC in Ando (2007), we find the bias corrected estimator of η to select an optimal model. We set $\boldsymbol{Z} = \boldsymbol{Y}$ and obtain an estimate of η as

$$\hat{\eta} = n^{-1} E_{\boldsymbol{\theta}|\boldsymbol{Y}} \{\log p(\boldsymbol{Y}|\boldsymbol{\theta})\} = n^{-1} \int \log p(\boldsymbol{Y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{Y}) d\boldsymbol{\theta}.$$

Because the same data $\boldsymbol{Y}$ are used to construct $p(\boldsymbol{Y}|\boldsymbol{\theta})$ and to evaluate η, we consider the bias of $\hat{\eta}$ relative to η, denoted by B_η, as follows:

$$B_\eta = E_{\boldsymbol{Y}}(\hat{\eta} - \eta) = n^{-1} \int \left[E_{\boldsymbol{\theta}|\boldsymbol{Y}} \{\log p(\boldsymbol{Y}|\boldsymbol{\theta})\} - E_{\boldsymbol{Z}} [E_{\boldsymbol{\theta}|\boldsymbol{Y}} \{\log p(\boldsymbol{Z}|\boldsymbol{\theta})\}] \right] g(\boldsymbol{Y}) d\boldsymbol{Y}.$$

If a consistent estimate of B_η, denoted by $\hat{B}_\eta$, exists, then a bias-corrected estimator of η is given by $\hat{\eta} - \hat{B}_\eta$. Thus, a Bayesian information criterion can be constructed as $\mathrm{IC}_B = E_{\boldsymbol{\theta}|\boldsymbol{Y}} \{-2 \log p(\boldsymbol{Y}|\boldsymbol{\theta})\} + 2n\hat{B}_\eta$. In particular, as shown in Theorem 3.5, the model complexity $\mathrm{MC}(I_S)$ is indeed a consistent estimator of nB_η. For ease of exposition, we assume $m_i = 1$, for $i = 1, \cdots, n$.

Theorem 3.5. *Suppose that Assumptions C1-C4 in the Appendix hold. Then, we have*

$$\begin{aligned} nB_\eta &= E_{\boldsymbol{Y}}\left[E_{\boldsymbol{\theta}|\boldsymbol{Y}}[\log\{p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\}] - \log\{p(\boldsymbol{Y}|\hat{\boldsymbol{\theta}})p(\hat{\boldsymbol{\theta}})\}\right] \\ &\quad + tr\{J_n^{-1}(\hat{\boldsymbol{\theta}})\tilde{K}_n(I_S|\hat{\boldsymbol{\theta}})\} + p/2 + o_p(1) \\ &= tr\{J_n^{-1}(\hat{\boldsymbol{\theta}})\tilde{K}_n(I_S|\hat{\boldsymbol{\theta}})\} + o_p(1), \end{aligned}$$

where $\tilde{K}_n(I_S|\hat{\boldsymbol{\theta}}) = \sum_{i=1}^n \{\partial_{\boldsymbol{\theta}} \log p_{[i]}(\boldsymbol{\theta}) + \partial_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta})/n\}^{\otimes 2}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$.

Assuming a uniform improper prior or a general prior with a very small value of $\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta})$, $\tilde{K}_n(I_S|\hat{\boldsymbol{\theta}}) \approx K_n(I_S|\hat{\boldsymbol{\theta}})$, and thus $nB_\eta = \mathrm{MC}(I_S) + o_p(1)$. Therefore, we propose a Bayesian Case-influence Information Criterion (BCIC) as follows:

$$\mathrm{BCIC} = -2E_{\boldsymbol{\theta}|\boldsymbol{Y}}\{\log p(\boldsymbol{Y}|\boldsymbol{\theta})\} + 2\mathrm{MC}(I_S). \tag{3.16}$$

We then choose the model that minimizes BCIC. The differences between BPIC, DIC, and BCIC are in the complexity terms. Since BCIC uses the complexity based on case influence, a model has a larger penalty than the other candidate models, if there exist more influential cases in the model compared to the other candidate models. Similar to DIC and BPIC, BCIC can be easily computed using MCMC samples from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{Y})$.

3.4 Theoretical Examples

In this section, we illustrate the proposed diagnostic measures for various regression models.

3.4.1 Normal Linear Models

We consider normal linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.17}$$

where $\boldsymbol{Y} = (y_1, \cdots, y_n)^T$ is an $n \times 1$ vector, $\boldsymbol{X}$ is an $n \times p$ covariate matrix with ith row $\boldsymbol{x}_i^T$, $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector, $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \tau^{-1}\mathbf{I})$, and $\tau = 1/\sigma^2$ is assumed known. Thus, we have $\boldsymbol{Y}|\boldsymbol{\beta}, \tau \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I})$. We consider a conjugate normal prior for $\boldsymbol{\beta}|\tau$ as $N_p(\boldsymbol{\mu}_0, \tau^{-1}\boldsymbol{\Sigma}_0)$. The posterior distribution for $\boldsymbol{\beta}|\tau$ based on the full data and the ith case deleted data, respectively, are given by

$$\boldsymbol{\beta}|\boldsymbol{Y} \sim N_p(\tilde{\boldsymbol{\beta}},\ \tau^{-1}(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}) \text{ and } \boldsymbol{\beta}|\boldsymbol{Y}_{[i]} \sim N_p(\tilde{\boldsymbol{\beta}}_{[i]},\ \tau^{-1}(\boldsymbol{X}_{[i]}^T\boldsymbol{X}_{[i]} + \boldsymbol{\Sigma}_0^{-1})^{-1}),$$

where $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0)$, $\tilde{\boldsymbol{\beta}}_{[i]} = (\boldsymbol{X}_{[i]}^T\boldsymbol{X}_{[i]} + \boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{X}_{[i]}^T\boldsymbol{Y}_{[i]} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0)$, $\boldsymbol{X}_{[i]}$ is $\boldsymbol{X}$ with $\boldsymbol{x}_i^T$ deleted, and $\boldsymbol{Y}_{[i]}$ is $\boldsymbol{Y}$ with $\boldsymbol{y}_i$ deleted. Note that $\boldsymbol{X}_{[i]}^T\boldsymbol{X}_{[i]} = \boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i^T$ and $\boldsymbol{X}_{[i]}^T\boldsymbol{Y}_{[i]} = \boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{x}_i\boldsymbol{y}_i$.

Here, we consider single case deletion for the clarity and ease of exposition in derivation of the diagnostic measures. For the K-L divergence, we can get an exact analytic form for $\mathrm{D}_\phi(i)$ (Cook and Weisberg 1982, p163, equation(4.3.4)) as follows:

$$\begin{aligned}\mathrm{D}_\phi(i) &= 0.5[\tau(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{[i]})^T(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{[i]}) - \tau(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{[i]})^T(\boldsymbol{x}_i\boldsymbol{x}_i^T)(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{[i]}) \\ &\quad - \log|1 - \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}\boldsymbol{x}_i| - \mathrm{tr}\{\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}\boldsymbol{x}_i\}].\end{aligned}$$

Letting $\boldsymbol{Q} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}\boldsymbol{X}^T$ with diagonal element $q_{ii} = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}\boldsymbol{x}_i$, we have

$$\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{[i]} = \frac{1}{1 - q_{ii}}(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}). \tag{3.18}$$

Thus, we obtain

$$\mathrm{D}_\phi(i) = 0.5\left\{\tau \cdot \frac{q_{ii}}{1-q_{ii}}(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})^T(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}) - \log(1-q_{ii}) - q_{ii}\right\}.$$

In model (3.17), we note that the posterior mode and posterior mean are the same (i.e. $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}_{[i]} = \tilde{\boldsymbol{\beta}}_{[i]}$), and $G_{\boldsymbol{\beta}}$=$W_{\boldsymbol{\beta}}$=$\tau(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})$. Therefore, CP($i$) and CM($i$) are exactly the same measures. By substituting (3.18) into the formula for CM(i), we obtain

$$\mathrm{CM}(i) = \tau\left(\frac{1}{1-q_{ii}}\right)^2 q_{ii}(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})^T(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}). \tag{3.19}$$

In addition, the result in (3.19) ensures the equivalence shown in Theorem 3.1 (d) since $\mathrm{D}_\phi(i) = 0.5\mathrm{CM}(i) - 0.5\{q_{ii}\mathrm{CM}(i) + \log(1-q_{ii}) + q_{ii}\}$. Since $p_i(\beta) = (2\pi)^{-1/2}\tau^{1/2}\exp\{-\frac{1}{2}\tau(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^T(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\}$ and $J_n(\boldsymbol{\beta}) = \tau(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})$, the approximations in Theorem 3.1 (a), (b), (c) for model (3.17) are given by

$$\begin{aligned}
\mathrm{D}_\phi(i) &= 0.5\ddot{\phi(1)}\tau q_{ii}(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})^T(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})\{1 + O_p(n^{-1})\},\\
\hat{\boldsymbol{\beta}}_{[i]} = \tilde{\boldsymbol{\beta}}_{[i]} &= \tilde{\boldsymbol{\beta}} - (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})\{1 + O_p(n^{-1})\},\\
\mathrm{AP}(i;\tilde{\boldsymbol{\beta}}) &= \tau q_{ii}(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})^T(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}).
\end{aligned}$$

Summing $\mathrm{AP}(i;\tilde{\boldsymbol{\beta}})$ for $i = 1, \cdots, n$ yields $\mathrm{MC}(I_S)$ and combining this with the posterior expectation of the deviance, $E_{\boldsymbol{\beta}|\boldsymbol{Y}}[-2\log p(\boldsymbol{Y}|\boldsymbol{\beta})] = n\log(2\pi) - n\log(\tau) + \tau(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) + \mathrm{tr}(\boldsymbol{Q})$, yields a closed form expression for BCIC. Moreover, we can easily extend the results to multiple case deletions using $p_S(\boldsymbol{\beta}) = \prod_{i=1, i\in S}^n (2\pi)^{-1/2}\tau^{1/2}\exp\{-\frac{1}{2}\tau(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^T(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\}$.

3.4.2 Linear Mixed Models

We consider the model of Laird and Ware (1982) as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad \text{for } i = 1, \cdots, n, \tag{3.20}$$

where $\boldsymbol{Y}_i$ is an $m_i \times 1$, $\boldsymbol{X}_i$ is an $m_i \times p$ matrix of fixed covariates, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{Z}_i$ is an $m_i \times q$ matrix of covariates for the $q \times 1$ vector of random effects $\boldsymbol{b}_i$, $\boldsymbol{b}_i \sim N_q(\boldsymbol{0}, \tau^{-1}\boldsymbol{D})$, $\boldsymbol{\epsilon}_i$ is an $m_i \times 1$ vector of random errors, $\boldsymbol{\epsilon}_i \sim N_{m_i}(\boldsymbol{0}, \tau^{-1}\mathbf{I}_{m_i})$, $\tau = 1/\sigma^2$, and $\boldsymbol{\epsilon}_i$ and $\boldsymbol{b}_i$ are independent. We can write this model as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$, where $\boldsymbol{Y} = (\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n)^T$, $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^T$, $\boldsymbol{Z} = \text{diag}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n)$, $\boldsymbol{b} = (\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n)^T$, and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_n)^T$. Thus $\boldsymbol{\epsilon} \sim N_N(\boldsymbol{0}, \tau^{-1}\mathbf{I}_N)$ and $\boldsymbol{b} \sim N_{nq}(\boldsymbol{0}, \tau^{-1}(\mathbf{I}_n \otimes \boldsymbol{D}))$, where $N = \sum_{i=1}^{n} m_i$, and $\otimes$ denotes kronecker product.

Known $\boldsymbol{D}$ and τ

If $\boldsymbol{D}$ and τ are known, we obtain closed form expressions for the diagnostic measures and the expressions are similar to Section 3.4.1. We note that upon integrating out $\boldsymbol{b}$, we have $\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{D}, \tau \sim N_N(\boldsymbol{X}\boldsymbol{\beta}, \tau^{-1}\boldsymbol{V})$, where $\boldsymbol{V} = \text{diag}(\boldsymbol{V}_1, \cdots, \boldsymbol{V}_n)$ with $\boldsymbol{V}_i = \mathbf{I}_{m_i} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T$. If we consider a normal prior for $\boldsymbol{\beta}$ as $N_p(\boldsymbol{\mu}_0, \tau^{-1}\boldsymbol{\Sigma}_0)$, the posterior distributions of $\boldsymbol{\beta}$ based on the full data and with the ith cluster deleted, are given by

$$\begin{aligned}
\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{D}, \tau &\sim N_p(\tilde{\boldsymbol{\beta}}, \tau^{-1}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}) \\
\boldsymbol{\beta}|\boldsymbol{Y}_{[i]}, \boldsymbol{D}, \tau &\sim N_p(\tilde{\boldsymbol{\beta}}_{[i]}, \tau^{-1}(\boldsymbol{X}_{[i]}^T\boldsymbol{V}_{[i]}^{-1}\boldsymbol{X}_{[i]} + \boldsymbol{\Sigma}_0^{-1})^{-1}),
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0)$ and $\tilde{\boldsymbol{\beta}}_{[i]} = (\boldsymbol{X}_{[i]}^T\boldsymbol{V}_{[i]}^{-1}\boldsymbol{X}_{[i]} + \boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{X}_{[i]}^T\boldsymbol{V}_{[i]}^{-1}\boldsymbol{Y}_{[i]} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0)$. Using similar calculations as in Section 3.4.1, we obtain

$\mathrm{D}_\phi(i)$ based on the K-L divergence as follows:

$$\mathrm{D}_\phi(i) = 0.5\left[\tau(\boldsymbol{Y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}})^T\{(\mathbf{I}_{m_i} - \boldsymbol{Q}_i)^{-1}\}^T\boldsymbol{V}_i^{-1}\boldsymbol{Q}_i(\boldsymbol{Y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}}) - \log|\mathbf{I}_{m_i} - \boldsymbol{Q}_i| - tr(\boldsymbol{Q}_i)\right],$$

where $\boldsymbol{Q}_i = \boldsymbol{X}_i(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}$. For model (3.20), we note that $\hat{\boldsymbol{\beta}}$=$\tilde{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}_{[i]}$=$\tilde{\boldsymbol{\beta}}_{[i]}$ and $J_N(\boldsymbol{\beta}) = G_{\boldsymbol{\beta}} = W_{\boldsymbol{\beta}} = \tau(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})$; therefore, we have

$$\mathrm{CP}(i) = \mathrm{CM}(i) = \tau\{(\mathbf{I}_{m_i} - \boldsymbol{Q}_i)^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}})\}^T\boldsymbol{V}_i^{-1}\boldsymbol{Q}_i(\mathbf{I}_{m_i} - \boldsymbol{Q}_i)^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}}). \quad (3.21)$$

In addition, since $p_i(\boldsymbol{\beta}) = (2\pi)^{-m_i/2}\tau^{m_i/2}|\boldsymbol{V}_i^{-1}|^{1/2}\exp[-\frac{1}{2}\tau(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})]$, the approximations in Theorem 3.1 are given by

$$\begin{aligned}
\mathrm{D}_\phi(i) &= 0.5\ddot{\phi}(1)\tau(\boldsymbol{Y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}})^T\boldsymbol{V}_i^{-1}\boldsymbol{Q}_i(\boldsymbol{Y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}})\{1 + O_p(N^{-1})\},\\
\hat{\boldsymbol{\beta}}_{[i]} = \tilde{\boldsymbol{\beta}}_{[i]} &= \tilde{\boldsymbol{\beta}} - (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}})\{1 + O_p(N^{-1})\},\\
\mathrm{AP}(i;\tilde{\boldsymbol{\beta}}) &= \tau(\boldsymbol{Y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}})^T\boldsymbol{V}_i^{-1}\boldsymbol{Q}_i(\boldsymbol{Y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}}).
\end{aligned}$$

Summing $\mathrm{AP}(i;\tilde{\boldsymbol{\beta}})$ yields $\mathrm{MC}(I_S)$ and combining this with $E_{\boldsymbol{\beta}|\boldsymbol{Y}}[-2\log P(\boldsymbol{Y}|\boldsymbol{\beta})] = N\log(2\pi) - N\log\tau + \log|\boldsymbol{V}| + \sum_{i=1}^n \tau(\boldsymbol{Y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}})^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}}) + \sum_{i=1}^n tr(\boldsymbol{Q}_i)$ yields BCIC for model (3.20). Moreover, we can easily extend the results to multiple cluster deletions using $p_S(\boldsymbol{\beta}) = \prod_{i=1, i\in S}^n p_i(\boldsymbol{\beta})$.

Unknown $\boldsymbol{D}$ and τ

When $\boldsymbol{D}$ and τ are unknown, we assume a joint prior of the form $p(\boldsymbol{\beta}, \tau, \boldsymbol{D}^{-1}) \propto p(\boldsymbol{\beta}|\tau)p(\tau)p(\boldsymbol{D}^{-1})$. The conjugate prior specifications are $\boldsymbol{\beta}|\tau \sim N_p(\boldsymbol{\mu}_0, \tau^{-1}\boldsymbol{\Sigma}_0)$, $\tau \sim Gamma(\delta_0/2, \gamma_0/2)$, and $\boldsymbol{D}^{-1} \sim Wishart_q(\nu_0, \boldsymbol{C}_0)$, where ν_0 is a scalar and $\boldsymbol{C}_0$ is a $q \times q$ positive definite matrix. Note that taking $\nu_0 = 0$, $\boldsymbol{C}_0^{-1} = \mathbf{0}$, $\boldsymbol{\Sigma}_0^{-1} = \mathbf{0}$, $\delta_0 = 0$, and $\gamma_0 = 0$ leads to a commonly used joint noninformative (and improper) prior specification. We can now write the joint posterior of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau, \boldsymbol{D})$ as $p(\boldsymbol{\beta}, \tau, \boldsymbol{D}^{-1}|\boldsymbol{Y}) \propto$

$p(\boldsymbol{Y}|\boldsymbol{\beta},\tau,\boldsymbol{D})p(\boldsymbol{\beta}|\tau)p(\tau)p(\boldsymbol{D}^{-1})$. Posterior samples of $(\boldsymbol{\beta},\tau,\boldsymbol{D})$ can be obtained for this model using MCMC methods. We note that each of the full conditional distributions have a closed form, so that Gibbs sampling is very efficient in this case.

The diagnostic measures can be computed by the formulas in Section 3.2.2 using MCMC samples of $(\boldsymbol{\beta},\tau,\boldsymbol{D}^{-1})$. Since the $\boldsymbol{Y}_i$s are independent, we have $p_S(\boldsymbol{\beta}) = \prod_{i=1,i\in S}^{n}(2\pi)^{-m_i/2}\tau^{m_i/2}|\boldsymbol{V}_i^{-1}|^{1/2}\exp\left\{-\frac{1}{2}\tau(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta})\right\}$, where $\boldsymbol{V}_i = \mathbf{I}_{m_i}+\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T$. Under the conjugate informative prior specification,

$$\begin{aligned} p(\boldsymbol{\beta},\tau,\boldsymbol{D}^{-1}|\boldsymbol{Y}) \;\propto\; & \prod_{i=1}^{n}\tau^{m_i/2}|\boldsymbol{V}_i^{-1}|^{1/2}\exp\left\{-\frac{1}{2}\tau(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta})\right\} \\ & \times\tau^{p/2}\exp\left\{-\frac{\tau}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right\}\times\tau^{\delta_0/2-1}\exp\left\{-\frac{\gamma_0}{2}\tau\right\} \\ & \times|\boldsymbol{D}^{-1}|^{(v_0-q-1)/2}\exp\left\{-\frac{1}{2}tr(\boldsymbol{C}_0^{-1}\boldsymbol{D}^{-1})\right\}, \end{aligned}$$

and therefore, $\partial_{\boldsymbol{\theta}}\log p(\boldsymbol{\beta},\tau,\boldsymbol{D}^{-1}|\boldsymbol{Y}) = \partial_{\boldsymbol{\theta}}\log p(\boldsymbol{Y}|\boldsymbol{\beta},\tau,\boldsymbol{D}) + \partial_{\boldsymbol{\theta}}\log p(\boldsymbol{\beta}|\tau)p(\tau)p(\boldsymbol{D}^{-1})$ and $\partial^2_{\boldsymbol{\theta}}\log p(\boldsymbol{\beta},\tau,\boldsymbol{D}^{-1}|\boldsymbol{Y}) = \partial^2_{\boldsymbol{\theta}}\log p(\boldsymbol{Y}|\boldsymbol{\beta},\tau,\boldsymbol{D}) + \partial^2_{\boldsymbol{\theta}}\log p(\boldsymbol{\beta}|\tau)p(\tau)p(\boldsymbol{D}^{-1})$, where $\boldsymbol{\theta}$ denotes $(\boldsymbol{\beta},\tau,\boldsymbol{D})$. The first and the second derivatives of $\log p(\boldsymbol{Y}|\boldsymbol{\beta},\tau,\boldsymbol{D})$ and $\log p(\boldsymbol{\beta}|\tau)p(\tau)p(\boldsymbol{D}^{-1})$ are given in the Appendix B. Using the MCMC posterior samples and the derivatives, we can compute $J_N(\boldsymbol{\theta})$, diagnostics measures, and their approximations, and BCIC.

3.4.3 Generalized Linear Models

Suppose $y_1,\cdots,y_n$ are independent, where y_i has a density in the exponential family indexed by the canonical parameter ψ_i and scale parameter τ. The joint density of $(y_1,\cdots,y_n)$ is given by

$$\prod_{i=1}^{n}p(y_i|\psi_i,\tau) = \prod_{i=1}^{n}\exp[\{y_i\psi_i - b(\psi_i)\}/a_i(\tau)+c(y_i,\tau)], \tag{3.22}$$

where the functions $b(\cdot)$ and $c(\cdot)$ determine a particular family in the class. Without loss of generality, we assume $a_i(\tau) = \tau^{-1}$ and τ is known. We consider a regression model via $\psi_i = \psi(\eta_i)$, $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, $i = 1, \cdots, n$, where $\boldsymbol{x}_i^T$ is ith row of the $n \times p$ covariate matrix $\boldsymbol{X}$, $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector, and $\psi(\cdot)$ is a monotone differentiable function. We consider three types of prior distributions for the GLM: i) improper uniform prior; ii) normal prior; and iii) conjugate prior. In this model, posterior samples of $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta}|\boldsymbol{Y})$ can be easily obtained using the Adaptive Rejection Metropolis Sampling (ARMS) algorithm (Gilks et al., 1995) to sample the full conditional distributions within the Gibbs sampling algorithm.

Computation of the proposed diagnostic measures can be achieved by the computational formulas as well as the approximation in 3.2.2. Since the y_i's are independent, $p_S(\boldsymbol{\beta}) = \prod_{i \in S} p(y_i|\boldsymbol{\beta}, \tau) = \prod_{i \in S} \exp[\tau\{y_i\psi(\boldsymbol{x}_i^T\boldsymbol{\beta}) - b(\psi(\boldsymbol{x}_i^T\boldsymbol{\beta}))\} + c(y_i, \tau)]$. The posterior mode, $\hat{\boldsymbol{\beta}}$ can be obtained by solving $\partial_{\boldsymbol{\beta}} \log p(\boldsymbol{Y}|\boldsymbol{\beta}) + \partial_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}) = \boldsymbol{0}$ and the posterior mean, $\tilde{\boldsymbol{\beta}}$ can be directly obtained from the posterior samples. $J_n(\boldsymbol{\beta}) = -\partial^2_{\boldsymbol{\beta}} \log p(\boldsymbol{Y}|\boldsymbol{\beta}) - \partial^2_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta})$ and $J_n(\boldsymbol{\beta})$ varies according to the different prior specifications. In model (3.22), we assume a canonical link for ease of exposition, so that $\psi_i = \eta_i$. Now, we have $\partial_{\boldsymbol{\beta}} \log p(\boldsymbol{Y}|\boldsymbol{\beta}) = \tau\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{\mu})$ and $\partial^2_{\boldsymbol{\beta}} \log p(\boldsymbol{Y}|\boldsymbol{\beta}) = -\tau\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}$, where $\boldsymbol{\mu} = (\partial_{\psi_1} b(\boldsymbol{x}_1^T\boldsymbol{\beta}), \cdots, \partial_{\psi_n} b(\boldsymbol{x}_n^T\boldsymbol{\beta}))^T$, $\boldsymbol{V} = \mathrm{diag}(\partial^2_{\psi_1} b(\boldsymbol{x}_1^T\boldsymbol{\beta}), \cdots, \partial^2_{\psi_n} b(\boldsymbol{x}_n^T\boldsymbol{\beta}))$. In the following, we examine approximations for the diagnostic measures under the different types of prior specifications for $\boldsymbol{\beta}$.

i) Uniform improper prior for $\boldsymbol{\beta}$:

We consider $p(\boldsymbol{\beta}) \propto 1$. Since $\partial_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}) = 0$, the posterior mode can be obtained by solving $\tau\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{0}$. In addition, we have $J_n(\boldsymbol{\beta}) = \tau\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}$. Thus, the

approximations are given by

$$\begin{aligned}
D_\phi(S) &= 0.5\ddot{\phi}(1)\tau\left\{\sum_{i=1,i\in S}^{n}(y_i-\hat{\mu}_i)\boldsymbol{x}_i^T\right\} \\
&\times (\boldsymbol{X}^T\hat{\boldsymbol{V}}\boldsymbol{X})^{-1}\left\{\sum_{i=1,i\in S}^{n}(y_i-\hat{\mu}_i)\boldsymbol{x}_i\right\}\{1+O_p(n^{-1})\}, \\
\hat{\boldsymbol{\beta}}_{[S]} &= \hat{\boldsymbol{\beta}}-(\boldsymbol{X}^T\hat{\boldsymbol{V}}\boldsymbol{X})^{-1}\left\{\sum_{i=1,i\in S}^{n}(y_i-\hat{\mu}_i)\boldsymbol{x}_i\right\}\{1+O_p(n^{-1})\}, \\
\tilde{\boldsymbol{\beta}}_{[S]} &= \tilde{\boldsymbol{\beta}}-(\boldsymbol{X}^T\hat{\boldsymbol{V}}\boldsymbol{X})^{-1}\left\{\sum_{i=1,i\in S}^{n}(y_i-\hat{\mu}_i)\boldsymbol{x}_i\right\}\{1+O_p(n^{-1})\}, \\
\mathrm{AP}(S;\tilde{\boldsymbol{\beta}}) &= \tau\left\{\sum_{i=1,i\in S}^{n}(y_i-\tilde{\mu}_i)\boldsymbol{x}_i^T\right\}(\boldsymbol{X}^T\tilde{\boldsymbol{V}}\boldsymbol{X})^{-1}\left\{\sum_{i=1,i\in S}^{n}(y_i-\tilde{\mu}_i)\boldsymbol{x}_i\right\}.
\end{aligned}$$

ii) Normal prior for $\boldsymbol{\beta}$:

We consider $p(\boldsymbol{\beta})\propto\exp\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\}$. We have $\partial_{\boldsymbol{\beta}}\log p(\boldsymbol{\beta})=-\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)$ and $\partial^2_{\boldsymbol{\beta}}\log p(\boldsymbol{\beta})=-\boldsymbol{\Sigma}_0^{-1}$. Thus, the posterior mode can be obtained by solving $\tau\boldsymbol{X}^T(\boldsymbol{Y}-\boldsymbol{\mu})-\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)=\boldsymbol{0}$ and we have $J_n(\boldsymbol{\beta})=\tau\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}+\boldsymbol{\Sigma}_0^{-1}$. The approximations are given by

$$\begin{aligned}
D_\phi(S) &= 0.5\ddot{\phi}(1)\tau^2\left\{\sum_{i=1,i\in S}^{n}(y_i-\hat{\mu}_i)\boldsymbol{x}_i^T\right\} \\
&\times (\tau\boldsymbol{X}^T\hat{\boldsymbol{V}}\boldsymbol{X}+\boldsymbol{\Sigma}_0^{-1})^{-1}\left\{\sum_{i=1,i\in S}^{n}(y_i-\hat{\mu}_i)\boldsymbol{x}_i\right\}\{1+O_p(n^{-1})\}, \\
\hat{\boldsymbol{\beta}}_{[S]} &= \hat{\boldsymbol{\beta}}-\tau(\tau\boldsymbol{X}^T\hat{\boldsymbol{V}}\boldsymbol{X}+\boldsymbol{\Sigma}_0^{-1})^{-1}\left\{\sum_{i=1,i\in S}^{n}(y_i-\hat{\mu}_i)\boldsymbol{x}_i\right\}\{1+O_p(n^{-1})\}, \\
\tilde{\boldsymbol{\beta}}_{[S]} &= \tilde{\boldsymbol{\beta}}-\tau(\tau\boldsymbol{X}^T\hat{\boldsymbol{V}}\boldsymbol{X}+\boldsymbol{\Sigma}_0^{-1})^{-1}\left\{\sum_{i=1,i\in S}^{n}(y_i-\hat{\mu}_i)\boldsymbol{x}_i\right\}\{1+O_p(n^{-1})\}, \\
\mathrm{AP}(S;\tilde{\boldsymbol{\beta}}) &= \tau^2\left\{\sum_{i=1,i\in S}^{n}(y_i-\tilde{\mu}_i)\boldsymbol{x}_i^T\right\}(\tau\boldsymbol{X}^T\hat{\boldsymbol{V}}\boldsymbol{X}+\boldsymbol{\Sigma}_0^{-1})^{-1}\left\{\sum_{i=1,i\in S}^{n}(y_i-\tilde{\mu}_i)\boldsymbol{x}_i\right\}.
\end{aligned}$$

iii) Conjugate prior for $\boldsymbol{\beta}$:

We consider the conjugate prior for $\boldsymbol{\beta}$ of Chen and Ibrahim (2003), given by $p(\boldsymbol{\beta}|a_0, \boldsymbol{y}_0, \tau) \propto \exp[a_0\tau\{\boldsymbol{y}_0^T\psi(\boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{J}^T b(\psi(\boldsymbol{X}\boldsymbol{\beta}))\}]$, where $\boldsymbol{J} = (1, \ldots, 1)^T$ is an $n \times 1$ vector of ones, and $\boldsymbol{y}_0$ and $a_0 > 0$ are the specified hyperparameters. Since we have $\partial_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}) = a_0\tau\boldsymbol{X}^T(\boldsymbol{y}_0 - \boldsymbol{\mu})$ and $\partial^2_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}) = -a_0\tau\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}$, the equation for solving for the posterior mode is $\tau\{\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu}) + a_0\boldsymbol{X}^T(\boldsymbol{y}_0 - \boldsymbol{\mu})\} = \boldsymbol{0}$ and we have $J_n(\boldsymbol{\beta}) = \tau(1 + a_0)\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}$. Thus, the approximations are given by

$$
\begin{aligned}
D_\phi(S) &= 0.5\ddot{\phi}(1)\tau(1 + a_0)^{-1} \\
&\times \left\{\sum_{i=1, i\in S}^{n} (y_i - \hat{\mu}_i)\boldsymbol{x}_i^T\right\} (\boldsymbol{X}^T\hat{\boldsymbol{V}}\boldsymbol{X})^{-1} \left\{\sum_{i=1, i\in S}^{n} (y_i - \hat{\mu}_i)\boldsymbol{x}_i\right\} \{1 + O_p(n^{-1})\}, \\
\hat{\boldsymbol{\beta}}_{[S]} &= \hat{\boldsymbol{\beta}} - (1 + a_0)^{-1}(\boldsymbol{X}^T\hat{\boldsymbol{V}}\boldsymbol{X})^{-1} \left\{\sum_{i=1, i\in S}^{n} (y_i - \hat{\mu}_i)\boldsymbol{x}_i\right\} \{1 + O_p(n^{-1})\}, \\
\tilde{\boldsymbol{\beta}}_{[S]} &= \tilde{\boldsymbol{\beta}} - (1 + a_0)^{-1}(\boldsymbol{X}^T\hat{\boldsymbol{V}}\boldsymbol{X})^{-1} \left\{\sum_{i=1, i\in S}^{n} (y_i - \hat{\mu}_i)\boldsymbol{x}_i\right\} \{1 + O_p(n^{-1})\}, \\
\mathrm{AP}(S; \tilde{\boldsymbol{\beta}}) &= \tau(1 + a_0)^{-1} \left\{\sum_{i=1, i\in S}^{n} (y_i - \tilde{\mu}_i)\boldsymbol{x}_i^T\right\} (\boldsymbol{X}^T\tilde{\boldsymbol{V}}\boldsymbol{X})^{-1} \left\{\sum_{i=1, i\in S}^{n} (y_i - \tilde{\mu}_i)\boldsymbol{x}_i\right\}.
\end{aligned}
$$

Summing $\mathrm{AP}(S; \tilde{\boldsymbol{\beta}})$ yields $\mathrm{MC}(I_S)$. Thus, for single case deletion, we obtain BCIC $= E_{\boldsymbol{\beta}|\boldsymbol{Y}}[-2\sum_{i=1}^{n}[\tau\{y_i\boldsymbol{x}_i^T\boldsymbol{\beta} - b(\boldsymbol{x}_i^T\boldsymbol{\beta})\} + c(y_i, \tau)]] + 2\mathrm{MC}(I_S)$.

3.4.4 Generalized Linear Mixed Models

Let y_{ij} denote the jth measurement on the ith subject. Suppose the sampling distribution of y_{ij}, $i = 1, \cdots, n$, $j = 1, \cdots, n_i$ is from an exponential family, so that

$$
p(y_{ij}|\theta_{ij}, \phi) = \exp\{\phi^{-1}(y_{ij}\theta_{ij} - a(\theta_{ij})) + c(y_{ij}, \phi)\}. \tag{3.23}
$$

Without loss of generality, we assume $\phi = 1$ for the logistic and Poisson regression models. In GLMM, the canonical parameter θ_{ij} is related to the covariates by $\theta(\theta_{ij}) =$

$\eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i$, where $\theta(\theta_{ij})$ is a monotonic function of θ_{ij}, $\boldsymbol{x}_{ij}^T$ is a $1 \times p$ vector of the jth row of $\boldsymbol{X}_i$, $\boldsymbol{z}_{ij}^T$ is a $1 \times q$ vector of the jth row of $\boldsymbol{Z}_i$, $\boldsymbol{\beta}$ is $p \times 1$ and $\boldsymbol{b}_i$ is $q \times 1$ random effects of the ith row of $\boldsymbol{b} = (\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n)^T$. We assume $\boldsymbol{b}_i \sim N_q(\boldsymbol{0}, \boldsymbol{D})$ and $\boldsymbol{b}_i$, $i = 1, \cdots, n$ are independent. Conditional on the random effects $\boldsymbol{b}_i$, the observation on the subject i are independent, thus the likelihood function for all n subjects is given by $p(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{b}) = \prod_{i=1}^{n}\prod_{j=1}^{n_i} p(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i)$, where $p(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i) = \exp\{y_{ij}\theta(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i) - a(\theta(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\boldsymbol{z}_{ij}\boldsymbol{b}_i))+c(y_{ij})\}$. The usual joint proper prior for $(\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{D}^{-1})$ is $p(\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{D}^{-1}) = p(\boldsymbol{\beta})p(\boldsymbol{b}|\boldsymbol{D}^{-1})p(\boldsymbol{D}^{-1})$, and we take $\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, $D^{-1} \sim W_q(\nu_0, C_0)$ and $\boldsymbol{b}|\boldsymbol{D} \sim N_{nq}(\boldsymbol{0}, (I_n \otimes \boldsymbol{D}))$. We write the kernel of the joint posterior of $(\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{D}^{-1})$ and run Gibbs sampler on the complete conditionals to obtain posterior samples of $(\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{D}^{-1})$. Since the complete conditionals do not have an analytic closed form for this model, we use ARMS within the Gibbs sampling algorithm. The joint posterior of $(\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{D}^{-1})$ can be written as $p(\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{D}^{-1}) \propto \{\prod_{i=1}^{n}\prod_{j=1}^{n_i} p(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i)p(\boldsymbol{b}_i)\}p(\boldsymbol{\beta})p(\boldsymbol{D}^{-1})$.

For this model, we can compute diagnostic measures using approximations in section 3.2.2 as well as $\mathrm{AP}(S; \tilde{\boldsymbol{\theta}})$ in equation (3.6). Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{D})$ then $p_S(\boldsymbol{\theta}) = \prod_{i \in S}\int_{\mathcal{R}^q}\prod_{j=1}^{n_i} p(y_{ij}|\boldsymbol{\theta}, \boldsymbol{b}_i)p(\boldsymbol{b}_i)d\boldsymbol{b}_i$. We can obtain $\partial_{\boldsymbol{\theta}} \log p_S(\boldsymbol{\theta})$ as sum of the ith component of $\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}|\boldsymbol{\theta})$ in S. The computation of $\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}|\boldsymbol{\theta})$ and $J_N(\boldsymbol{\theta}) = -\partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{Y}) = -\partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}|\boldsymbol{\theta}) - \partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta})$ can be done via the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). In detail, $\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}|\boldsymbol{\theta})$ can be evaluated by $E_{\boldsymbol{b}}[\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}, \boldsymbol{b}|\boldsymbol{\theta})|\boldsymbol{Y}]$ and $\partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}|\boldsymbol{\theta})$ can be computed by the Louis's method (Louis, 1982) as follows:

$$\partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}|\boldsymbol{\theta}) = \ddot{Q} + E_{\boldsymbol{b}}[\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}, \boldsymbol{b}|\boldsymbol{\theta})\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}, \boldsymbol{b}|\boldsymbol{\theta})^T|\boldsymbol{Y}] - \dot{Q}\dot{Q}^T, \tag{3.24}$$

where $\ddot{Q} = E_{\boldsymbol{b}}[\partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}, \boldsymbol{b}|\boldsymbol{\theta})|\boldsymbol{Y}]$ and $\dot{Q} = E_{\boldsymbol{b}}[\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}, \boldsymbol{b}|\boldsymbol{\theta})|\boldsymbol{Y}]$. In numerical examples, $[J_N(\tilde{\boldsymbol{\theta}})]^{-1}$ can be estimated by the empirical posterior covariance matrix, obtained from the MCMC samples.

3.5 Illustrative Examples

3.5.1 Generalized Linear Models: Binary Data

We first illustrate our methodology with a logistic regression example. We considered data on 200 men taken from the Los Angeles Heart Study conducted under the supervision of John M. Chapman (Dixon and Massey, 1983). The response variable is the occurrence or nonoccurrence of a coronary incident in the previous ten years. Of the 200 cases, 26 had coronary incidents and the dataset contains six other covariates: Age (x_1) (mean= 42.56, sd=11.65), Systolic blood pressure (x_2) (mean=121.64, sd=16.70), Diastolic blood pressure (x_3) (mean=81.59, sd=9.99), Cholesterol (x_4) (mean=285.11, sd=65.04), Height (x_5) (mean=65.58, sd=2.5) and Weight (x_6) (mean=165.19, sd=24.94). The logistic regression frequentist analysis of these data has been carried out by Christensen (1997). Here, we illustrate the proposed Bayesian diagnostic methods under both a uniform improper prior and normal priors for $\boldsymbol{\beta}$.

The model is given by $\log(p_i/(1-p_i)) = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4}+\beta_5 x_{i5}+\beta_6 x_{i6}$, where p_i is the probability of the occurrence of a coronary incident for the ith case for $i = 1, \cdots, 200$. For the normal prior for $\boldsymbol{\beta}$, we took $\boldsymbol{\beta} \sim N(\mathbf{0}, \kappa(\boldsymbol{X}^T\boldsymbol{X})^{-1})$, and considered several values of κ including κ=1, 3, 10 and 100. The posterior samples were obtained using Adaptive Rejection Sampling (ARS) within Gibbs (Gilks and Wild, 1992) and 40,000 MCMC posterior samples were used in the analysis after burn-in. For numerical stability in the MCMC sampling, we standardized all of the covariates. The posterior means (standard deviations) for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$ were, respectively, given by: -2.375 (0.289), 0.559 (0.285), 0.113 (0.356), -0.069 (0.400), 0.433 (0.245), -0.196 (0.274) and 0.528 (0.258).

To examine the performance of the proposed diagnostic measures, we computed the K-L divergence ($\phi(u) = -\log(u)$), denoted by KL, and the AP statistic. The posterior

TABLE 3.1: *Chapman data.* Case influence diagnostics based on the uniform improper prior for $\boldsymbol{\beta}$

Single Case Deletion					Simultaneous Two Case Deletion					
Case ID	KL	Cal.	AP	p-value	Case ID	Case ID	KL	Cal.	AP	p-value
86	0.202	0.788	0.404	0.027	86	192	0.638	0.924	1.276	0.003
151	0.191	0.782	0.382	0.075	41	126	0.619	0.921	1.238	0.060
192	0.179	0.774	0.358	0.090	129	192	0.579	0.914	1.158	0.015
41	0.177	0.773	0.355	0.401	48	151	0.568	0.912	1.136	0.017
126	0.166	0.766	0.331	0.130	86	129	0.558	0.910	1.117	0.004
48	0.150	0.755	0.300	0.185	48	192	0.529	0.904	1.057	0.019
129	0.143	0.749	0.286	0.134	86	151	0.510	0.900	1.021	0.002
5	0.123	0.734	0.246	0.451	151	159	0.490	0.895	0.979	0.005
21	0.108	0.720	0.216	0.079	86	184	0.469	0.890	0.938	0.004
159	0.106	0.718	0.212	0.054	86	159	0.453	0.886	0.906	0.002

Cal. denotes calibration of KL computed by the methods in McCulloch (1989) and Chapter 2.

Cal. close to 1 implies an influential observation.

predictive p-value of AP was computed as described in Section 3.2.2 with M=25 and M=10 for single case deletion and two case deletion, respectively. The changes in the posterior estimates across the cases were computed as well. Tables 3.1 and 3.2 show the top ten most influential cases based on uniform and normal priors, respectively, for single case deletion. We observe from Table 3.1 that case 86 (KL=0.202, AP=0.404) is identified as the most influential case followed by cases 151, 192 and 41 for the uniform prior. Under the normal prior with κ=10 and κ=100, case 41 is identified as the most influential, whereas for the normal prior with κ=1 and κ=3, essentially no influential cases are identified (Figure 3.1). This is due to the fact that the prior is very informative and therefore dominates the likelihood. When κ gets large, the normal prior becomes more noninformative and thus yields similar results to the uniform prior. The changes in the posterior estimates also describe the influence of the identified cases very well (results not shown for brevity). The Bayesian model complexity measure MC(I_S)

TABLE 3.2: *Chapman data.* Case influence diagnostics based on the normal prior for $\boldsymbol{\beta}$

Single Case Deletion							
κ=10				κ=100			
Case ID	KL	Cal.	AP	Case ID	KL	Cal.	AP
41	0.067	0.677	0.134	41	0.150	0.754	0.299
5	0.055	0.661	0.110	151	0.139	0.747	0.279
19	0.050	0.654	0.099	192	0.122	0.732	0.243
151	0.048	0.651	0.096	126	0.121	0.732	0.242
126	0.043	0.644	0.087	86	0.121	0.732	0.242
48	0.041	0.641	0.083	48	0.111	0.724	0.223
192	0.039	0.637	0.078	5	0.105	0.718	0.210
113	0.037	0.633	0.073	129	0.100	0.713	0.200
129	0.034	0.628	0.068	19	0.088	0.701	0.177
111	0.032	0.624	0.064	21	0.071	0.682	0.143
86	0.031	0.623	0.062	42	0.071	0.682	0.142

is 6.82 with a p-value=0.347, which means that the model fits the data reasonably well although the data contains some influential cases. Moreover, MC(I_S) is close to p_D=7.04 as well as number of parameters p=7. After an investigation as to the reasons why these identified cases were more influential than other cases, we found that one or two of the covariate values were extreme, or that a coronary incident occurred for the cases having covariate values corresponding to those at lower risk of a coronary incident. For example, case 86 has low values for the covariates, age, cholesterol level, and weight corresponding to a low risk of a coronary incident (age (x_1)=34, cholesterol(x_4)=214 and weight (x_6)=139), but a coronary incident had occurred for this case. Case 41 has an exceptionally high cholesterol value (x_4=520), and case 151 has a low weight (x_6=128), which is the smallest weight among those that had coronary incidents.

We also illustrate the performance of the proposed diagnostic measures for multiple

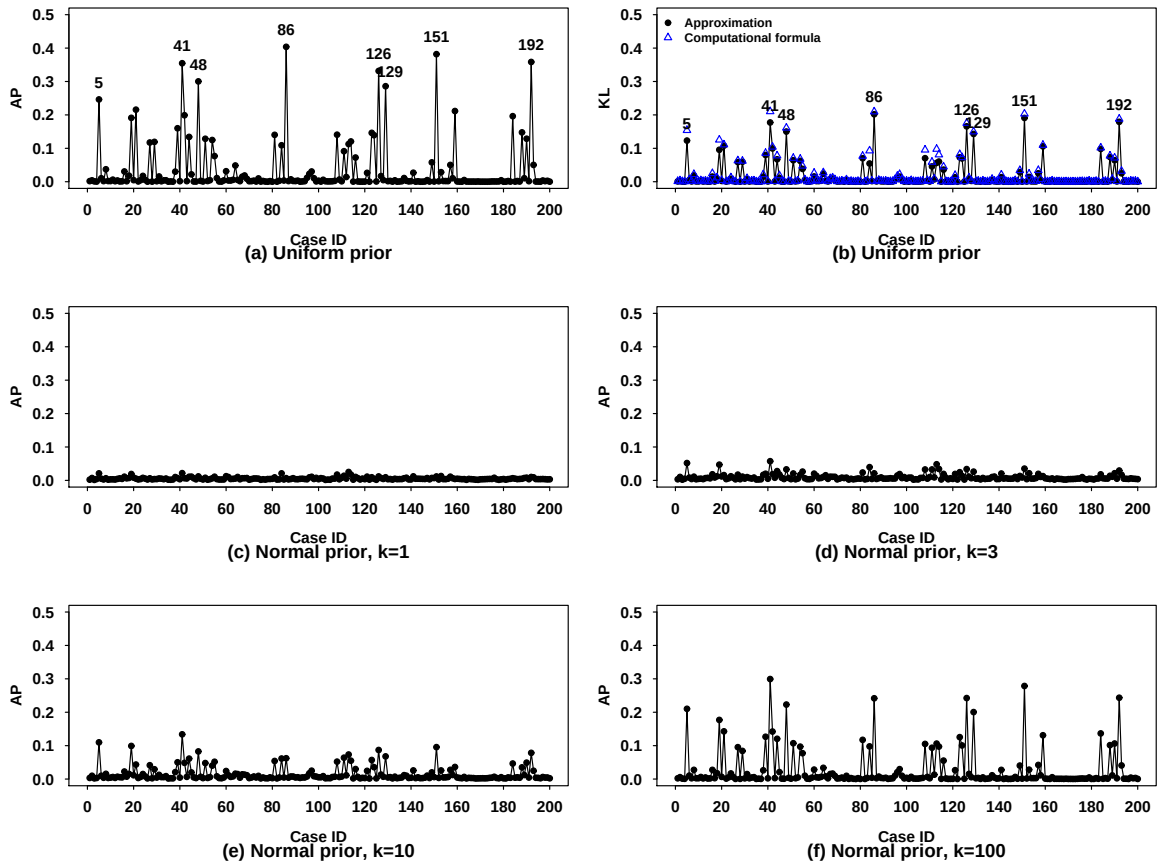

FIGURE 3.1: *Chapman data.* Case influence diagnostics, single case deletion: (a) AP based on the uniform improper prior; (b) KL based on the uniform improper prior. KL is computed using the computational formula and the approximation evaluated at the posterior mean. The results from the two methods agree well; (c), (d), (e) and (f) AP based on normal priors for κ=1, 3, 10 and 100, respectively.

case deletion by deleting two cases simultaneously. Table 3.1 shows top 10 most influential pairs of cases based on the uniform improper prior for $\boldsymbol{\beta}$. We observe large AP values and small p-values for the identified pairs. Moreover, most of the pairs consisted of the cases identified in single case deletion. We visualized the diagnostic measure, AP, for the simultaneous two case deletion scheme using a scatter plot in 2-dimensional space as well as a colored-surface plot in 3-dimensional space (Figure 3.2 (a), (b)). Note that the colors represent the magnitude of AP in both plots and the size of the symbol is proportional to the magnitude in the scatter plot. Mosaic patterns are found for larger values of AP in both plots, when one of the cases in the pairs is influential in

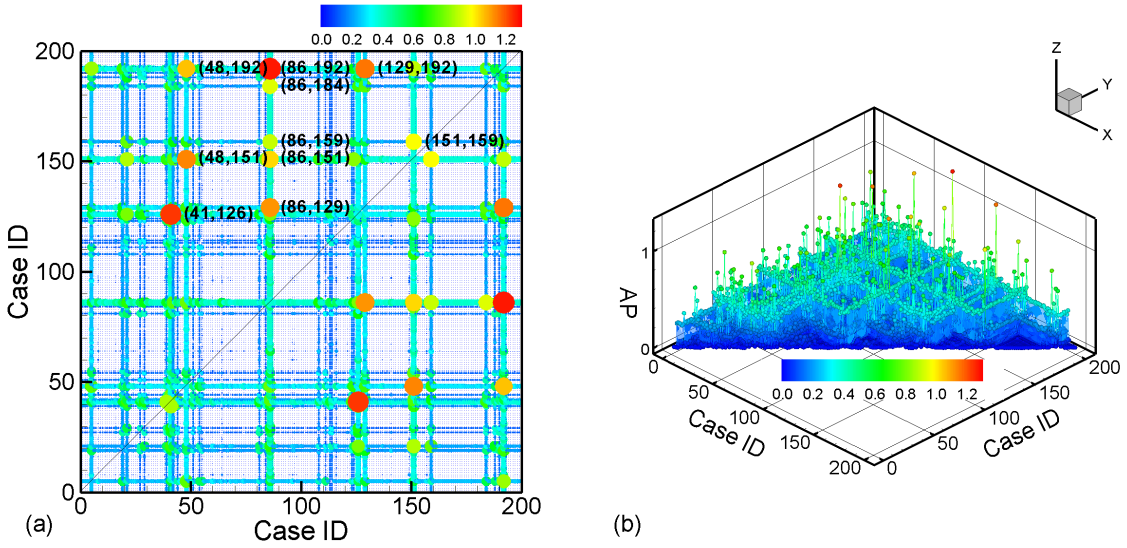

FIGURE 3.2: *Chapman data.* Case influence diagnostics based on the uniform improper prior, two case deletion: (a) 2-D scatter plot; (b) 3-D surface plot

single case deletion. The sets of case deletion pairs for the top 10 largest AP values are labeled in the scatter plot. Interestingly enough, the two case deletion scheme where each single case deletion has a large value of AP in Figure 3.1, magnify the values of AP. This indicates that the proposed diagnostic measure captures influential cases quite well whether in a single or simultaneous case deletion schemes.

TABLE 3.3: *Chapman data.* Information criteria for the top five models selected by BCIC based on the uniform improper prior for $\boldsymbol{\beta}$

Model No.	Covariate	MC_n	BCIC	AIC	BIC	p_D	DIC	$n\hat{b}_\beta$	BPIC
31	x_1, x_6	2.80	147.42	144.77	154.66	3.01	144.84	2.68	147.19
27	x_1, x_4, x_6	4.20	148.03	143.52	156.72	4.02	143.64	3.98	147.58
32	x_1	1.89	148.54	146.74	153.34	2.00	146.77	1.83	148.42
29	x_1, x_5, x_6	3.69	149.52	146.05	159.25	4.01	146.15	3.50	149.14
28	x_1, x_4	3.26	149.53	145.93	155.83	3.02	146.03	3.12	149.24

$n\hat{b}_\beta$ is the estimated asymptotic bias of the predictive discrepancy measures (Ando, 2007).

Now, we illustrate model selection using BCIC. We fit $2^6 = 64$ models; the full model and reduced models, with each model having an intercept term. Results for both the

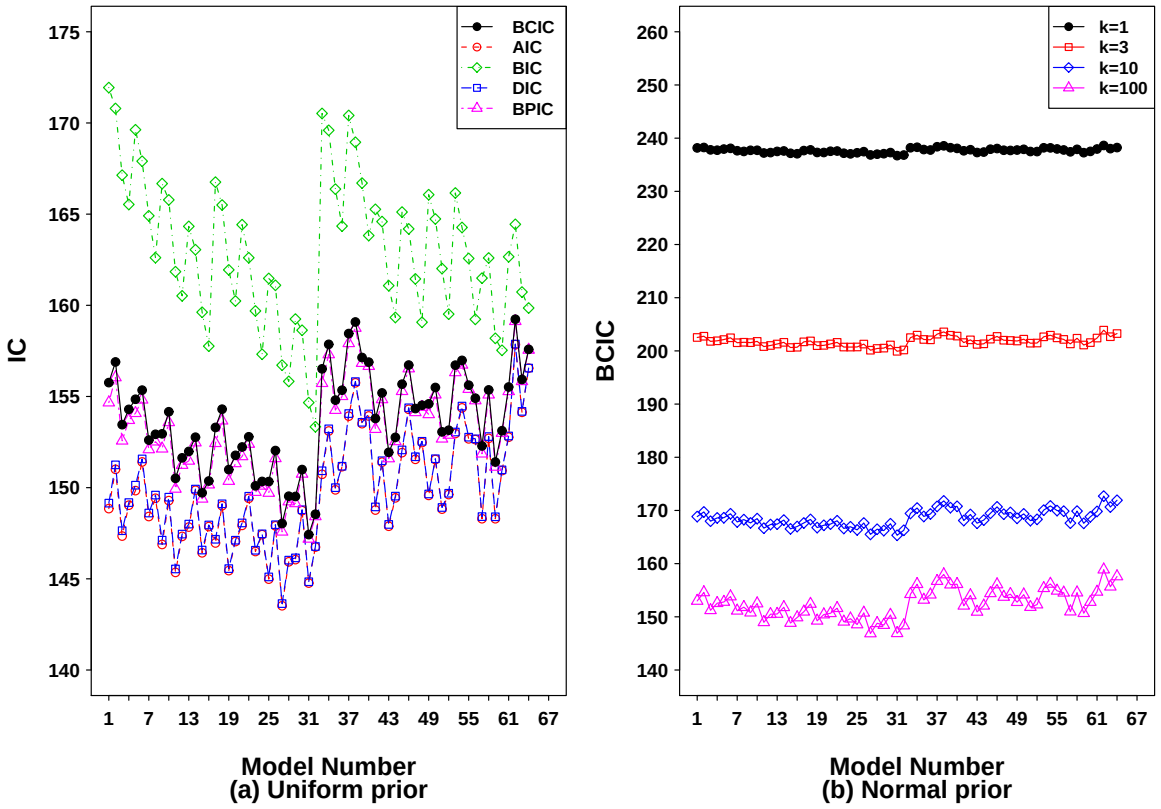

FIGURE 3.3: *Chapman data.* Information criteria: (a) comparison of BCIC with other information criteria based on the uniform improper prior for β; (b) behavior of BCIC for the different choices of κ based on the normal prior for β

uniform prior and normal prior are presented. Also, for each model, AIC, BIC, DIC and BPIC were computed for comparison purposes. Table 3.3 presents results based on the uniform prior. Form the results in Table 3.3, we can see that model (x_1, x_6) is the best fitting model by BCIC. The top four models selected by BCIC are (x_1, x_4, x_6), (x_1), (x_1, x_5, x_6) and (x_1, x_4). Moreover, these models are also selected by BPIC among the top five best fitting models. AIC and DIC selected (x_1, x_4, x_6) as the best model and (x_1, x_6) as the second best fitting model, whereas BIC selected (x_1) as the best and (x_1, x_6) as the second best fitting model based on the uniform prior (Figure 3.3 (a)). Figure 3.3 (b) shows that the behavior of BCIC is similar for the different choices of κ, however, the magnitude of the variation across the models gets smaller as κ gets

smaller. Overall, we can see that BCIC selects model (x_1, x_6) as the best fitting model for both the uniform and normal prior with sufficiently large κ. Note that the model number is defined as $64 - (2^5 I(x_1) + 2^4 I(x_2) + 2^3 I(x_3) + 2^2 I(x_4) + 2^1 I(x_5) + I(x_6))$, where $I(x_j)$ is 1 if a model includes covariate x_j, $j = 1, \cdots, 6$ and 0 otherwise.

3.5.2 Generalized Linear Mixed Models: Longitudinal Data

We consider the data from a clinical trial of 59 epileptics presented in Table 2 of Thall and Vail (1990). These epileptic patients were randomized to receive either the antiepileptic drug progabide (Trt=1) or a placebo (Trt=0) as an adjuvant to standard chemotherapy. Each patient reported the number of seizures that occurred over the previous 2 weeks at each of four successive postrandomization clinic visits. The other covariates measured were 8-week prerandomization seizure count and patient age in years. Frequentist analyses of these data as well as diagnostics have been done by many researchers (Thall and Vail, 1990; Breslow and Clayton, 1993; Zhu and Lee, 2001). In this analysis, we fit a model similar to that of Breslow and Clayton (1993) (Model IV) and Zhu and Lee (2001) within the Bayesian paradigm and illustrate the performance of the proposed Bayesian diagnostics.

The response variable y_{ij}, the seizure count for patient i on the jth visit, is assumed to be conditionally Poisson distributed with mean μ_{ij} such that

$$\log(\mu_{ij}(\boldsymbol{b}_i)) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + b_{i1} + b_{i2}\text{Visit}_j/10, \tag{3.25}$$

where the covariate $\boldsymbol{x}_{ij}$ include the intercept term, the logarithm of $\frac{1}{4}$ the number of baseline seizure count (Base), treatment (Trt), an interaction between baseline seizure count and treatment (Base$\times$Trt), the logarithm of age (Age) and a variable Visit_j for each of the four clinic visits coded as (-3, -1, 1, 3); $\boldsymbol{b}_i = (b_{i1}, b_{i2})^T$ are normally distributed random effects. We assumed that b_{i1} and b_{i2} are independent, $b_{i1} \sim N(0, \tau_1^{-1})$

TABLE 3.4: *Epileptic data.* Case influence diagnostics

Single Patient Deletion		Simultaneous Two Patient Deletion		
Patient ID	AP	Patient ID	Patient ID	AP
135	17.629	135	227	40.683
227	5.469	135	126	40.200
126	5.028	135	116	28.139
232	3.040	135	143	27.753
225	3.030	135	217	25.281
112	2.635	135	225	24.328

and $b_{i2} \sim N(0, \tau_2^{-1})$. We choose noninformative prior distributions for $\boldsymbol{\beta}$, τ_1 and τ_2 as $\boldsymbol{\beta} \sim N_6(\mathbf{0}, 10^6\mathbf{I})$, $\tau_1 \sim \text{Gamma}(10^{-4}, 10^{-4})$ and $\tau_2 \sim \text{Gamma}(10^{-4}, 10^{-4})$, respectively, where $\text{Gamma}(\alpha, \lambda)$ denotes the gamma distribution with mean α/λ ($\alpha > 0, \lambda > 0$). Posterior samples were obtained using WinBUGS (Spiegelhalter et al., 2003). We used every 25th sample after burn-in to reduce autocorrelations and yield better convergence results, and the results are based on 20,000 samples. The posterior means (standard deviations) for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \tau_1, \tau_2)$ were, respectively, given by: -1.363 (1.257), 0.8945 (0.1378), -0.9044 (0.4198), 0.3234 (0.2126), 0.4688 (0.3691), -0.2621 (0.1592), 3.6 (0.8755) and 2.225 (1.492).

Table 3.4 presents the patients (pairs of patients) having larger AP values compared to AP values of the other patients, obtained from deleting a single patient (cluster) as well as deleting two patients simultaneously. A deletion of a patient here means that we delete an entire cluster, that is we delete the observations for all four time points for that patients. For single cluster deletion, we observed that patient 135, 227, 126, 232, 225 and 112 were identified as influential (Table 3.4, Figure 3.4). Among the identified patients, patient 135 was highly influential. Moreover, pairs involving patient 135 were influential for deleting two patients simultaneously (Figure 3.5). Breslow and Clayton (1993) pointed out that patient 135 has a marked improvement over time after an

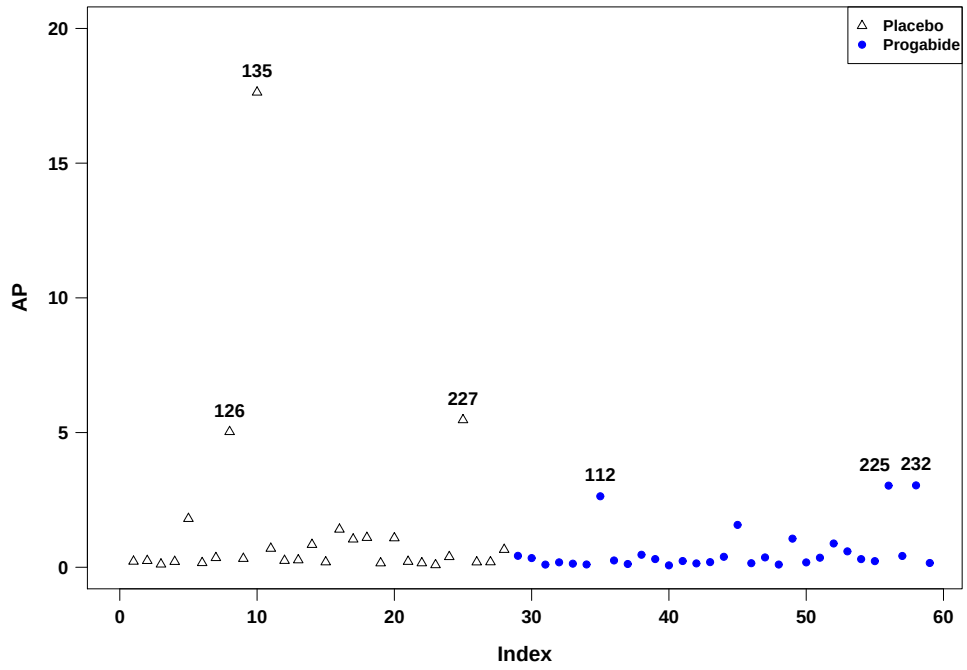

FIGURE 3.4: *Epileptic data.* Index plot of case influence diagnostics for deleting a single patient at a time. Indices on the horizontal axis correspond to the order of the patients in Table 2 of Thall and Vail (1990). IDs for the influential patients are indicated in the plot.

initially high seizure rate. Patients 227, 225, 112 have the highest overall count levels relative to their covariate values, and patient 232 has especially low or zero counts. In addition, some of these patients were regarded as outliers (Thall and Vail, 1990) and identified as influential by local influence measures (Zhu and Lee, 2001). Our Bayesian analysis confirms the findings of Breslow and Clayton (1993) as well as the others. In addition, we identified patient 126 who also had a marked improvement over time after an initially high seizure rate, in spite of the fact that the patient was on the placebo arm and is of old age compared to the other patients. We note that the Bayesian model complexity MC(I_S) is 58.98 for single cluster (patient) deletion.

3.6 Discussion

We have developed a general framework for evaluating Bayesian case influence measures based on case deletion for general parametric models. We have derived approximations to these case influence measures and proposed a calibration method under the deletion of a small (or large) number of observations. We have showed that these case influence measures are also associated with model complexity. When the number of observations in each set is large, we have shown that it is advantageous to use Cook's posterior mode and posterior mean distance for diagnostic purposes. The analytic forms for the proposed measures were derived for linear models, generalized linear models, normal mixed models and generalized linear mixed models. Future work includes developing a general framework for Bayesian diagnostic methods in semiparametric model.

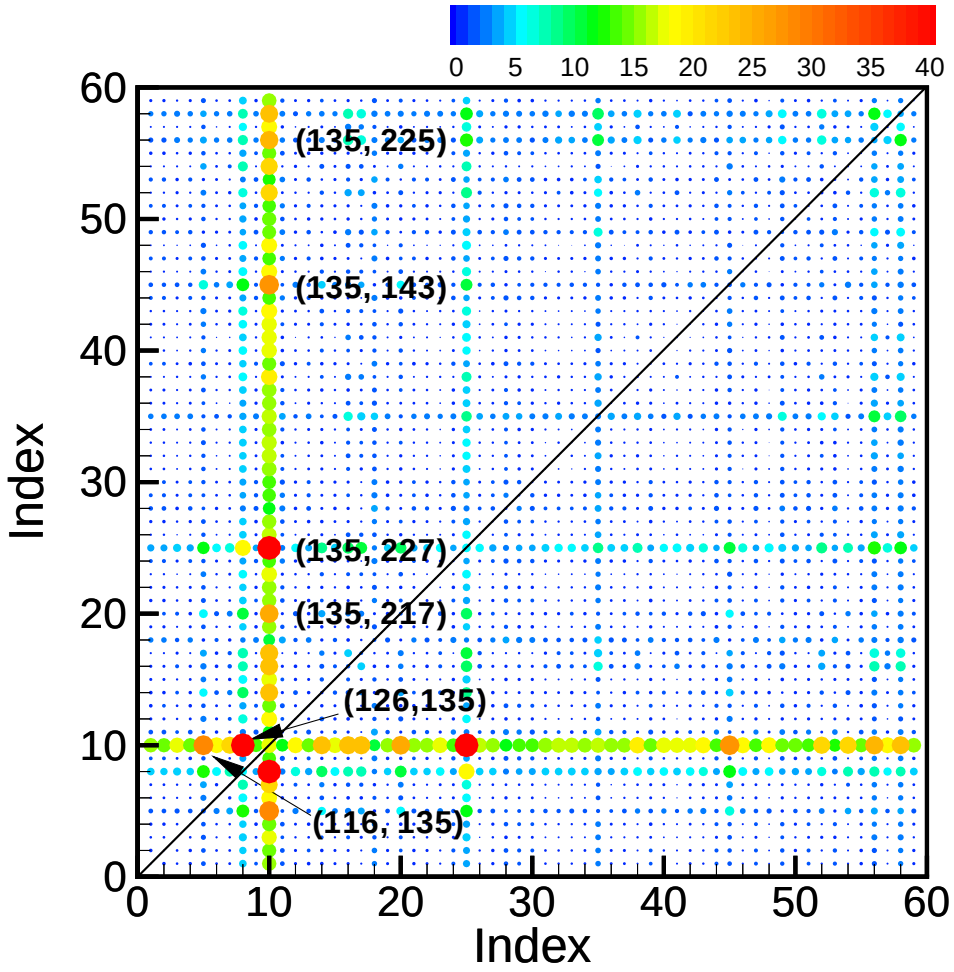

FIGURE 3.5: *Epileptic data.* 2-D scatter plot of case influence diagnostics for deleting two patients simultaneously. Indices on the horizontal and vertical axis correspond to the order of the patient in Table 2 of Thall and Vail (1990). IDs for the sets of influential patients are indicated in the plot.

CHAPTER 4

SCALED COOK'S DISTANCE

4.1 Introduction

In influence analysis, a set of observations is flagged as 'influential' if its removal from the data set produces a significant difference in the parameter estimate. Since the seminal work of Cook (1977) on Cook's distance in linear regression, considerable research has been devoted to developing deletion diagnostics including Cook's distance for detecting influential observations (or clusters) in various statistical models (Cook, 1977; Cook and Weisberg, 1982; Chatterjee and Hadi, 1988; Andersen, 1992; Davison and Tsai, 1992; Wei, 1998; Haslett, 1999; Zhu et al., 2001; Fung et al., 2002). Moreover, Cook's distance has been widely used in statistical practice and can be calculated in popular statistical software, such as SAS and R.

A fundamental issue of Cook's distance is that "size matters", that is Cook's distance is a monotonic function of the size of the perturbation. This issue has been largely neglected in the literature. The size matters issue persists in any deletion diagnostic, because the size of the deletion diagnostic is associated with the size of the perturbation. Although Critchley et al. (2001) have systematically addressed the size matters issue in deletion diagnostics for a simple data structure, such as one sample problems,

extending their method to complex data structures, such as longitudinal data, and general parametric models represents new theoretical and computational challenges. The issue that size matters, however, is central to the development of deletion diagnostics in complex models and data structures, because arbitrarily perturbing a model may lead to inappropriate inference about influential observations of a large effect. Consider two possibly overlapping subsets I_1 and I_2 and size$(I_1) \geq$ size(I_2). If CD$(I_1) \leq$ CD(I_2), then it is reasonable to regard I_2 to be influential, where CD$(\cdot)$ denotes Cook's distance. However, when CD$(I_1) >$ CD(I_2) is true, it is very difficult to compare the influential levels of I_1 and I_2, because a larger perturbation typically implies a larger influential measure. In particular, the issue of size arises often in assessing influential clusters and families in longitudinal and family studies, because cluster size (or family size) can vary significantly across all clusters (or families) and deleting a larger cluster may have a higher probability of having a larger influence.

The aim of this paper is to develop a scaled version of Cook's distance to address the size issue for deletion diagnostics in general parametric models. Our scaled Cook's distance properly accounts for the size of a perturbation and the fitted model to the data. In Section 4.2, we review Cook's distance and the issue of size of a perturbation. We develop several scaled Cook's distances to address the size issue in Cook's distance. We illustrate our development with linear regression, generalized linear models, linear mixed models, and generalized linear mixed models. In Section 4.3, we analyze two datasets using the proposed scaled Cook's distance. We give some final remarks in Section 4.4.

4.2 Scaled Cook's Distance

4.2.1 Cook's Distance

Consider a probability function $p(\mathbf{Y}|\theta)$ for a random vector $\mathbf{Y}^T = (Y_1^T, \cdots, Y_n^T)$, where $\theta = (\theta_1, \cdots, \theta_q)^T$ is a $q \times 1$ vector in an open subset Θ of R^q and $Y_i = (y_{i,1}, \cdots, y_{i,m_i})$, in which the dimension of Y_i, denoted by m_i, may vary across all i. For instance, in longitudinal studies, if our interest focuses on detecting influential clusters, then Y_i includes all responses and covariates of interest in the ith cluster. Thus, the number of observations in the ith cluster may vary significantly across clusters. However, if our interest is to detect influential observations for longitudinal studies, then Y_i can be data observed from a particular time point from a subject. Another example is a family study, in which the number of subjects in the ith family can vary across families.

Cook's distance and many other deletion diagnostics measure the distance between the maximum likelihood estimators of θ with and without Y_i (Cook and Weisberg, 1982; Cook, 1977). A subscript '[I]' denotes the relevant quantity with all observations in I deleted. For instance, if $I = \{i\}$, then $\mathbf{Y}_{[i]}$ is the corresponding observed data with all of the components of Y_i deleted. We define the maximum likelihood estimators of θ for the full sample $\mathbf{Y}$ and a subsample $\mathbf{Y}_{[i]}$ as

$$\hat{\theta} = \text{argmax}_\theta \log p(\mathbf{Y}|\theta) \quad \text{and} \quad \hat{\theta}_{[i]} = \text{argmax}_\theta \log p(\mathbf{Y}_{[i]}|\theta), \tag{4.1}$$

respectively. Cook's distance for $\{i\}$, denoted by CD(i), can be defined as follows:

$$\text{CD}(\{i\}) = (\hat{\theta}_{[i]} - \hat{\theta})^T G_\theta (\hat{\theta}_{[i]} - \hat{\theta}), \tag{4.2}$$

where G_θ is chosen to be a positive definite matrix. For instance, G_θ can be $-\partial_\theta^2 \log p(\mathbf{Y}|\hat{\theta})$, where ∂_θ^2 represents the second-order derivative with respect to θ.

We can easily generalize the above Cook's distance to quantify the effects of a subset of observations I on the parameter estimate. Let $\hat{\theta}_{[I]}$ be the maximum likelihood estimator of θ for $p(\mathbf{Y}_I|\theta)$, where $\mathbf{Y}_{[I]}$ is a subsample of $\mathbf{Y}$ with $\{Y_i : i \in I\}$ deleted. Similar to (4.2), Cook's distance for assessing the subset of observations in I is defined as

$$\mathrm{CD}(I) = (\hat{\theta}_{[I]} - \hat{\theta})^T G_\theta (\hat{\theta}_{[I]} - \hat{\theta}). \tag{4.3}$$

We can use the values of $\mathrm{CD}(I)$ to assess the influential level of the subset I. For instance, for $I = \{i\}$, the ith observation is flagged as influential if the value of $\mathrm{CD}(\{i\})$ is relatively large compared with other $\mathrm{CD}(\{j\})$ for $j \neq i$. Similarly, we can regard I as influential if the value of $\mathrm{CD}(I)$ is relatively large compared with other $\mathrm{CD}(J)$ for all J having a similar structure to I. Moreover, we may determine the magnitude of $\mathrm{CD}(I)$ based on critical points of the χ^2 distribution (Cook & Weisberg, 1982, p.183).

To have a better understanding of Cook's distance, we consider the following example in longitudinal data.

Example 1. The Yale infant growth data were collected to study whether cocaine exposure during pregnancy may lead to the maltreatment of infants after birth such as physical and sexual abuse. The total 298 children were recruited from two subject groups (cocaine exposed group and unexposed group). The key feature of this database is that different children had different numbers and patterns of visits during the study period (Wasserman and Leventhal, 1993; Stier et al., 1993). The total number of data points is $\sum_{i=1}^{n} m_i = 3176$, whereas m_i varies from 2 to 30.

Following Zhang (1999) and Zhu et al. (2007), we consider a linear mixed model with a compound symmetry covariance structure as follows: $y_{ij} = \mathbf{x}_{ij}^T \beta + \epsilon_{ij}$, where $\mathbf{x}_{ij} = (1, d, (d-120)^+, (d-200)^+, (g_a-28)^+, d(g_a-28)^+, (d-60)^+(g_a-28)^+, (d-490)^+(g_a-28)^+, sd, s(d-120)^+)^T$, in which d and g_a are the age of visit and gestational age, respectively, and s is the indicator for gender, with one for a girl and zero for a

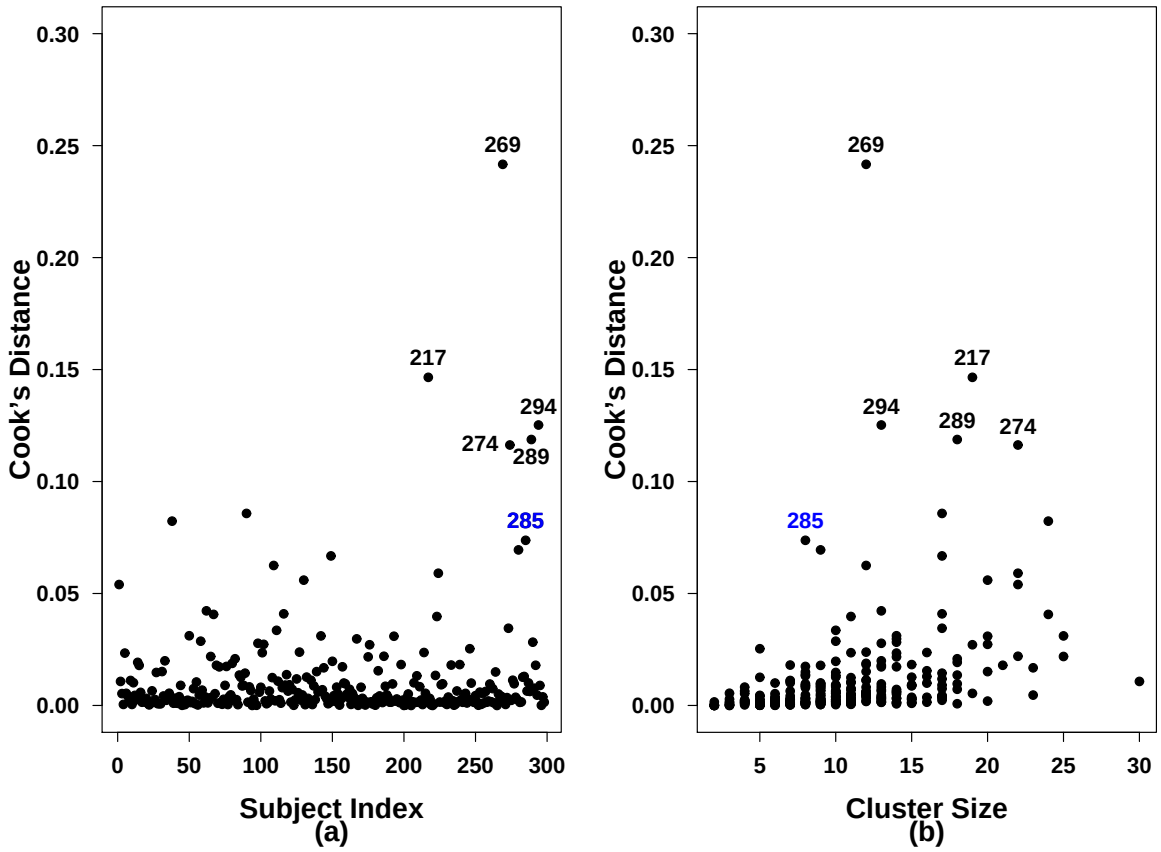

FIGURE 4.1: Yale infant growth data: (a) the index plot of Cook's distance; (b) cluster size versus Cook's distance.

boy. In addition, we assume $\epsilon_i \sim N_{m_i}[\mathbf{0}, \sigma^2 R_i]$ and consider a compound symmetry covariance structure for R_i. More details regarding this example are given in Section 4.3.2.

By using PROC MIXED (SAS 9.1, Cary, NC), we calculated Cook's distance for each child, which relates more to the detection of influential clusters (Banerjee and Frees, 1997). We obtained a strong correlation 0.363 between Cook distance and the cluster size with a $p-$value smaller than 1.03×10^{-10}. This indicates that the bigger cluster size, the larger the Cook distance measure. Figure 4.1 reveals five influential subjects $217, 269, 274, 289$, and 294, whose $(\text{CD}(i), m_i)$s are, respectively, given by

(0.147, 19), (0.242, 12), (0.116, 22), (0.119, 18), and (0.125, 13). Comparing subjects 269 and 274, we observe that $m_{269} = 12 < m_{274} = 22$, but CD(269) is much larger than CD(274). It is reasonable to claim that subject 269 is more influential than subject 274. Further, subject 269 is more influential than subject 294, because $m_{269} \approx m_{294}$ and CD(269) > CD(294). However, comparing subjects 274 and 285, we observe $(m_{285}, \text{CD}(285)) = (8, 0.074)$ and $(m_{274}, \text{CD}(274)) = (22, 0.116)$. Because m_{274} is much larger than m_{285}, it is difficult to claim that subject 274 is more influential than subject 285. This example illustrates the difficulty in comparing the Cook's distances across subsets of different sizes.

4.2.2 Size Matters

Based on the above analyses of the Yale infant growth data, we know that Cook's distance can be represented as follows:

$$\text{CD}(I) = F(\mathcal{P}(I|\mathcal{M}), \mathcal{M}, \mathcal{D}), \tag{4.4}$$

where $\mathcal{P}(I|\mathcal{M}), \mathcal{M}$, and $\mathcal{D}$, respectively, represent the size of the perturbation, the fitted model, and the dataset at hand and $F(\cdot)$ denotes a nonlinear function. The representation (4.4) reflects the fact that the influence level of the subset depends critically on the fitted model to the data and the amount of perturbation under $\mathcal{M}$. For a given $\mathcal{M}$, $\mathcal{P}(I|\mathcal{M})$, which is a function mapping from a subset I to a nonnegative number, quantifies the degree of perturbation introduced by deleting the subset I for the fitted $\mathcal{M}$ independent of the data. For instance, for the one sample problem, Critchley et al. (2001) use the Euclidean geometry of P^n and associated geodesics to quantify the size of the perturbation. Specifically, in this case, $\mathcal{P}(I|\mathcal{M})$ is defined as the geodesic distance between the null perturbation and the probability vector corresponding to deleting the subset I. However, there is no explicit expression for $\mathcal{P}(I|\mathcal{M})$ in relatively

complex data structures, such as time series data and longitudinal data.

EXAMPLE 2. To have a better understanding of Cook's distance, we consider the linear regression model $\mathbf{y} = \mathbf{X}\beta + \epsilon$, where $\mathbf{y} = (y_1, \cdots, y_n)^T$, $\mathbf{X}$ is $n \times p$ of rank p, and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T \sim N(\mathbf{0}, \sigma^2 I_n)$. Recall that $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and $H = (h_{ij}) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, where $\mathbf{y} = (y_1, \cdots, y_n)^T$. Cook's distance (Cook, 1977) for the ith point $(y_i, \mathbf{x}_i)$ is given by

$$\mathrm{CD}(\{i\}) = \frac{(\hat{\beta} - \hat{\beta}_{[i]})^T X^T X(\hat{\beta} - \hat{\beta}_{[i]})}{p\hat{\sigma}^2} = \frac{\sigma^2}{p\hat{\sigma}^2} t_i^2 \frac{h_{ii}}{1 - h_{ii}}, \tag{4.5}$$

where $\hat{\sigma}^2$ is a consistent estimator of σ^2, $t_i = \hat{e}_i/(\sigma\sqrt{1 - h_{ii}})$ and $\hat{\beta}_{[i]} = \hat{\beta} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\hat{e}_i/(1 - h_{ii})$, in which $\hat{e}_i = y_i - \mathbf{x}_i^T\hat{\beta}$. Clearly, Cook's distance involves the joint effect of two components:

$$\mathrm{CD}(\{i\}) = \text{Effect of deleting } \mathbf{x}_i \oplus \text{Effect of deleting } y_i \text{ given } \mathbf{x}_i,$$

where $\oplus$ denotes a joint effect. It is natural to think that the size of the perturbation for deleting different $(y_i, \mathbf{x}_i)$ should equal each other. For diagnostic purposes, if the true data generator is the same as the fitted model, then $\mathrm{CD}(\{i\})$ should be comparable regardless of i. Specifically, if $\epsilon \sim N(0, \sigma^2)$, then t_i^2 follows the $\chi^2(1)$ distribution for all i. To eliminate the variation of $\mathbf{x}_i$, we may assume that $\mathbf{x}_i$ follows the same distribution. Therefore, all $\mathrm{CD}(\{i\})$ are truly comparable, because they follow the same distribution under the fitted model.

We consider deleting multiple observations in the linear model. Cook's distance for deleting the subset I with size(I)= m is given by

$$\mathrm{CD}(I) = \frac{(\hat{\beta} - \hat{\beta}_{[I]})^T X^T X(\hat{\beta} - \hat{\beta}_{[I]})}{p\hat{\sigma}^2} = \frac{1}{p\hat{\sigma}^2}\hat{\mathbf{e}}_I(\mathbf{I}_m - H_I)^{-1}H_I(\mathbf{I}_m - H_I)^{-1}\hat{\mathbf{e}}_I, \tag{4.6}$$

where $\hat{\mathbf{e}}_I$ is an $m \times 1$ vector containing all $\hat{e}_i$ for $i \in I$, $\mathbf{I}_m$ is an $m \times m$ identity matrix,

and $H_I = \mathbf{X}_I(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_I^T$, in which $\mathbf{X}_I$ is an $m \times p$ matrix whose rows are $\mathbf{x}_i^T$ for all $i \in I$. Compared with the deletion of a single case, deleting multiple observations in the subset I introduces a larger size of the perturbation. Furthermore, we can show that the stochastic relationship between CD(I_1) and CD(I_2) for any two subsets I_1 and I_2 is as follows:

Theorem 4.1. *For the standard linear model, where* $\mathbf{y} = \mathbf{X}\beta + \epsilon$, $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$, *we have the following results:*

(a) For any $I_2 \subset I_1$, *CD*(I_1) *is stochastically larger than CD*(I_2) *for any fixed* $\mathbf{X}$;

(b) If H_I *and* $H_{I'}$ *follow the same distribution for any* I *and* I' *with size*(I) =*size*(I'), *then CD*(I) *and CD*(I') *follow the same distribution;*

(c) Under the same assumptions of Theorem 1 (b), CD(I_1) *is stochastically larger than CD*(I_2) *for any two subsets* I_2 *and* I_1 *with size*(I_1) >*size*(I_2).

Proof of Theorem 4.1.

(a) Let $I_3 = I_1/I_2$, I_1 is a union of two disjoint sets I_3 and I_2. Without loss of generality, H_{I_1} can be decomposed as

$$H_{I_1} = \mathbf{X}_{I_1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{I_1}^T = \begin{pmatrix} \mathbf{X}_{I_2}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{I_2}^T & \mathbf{X}_{I_2}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{I_3}^T \\ \mathbf{X}_{I_3}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{I_2}^T & \mathbf{X}_{I_3}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{I_3}^T \end{pmatrix}.$$

Let $\lambda_{1,1} \geq \cdots \geq \lambda_{1,m_1} \geq 0$ and $\lambda_{2,1} \geq \cdots \geq \lambda_{2,m_2} \geq 0$ be ordered eigenvalues of H_{I_1} and H_{I_2}, respectively, where $m_k = \text{size}(I_k)$ for $k = 1, 2$. It follows from Wielandt's eigenvalue inequality (Eaton and Tyler, 1991) that $\lambda_{1,i} \geq \lambda_{2,i}$ for all $i = 1, \cdots, m_2$. For $k = 1, 2$, we define $\Gamma_k\Lambda_k\Gamma_k^T$ as the spectral decomposition of H_{I_k} and $\mathbf{h}_k = (\mathbf{I}_{m_k} - \Lambda_k)^{-1/2}\Gamma_k^T\hat{\mathbf{e}}_{I_k} = (h_{k,1}, \cdots, h_{k,m_k})^T$, where Γ_k is the orthnormal matrix

and $\Lambda_k = \text{diag}(\lambda_{k,1}, \cdots, \lambda_{k,m_k})$. It can be shown that for $k = 1, 2$,

$$\mathbf{h}_k \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{m_k}) \text{ and } \text{CD}(I_k) = \frac{1}{p\hat{\sigma}^2} \sum_{j=1}^{m_k} \frac{\lambda_{k,j}}{1-\lambda_{k,j}} h_{k,j}^2.$$

Since $f(x) = x/(1-x)$ is an increasing function of $x \in (0,1)$, this completes the proof of Theorem 4.1 (a).

(b) Note that $\text{CD}(I) = (p\hat{\sigma}^2)^{-1} \sum_{j=1}^{m} \lambda_j (1-\lambda_j)^{-1} h_j^2$, where $\text{size}(I) = m$, λ_j are the eigenvalues of H_I and $\mathbf{h} = (h_1, \cdots, h_m)^T \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_m)$. Since

$$p(\mathbf{h}, \lambda_1, \cdots, \lambda_m) = p(\mathbf{h}|\lambda_1, \cdots, \lambda_m) p(\lambda_1, \cdots, \lambda_m) = p(\mathbf{h}) p(\lambda_1, \cdots, \lambda_m),$$

$\mathbf{h}$ and $\lambda = (\lambda_1, \cdots, \lambda_m)$ are independent. Moreover, the distribution of λ is uniquely determined by H_I. Combining $\mathbf{h} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_m)$ with the assumptions of Theorem 4.1 (b) yields that $\text{CD}(I)$ and $\text{CD}(I')$ follow the same distribution when $\text{size}(I)$=$\text{size}(I')$.

(c) We can always choose a I_2' such that $\text{size}(I_2')$=$\text{size}(I_2)$ and $I_1 \subset I_2'$. Combining Theorem 4.1 (a) and (b), we can then complete the proof of Theorem 4.1 (c).

Theorem 4.1 shows that for the standard linear model, a larger perturbation can cause a larger effect. Theorem 4.1 (a) shows that if the covariates in the design matrix $\mathbf{X}$ are treated as fixed, Cook's distances for two nested subsets satisfy the stochastic ordering property. Theorem 4.1 (b) and (c) indicates that H_I and $H_{I'}$ follow the same distribution even for any non-nested subsets I and I' with $\text{size}(I)$=$\text{size}(I')$, and the Cook's distances for any two subsets satisfy the stochastic ordering property. Generally, if X_I follows the same distribution for different I, then H_I and $H_{I'}$ follow the same distribution for any I and I' with $\text{size}(I)$=$\text{size}(I')$.

According to Theorem 4.1, it is natural to use the stochastic order to stochastically quantify the positive association between the degree of the perturbation and the amount of the effect. Specifically, we consider two possibly overlapping subsets I_1 and I_2 with

$\mathcal{P}(I_1|\mathcal{M}) > \mathcal{P}(I_2|\mathcal{M})$. Although, for a fixed data set $\mathcal{D}$, $\mathrm{CD}(I_1)$ may not be greater than $\mathrm{CD}(I_2)$, $\mathrm{CD}(I_1)$, as a random variable, should be *stochastically larger* than $\mathrm{CD}(I_2)$. We make the following assumption:

Assumption A1. For any two subsets I_1 and I_2 with $\mathcal{P}(I_1|\mathcal{M}) > \mathcal{P}(I_2|\mathcal{M})$,

$$\mathrm{P}(\mathrm{CD}(I_1) > t|\mathcal{M}) \geq \mathrm{P}(\mathrm{CD}(I_2) > t|\mathcal{M}) \tag{4.7}$$

holds for any $t > 0$, where the probability is taken with respect to the fitted model $\mathcal{M}$.

Assumption A1 is essentially saying that if the fitted model $\mathcal{M}$ is the true data generator, $\mathrm{CD}(I_1)$ stochastically dominates $\mathrm{CD}(I_2)$ whenever $\mathcal{P}(I_1|\mathcal{M}) > \mathcal{P}(I_2|\mathcal{M})$. We can now obtain the following theorem.

Theorem 4.2. *Under Assumption A1, Cook's distance satisfies that for any two subsets I_1 and I_2 with $\mathcal{P}(I_1|\mathcal{M}) > \mathcal{P}(I_2|\mathcal{M})$,*

$$E[h(CD(I_1))|\mathcal{M}] \geq E[h(CD(I_2))|\mathcal{M}] \tag{4.8}$$

holds for all increasing functions $h(\cdot)$. In particular, we have $E[CD(I_1)|\mathcal{M}] \geq E[CD(I_2)|\mathcal{M}]$ and $Q_{CD(I_1)}(\alpha|\mathcal{M})$ is greater than the α-quantile of $Q_{CD(I_2)}(\alpha|\mathcal{M})$ for any $\alpha \in [0,1]$, where $Q_{CD(I)}(\alpha|\mathcal{M})$ denotes the $\alpha-$quantile of the distribution of $CD(I)$ for any subset I.

Proof of Theorem 4.2. Theorem 4.2 follows directly from the definition of stochastic order. We omit the details here.

Theorem 4.2 formally characterizes the size matters issue of Cook's distance. Therefore, for any two subsets I_1 and I_2 with $\mathcal{P}(I_1|\mathcal{M}) > \mathcal{P}(I_2|\mathcal{M})$, $\mathrm{CD}(I_1)$ has high probability of being greater than $\mathrm{CD}(I_2)$. Thus, it is reasonable to regard I_2 to be influential when $\mathrm{CD}(I_1) << \mathrm{CD}(I_2)$, whose probability is small. However, in general, Cook's distance for subsets with different sizes are not directly comparable, since the scale of

Cook's distance depends on the size of the perturbation.

4.2.3 Scaled Cook's Distance

We focus on developing several corrections of size for Cook's distance for detecting influential subsets. Let $\mathrm{SCD}(I)$ denote a scaled Cook's distance. Consider K_0 features (e.g., mean, variance, median) of $\mathrm{SCD}(I)$, denoted by $\{S_k[\mathrm{SCD}(I)] : k = 1, \cdots, K_0\}$, when the fitted model is the true data generator. One type of correction for size is a feature-matching condition defined as follows:

Feature-Matching Condition: $S_k[\mathrm{SCD}(I_1)] = S_k[\mathrm{SCD}(I_2)]$ holds for all k and any two subsets I_1 and I_2, when the fitted model $\mathcal{M}$ is true.

The key features that we will consider below mainly include the mean and the median. By choosing either the mean or median, we can at least ensure that the centers of the scaled Cook's distances for different subsets are the same. Therefore, for any two subsets I_1 and I_2, the probability of observing the events $\mathrm{SCD}(I_1) > \mathrm{SCD}(I_2)$ and $\mathrm{SCD}(I_1) < \mathrm{SCD}(I_2)$ should be reasonably close to each other. Thus, the $\mathrm{SCD}(I)$ are roughly comparable.

We introduce two scaled Cook's distance measures as follows.

Definition 4.1. *The scaled Cook's distance for matching the mean and the median are, respectively, defined as*

$$SCD_1(I) = \frac{CD(I)}{E[CD(I)|\mathcal{M}]} \quad \textit{and} \quad SCD_2(I) = \frac{CD(I)}{Q_{CD(I)}(0.5|\mathcal{M})}. \tag{4.9}$$

It can be shown that $E[\mathrm{SCD}_1(I)|\mathcal{M}] = 1$ and $Q_{\mathrm{SCD}_2(I)}(0.5|\mathcal{M}) = 1$ hold for every subset I. Thus, we can use $\mathrm{SCD}_1(I)$ and $\mathrm{SCD}_2(I)$ to evaluate the influential level of different subsets I. A large value of $\mathrm{SCD}_1(I)$ (or $\mathrm{SCD}_2(I)$) indicates a large influence of the subset I, whereas a small value indicates a small influence.

The next problem is to compute $E[\mathrm{CD}(I)|\mathcal{M}]$ and $Q_{\mathrm{CD}_{(I)}}(0.5|\mathcal{M})$ for each subset I under the fitted model. Although the postulated model $p(\mathbf{Y}|\theta)$ may not represent the true data generator, we may find an 'optimal' model $p(\mathbf{Y}|\hat{\theta}) \in \mathcal{M}$ using the observed data (White, 1982). Then, based on $p(\mathbf{Y}|\hat{\theta})$, we use resampling methods (e.g., parametric bootstrap) or asymptotic methods to approximate $E[\mathrm{CD}(I)|\mathcal{M}]$ and $Q_{\mathrm{CD}_{(I)}}(0.5|\mathcal{M})$. In the following, we will derive the scaled Cook's distance for generalized linear models.

EXAMPLE 3. We consider Cook's distance in generalized linear models (McCullagh and Nelder, 1989) as follows. Suppose that the components of $\mathbf{y} = (y_1, \cdots, y_n)^T$ are mutually independent, and the conditional density of y_i given $\mathbf{x}_i$ is given by

$$p(y_i|\mathbf{x}_i, \tau) = \exp\left\{a_i^{-1}(\tau)[y_i\eta_i(\beta) - b(\eta_i(\beta))] + c(y_i, \tau)\right\}, \tag{4.10}$$

where $a_i(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ are known functions, $\eta_i = \eta(\mu_i)$ and $\mu_i(\beta) = g(\mathbf{x}_i^T\beta)$, in which $g(\cdot)$ is a known monotonic function and twice continuously differentiable and $\beta = (\beta_1, \ldots, \beta_p)^T$. Throughout the example, the parameter of interest is β and τ is a nuisance parameter and is fixed at $\hat{\tau}$. Let $V(\beta) = \mathrm{diag}(\ddot{b}(\eta_1(\beta)), \cdots, \ddot{b}(\eta_n(\beta)))$ and $D(\beta)^T = (\partial_\beta\mu_1(\beta), \cdots, \partial_\beta\mu_n(\beta))$, where ∂_β denotes differentiation with respect to β and $\ddot{b}(\eta)$ denotes the second derivative of $b(\eta)$ with respect to η. Using a first-order approximation, we can show that Cook's distance for deleting subset I is given by

$$\mathrm{CD}(I) \approx \mathrm{CD}(I)^1 = \frac{1}{p\hat{\sigma}^2}\hat{\mathbf{e}}^T\hat{V}^{-1/2}U_I(\mathbf{I}_m - \hat{H}_I)^{-1}\hat{H}_I(\mathbf{I}_m - \hat{H}_I)^{-1}U_I^T\hat{V}^{-1/2}\hat{\mathbf{e}}, \tag{4.11}$$

where $\hat{D} = D(\hat{\beta})$, $\hat{V} = V(\hat{\beta})$, $\hat{\mathbf{e}}$ is an $n \times 1$ vector containing all $\hat{e}_i = y_i - \mu_i(\hat{\beta})$, and $\hat{H}_I = \tilde{\mathbf{X}}_I(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}_I^T$. In addition, $\tilde{\mathbf{X}} = \hat{V}^{-1/2}\hat{D}$ and $\tilde{\mathbf{X}}_I$ is an $m \times p$ matrix containing the ith row of $\tilde{\mathbf{X}}$ for all $i \in I$, and $U_I = (\mathbf{u}_{i_1}, \cdots, \mathbf{u}_{i_m})$, in which $i_k \in I$ and $\mathbf{u}_{i_k}$ is an $n \times 1$ vector with the i_kth element equal to 1 and zero otherwise. Since $\hat{\sigma}^{-2}$ appears in all $\mathrm{CD}(I)$, we can fix $\hat{\sigma}$ from here on.

For generalized linear models, we can calculate the scaled Cook's distance and thus obtain the following theorem.

Theorem 4.3. *For the generalized linear model (4.10), we have the following results:*
(a) $CD(I)^1$ *can be approximated by*

$$\frac{\mathbf{e}_*^T V_*^{-1/2}(\mathbf{I}_n - H_*)U_I(\mathbf{I}_m - H_{*,I})^{-1}H_{*,I}(\mathbf{I}_m - H_{*,I})^{-1}U_I^T(\mathbf{I}_n - H_*)V_*^{-1/2}\mathbf{e}_*}{p\hat{\sigma}^2}, \quad (4.12)$$

where $\mathbf{e}_* = (e_{1*}, \cdots, e_{n*})^T$ *and* $e_{i*} = y_i - \mu_i(\beta_*)$, $D_* = D(\beta_*)$, $V_* = V(\beta_*)$, $H_* = \mathbf{X}_*(\mathbf{X}_*^T\mathbf{X}_*)^{-1}\mathbf{X}_*^T$, $\mathbf{X}_* = V_*^{-1/2}D_*$, *and* $H_{*,I} = U_I^T H_* U_I$, *in which* β_* *is the true value of* β.
(b) $pE[CD(I)^1|\mathcal{M}] \approx E\{tr[(\mathbf{I}_m - H_{*,I})^{-1}]|\mathcal{M}\} - m = \sum_{j=1}^m E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - m$, *where* $\lambda_{I,1} \geq \cdots \lambda_{I,m} \geq 0$ *are the ordered eigenvalues of* $H_{*,I}$. *Moreover, if* $m \geq p$, *then*

$$\sum_{j=1}^m E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - m = \sum_{j=1}^p E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - p. \quad (4.13)$$

(c) If the $\mathbf{x}_i$ *are independently and identically distributed with* $E[||\ddot{b}(\eta(\mathbf{x},\beta))^{-1/2}\partial_\beta\mu(\mathbf{x},\beta)||_2^{1+s}] < \infty$, *in which* $s > 0$, *then* $\lambda_{I,j} - m/n = o_p(1)$ *for* $j \leq p$ *as* $m \to \infty$ *and* $m/n \to \gamma \in [0,1)$.

Proof of Theorem 4.3.
(a) Let $\mu(\beta) = (\mu_1(\beta), \cdots, \mu_n(\beta))^T$. If the model $\mathcal{M}$ is true, then $(\hat{\beta} - \beta_*) = (D_*^T V_*^{-1} D_*)^{-1} D_*^T V_*^{-1}\mathbf{e}_* + o_p(n^{-1/2})$. Thus, it can be shown that

$$V_*^{-1/2}\hat{\mathbf{e}} = V_*^{-1/2}[\mathbf{y} - \mu(\beta_*) + \mu(\beta_*) - \mu(\hat{\beta})] = V_*^{-1/2}[\mathbf{e}_* - D_*(\hat{\beta} - \beta_*)] \approx (\mathbf{I}_n - H_*)V_*^{-1/2}\mathbf{e}_*,$$

where $\mathbf{e}_* = \mathbf{y} - \mu(\beta_*)$. This yields Theorem 4.3 (a).
(b) Since $E[\mathbf{e}_*^{\otimes 2}|\mathcal{M}] = \sigma^2 V_*$, we have

$$E[\mathrm{CD}_*(I)|\mathcal{M}] = p^{-1}E\{\mathrm{tr}[H_{*,I}(\mathbf{I}_m - H_{*,I})^{-1}]|\mathcal{M}\} = p^{-1}E\{\mathrm{tr}[(\mathbf{I}_m - H_{*,I})^{-1}]|\mathcal{M}\} - p^{-1}m.$$

Since H_* only has p non-zero eigenvalues and $H_{*,I}$ is a submatrix of H_*, it follows from Wielandt's eigenvalue inequality that $\lambda_{I,1} \geq \cdots \geq \lambda_{I,p} \geq 0 = \lambda_{I,p+1} = \cdots = \lambda_{I,m}$ for $m \geq p$. This yields Theorem 4.3 (b).

(c) Note that the matrices $H_{*,I}$ and $(\mathbf{X}_*^T\mathbf{X}_*)^{-1}\mathbf{X}_{*,I}^T\mathbf{X}_{*,I}$ have the same set of nonzero eigenvalues. Since $n^{-1}\mathbf{X}_*^T\mathbf{X}_*$ and $m^{-1}\mathbf{X}_{*,I}^T\mathbf{X}_{*,I}$ converge to the same matrix almost surely, $mn^{-1}[(n^{-1}\mathbf{X}_*^T\mathbf{X}_*)^{-1}m^{-1}\mathbf{X}_{*,I}^T\mathbf{X}_{*,I} - \mathbf{I}_p]$ should be close to $\mathbf{0}$ as $n, m \to \infty$. This completes the proof of Theorem 4.3 (c).

Theorem 4.3 has several important implications for generalized linear models. Theorem 4.3 (a) characterizes the stochastic behavior of $\mathrm{CD}(I)^1$, which depends on both the responses and the covariates in the set I. To ensure that $E[\mathrm{CD}(I)|\mathcal{M}]$ and $Q_{\mathrm{CD}_{(I)}}(0.5|\mathcal{M})$ depend only on the size of the perturbation, not the set I itself, we need to bootstrap the randomness in both the responses and the covariates. Specifically, we can generate a new set of responses from the fitted model and draw an I_s at random from the original covariate data without (or with) replacement, where size(I_s)=size(I). Then, we calculate $\mathrm{CD}(I_s)$ based on the bootstrapped data for $s = 1, \cdots, S$ and use their sample median to approximate $Q_{\mathrm{CD}_{(I)}}(0.5|\mathcal{M})$. Theorem 4.3 (b) gives an approximation of $E[\mathrm{CD}(I)^1|\mathcal{M}]$. We can draw a sample of sets $\{I_s : s = 1, \cdots, S\}$ of size(I) at random from the original covariate data without (or with) replacement and approximate $E\{\mathrm{tr}[(\mathbf{I}_m - H_{*,I})^{-1}]|\mathcal{M}\}$ by using $\sum_{s=1}^{S} \mathrm{tr}[(\mathbf{I}_m - H_{*,I_s})^{-1}]/S$. Moreover, it should be noted that $\sum_{j=1}^{m} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - m$ increases with the size of I even for $m \geq p$. Theorem 4.3 (c) shows the asymptotic consistency of $\lambda_{I,j}$ for $j \leq p$. As $m/n \to \gamma \in [0, 1)$, $\sum_{j=1}^{p} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - p$ converges to $p\gamma/(1-\gamma)$.

We consider the general linear model with correlated errors (LMCE).

Example 4. Consider the LMCE given by $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, where $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{R})$. By choosing various $\mathbf{R}$'s, LMCE includes the linear model with independent data, the multivariate linear model, time series models, geostatistical models, and mixed effect models as special cases (Haslett, 1999). Similar to Haslett (1999), we fix $\mathbf{R}$ at an

appropriate estimate $\hat{\mathbf{R}}$ throughout the example. We can calculate the generalized least squares estimator $\hat{\beta} = (\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{Y} = \mathbf{B}\mathbf{Y}$, $\text{var}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X})^{-1}$, and $\hat{\sigma}^2 = \mathbf{Y}^T\mathbf{Q}\mathbf{Y}/(n-p) = \hat{\mathbf{e}}^T\mathbf{R}^{-1}\hat{\mathbf{e}}/(n-p)$, where $\mathbf{Q} = \mathbf{R}^{-1} - \mathbf{H}$, $\hat{\mathbf{e}} = \mathbf{R}\mathbf{Q}\mathbf{Y}$, and $\mathbf{H} = \mathbf{R}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{R}^{-1}$. It has been shown in Haslett (1999) that Cook's distance for deleting the subset I is given by

$$\text{CD}(I) = \frac{1}{p\hat{\sigma}^2}\epsilon^T\mathbf{Q}U_I\mathbf{Q}_{II}^{-1}(\mathbf{R}^{II} - \mathbf{Q}_{II})\mathbf{Q}_{II}^{-1}U_I^T\mathbf{Q}\epsilon, \tag{4.14}$$

where $\mathbf{Q}_{II}$ is the (I, I) subset of $\mathbf{Q}$ and $\mathbf{R}^{II}$ is the (I, I) subset of $\mathbf{R}^{-1}$. With some algebraic calculation, it can be shown that

$$E[\text{CD}(I)|\mathcal{M}] \approx E[\text{tr}(\mathbf{Q}_{II}^{-1}\mathbf{R}^{II})|\mathcal{M}] - m = \sum_{j=1}^{m} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - m, \tag{4.15}$$

where $\lambda_{I,1} \geq \cdots \geq \lambda_{I,m}$ are the ordered eigenvalues of $(\mathbf{R}^{II})^{-1/2}\mathbf{H}_{II}(\mathbf{R}^{II})^{-1/2}$, in which $\mathbf{H}_{II}$ is the (I, I) subset of $\mathbf{H}$. Similar to Theorem 4.3 (b), when $m \geq p$, the right-hand side of (4.15) reduces to $\sum_{j=1}^{p} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}] - p$. In many scenarios such as the multivariate linear model, we can follow the strategies in Example 3 to approximate $\sum_{j=1}^{m} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}]$. However, for time series data, since the elements in $\mathbf{X}$ are responses in an autoregressive (AR(p)) model, we can use the parametric bootstrap method to generate random samples from the fitted model and then approximate $\sum_{j=1}^{m} E[(1-\lambda_{I,j})^{-1}|\mathcal{M}]$.

4.2.4 Conditional Scaled Cook's Distance

In some statistical problems, it may be better to perform influence analysis while fixing some covariates of interest, such as measurement time. For instance, in longitudinal data, since different subjects can have different numbers of measurements and measurement times, which are not covariates of interest in influence analysis, it may be better

to eliminate their effect in calculating Cook's distance. To eliminate the effect of some fixed covariates, we introduce two conditional scaled Cook's distances as follows.

Definition 4.2. *The conditional scaled Cook's distance (CSCD) for matching the mean and the median are, respectively, defined as*

$$CSCD_1(I,\mathbf{Z}) = \frac{CD(I)}{E[CD(I)|\mathcal{M},\mathbf{Z}]} \quad and \quad CSCD_2(I,\mathbf{Z}) = \frac{CD(I)}{Q_{CD(I)}(0.5|\mathcal{M},\mathbf{Z})}, \tag{4.16}$$

where $\mathbf{Z}$ *is the set of some fixed covariates in* $\mathbf{Y}$ *and the expectation and quantiles are taken with respect to the fitted model* $\mathcal{M}$ *given* $\mathbf{Z}$*.*

We can show that $E[\mathrm{CSCD}_1(I,\mathbf{Z})|\mathcal{M},\mathbf{Z}] = 1$ and $Q_{\mathrm{CSCD}_2(I,\mathbf{Z})}(0.5|\mathcal{M},\mathbf{Z}) = 1$ hold for every subset I given $\mathbf{Z}$. Thus, these conditional scaled Cook's distances can be used to evaluate the influential level of different subsets I given $\mathbf{Z}$. Similar to $\mathrm{SCD}_1(I)$ and $\mathrm{SCD}_2(I)$, a large value of $\mathrm{CSCD}_1(I,\mathbf{Z})$ (or $\mathrm{CSCD}_2(I,\mathbf{Z})$) indicates a large influence of the subset I after controlling for $\mathbf{Z}$. It should be noted that because $\mathbf{Z}$ is fixed, the $\mathrm{CSCD}_k(I,\mathbf{Z})$ do not reflect the influential level of $\mathbf{Z}$ and the $\mathrm{CSCD}_k(I,\mathbf{Z})$ may vary across different $\mathbf{Z}$.

For generalized linear models, we can fix all covariates and then calculate CSCDs. First, $pE[\mathrm{CD}(I)^1|\mathcal{M},\mathbf{Z}] \approx \mathrm{tr}[(\mathbf{I}_m - H_{*,I})^{-1}] - m = \sum_{j=1}^{m}(1-\lambda_{I,j})^{-1} - m$, which reduces to $\sum_{j=1}^{p}(1-\lambda_{I,j})^{-1} - p$ for $m \geq p$. Then, the conditional scaled Cook's distance $\mathrm{CSCD}_1(I,\mathbf{X})$ is given by

$$\mathrm{CSCD}_1(I,\mathbf{X}) \approx \frac{\hat{\mathbf{e}}^T\hat{V}^{-1/2}U_I(\mathbf{I}_m - \hat{H}_I)^{-1}\hat{H}_I(\mathbf{I}_m - \hat{H}_I)^{-1}U_I^T\hat{V}^{-1/2}\hat{\mathbf{e}}}{[\sum_{j=1}^{m}(1-\lambda_{I,j})^{-1} - m]}. \tag{4.17}$$

To approximate $Q_{\mathrm{CD}_{(I)}}(0.5|\mathcal{M},\mathbf{Z})$, we can generate responses from the fitted model and then substitute them into (4.12) to obtain a sample of simulated $\mathrm{CD}(I)$ given the covariates. Finally, we can use the empirical median of the simulated $\mathrm{CD}(I)$ to approximate $Q_{\mathrm{CD}_{(I)}}(0.5|\mathcal{M},\mathbf{Z})$ and calculate $\mathrm{CSCD}_2(I,\mathbf{Z})$.

Let's consider cluster deletion in generalized linear mixed models (GLMM).

EXAMPLE 5. Consider a dataset that is composed of a response y_{ij}, covariate vectors $\mathbf{x}_{ij}(p \times 1)$ and $\mathbf{z}_{ij}(p_1 \times 1)$ for observations $j = 1, \cdots, m_i$ within clusters $i = 1, \cdots, n$. The GLMM assumes that conditional on a $p_1 \times 1$ random variable $\mathbf{b}_i$, y_{ij} follows an exponential family distribution of the form (McCullagh and Nelder, 1989)

$$p(y_{ij}|\mathbf{b}_i) = \exp[a_{ij}(\tau)^{-1}\{y_{ij}\eta_{ij} - b(\eta_{ij})\} + c(y_{ij}, \tau)], \tag{4.18}$$

where $\eta_{ij} = k(\mathbf{x}_{ij}^T\beta + \mathbf{z}_{ij}^T\mathbf{b}_i)$ in which $\beta = (\beta_1, \cdots, \beta_p)^T$ and $k(\cdot)$ is a known continuously differentiable function. The distribution of $\mathbf{b}_i$ is assumed to be $N(\mathbf{0}, \Sigma)$, where $\Sigma = \Sigma(\gamma)$ depends on a $p_2 \times 1$ vector γ of unknown variance components. For simplicity, we fix (γ, τ) at an appropriate estimate $(\hat{\gamma}, \hat{\tau})$ throughout the example.

We focus here on cluster deletion in GLMMs. After some calculations, the first order approximation of Cook's distance for deleting the ith cluster is given by

$$\mathrm{CD}(I_i)^1 = \partial_\beta \ell_i(\hat{\beta})^T[\mathbf{F}_n(\hat{\beta}) - \mathbf{f}_i(\hat{\beta})]^{-1}\mathbf{F}_n(\hat{\beta})[\mathbf{F}_n(\hat{\beta}) - \mathbf{f}_i(\hat{\beta})]^{-1}\partial_\beta \ell_i(\hat{\beta}), \tag{4.19}$$

where $I_i = \{(i,1), \cdots, (i, m_i)\}$, $\ell_i(\beta)$ is the log-likelihood function for the ith cluster, $\mathbf{f}_i(\beta) = -\partial_\beta^2 \ell_i(\beta)$ and $\mathbf{F}_n(\beta) = \sum_{i=1}^n \mathbf{f}_i(\beta)$. Note that $\partial_\beta \ell_i(\hat{\beta}) \approx \{\mathbf{I}_p - \mathbf{f}_i(\hat{\beta})[\mathbf{F}_n(\beta_*)]^{-1}\}\partial_\beta \ell_i(\beta_*) + \mathbf{f}_i(\hat{\beta})[F_n(\beta_*)]^{-1}\sum_{j \neq i} \partial_\beta \ell_j(\beta_*)$. Then, conditional on all the covariates and $\{m_1, \cdots, m_n\}$ in $\mathbf{Z}$, we can show that if the fitted model is true, then $E[\mathrm{CD}(I_i)^1|\mathcal{M}, \mathbf{Z}]$ can be approximated by $\mathrm{tr}\{(E[\mathbf{F}_n(\hat{\beta})|\mathcal{M}, \mathbf{Z}] - E[\mathbf{f}_i(\hat{\beta})|\mathcal{M}, \mathbf{Z}])^{-1}E[\mathbf{f}_i(\hat{\beta})|\mathcal{M}, \mathbf{Z}]\}$. Thus, the conditional scaled Cook's distance $\mathrm{CSCD}_1(I_i, \mathbf{Z})$ is given by

$$\mathrm{CSCD}_1(I_i, \mathbf{Z}) \approx \frac{\partial_\beta \ell_i(\hat{\beta})^T[\mathbf{F}_n(\hat{\beta}) - \mathbf{f}_i(\hat{\beta})]^{-1}\mathbf{F}_n(\hat{\beta})[\mathbf{F}_n(\hat{\beta}) - \mathbf{f}_i(\hat{\beta})]^{-1}\partial_\beta \ell_i(\hat{\beta})}{\mathrm{tr}\{(E[\mathbf{F}_n(\hat{\beta})|\mathcal{M}, \mathbf{Z}] - E[\mathbf{f}_i(\hat{\beta})|\mathcal{M}, \mathbf{Z}])^{-1}E[\mathbf{f}_i(\hat{\beta})|\mathcal{M}, \mathbf{Z}]\}}. \tag{4.20}$$

To approximate $Q_{\mathrm{CD}_{(I_i)}}(0.5|\mathcal{M}, \mathbf{Z})$, we can generate responses from the fitted GLMM

and then substitute them into (4.19) to obtain a sample of simulated $\mathrm{CD}(I_i)^1$ given $\mathbf{Z}$. Finally, we can use the empirical median of the simulated $\mathrm{CD}(I_1)^1$ to approximate $Q_{\mathrm{CD}_{(I_i)}}(0.5|\mathcal{M},\mathbf{Z})$ and calculate $\mathrm{CSCD}_2(I_i,\mathbf{Z})$.

Finally, we consider a large class of parametric models for both independent and dependent data and develop the associated (conditional) scaled Cook's distance. Let $p(\mathbf{Y}_{[I]},\theta)$ be the probability function for the full data with all observations in the set I deleted. Then, $p(\mathbf{Y}_I|\mathbf{Y}_{[I]},\theta)$ is the conditional distribution of $\mathbf{Y}_I$ given $\mathbf{Y}_{[I]}$. We obtain the following theorem.

Theorem 4.4. *If Assumptions C1-C4 in the Appendix hold and $m(I)/n \to \gamma \in [0,1)$, where $m(I)$ denotes the size of I, then we have the following results:*
(a) CD(I) can be approximated by

$$\partial_\theta \log p(\mathbf{Y}_I|\mathbf{Y}_{[I]},\hat{\theta})^T[\mathbf{F}_n(\hat{\theta})-\mathbf{f}_I(\hat{\theta})]^{-1}\mathbf{F}_n(\hat{\theta})[\mathbf{F}_n(\hat{\theta})-\mathbf{f}_I(\hat{\theta})]^{-1}\partial_\theta \log p(\mathbf{Y}_I|\mathbf{Y}_{[I]},\hat{\theta}), \quad (4.21)$$

where $\mathbf{F}_n(\theta)=-\partial_\theta^2 \log p(\mathbf{Y},\theta)$ and $\mathbf{f}_I(\theta)=-\partial_\theta^2 \log p(\mathbf{Y}_I|\mathbf{Y}_{[I]},\theta)$;
(b) $E[CD(I)|\mathcal{M}] \approx tr(\{E[\mathbf{F}_n(\hat{\theta})|\mathcal{M}]-E[\mathbf{f}_I(\hat{\theta})|\mathcal{M}]\}^{-1}E[\mathbf{f}_I(\hat{\theta})|\mathcal{M}])$;
(c) $E[CD(I)|\mathcal{M},\mathbf{Z}] \approx tr(\{E[\mathbf{F}_n(\hat{\theta})|\mathcal{M},\mathbf{Z}]-E[\mathbf{f}_I(\hat{\theta})|\mathcal{M},\mathbf{Z}]\}^{-1}E[\mathbf{f}_I(\hat{\theta})|\mathcal{M},\mathbf{Z}])$.

Proof of Theorem 4.4.
(a) It follows from a Taylor's series expansion and assumption (C2) that

$$\partial_\theta \log p(\mathbf{Y}_{[I]},\hat{\theta}_{[I]}) = \mathbf{0} = \partial_\theta \log p(\mathbf{Y}_{[I]},\hat{\theta}) + \partial_\theta^2 \log p(\mathbf{Y}_{[I]},\tilde{\theta})(\hat{\theta}_{[I]}-\hat{\theta}),$$

where $\tilde{\theta}=t\hat{\theta}_{[I]}+(1-t)\hat{\theta}$ for $t\in[0,1]$. Combining this with Assumption (C3) and the fact that $\partial_\theta \log p(\mathbf{Y},\hat{\theta}) = \partial_\theta \log p(\mathbf{Y}_{[I]},\hat{\theta}) + \partial_\theta \log p(\mathbf{Y}_I|\mathbf{Y}_{[I]},\hat{\theta}) = \mathbf{0}$, we get

$$\begin{aligned}\hat{\theta}_{[I]}-\hat{\theta} &= [-\partial_\theta^2 \log p(\mathbf{Y}_{[I]},\hat{\theta})]^{-1}\partial_\theta \log p(\mathbf{Y}_{[I]},\hat{\theta})[1+o_p(1)] \\ &= -[-\partial_\theta^2 \log p(\mathbf{Y}_{[I]},\hat{\theta})]^{-1}\partial_\theta \log p(\mathbf{Y}_I|\mathbf{Y}_{[I]},\hat{\theta})[1+o_p(1)]. \quad (4.22)\end{aligned}$$

Substituting (4.22) into $\mathrm{CD}(I) = (\hat{\theta}_{[I]} - \hat{\theta})^T \mathbf{F}_n(\hat{\theta})(\hat{\theta}_{[I]} - \hat{\theta})$, completes the proof of Theorem 4.4 (a).

(b) It follows from Assumptions (C1)-(C3) that

$$\begin{aligned} \hat{\theta} - \theta_* &= \mathbf{F}_n(\theta_*)^{-1} \partial_\theta \log p(\mathbf{Y}, \theta_*)[1 + o_p(1)] \\ &= \mathbf{F}_n(\theta_*)^{-1} [\partial_\theta \log p(\mathbf{Y}_{[I]}, \theta_*) + \partial_\theta \log p(\mathbf{Y}_I | \mathbf{Y}_{[I]}, \theta_*)][1 + o_p(1)]. \end{aligned}$$

Let $J_I(\theta) = \partial_\theta \log p(\mathbf{Y}_I | \mathbf{Y}_{[I]}, \theta)$. Using a Taylor's series expansion along with Assumptions (C3) and (C4), we get

$$\begin{aligned} & J_I(\hat{\theta}) \approx J_I(\theta_*) - \mathbf{f}_I(\theta_*)(\hat{\theta} - \theta_*) \approx J_I(\theta_*) - E[\mathbf{f}_I(\theta_*)|\mathcal{M}](\hat{\theta} - \theta_*) \qquad (4.23) \\ = \; & \{\mathbf{I}_p - E[\mathbf{f}_I(\theta)|\mathcal{M}]\mathbf{F}_n(\theta_*)^{-1}\} J_I(\theta_*) - E[\mathbf{f}_I(\theta)|\mathcal{M}]\mathbf{F}_n(\theta_*)^{-1} \partial_\theta \log p(\mathbf{Y}_{[I]}, \theta_*). \end{aligned}$$

Since $E[J_I(\theta_*)\partial_\theta \log p(\mathbf{Y}_{[I]}, \theta_*)|\mathcal{M}] = \mathbf{0}$,

$$E[J_I(\hat{\theta}) J_I(\hat{\theta})^T | \mathcal{M}] \approx E[\mathbf{f}_I(\theta_*)|\mathcal{M}]\mathbf{F}_n(\theta_*)^{-1}\{\mathbf{F}_n(\theta_*) - E[\mathbf{f}_I(\theta_*)|\mathcal{M}]\}.$$

It follows from Assumption (C3) that for θ in a neighborhood of θ_*, $\mathbf{F}_n(\theta)$ and $\mathbf{F}_n(\theta_*) - \mathbf{f}_I(\theta)$ can be replaced by $E[\mathbf{F}_n(\theta)|\mathcal{M}]$ and $E[\mathbf{F}_n(\theta_*) - \mathbf{f}_I(\theta)|\mathcal{M}]$, respectively, which completes the proof of Theorem 4.4 (b).

(c) Similar to Theorem 4.4 (b), we can prove Theorem 4.4 (c).

Theorem 4.4 has several important implications for general parametric models. Theorem 4.4 (a) establishes the first order approximation of Cook's distance for a large class of parametric models for both dependent and independent data. Theorem 4.4 (b) and (c) give an approximation of $E[\mathrm{CD}(I)|\mathcal{M}]$ and $E[\mathrm{CD}(I)|\mathcal{M}, \mathbf{Z}]$, respectively. Based on these results, we can construct the scaled (or conditional scaled) Cook's distance measure.

4.3 Illustrative Examples

4.3.1 Finney Data

We consider a logistic regression model using data from Finney (1947). The data were obtained to study the effect of the rate and volume of air inspired on a transient vaso-constriction in the skin of the digits. There are 39 observations in the data set. The response variable is occurrence or nonoccurrence of vaso-constriction. These data have been considered in many papers (Pregibon, 1981; Ibrahim and Laud, 1991; Chen and Ibrahim, 2003).

We fitted the logistic regression model with $\text{logit}(p) = \beta_0 + \beta_1 \log(\text{rate}) + \beta_2 \log(\text{volume})$, where p is the probability of the occurrence of vaso-constriction. We computed CD, SCD_1, SCD_2, CSCD_1 and CSCD_2 as described in Section 4.2. Specifically, CD was computed using the first order approximation in (11), and SCD_1, SCD_2, CSCD_1 and CSCD_2 were computed using 500 bootstrap samples. We considered single case deletion as well as multiple case deletion by deleting two and three cases simultaneously.

For single case deletion, the 4th and the 18th observations are identified as influential cases by all of the proposed Cook's distance measures (Figure 4.2). For simultaneous two case deletion, the pairs involving either the 4th or the 18th case become influential and mosaic patterns are found for those pairs (Figure 4.3). Although similar mosaic patterns are found in Figure 4.3, the influence of the pair (4, 18) is dominant compared to those of the other pairs when influence is measured by CD, SCD_1 and SCD_2. On the other hand, CSCD_1 and CSCD_2 identified more pairs involving either the 4th or the 18th case other than (4, 18). In addition, the pair (4, 18) was identified as the most influential set by CD, SCD_1, SCD_2 and CSCD_1 (CD=1.856, SCD_1=24.506, SCD_2=73.926 and CSCD_1=19.208); however, CSCD_2 identified pair (4, 32) as the most influential set (CSCD_2=153.173), (4, 10) as the second most influential set (CSCD_2=144.611) and (4,

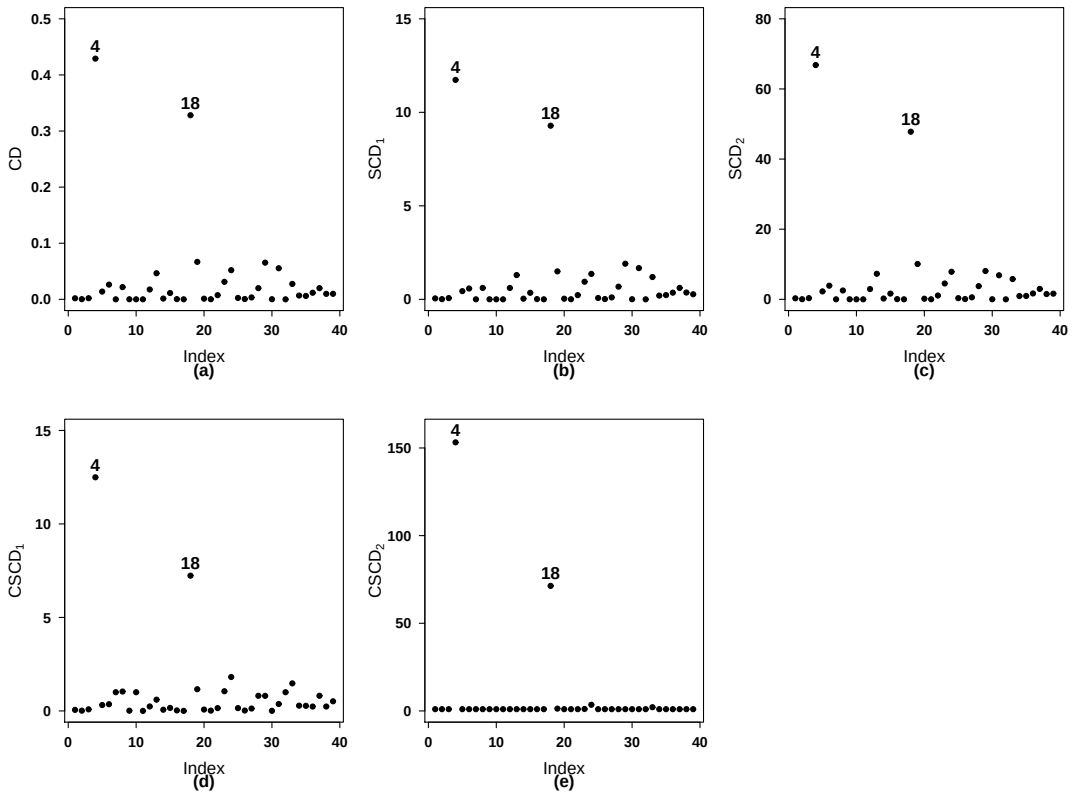

FIGURE 4.2: *Finney data.* Index plots for single case deletion; (a) CD, (b) SCD_1, (c) SCD_2, (d) $CSCD_1$ and (e) $CSCD_2$.

18) as the 6th most influential set ($CSCD_2$=103.587). The pair (4, 32) was the second most influential set by $CSCD_1$ ($CSCD_1$=12.493) but not a highly influential set by the other measures. We observe a similar phenomenon for simultaneous three case deletion (Figure 4.4). Although the sets identified by CD, SCD_1, SCD_2 have similar patterns among those measures, $CSCD_1$ and $CSCD_2$ identified more influential sets. Specifically, the most influential set is (4, 18, 29) by using CD, SCD_1 and SCD_2 (CD=2.409, SCD_1=19.836 and SCD_2=55.771), whereas (4, 18, 32) is most influential by $CSCD_1$ ($CSCD_1$=19.208) and (4, 10, 32) is most influential by $CSCD_2$ ($CSCD_2$=144.601).

Figure 4.5 shows the relationship between the size of the deletion and the performance of the scaled Cook's distance. For the unscaled Cook's distance, as the number of deleted cases increases, the density of CD shifts toward larger values. For the scaled and conditional scaled Cook's distance, the densities are not shifted and thus the modes

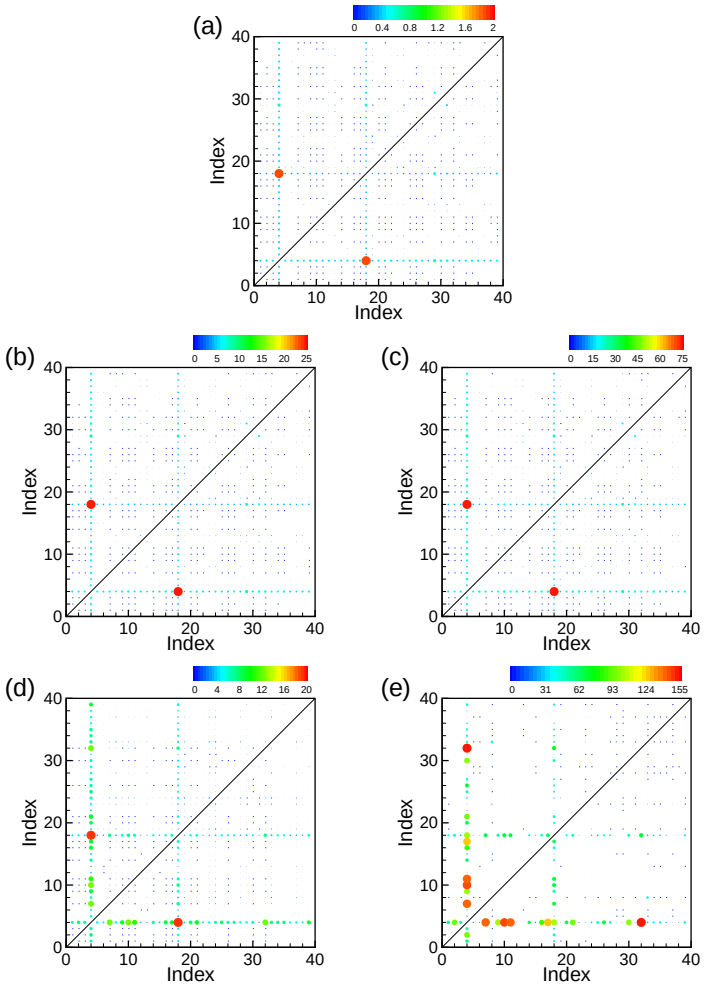

FIGURE 4.3: *Finney data.* 2-D scatter plots for simultaneous two case deletion; (a) CD, (b) SCD_1, (c) SCD_2, (d) $CSCD_1$ and (e) $CSCD_2$. Note that the colors represent the magnitude of each Cook's distance and the size of the symbol is proportional to the magnitude.

of the density occur at the same location for single, two and three case deletion. We can also observe that $CSCD_2$ is robust to the size of deletion. This indicates that the proposed scaled and conditional scaled Cook's distance eliminate the size issue altogether.

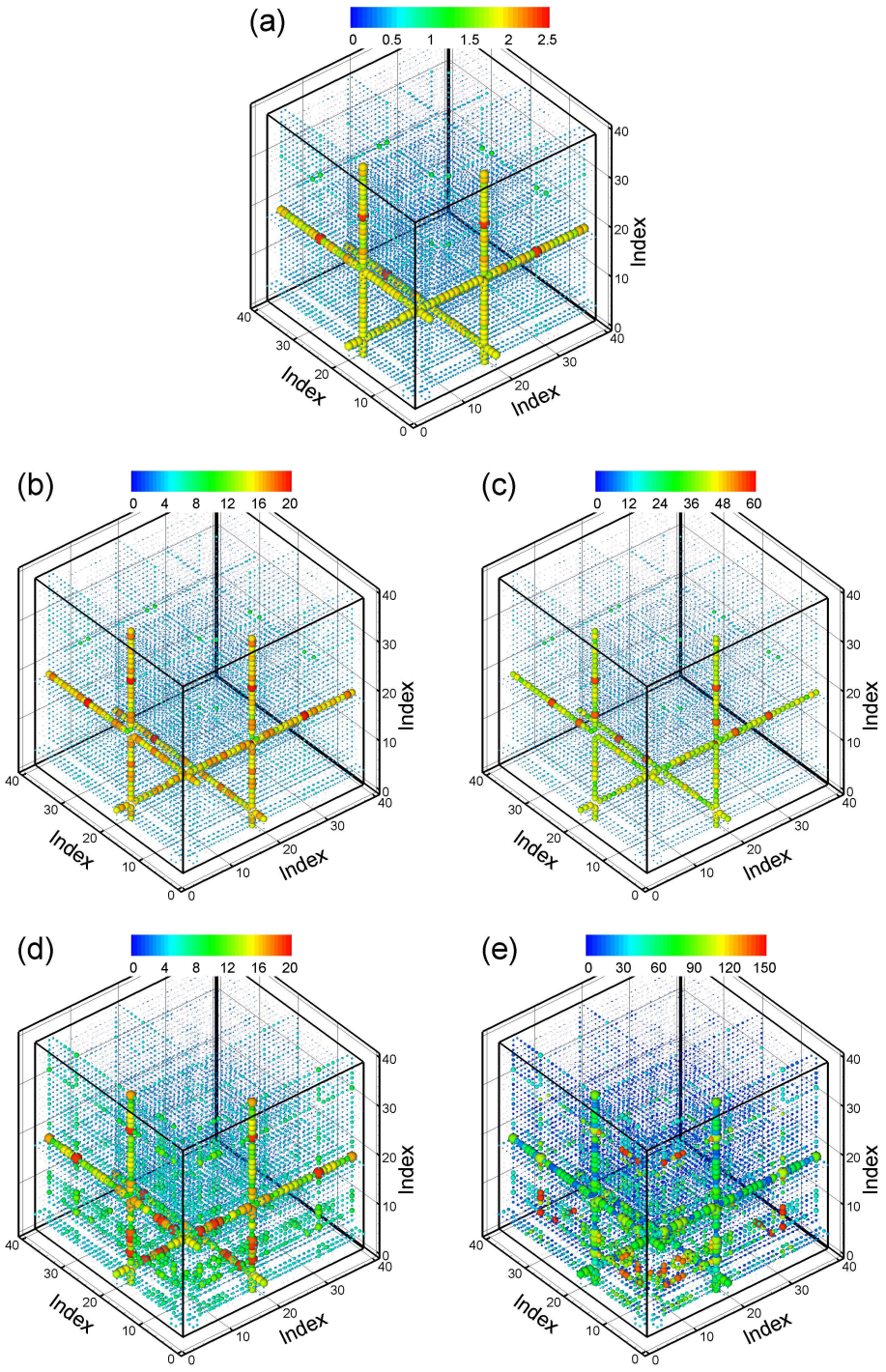

FIGURE 4.4: *Finney data.* 3-D scatter plots for simultaneous three case deletion; (a) CD, (b) SCD_1, (c) SCD_2, (d) $CSCD_1$ and (e) $CSCD_2$. Note that the colors represent the magnitude of each Cook's distance and the size of the symbol is proportional to the magnitude.

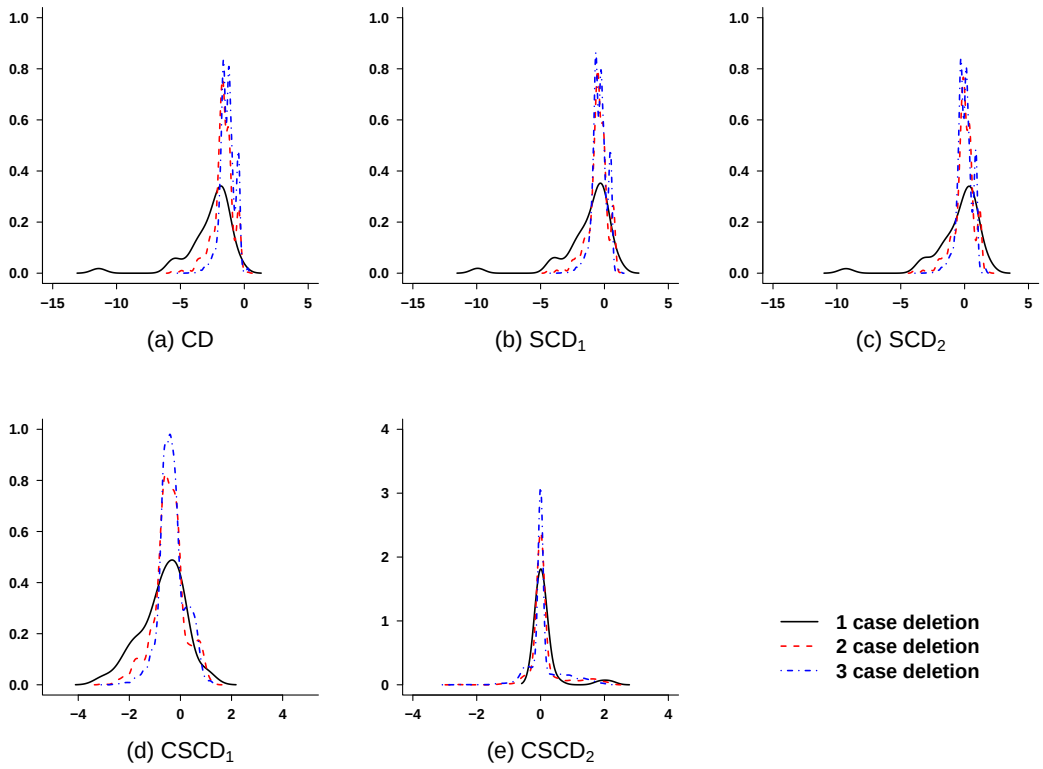

FIGURE 4.5: *Finney data.* Density plots in $\log_{10}$-scale of (a) CD, (b) SCD_1, (c) SCD_2, (d) CSCD_1 and (e) CSCD_2.

4.3.2 Yale Infant Growth Data

We computed CD using the first order approximation in (4.19), CSCD_1 and CSCD_2 using 100 bootstrap samples as described in Section 4.2.4. We consider single subject deletion as well as simultaneous two subject deletion. A deletion of a subject here means that we delete the observations for all of the visits for that subject. Subjects 269, 217, 294, 289, 274, 90, 38, 285, 280 and 149 are identified as the top 10 most influential observations by CD, whereas subjects 217, 274, 246, 109, 90, 289, 294, 269, 38 and 149 are identified using the conditional scaled Cook's distance (Table 4.1 and Figure 4.6). Specifically, subject 269 is the most influential according to CD, whereas it is less influential according to the conditional scaled Cook's distance. On the other hand, subject 246 is not influential according to CD (CD=0.253), but becomes influential

TABLE 4.1: *Yale infant growth data.* Top 10 influential subjects for single case deletion with compound symmetry model.

ID	m_i	CD	ID	m_i	$CSCD_1$	ID	m_i	App. $CSCD_1$	ID	m_i	$CSCD_2$
269	12	2.416	274	22	24.911	274	22	23.769	217	19	30.196
217	19	1.465	217	19	23.734	217	19	23.544	274	22	27.809
294	13	1.252	246	5	19.801	109	12	19.584	246	5	25.264
289	18	1.188	90	17	18.151	90	17	18.875	109	12	24.761
274	22	1.163	109	12	18.058	294	13	18.653	90	17	22.526
90	17	0.858	149	17	17.094	246	5	17.168	289	18	20.682
38	24	0.823	294	13	16.904	149	17	16.974	294	13	20.324
285	8	0.738	289	18	16.537	289	18	16.732	269	12	19.835
280	9	0.695	38	24	16.386	269	12	15.943	38	24	18.719
149	17	0.668	269	12	14.168	38	24	15.893	149	17	18.696

Note that m_i represents cluster size and App. $CSCD_1$ is computed by equation (4.20).

after eliminating the effect of the cluster size (Table 4.1). The relationship between Cook's distance and cluster size indicates that the influential observations have larger CD values for larger cluster sizes (Figure 4.7 (a)), while cluster size does not affect the magnitude of the conditional scaled Cook's distance (Figure 4.7 (b)-(d)). Thus, we expect uniform patterns across cluster size in the conditional scaled Cook's distance. For simultaneous two subject deletion, the sets involving subject 269 were identified as influential using CD, whereas the conditional scaled Cook's distance detects the sets involving subjects 217 and 274 as most influential compared to other sets (Figure 4.8).

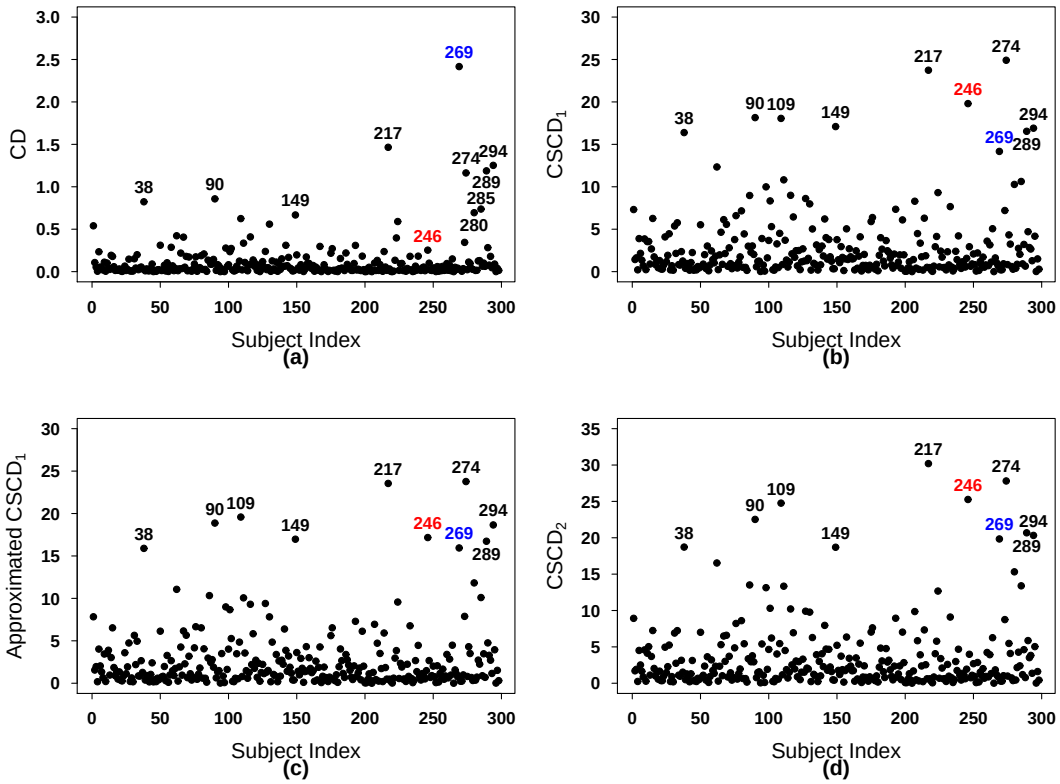

FIGURE 4.6: *Yale infant growth data.* Index plots for single subject deletion with compound symmetry model; (a) CD, (b) $CSCD_1$, (c) Approximation of $CSCD_1$ is computed by equation (4.20) and (d) $CSCD_2$.

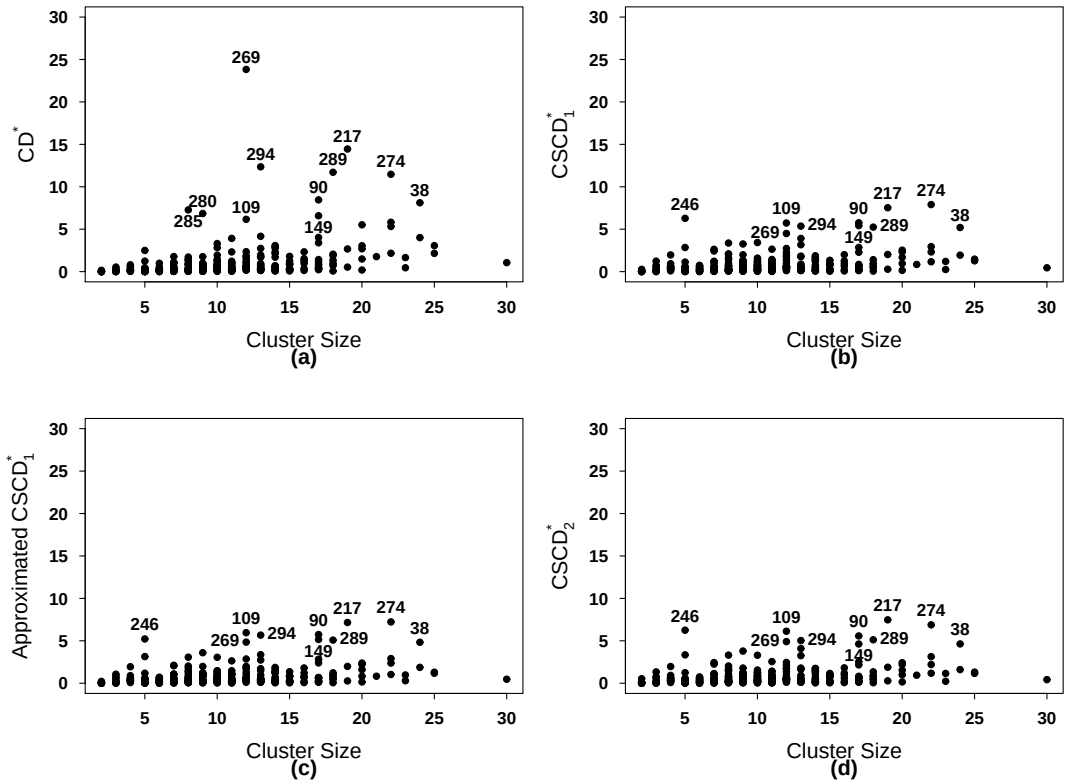

FIGURE 4.7: *Yale infant growth data.* Cluster size versus Cook's distance for single subject deletion with compound symmetry model; (a) CD, (b) CSCD_1, (c) Approximation of CSCD_1 is computed by equation (4.20) and (d) CSCD_2. Note that the superscript * indicates that each Cook's distance is divided by its 75^{th} percentile to compare the relative sizes between the Cook's distances. The reason we choose the 75^{th} percentile instead of the maximum is that the maximum Cook's distance is not robust to influential observations.

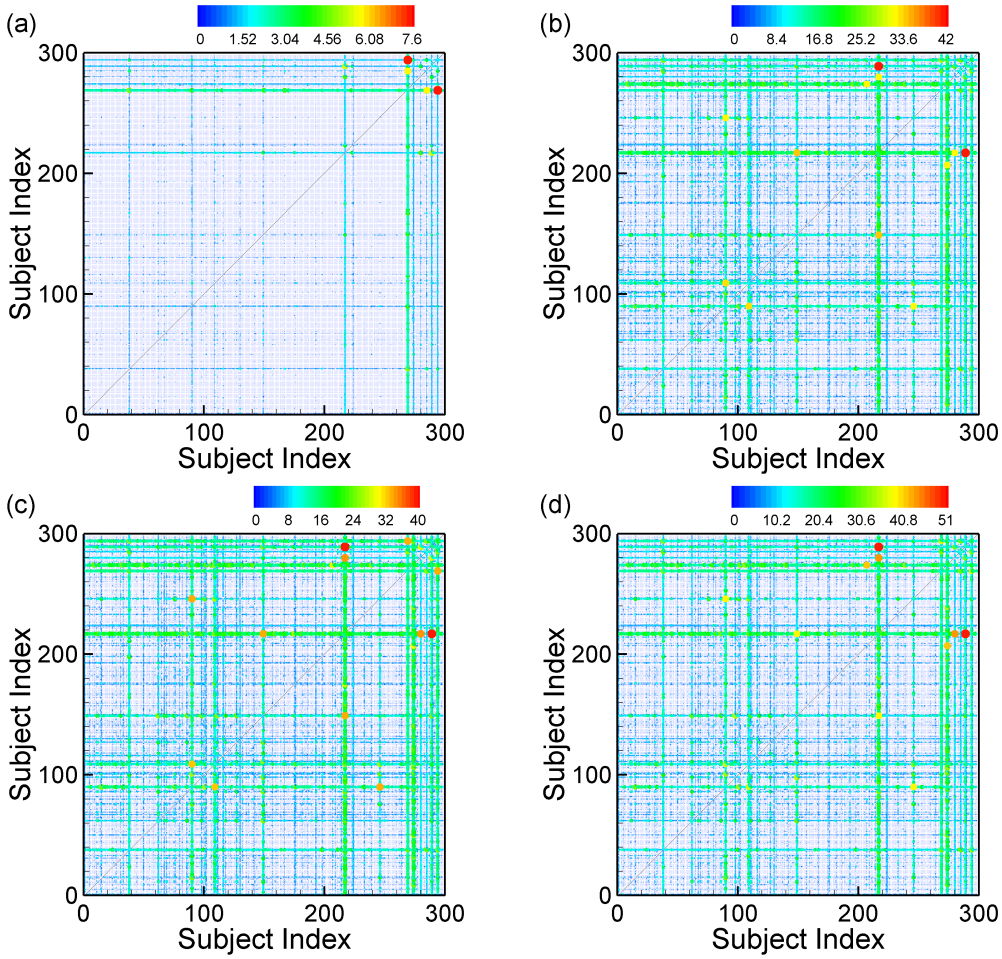

FIGURE 4.8: *Yale infant growth data.* 2-D scatter plots for simultaneous two subjects deletion with compound symmetry model; (a) CD, (b) $CSCD_1$, (c) Approximation of $CSCD_1$ is computed by equation (4.20) and (d) $CSCD_2$. Note that the colors represent the magnitude of each Cook's distance and the size of the symbol is proportional to the magnitude.

4.4 Discussion

We have developed a scaled Cook's distance to address the issue of size matters for deletion diagnostics in general parametric models. We have used stochastic ordering to quantify the relationship between the size of the perturbation and the amount of the perturbation in Cook's distance. We have shown that the scaled Cook's distance provide important information about outliers and influential observations for a fitted model to a given data set. We have illustrated our development with linear regression, generalized linear models, linear mixed models, and generalized linear mixed models. We have analyzed two datasets using the scaled Cook's distance measure. Future work includes developing Bayesian analog's to Cook's scaled distance measure and developing such a methodology for other types of models, such as survival models and models with missing covariate data.

CHAPTER 5

DISCUSSION

We have proposed diagnostic methods for assessing the influence of observations on model fit and complexity for Bayesian regression models. These models include linear models, mixed models, generalized linear models, generalized linear mixed models and survival models. The proposed methods provide efficient computational formula and nice statistical properties such as approximations and asymptotic equivalence. We can view $\mathrm{D}_\phi(i)$ as a Bayesian analogue of the likelihood distance (Cook and Weisberg, 1982) and Cook's posterior mean distance (or Cook's posterior mode distance) as a Bayesian analogue of Cook's distance (Cook, 1977). In the frequentist paradigm, there is an asymptotic equivalence between the likelihood distance and Cook's distance. Analogously, we showed that $\mathrm{D}_\phi(i)$ and Cook's posterior mean distance are asymptotically equivalent in the Bayesian paradigm. The implementation of the proposed diagnostic measures does not involve intensive computation, and only requires MCMC samples from the full posterior distribution.

We have also presented a scaled Cook's distance to address the issue of size matters for the deletion diagnostics in general parametric models. The proposed scaled and conditional scaled Cook's distance eliminate the size issues when we consider deleting a set of observations, and particularly useful for analyzing longitudinal data with different cluster size.

The issue of what to do in a statistical analysis once an influential observation has been detected is a huge issue with no easy answer. Most researchers in this area recommend that i) analyses with and without the influential case should be clearly reported, indicating differences in point and interval estimates, as well as variance estimates, ii) If one seeks remedies to the problem, three strategies are typically mentioned: one can transform the data, re-parameterize the model, or fit a new model all together. Remedies for influential observations is a very large research area on its own.

APPENDIX A

Proofs in Chap. 2

In this subsection, we provide the proof of equations in Sections 2.2.1, 2.3.2 and 2.4.1.

Proof of equation (2.2) in Section 2.2.1:

$$\begin{aligned} K(P, P_{-i}) &= \int p(\boldsymbol{\beta}|D) \log \left\{ \frac{L(\boldsymbol{\beta}|D)\pi(\boldsymbol{\beta})/C}{L(\boldsymbol{\beta}|D_{-i})\pi(\boldsymbol{\beta})/C_{-i}} \right\} d\boldsymbol{\beta} \\ &= \int p(\boldsymbol{\beta}|D) \log \left\{ \frac{C_{-i}}{C} \right\} d\boldsymbol{\beta} + \int p(\boldsymbol{\beta}|D) \log \left\{ \frac{L(\boldsymbol{\beta}|D)}{L(\boldsymbol{\beta}|D_{-i})} \right\} d\boldsymbol{\beta} \\ &= \log E_{\boldsymbol{\beta}} \left[\frac{L(\boldsymbol{\beta}|D_{-i})}{L(\boldsymbol{\beta}|D)} \middle| D \right] + E_{\boldsymbol{\beta}} \left[\log \left\{ \frac{L(\boldsymbol{\beta}|D)}{L(\boldsymbol{\beta}|D_{-i})} \right\} \middle| D \right], \end{aligned} \tag{A.1}$$

where $C = \int L(\boldsymbol{\beta}|D)\pi(\boldsymbol{\beta})d\boldsymbol{\beta}$ and $C_{-i} = \int L(\boldsymbol{\beta}|D_{-i})\pi(\boldsymbol{\beta})d\boldsymbol{\beta}$. Moreover,

$$\log \left\{ \frac{C_{-i}}{C} \right\} = \log \left\{ \int \frac{L(\boldsymbol{\beta}|D_{-i})}{L(\boldsymbol{\beta}|D)} \times \frac{L(\boldsymbol{\beta}|D)\pi(\boldsymbol{\beta})}{C} \, d\boldsymbol{\beta} \right\} = \log E_{\boldsymbol{\beta}} \left[\frac{L(\boldsymbol{\beta}|D_{-i})}{L(\boldsymbol{\beta}|D)} \middle| D \right]. \tag{A.2}$$

Proof of equation (2.4) in Section 2.2.1:

$$\begin{aligned} K(P_1, P_{1,-i}) &= \int p_1(\boldsymbol{\beta_1}|D) \log \left\{ \frac{\int L(\boldsymbol{\beta}|D)\pi(\boldsymbol{\beta})/C \, d\boldsymbol{\beta_2}}{\int L(\boldsymbol{\beta}|D_{-i})\pi(\boldsymbol{\beta})/C_{-i} \, d\boldsymbol{\beta_2}} \right\} d\boldsymbol{\beta_1} \\ &= \int p_1(\boldsymbol{\beta_1}|D) \log \left\{ \frac{C_{-i}}{C} \right\} d\boldsymbol{\beta_1} + \int p_1(\boldsymbol{\beta_1}|D) \log \frac{\int L(\boldsymbol{\beta}|D)\pi(\boldsymbol{\beta})d\boldsymbol{\beta_2}}{\int \frac{L(\boldsymbol{\beta}|D_{-i})}{L(\boldsymbol{\beta}|D)} L(\boldsymbol{\beta}|D)\pi(\boldsymbol{\beta})d\boldsymbol{\beta_2}} d\boldsymbol{\beta_1} \\ &= \log \left\{ \frac{C_{-i}}{C} \right\} - \int p_1(\boldsymbol{\beta_1}|D) \log \left[\int \frac{L(\boldsymbol{\beta}|D_{-i})}{L(\boldsymbol{\beta}|D)} \times \frac{L(\boldsymbol{\beta}|D)\pi(\boldsymbol{\beta})/C}{\int L(\boldsymbol{\beta}|D)\pi(\boldsymbol{\beta})/C d\boldsymbol{\beta_2}} d\boldsymbol{\beta_2} \right] d\boldsymbol{\beta_1} \\ &= \log E_{\boldsymbol{\beta}} \left[\frac{L(\boldsymbol{\beta}|D_{-i})}{L(\boldsymbol{\beta}|D)} \middle| D \right] - E_{\boldsymbol{\beta_1}} \left[\log \int \frac{L(\boldsymbol{\beta}|D_{-i})}{L(\boldsymbol{\beta}|D)} p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, D) d\boldsymbol{\beta_2} \middle| D \right], \end{aligned} \tag{A.3}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta_1}, \boldsymbol{\beta_2})$ and $p(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}, D) = p(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}|D) / \int p(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}|D) d\boldsymbol{\beta_2}$.

Proof of equation (2.9) in Section 2.3.1:

Following a similar justification as in Sinha et al. (2003), we have

$$\begin{aligned} & P(\boldsymbol{Y} > \boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}, H_0)_{k\neq i} \\ &= \exp\left\{-\sum_{k=1,k\neq i}^{n} H_0(y_k)\exp(\boldsymbol{x}_k'\boldsymbol{\beta})\right\} \\ &= \exp\left[-\left\{\sum_{k=1}^{i-1} h_k(A_k - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})) + h_i(A_i - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})) + \sum_{k=i+1}^{n} h_k A_k\right\}\right], \end{aligned} \tag{A.4}$$

where $A_k = \sum_{l\in\mathcal{R}(y_k)} \exp(\boldsymbol{x}_l'\boldsymbol{\beta})$ and $\mathcal{R}(y_k) = \{l : y_l \geq y_k\}$.

Letting E_{GP} denote expectation with respect to the gamma process prior, we have

$$\begin{aligned} & E_{GP}\{P(\boldsymbol{Y} > \boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}, H_0)_{k\neq i}\} \\ &= \prod_{k=1}^{i-1}\left(\frac{c}{c + A_k - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right)^{ch_{0k}} \left(\frac{c}{c + A_i - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right)^{ch_{0i}} \prod_{k=i+1}^{n}\left(\frac{c}{c + A_k}\right)^{ch_{0k}} \\ &= \prod_{k=1}^{i-1} \exp\left[cH^*(y_k)\log\left\{\frac{c + A_{k+1} - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{c + A_k - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}\right] \\ &\times \prod_{k=i+1}^{n} \exp\left[cH^*(y_k)\log\left\{\frac{c + A_{k+1}}{c + A_k}\right\}\right] \\ &= \prod_{k=1}^{i-1} \exp\left[cH^*(y_k)\log\left\{1 - \frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_k - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}\right] \\ &\times \prod_{k=i+1}^{n} \exp\left[cH^*(y_k)\log\left\{1 - \frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta})}{c + A_k}\right\}\right], \end{aligned} \tag{A.5}$$

where $h_{0k} = H^*(y_k) - H^*(y_{k-1})$.

Now, we write $L(\boldsymbol{\beta}|D_{-i})$ as $L(\boldsymbol{\beta}|D_{-i}) = \{\prod_{k=1}^{i-1} L_{k,-i}(\boldsymbol{\beta}|D)\} \cdot \{\prod_{k=i+1}^{n} L_k(\boldsymbol{\beta}|D)\}$. Then, the likelihood function with the ith subject deleted can be obtained as in equation (9).

Proof of equations (2.10) and (2.13) in Section 2.3.2:

The ratio of the likelihood for full data and the data without the ith subject is written

as

$$\frac{L(\boldsymbol{\beta}|D)}{L(\boldsymbol{\beta}|D_{-i})} = \frac{\prod_{k=1}^{n} L_k(\boldsymbol{\beta}|D)}{\prod_{k=1}^{i-1} L_{k,-i}(\boldsymbol{\beta}|D) \prod_{k=i+1}^{n} L_k(\boldsymbol{\beta}|D)} = g_i(\boldsymbol{\beta}) L_i(\boldsymbol{\beta}|D), \tag{A.6}$$

where $g_i(\boldsymbol{\beta}) = \prod_{k=1}^{i-1} L_k(\boldsymbol{\beta}|D) / \prod_{k=1}^{i-1} L_{k,-i}(\boldsymbol{\beta}|D)$.

We compute CPO as follows:

$$\begin{aligned} CPO_i &= \int L_i(\boldsymbol{\beta}|D) p(\boldsymbol{\beta}|D_{-i}) d\boldsymbol{\beta} \\ &= \int L_i(\boldsymbol{\beta}|D) \frac{\frac{L(\boldsymbol{\beta}|D_{-i})}{L(\boldsymbol{\beta}|D)} L(\boldsymbol{\beta}|D)\pi(\boldsymbol{\beta})/C}{\int \frac{L(\boldsymbol{\beta}|D_{-i})}{L(\boldsymbol{\beta}|D)} L(\boldsymbol{\beta}|D)\pi(\boldsymbol{\beta})/C \; d\boldsymbol{\beta}} \, d\boldsymbol{\beta} \\ &= \frac{\int L_i(\boldsymbol{\beta}|D)\{g_i(\boldsymbol{\beta}) L_i(\boldsymbol{\beta}|D)\}^{-1} p(\boldsymbol{\beta}|D) d\boldsymbol{\beta}}{\int \{g_i(\boldsymbol{\beta}) L_i(\boldsymbol{\beta}|D)\}^{-1} p(\boldsymbol{\beta}|D) d\boldsymbol{\beta}} \\ &= \frac{E_{\boldsymbol{\beta}}[\{g_i(\boldsymbol{\beta})\}^{-1}|D]}{E_{\boldsymbol{\beta}}[\{g_i(\boldsymbol{\beta}) L_i(\boldsymbol{\beta}|D)\}^{-1}|D]} \; . \end{aligned} \tag{A.7}$$

Thus, we have $E_{\boldsymbol{\beta}}[\{g_i(\boldsymbol{\beta}) L_i(\boldsymbol{\beta}|D)\}^{-1}|D] = E_{\boldsymbol{\beta}}[\{g_i(\boldsymbol{\beta})\}^{-1}|D]/CPO_i$ and applying these results to (A.1) completes the proof.

Proof of equations (2.45) in Section 2.5.2 :

$$\begin{aligned} CPO_i &= p(z_i|D_{-i})|_{z_i \in I_a} = \iint p(z_i|\boldsymbol{\beta}, \boldsymbol{h}) p(\boldsymbol{\beta}, \boldsymbol{h}|D_{-i}) \, d\boldsymbol{h} \, d\boldsymbol{\beta} \Big|_{z_i \in I_a} \\ &= \iint p(z_i|\boldsymbol{\beta}, \boldsymbol{h}) \frac{\frac{L(\boldsymbol{\beta}, \boldsymbol{h}|D_{-i})}{L(\boldsymbol{\beta}, \boldsymbol{h}|D)} L(\boldsymbol{\beta}, \boldsymbol{h}|D)\pi(\boldsymbol{h})\pi(\boldsymbol{\beta})}{\iint \frac{L(\boldsymbol{\beta}, \boldsymbol{h}|D_{-i})}{L(\boldsymbol{\beta}, \boldsymbol{h}|D)} L(\boldsymbol{\beta}, \boldsymbol{h}|D)\pi(\boldsymbol{h})\pi(\boldsymbol{\beta}) \, d\boldsymbol{h} \, d\boldsymbol{\beta}} \, d\boldsymbol{h} \, d\boldsymbol{\beta} \Big|_{z_i \in I_a} \\ &= \frac{\iint p(z_i|\boldsymbol{\beta}, \boldsymbol{h}) g_i(\boldsymbol{\beta}, h_a)^{-1} p(\boldsymbol{\beta}, \boldsymbol{h}|D) \, d\boldsymbol{h} \, d\boldsymbol{\beta}|_{z_i \in I_a}}{\iint g_i(\boldsymbol{\beta}, h_a)^{-1} p(\boldsymbol{\beta}, \boldsymbol{h}|D) \, d\boldsymbol{h} \, d\boldsymbol{\beta}}, \end{aligned} \tag{A.8}$$

where $L(\boldsymbol{\beta}, \boldsymbol{h}|D)/L(\boldsymbol{\beta}, \boldsymbol{h}|D_{-i}) = (1-\delta_i)(1-h_a)^{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})} + \delta_i \left\{1 - (1-h_a)^{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}$ and we define $L(\boldsymbol{\beta}, \boldsymbol{h}|D)/L(\boldsymbol{\beta}, \boldsymbol{h}|D_{-i}) = g_i(\boldsymbol{\beta}, h_a)$. Note that $p(z_i|\boldsymbol{\beta}, \boldsymbol{h})|_{z_i \in I_a} = g_i(\boldsymbol{\beta}, h_a)$ for beta process model with grouped survival data. Therefore, we have $CPO_i = [E_{\boldsymbol{\beta}, \boldsymbol{h}}[\{g_i(\boldsymbol{\beta}, h_a)\}^{-1}|D]]^{-1}$.

APPENDIX B

Assumptions and Proofs in Chap. 3

Assumptions:

We define $F_N(\boldsymbol{\theta}) = \partial_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{Y})$ and $F_{N,[S]}(\boldsymbol{\theta}) = \partial_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{Y}_{[S]})$. Even though $p(\boldsymbol{Y}|\boldsymbol{\theta})$ may be misspecified, the posterior mode $\hat{\boldsymbol{\theta}}$ converges to the $\boldsymbol{\theta}_{n*}$ that minimizes $E\{-\log p(\boldsymbol{\theta}|\boldsymbol{Y})\}$, where the expectation is taken with respect to the true distribution of $\boldsymbol{Y}$; see for example, Bunke and Milhaud (1998). For simplicity, we further assume that $\boldsymbol{\theta}_{n*} = \boldsymbol{\theta}_*$ for all n. We use $||\cdot||$ to denote the Euclidean norm of a vector or a matrix and use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote the largest and smallest eigenvalues of a symmetric matrix A, respectively. We use the mathematical symbols (e.g., $O(N^{-1})$) and the stochastic-order symbols including $O_p(1)$, $o_p(1)$, and $O_p(N^{-1})$ throughout.

The following assumptions are needed to facilitate the technical details, although they are not the weakest possible conditions. Because we develop all results for general parametric models, we only assume several high-level assumptions as follows.

Assumption C1. $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{[S]}$ for all S are consistent estimates of $\boldsymbol{\theta}_*$, an interior point of Θ.

Assumption C2. Let $\Delta(\boldsymbol{\theta}) = \boldsymbol{\theta} - \boldsymbol{\theta}_*$ and suppose

$$\begin{aligned}
\log p(\boldsymbol{\theta}|\boldsymbol{Y}) &= \log p(\boldsymbol{\theta}_*|\boldsymbol{Y}) + \Delta(\boldsymbol{\theta})^T F_N(\boldsymbol{\theta}_*) - 0.5\Delta(\boldsymbol{\theta})^T J_N(\boldsymbol{\theta}_*)\Delta(\boldsymbol{\theta})[1 + o_p(1)] \quad \text{and} \\
\log p(\boldsymbol{\theta}|\boldsymbol{Y}_{[S]}) &= \log p(\boldsymbol{\theta}_*|\boldsymbol{Y}_{[S]}) + \Delta(\boldsymbol{\theta})^T F_{N,[S]}(\boldsymbol{\theta}_*) - 0.5\Delta(\boldsymbol{\theta})^T J_{N,[S]}(\boldsymbol{\theta}_*)\Delta(\boldsymbol{\theta})[1 + o_p(1)]
\end{aligned}$$

uniformly for all $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_*, \delta_0/\sqrt{N}) = \{\boldsymbol{\theta} : \sqrt{N}||\boldsymbol{\theta} - \boldsymbol{\theta}_*|| \leq \delta_0\}$. Moreover, $\max_{S \in I_S} \sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in B(\boldsymbol{\theta}_*, N^{-1/2}\delta_0)} ||J_{N,[S]}(\boldsymbol{\theta}) - J_{N,[S]}(\boldsymbol{\theta}')|| = o_p(N)$, $N^{-1/2}F_N(\boldsymbol{\theta}_*) = O_p(1)$,

$N^{-1/2}F_{N,[S]}(\boldsymbol{\theta}_*) = O_p(1)$,

$$\begin{aligned} 0 &< \inf_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*,\delta_0 N^{-1/2})} \lambda_{\min}(n^{-1}J_N(\boldsymbol{\theta})) \leq \sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*,\delta_0 N^{-1/2})} \lambda_{\max}(N^{-1}J_N(\boldsymbol{\theta})) < \infty, \text{ and} \\ 0 &< \min_{S\in I_S} \inf_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*,\delta_0 N^{-1/2})} \lambda_{\min}(N^{-1}J_{N,[S]}(\boldsymbol{\theta})) \\ &\leq \max_{S\in I_S} \sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*,\delta_0 N^{-1/2})} \lambda_{\max}(N^{-1}J_{N,[S]}(\boldsymbol{\theta})) < \infty. \end{aligned}$$

Assumption C3. Assume that for small $\delta_0 > 0$, if $N_S \leq N_0$, a fixed constant, then

$$\max_{S\in I_S} \sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*,\delta_0)} ||\partial_{\boldsymbol{\theta}} \log p_S(\boldsymbol{\theta})|| = O_p(1) \text{ and } \max_{S\in I_S} \sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*,\delta_0)} ||\partial^2_{\boldsymbol{\theta}} \log p_S(\boldsymbol{\theta})|| = o_p(N).$$

Assumption C3'. Assume that for small $\delta_0 > 0$, if $N_S \to \infty$, then

$$\begin{aligned} &\max_{S\in I_S} \sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*,\delta_0/\sqrt{N})} ||\partial_{\boldsymbol{\theta}} \log p_S(\boldsymbol{\theta})|| = O_p(\sqrt{N_S}) \text{ and} \\ &\max_{S\in I_S} \sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*,\delta_0)} ||\partial^k_{\boldsymbol{\theta}} \log p_S(\boldsymbol{\theta})|| = O_p(N_S) \text{ for } k = 0, \cdots, 5. \end{aligned}$$

Assumption C4. $\log p(\boldsymbol{\theta}|\boldsymbol{Y})$ and $\log p(\boldsymbol{\theta}|\boldsymbol{Y}_{[S]})$ for all $S \in I_S$ are Laplace regular (Kass et al., 1990).

Assumption C5. $\lim_{N_{I_S}\to\infty} N_{I_S}^{-1} E[K_N(I_S|\boldsymbol{\theta}_*)] = K_*(I_S)$ and $\lim_{N\to\infty} N^{-1}E[J_N(\boldsymbol{\theta}_*)] = J_*$, where the expectation is taken with respect to the true data generator. Moreover, for a small $\delta_0 > 0$, we have

$$\begin{aligned} &\sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*,\delta_0)} ||K_N(I_S|\boldsymbol{\theta}) - E[K_N(I_S|\boldsymbol{\theta})]|| = o_p(1) \text{ and} \\ &\sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*,\delta_0)} ||J_N(I_S|\boldsymbol{\theta}) - E[J_N(I_S|\boldsymbol{\theta})]|| = o_p(1). \end{aligned}$$

Assumption C6. Each component of $N_{I_S}^{-1}\sqrt{N}\{K_N(I_S|\boldsymbol{\theta}_*) - E[K_N(I_S|\boldsymbol{\theta}_*)]\}$ is asymptotically normal.

Remarks: Assumptions C1 and C2 are very general conditions and have been widely

used to examine the asymptotic properties of the extremum estimator, such as the maximum likelihood estimate in general parametric models such as time series models (Andrews, 1999). Sufficient conditions of Assumptions C1 and C2 have been extensively discussed in the literature (Andrews, 1999). Assumptions C3 and C3' are needed to examine the asymptotic properties of the three case influence measures for each $S \in I_S$. Most models with a smooth likelihood automatically satisfy Assumptions C3 and C3'. Assumption C4 is needed to use the Laplace approximation formula (Kass et al., 1990; Tierney et al., 1989). Assumption C5 is ensured by the law of large numbers (van der Vaart and Wellner, 1996). Assumption C6 is usually ensured by central limit theory. Recall that $p_S(\boldsymbol{\theta}) = p(\boldsymbol{Y}_S|\boldsymbol{Y}_{[S]}, \boldsymbol{\theta})$. If $p_S(\boldsymbol{\theta})$ only depends on few observations in $\boldsymbol{Y}_{[S]}$, then we can apply the theory of U-statistics to establish Assumption C6 (van der Vaart, 1998).

Proof of Theorem 3.1:

For notational simplicity, we temporarily assume that the dimension of $\boldsymbol{\theta}$ is 1. The proof of Theorem 3.1 (a) consists of four steps as follows:

In Step 1, we approximate $\log p(\boldsymbol{\theta}|\boldsymbol{Y})$ using $\log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) - 0.5N(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^2\partial^2_{\boldsymbol{\theta}}h(\hat{\boldsymbol{\theta}})$, where $h(\boldsymbol{\theta}) = -N^{-1}\log p(\boldsymbol{\theta}|\boldsymbol{Y})$. Then, we can use the formula for the Laplace approximation to integrals (eqn (2.6) in Tierney et al. (1989)) to obtain

$$\begin{aligned} E_{\theta|Y}[\phi\left(R_{[S]}(\boldsymbol{\theta})\right)] &= \phi_{[S]}(\hat{\boldsymbol{\theta}}) + 0.5N^{-1}[\partial^2_{\boldsymbol{\theta}}h(\hat{\boldsymbol{\theta}})]^{-1}\{\partial^2_{\boldsymbol{\theta}}\phi_{[S]}(\hat{\boldsymbol{\theta}}) \\ &\quad -[\partial^2_{\boldsymbol{\theta}}h(\hat{\boldsymbol{\theta}})]^{-1}\partial^3_{\boldsymbol{\theta}}h(\hat{\boldsymbol{\theta}})[\partial_{\boldsymbol{\theta}}\phi_{[S]}(\hat{\boldsymbol{\theta}})]\} + O(N^{-2}), \end{aligned} \tag{B.1}$$

where $\phi_{[S]}(\hat{\boldsymbol{\theta}}) = \phi\left(R_{[S]}(\hat{\boldsymbol{\theta}})\right)$.

In Step 2, we note that

$$R_{[S]}(\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta}|\boldsymbol{Y}_{[S]})}{p(\boldsymbol{\theta}|\boldsymbol{Y})} = \frac{p(\boldsymbol{Y}_{[S]}|\boldsymbol{\theta})}{p(\boldsymbol{Y}|\boldsymbol{\theta})} \times \frac{\int p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int p(\boldsymbol{Y}_{[S]}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{\mathrm{CPO}_S}{p_S(\boldsymbol{\theta})}, \tag{B.2}$$

where $\text{CPO}_S = \int p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} / \int p(\boldsymbol{Y}_{[S]}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$. Again, the formula for the Laplace approximation to integrals (eqn (2.6) in Tierney et al. (1989)) leads to

$$R_{[S]}(\hat{\theta}) = \frac{\left(E_{\theta|Y}\{[p_S(\hat{\boldsymbol{\theta}})]^{-1}\}\right)^{-1}}{p_S(\hat{\boldsymbol{\theta}})} = 1 - N^{-1}[\partial^2_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}})]^{-1}\{[\partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})]^2 \quad \text{(B.3)}$$
$$-0.5[p_S(\hat{\boldsymbol{\theta}})]^{-1}\partial^2_{\boldsymbol{\theta}} p_S(\hat{\boldsymbol{\theta}}) + 0.5[\partial^2_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}})]^{-1}\partial^3_{\boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}})\partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})\} + O_p(N^{-2}).$$

In Step 3, differentiating $\phi_{[S]}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and using a Taylor's series expansion, we can show that

$$\begin{aligned}
\partial_{\boldsymbol{\theta}}\phi_{[S]}(\hat{\boldsymbol{\theta}}) &= \partial_u \phi(R_{[S]}(\boldsymbol{\theta}))\partial_{\boldsymbol{\theta}} R_{[S]}(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -\dot{\phi}(1)\partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}}) + O_p(N^{-1}) \text{ and} \\
\partial^2_{\boldsymbol{\theta}}\phi_{[S]}(\hat{\boldsymbol{\theta}}) &= \ddot{\phi}(1)[\partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})]^2 - \dot{\phi}(1)\partial^2_{\boldsymbol{\theta}} p_S(\hat{\boldsymbol{\theta}})[p_S(\hat{\boldsymbol{\theta}})]^{-1} \\
&\quad +2\dot{\phi}(1)[\partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})]^2 + O_p(N^{-1}), \qquad \text{(B.4)}
\end{aligned}$$

where $\dot{\phi}(1) = \partial_u \phi(u)|_{u=1}$.

In Step 4, the Taylor's series expansion yields $\phi(R_{[S]}(\hat{\boldsymbol{\theta}})) = \dot{\phi}(1)[R_{[S]}(\hat{\boldsymbol{\theta}}) - 1] + O_p(N^{-2})$. Then, we combine the above results in Steps 2-4 to complete the proof of Theorem 3.1 (a).

The proof of Theorem 3.1 (b) consists of two steps as follows:

In Step 1, using Assumptions C1 and C2, we can show (Andrews, 1999) that

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_* = O_p(N^{-1/2}) \text{ and } \hat{\boldsymbol{\theta}}_{[S]} - \boldsymbol{\theta}_* = O_p(N^{-1/2}) \text{ for all } S \in I_S.$$

Thus, $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{[S]} = O_p(N^{-1/2})$.

In Step 2, expanding $\partial_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}_{[S]}|\boldsymbol{Y}_{[S]})$ at $\hat{\boldsymbol{\theta}}$ yields

$$0 = \partial_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}_{[S]}) + \partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}_{[S]})(\hat{\boldsymbol{\theta}}_{[S]} - \hat{\boldsymbol{\theta}})[1 + o_p(1)],$$

Since $0 = \partial_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}_{[S]}) + \partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})$ and $\partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}_{[S]}) = \partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) - \partial^2_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})$, we can use Assumption C3 to obtain

$$0 = \partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}}) + \partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})(\hat{\boldsymbol{\theta}}_{[S]} - \hat{\boldsymbol{\theta}})[1 + o_p(1)],$$

which leads to Theorem 3.1 (b).

To prove Theorem 3.1 (c), we use the formula for the Laplace approximation to integrals (eqn (2.6) in Tierney et al. (1989)) to get

$$\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + 0.5[\partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})]^{-2}[\partial^3_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})] + O(N^{-2}).$$

Furthermore, for each S, we also use the formula for the Laplace approximation to integrals to get

$$\tilde{\boldsymbol{\theta}}_{[S]} = \hat{\boldsymbol{\theta}} + [\partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})]^{-1}[\partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})] + 0.5[\partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})]^{-2}[\partial^3_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})] + O(N^{-2}).$$

Thus, subtracting $\tilde{\boldsymbol{\theta}}_{[S]}$ from $\tilde{\boldsymbol{\theta}}$, we obtain the proof Theorem 3.1 (c).

Combining Theorem 3.1 (b) and (c) leads to Theorem 3.1 (d).

Proof of Theorem 3.2:

By following the proof of Theorem 3.1 (b), we can show that $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{[S]} = O_p(N^{-1/2})$, $O_p(N^{1/2}) = -\partial_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}_{[S]}) = \partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})$ and $\partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}_{[S]}) = \partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) - \partial^2_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})$. Thus, we have

$$\hat{\boldsymbol{\theta}}_{[S]} = \hat{\boldsymbol{\theta}} - [J_{N,[S]}(\hat{\boldsymbol{\theta}})]^{-1} \partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})[1 + o_p(1)].$$

If $N_S/N \to 0$, then $J_{N,[S]}(\hat{\boldsymbol{\theta}}) = J_N(\hat{\boldsymbol{\theta}})[1 + O_p(N_S/N)] = J_N(\hat{\boldsymbol{\theta}})[1 + o_p(1)]$ and thus we can replace $J_{N,[S]}(\hat{\boldsymbol{\theta}})$ by $J_N(\hat{\boldsymbol{\theta}})$. Otherwise, we cannot replace $J_{N,[S]}(\hat{\boldsymbol{\theta}})$ by $J_N(\hat{\boldsymbol{\theta}})$ when $N_S/N \to \gamma < 1$. This completes the proof of Theorem 3.2 (a).

To prove Theorem 3.2 (b), we use the formula for the Laplace approximation to integrals (eqn (2.6) in Tierney et al. (1989)) to get

$$\begin{aligned}\tilde{\boldsymbol{\theta}} &= \hat{\boldsymbol{\theta}} + 0.5[\partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})]^{-2}[\partial^3_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})] + O_p(N^{-2}) \\ \tilde{\boldsymbol{\theta}}_{[S]} &= \hat{\boldsymbol{\theta}}_{[S]} + 0.5[\partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}_{[S]})]^{-2}[\partial^3_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}_{[S]})] + O_p(N^{-2}).\end{aligned}$$

It follows from a Taylor's series expansion that

$$\begin{aligned}&\Delta_1 = [\partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})]^{-2}[\partial^3_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})] - [\partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}_{[S]})]^{-2}[\partial^3_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}_{[S]})] \\ &= [\partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})]^{-2}[\partial^3_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) - \partial^3_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}_{[S]})] \\ &+ \left\{[\partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})]^{-2} - [\partial^2_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}_{[S]})]^{-2}\right\}[\partial^3_{\boldsymbol{\theta}} \log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}_{[S]})] = O_p(N_S/N^2),\end{aligned}$$

which yields $\Delta_1 = o_p(1/N)$ as $N_S/N \to 0$ and $\Delta_1 = O_p(1/N)$ as $N_S/N \to \gamma < 1$. Therefore, we have $\Delta_1 = o_p(1)[J_{N,[S]}(\hat{\boldsymbol{\theta}})]^{-1}\partial_{\boldsymbol{\theta}} \log p_S(\hat{\boldsymbol{\theta}})$ and $\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_{[S]} = (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{[S]})[1 + o_p(1)]$.

Theorem 3.2 (c) easily follows from Theorems 3.2 (a) and (b), and therefore, we omit all details. To prove Theorem 3.2 (d), we follow the proof of Theorem 3.1 (a). Equation (B.1) is still valid, but we cannot directly apply equation (2.6) in Tierney et al. (1989) to get (B.2) since $\log p_S(\boldsymbol{\theta})$ depends on N. Instead, we use the basic Laplace approximation to integrals and equation (2.4) in Tierney et al. (1989) to get

$$R_{[S]}(\hat{\boldsymbol{\theta}}) = \frac{\text{CPO}_S}{p_S(\hat{\boldsymbol{\theta}})} = A_S[1 + O_p(\frac{1}{N})]. \tag{B.5}$$

If $N_S/N \to \gamma \in (0, 1)$, then A_S does not converge to 1 in probability, since $\sigma^2 = \sigma^2_S + O(1)$ and $p(\boldsymbol{Y}_{[S]}|\hat{\boldsymbol{\theta}})p(\hat{\boldsymbol{\theta}}) = p(\boldsymbol{Y}_{[S]}|\hat{\boldsymbol{\theta}}_S)p(\hat{\boldsymbol{\theta}}_S)[1 + O_p(1)]$. We get

$$E_{\theta|Y}[\phi\left(R_{[S]}(\boldsymbol{\theta})\right)] = \phi(A_S) + \dot{\phi}(A_S)[R_{[S]}(\hat{\boldsymbol{\theta}}) - A_S] + O_p(N^{-1}) = \phi(A_S) + O_p(N^{-1}).$$

Thus, since $\phi(A_S) \neq \phi(1) = 0$, $\phi(A_S)$ dominates the asymptotic expansion of $E_{\theta|Y}[\phi\left(R_{[S]}(\boldsymbol{\theta})\right)]$.

If $N_S/N \to 0$, then it follows from a Taylor's series expansion that

$$\frac{\sigma}{\sigma_{[S]}} = [J_N(\hat{\boldsymbol{\theta}})]^{-1/2}[J_{N,[S]}(\hat{\boldsymbol{\theta}}_S)]^{1/2} = 1 + 0.5[J_N(\hat{\boldsymbol{\theta}})]^{-1}J_{N,[S]}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_S - \hat{\boldsymbol{\theta}}) + O_p(\frac{N_S}{N}), \text{ and}$$

$$\log p(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}_{[S]}) - \log p(\hat{\boldsymbol{\theta}}_S|\boldsymbol{Y}_{[S]}) = -0.5N(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_S)^2 J_{N,[S]}(\hat{\boldsymbol{\theta}}_S) + O_p(\frac{N_S^{3/2}}{N^2}) = O_p(\frac{N_S}{N}).$$

These yield that

$$A_S = 1 + 0.5[J_N(\hat{\boldsymbol{\theta}})]^{-1}[\partial_{\boldsymbol{\theta}} J_{N,[S]}(\hat{\boldsymbol{\theta}})](\hat{\boldsymbol{\theta}}_S - \hat{\boldsymbol{\theta}}) + O_p(\frac{N_S}{N}).$$

Finally, we can show that

$$\begin{aligned} E_{\theta|Y}[\phi\left(R_{[S]}(\boldsymbol{\theta})\right)] &= \phi(A_S) + O_p(N^{-1}) = \dot{\phi}(1)(A_S - 1) + O_p(N^{-1}) \\ &= 0.5\dot{\phi}(1)[J_N(\hat{\boldsymbol{\theta}})]^{-1}[\partial_{\boldsymbol{\theta}} J_{N,[S]}(\hat{\boldsymbol{\theta}})](\hat{\boldsymbol{\theta}}_S - \hat{\boldsymbol{\theta}}) + O_p(\frac{N_S}{N}) = O_p(\frac{\sqrt{N_S}}{N}), \end{aligned}$$

which completes the proof of Theorem 3.2 (d).

Proof of Theorem 3.3:

Theorem 3.3 (a) follows directly from Assumptions C1 and C5. Theorem 3.3 (b) follows from Assumptions C5 and C6.

Proof of Theorem 3.4:

To prove Theorem 3.4, we expand $\log p_S(\tilde{\boldsymbol{\theta}}_{[S]})$ at $\tilde{\boldsymbol{\theta}}$ for each S and obtain

$$\sum_{S \in I_S} \log p_S(\tilde{\boldsymbol{\theta}}_{[S]}) = \sum_{S \in I_S} \log p_S(\tilde{\boldsymbol{\theta}}) + \sum_{S \in I_S} \partial_{\boldsymbol{\theta}} \log p_S(\tilde{\boldsymbol{\theta}})^T \Delta_S[1 + o_p(1)],$$

where $\Delta_S = \tilde{\boldsymbol{\theta}}_{[S]} - \tilde{\boldsymbol{\theta}}$. It follows from Theorem 3.1 (c) that

$$\sum_{S \in I_S} \log p_S(\tilde{\boldsymbol{\theta}}_{[S]}) = \sum_{S \in I_S} \log p_S(\tilde{\boldsymbol{\theta}}) - \sum_{S \in I_S} [\partial_{\boldsymbol{\theta}} \log p_S(\tilde{\boldsymbol{\theta}})]^T [J_n(\tilde{\boldsymbol{\theta}})]^{-1} \partial_{\boldsymbol{\theta}} \log p_S(\tilde{\boldsymbol{\theta}})[1 + o_p(1)],$$

which yields Theorem 3.4.

Proof of Theorem 3.5:

The proof of Theorem 3.5 consists of four steps as follows:

In Step 1, following Lemma A1 of (Ando, 2007), we can show that

$$J_n(\boldsymbol{\theta}_*)^{-1}\tilde{K}_n(I_S|\boldsymbol{\theta}_*)^{-1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*) \to^d N(\mathbf{0},\mathbf{I}_p),$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix.

In Step 2, following Lemma A2 of (Ando, 2007), we can show that

$$E_{\boldsymbol{Y}}[E_{\boldsymbol{\theta}|\boldsymbol{Y}}\{(\boldsymbol{\theta}-\boldsymbol{\theta}_*)^{\otimes 2}] = J_n(\boldsymbol{\theta}_*)^{-1} + \{J_n(\boldsymbol{\theta}_*)\}^{-1}\tilde{K}_n(I_S|\boldsymbol{\theta}_*)\{J_n(\boldsymbol{\theta}_*)\}^{-1}.$$

In Step 3, we decompose B_η as a sum of E_1, E_2, and E_3, where

$$\begin{aligned}
E_1 &= E_{\boldsymbol{Y}}(n^{-1}E_{\boldsymbol{\theta}|\boldsymbol{Y}}\log p(\boldsymbol{Y}|\boldsymbol{\theta}) - n^{-1}\log\{p(\boldsymbol{Y}|\boldsymbol{\theta}_*)p(\boldsymbol{\theta}_*)\}),\\
E_2 &= E_{\boldsymbol{Y}}(n^{-1}\log\{p(\boldsymbol{Y}|\boldsymbol{\theta}_*)p(\boldsymbol{\theta}_*)\} - n^{-1}E_{\boldsymbol{Z}}[\log\{p(\boldsymbol{Z}|\boldsymbol{\theta}_*)p(\boldsymbol{\theta}_*)\}]),\\
E_3 &= E_{\boldsymbol{Y}}\left(n^{-1}E_{\boldsymbol{Z}}[\log\{p(\boldsymbol{Z}|\boldsymbol{\theta}_*)p(\boldsymbol{\theta}_*)\}] - n^{-1}E_{\boldsymbol{Z}}\{E_{\boldsymbol{\theta}|\boldsymbol{Y}}\log p(\boldsymbol{Z}|\boldsymbol{\theta})\}\right).
\end{aligned}$$

E_2 equals zero. Following the arguments in Theorem 1 of Ando (2007), we can do a second order expansion of $\log\{p(\boldsymbol{Y}|\boldsymbol{\theta}_*)p(\boldsymbol{\theta}_*)\}$ at $\hat{\boldsymbol{\theta}}$ to obtain

$$E_1 = n^{-1}E_{\boldsymbol{Y}}[E_{\boldsymbol{\theta}|\boldsymbol{Y}}\{\log p(\boldsymbol{Y}|\boldsymbol{\theta})\} - \log\{p(\boldsymbol{Y}|\hat{\boldsymbol{\theta}})p(\hat{\boldsymbol{\theta}})\}] + 0.5n^{-1}\mathrm{tr}\{J_n^{-1}(\boldsymbol{\theta}_*)\tilde{K}_n(I_S\boldsymbol{\theta}_*)\} + o_p(1).$$

E_3 can be written as

$$E_{\boldsymbol{Y}}\left(n^{-1}E_{\boldsymbol{Z}}[\log\{p(\boldsymbol{Z}|\boldsymbol{\theta}_*)p(\boldsymbol{\theta}_*)\}] - n^{-1}E_{\boldsymbol{Z}}[E_{\boldsymbol{\theta}|\boldsymbol{Y}}\log\{p(\boldsymbol{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})\}]\right) + E_{\boldsymbol{Y}}\{E_{\boldsymbol{\theta}|\boldsymbol{Y}}\log p(\boldsymbol{\theta})\}.$$

Combining Step 2 and the Taylor expansion of $\log\{p(\boldsymbol{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})\}$, we get

$$E_3 = 0.5n^{-1}\mathrm{tr}\{J_n^{-1}(\boldsymbol{\theta}_*)\tilde{K}_n(I_S|\boldsymbol{\theta}_*)\} + 0.5n^{-1}p + E_{\boldsymbol{Y}}\{E_{\boldsymbol{\theta}|\boldsymbol{Y}}\log p(\boldsymbol{\theta})\} + o_p(1).$$

In Step 4, we use the formula of the Laplace approximation to integrals to obtain

$$n^{-1}E_{\boldsymbol{\theta}|\boldsymbol{Y}}[\log\{p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\}] = n^{-1}\log\{p(\boldsymbol{Y}|\hat{\boldsymbol{\theta}})p(\hat{\boldsymbol{\theta}})\} - 0.5n^{-1}p + O(n^{-2}).$$

Combining Steps 3 and 4, we obtain the proof of Theorem 3.5.

Partial derivatives of $\log p(\boldsymbol{Y}|\boldsymbol{\beta},\tau,\boldsymbol{D})$ ***and*** $\log p(\boldsymbol{\beta}|\tau)p(\tau)p(\boldsymbol{D}^{-1})$ ***in Section 3.4.2***:

$$\begin{aligned}
\partial_{\boldsymbol{\beta}}\log p(\boldsymbol{Y}|\boldsymbol{\beta},\tau,\boldsymbol{D}) &= \sum_{i=1}^{n}\tau\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}),\\
\partial_{\tau}\log p(\boldsymbol{Y}|\boldsymbol{\beta},\tau,\boldsymbol{D}) &= \sum_{i=1}^{n}\left\{\frac{m_i}{2\tau} - \frac{1}{2}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right\},\\
\partial_{\boldsymbol{D}}\log p(\boldsymbol{Y}|\boldsymbol{\beta},\tau,\boldsymbol{D}) &= \sum_{i=1}^{n}\Big[-\frac{1}{2}\mathrm{vec}(\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i)\\
&\quad + \frac{\tau}{2}\mathrm{vec}\{\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i\}\Big],
\end{aligned}$$

$$\begin{aligned}
\partial^2_{\boldsymbol{\beta}}\log p(\boldsymbol{Y}|\boldsymbol{\beta},\tau,\boldsymbol{D}) &= \sum_{i=1}^{n} -\tau\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i,\\
\partial^2_{\tau}\log p(\boldsymbol{Y}|\boldsymbol{\beta},\tau,\boldsymbol{D}) &= \sum_{i=1}^{n} -\frac{m_i}{2\tau^2},\\
\partial^2_{\boldsymbol{D}}\log p(\boldsymbol{Y}|\boldsymbol{\beta},\tau,\boldsymbol{D}) &= \sum_{i=1}^{n}\frac{1}{2}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i\\
&\quad - \tau\{\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i\},
\end{aligned}$$

$$
\begin{aligned}
\partial_{\tau}\partial_{\boldsymbol{\beta}} \log p(\boldsymbol{Y}|\boldsymbol{\beta},\tau,\boldsymbol{D}) &= \sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \\
\partial_{\boldsymbol{D}}\partial_{\boldsymbol{\beta}} \log p(\boldsymbol{Y}|\boldsymbol{\beta},\tau,\boldsymbol{D}) &= \sum_{i=1}^{n} -\tau\{\boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1}\boldsymbol{X}_i \otimes \boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\}, \\
\partial_{\boldsymbol{D}}\partial_{\tau} \log p(\boldsymbol{Y}|\boldsymbol{\beta},\tau,\boldsymbol{D}) &= \sum_{i=1}^{n} \frac{1}{2}\mathrm{vec}\{\boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T \boldsymbol{V}_i^{-1}\boldsymbol{Z}_i\},
\end{aligned}
$$

$$
\begin{aligned}
\partial_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}|\tau)p(\tau)p(\boldsymbol{D}^{-1}) &= -\tau\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0), \\
\partial_{\tau} \log p(\boldsymbol{\beta}|\tau)p(\tau)p(\boldsymbol{D}^{-1}) &= (\frac{p}{2} + \frac{\delta_0}{2} - 1)\frac{1}{\tau} - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0) - \frac{\gamma_0}{2}, \\
\partial_{\boldsymbol{D}} \log p(\boldsymbol{\beta}|\tau)p(\tau)p(\boldsymbol{D}^{-1}) &= -\frac{1}{2}(v_0 - q - 1)\mathrm{vec}(\boldsymbol{D}^{-1}) + \frac{1}{2}\mathrm{vec}\{\boldsymbol{D}^{-1}(\boldsymbol{C}_0^{-1})^T\boldsymbol{D}^{-1}\},
\end{aligned}
$$

$$
\begin{aligned}
\partial^2_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}|\tau)p(\tau)p(\boldsymbol{D}^{-1}) &= -\tau\boldsymbol{\Sigma}_0^{-1}, \\
\partial^2_{\tau} \log p(\boldsymbol{\beta}|\tau)p(\tau)p(\boldsymbol{D}^{-1}) &= -(\frac{p}{2} + \frac{\delta_0}{2} - 1)\frac{1}{\tau^2}, \\
\partial^2_{\boldsymbol{D}} \log p(\boldsymbol{\beta}|\tau)p(\tau)p(\boldsymbol{D}^{-1}) &= (\boldsymbol{D}^{-1} \otimes \boldsymbol{D}^{-1})\left[\frac{1}{2}(v_0 - q - 1)\mathbf{I}_{qq} - \{\mathbf{I}_q \otimes (\boldsymbol{C}_0^{-1})^T\boldsymbol{D}^{-1}\}\right], \\
\partial_{\tau}\partial_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}|\tau)p(\tau)p(\boldsymbol{D}^{-1}) &= -\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0), \\
\partial_{\boldsymbol{D}}\partial_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}|\tau)p(\tau)p(\boldsymbol{D}^{-1}) &= \mathbf{0}, \\
\partial_{\boldsymbol{D}}\partial_{\tau} \log p(\boldsymbol{\beta}|\tau)p(\tau)p(\boldsymbol{D}^{-1}) &= \mathbf{0}.
\end{aligned}
$$

Note that for any matrix $\mathbf{A}(p \times q)$, $\mathrm{vec}(\mathbf{A})$ is a column vector which is formed by stacking the rows of $\mathbf{A}$ sequentially (Lee, 2007).

APPENDIX C

Assumptions in Chap. 4

The following assumptions are needed to facilitate the technical details, although they are not the weakest possible conditions. Because we develop all results for general parametric models, we only assume several high-level assumptions as follows.

Assumption C1. $\hat{\theta}_{[I]}$ for any I is a consistent estimate of θ_*, an interior point of Θ.

Assumption C2. All $p(\mathbf{Y}_{[I]}, \theta)$ are three times continuously differentiable on Θ and satisfy

$$\log p(\mathbf{Y}_{[I]}, \theta) = \log p(\mathbf{Y}_{[I]}, \theta_*) + \Delta(\theta)^T J_{n,[I]}(\theta_*) - 0.5\Delta(\theta)^T \mathbf{F}_{n,[I]}(\theta_*)\Delta(\theta) + R_{[I]}(\theta),$$

in which $|R_{[I]}(\theta)| = o_p(1)$ uniformly for all $\theta \in B(\theta_*, \delta_0 n^{-1/2}) = \{\theta : \sqrt{n}||\theta - \theta_*|| \leq \delta_0\}$, where $\Delta(\theta) = \theta - \theta_*$ and $J_{n,[I]}(\theta) = \partial_\theta \log p(\mathbf{Y}_{[I]}, \theta)$ and $\mathbf{F}_{n,[I]}(\theta_*) = \partial_\theta^2 \log p(\mathbf{Y}_{[I]}, \theta)$.

Assumption C3. For any set I and $\mathbf{Z}$, $\sup_{\theta,\theta' \in B(\theta_*, n^{-1/2}\delta_0)} n^{-1/2} J_{n,[I]}(\theta) = O_p(1)$,

$$\sup_{\theta \in B(\theta_*, n^{-1/2}\delta_0)} ||\mathbf{F}_{n,[I]}(\theta) - E[\mathbf{F}_I(\theta)|\mathcal{M}, \mathbf{Z}]|| = o_p(1),$$

$$\sup_{\theta,\theta' \in B(\theta_*, n^{-1/2}\delta_0)} n^{-1}||\mathbf{F}_{n,[I]}(\theta) - \mathbf{F}_{n,[I]}(\theta')|| = O_p(1),$$

and $0 < \inf_{\theta \in B(\theta_*, \delta_0 n^{-1/2})} \lambda_{\min}(n^{-1}\mathbf{F}_{n,[I]}(\theta)) \leq \sup_{\theta \in B(\theta_*, \delta_0 n^{-1/2})} \lambda_{\max}(n^{-1}\mathbf{F}_{n,[I]}(\theta)) < \infty$.

Assumption C4. For any set I and $\mathbf{Z}$,

$$\sup_{\theta \in B(\theta_*, n^{-1/2}\delta_0)} J_I(\theta) = O_p(\sqrt{m(I)}), \qquad \sup_{\theta \in B(\theta_*, n^{-1/2}\delta_0)} ||\mathbf{f}_I(\theta)|| = O_p(m(I)),$$
$$\sup_{\theta \in B(\theta_*, n^{-1/2}\delta_0)} ||\mathbf{f}_I(\theta) - E[\mathbf{f}_I(\theta)|\mathcal{M}, \mathbf{Z}]|| = O_p(\sqrt{m(I)}).$$

Remarks: Assumptions C1-C4 are very general conditions and are generalizations of some higher level conditions for the extremum estimator, such as the maximum likelihood estimate, given in Andrews (1999). Sufficient conditions of Assumptions C2-C4 have been extensively discussed in the literature (Andrews, 1999; Zhu and Zhang, 2006). Moreover, for simplicity, we use the rates n and $m(I)$ in Assumptions C2-C4, which can be modified to accommodate more intricate examples in Andrews (1999) and Zhu and Zhang (2006).

BIBLIOGRAPHY

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csáki, F., editors, *Second International Symposium in Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest.

Andersen, E. B. (1992). Diagnostics in categorical data analysis. *Journal of the Royal Statistical Society, Series B: Methodological*, 54:781–791.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayes models. *Biometrika*, 94:443–458.

Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica*, 67:1341–1383.

Banerjee, M. and Frees, E. W. (1997). Influence diagnostics for linear longitudinal models. *Journal of the American Statistical Association*, 92:999–1005.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons.

Blyth, S. (1994). Local divergence and association (Corr: 95V82 p667). *Biometrika*, 81:579–584.

Bradlow, E. T. and Zaslavsky, A. M. (1997). Case influence analysis in Bayesian inference. *Journal of Computational and Graphical Statistics*, 6:314–331.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25.

Bunke, O. and Milhaud, X. (1998). Asymptotic behavior of Bayes estimates under possibly incorrect models. *The Annals of Statistics*, 26(2):617–644.

Carlin, B. P. and Polson, N. G. (1991). An expected utility approach to influence diagnostics. *Journal of the American Statistical Association*, 86:1013–1021.

Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression*. John Wiley & Sons.

Chen, M.-H., Dey, D. K., and Ibrahim, J. G. (2004). Bayesian criterion based model assessment for categorical data. *Biometrika*, 91(1):45–63.

Chen, M.-H. and Ibrahim, J. G. (2003). Conjugate priors for generalized linear models. *Statistica Sinica*, 13:461–476.

Christensen, R. (1997). *Log-linear Models and Logistic Regression*. Springer-Verlag Inc.

Christensen, R., Pearson, L. M., and Johnson, W. (1992). Case-deletion diagnostics for mixed models. *Technometrics*, 34:38–45.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19:15–18.

Cook, R. D. (1986). Assessment of local influence (with Discussion). *Journal of the Royal Statistical Society, Series B: Methodological*, 48:133–169.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall Ltd.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B: Methodological*, 34:187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62:269–276.

Critchley, F., Atkinson, R. A., Lu, G., and Biazi, E. (2001). Influence analysis based on the case sensitivity function. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 63(2):307–323.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318.

Davison, A. C. and Tsai, C.-L. (1992). Regression model diagnostics. *International Statistical Review*, 60:337–353.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (C/R: P22-37). *Journal of the Royal Statistical Society, Series B: Methodological*, 39:1–22.

Dixon, W. J. and Massey, F. J. (1983). *Introduction to Statistical Analysis*. McGraw Hill Book Co.

Eaton, M. L. and Tyler, D. E. (1991). On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *The Annals of Statistics*, 19:260–271.

Escobar, L. A. and Meeker, W. Q. (1992). Assessing influence in regression analysis with censored data. *Biometrics*, 48:507–528.

Finney, D. J. (1947). The estimation form individual records of the relationship between dose and quantal response. *Biometrika*, 34:320–334.

Fung, W.-K., Zhu, Z.-Y., Wei, B.-C., and He, X. (2002). Influence diagnostics and outlier tests for semiparametric mixed models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 64(3):565–579.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70:320–328.

Geisser, S. (1993). *Predictive Inference: an Introduction*. London: Chapman & Hall Ltd.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection (Corr: V75 p765). *Journal of the American Statistical Association*, 74:153–160.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions, with implementation via sampling-based methods (Disc: P160-167). In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 4*, pages 147–159. Oxford: Oxford University Press.

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, 85:1–11.

Gelman, A., Carlin, J. B., Stern, H., and Rubin, D. B. (2003). *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (Disc: P760-807). *Statistica Sinica*, 6:733–760.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (Disc: P189-193). In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 4*, pages 169–188. Oxford: Oxford University Press.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling (Corr: 97V46 p541-542 with R. M. Neal). *Applied Statistics*, 44:455–472.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41:337–348.

Haslett, J. (1999). A simple derivation of deletion diagnostic results for the general linear model with correlated errors. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 61:603–609.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, 18:1259–1294.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001a). *Bayesian Survival Analysis*. New York: Springer-Verlag Inc.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001b). Criterion-based methods for Bayesian model assessment. *Statistica Sinica*, 11(2):419–443.

Ibrahim, J. G. and Laud, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffreys's prior. *Journal of the American Statistical Association*, 86:981–986.

Ibrahim, J. G. and Laud, P. W. (1994). A predictive approach to the analysis of designed experiments. *Journal of the American Statistical Association*, 89:309–319.

Johnson, W. (1985). Influence measures for logistic regression: Another point of view. *Biometrika*, 72:59–65.

Johnson, W. and Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *Journal of the American Statistical Association*, 78:137–144.

Johnson, W. and Geisser, S. (1985). Estimative influence measures for the multivariate general linear model. *Journal of Statistical Planning and Inference*, 11:33–56.

Kalbfleisch, J. D. (1978). Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B: Methodological*, 40:214–221.

Kass, R. E., Kadane, J. B., and Tierney, L. (1990). Asymptotic evaluation of integrals arising in Bayesian inference. In Page, C. and LePage, R., editors, *Computing Science and Statistics: Proceedings of the Symposium on the Interface*, pages 38–42. Springer-Verlag Inc.

Kass, R. E., Tierney, L., and Kadane, J. B. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika*, 76:663–674.

Kirkwood, J., Ibrahim, J., Sondak, V., Richards, J., Flaherty, L., Ernstoff, M., Smith, T., Rao, U., Steele, M., and Blum, R. (2000). High- and low-dose interferon alfa-2b in high-risk melanoma: First analysis of intergroup trial e1690/s9111/c9190. *Journal of Clinical Oncology*, 18:2444–2458.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, 83:875–890.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:963–974.

Lee, S.-Y. (2007). *Structural Equation Modelling: A Bayesian Approach*. Wiley.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B: Methodological*, 44:226–233.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall Ltd.

McCulloch, R. E. (1989). Local model influence. *Journal of the American Statistical Association*, 84:473–478.

Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika*, 69:521–531.

Murata, N., Yoshizawa, S., and Amari, S. (1994). Network information criterion determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, 5:865–872.

Peng, F. and Dey, D. K. (1995). Bayesian analysis of outlier problems using divergence measures. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 23:199–213.

Pettit, L. I. (1986). Diagnostics in Bayesian model choice. *The Statistician: Journal of the Institute of Statisticians*, 35:183–190.

Pettitt, A. N. and Daud, I. B. (1989). Case-weighted measures of influence for proportional hazards regression. *Applied Statistics*, 38:51–67.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, 9:705–724.

Preisser, J. S. and Qaqish, B. F. (1996). Deletion diagnostics for generalised estimating equations. *Biometrika*, 83:551–562.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Sargent, D. J. (1998). A general framework for random effects survival analysis in the Cox proportional hazards setting. *Biometrics*, 54:1486–1497.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Sinha, D. (1997). Time-discrete beta-process model for interval-censored survival data. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 25:445–456.

Sinha, D., Ibrahim, J. G., and Chen, M.-H. (2003). A Bayesian justification of Cox's partial likelihood. *Biometrika*, 90:629–641.

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

Spiegelhalter, D. J., Thomas, A., Best, N., and Lunn, D. (2003). *WinBUGS User Manual version 1.4*. MRC Biostatistics Unit and Imperial College. Avaliable from http://www.mrc-bsu.cam.ac.uk/bugs.

Stier, D. M., Leventhal, J. M., Berg, A. T., Johnson, L., and Mezger, J. (1993). Are children born to young mothers at increased risk of maltreatment. *Pediatrics*, 91(3):642–648.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion) (Corr: 76V38 p102). *Journal of the Royal Statistical Society, Series B: Methodological*, 36:111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B: Methodological*, 39:44–47.

Stone, M. (2002). In discussion of "Bayesian measure of model complexity and fit" by Spiegelhalter, D. J., Best, N.G., Carlin, B.P., and van der Linde. *Journal of the Royal Statistical Society, Series B: Methodological*, 64:621.

Takeuchi, K. (1976). Distribution of information statistics and critera for adequacy of models (in Japanese). *Mathematical Sciences*, 153:12–18.

Thall, P. F. and Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46:657–671.

Thomas, W. and Cook, R. D. (1989). Assessing influence on regression coefficients in generalized linear models. *Biometrika*, 76:741–749.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, 84:710–716.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag Inc.

Wasserman, D. and Leventhal, J. (1993). Maltreatment of Children Born to Cocaine-Dependent Mothers. *American Journal of Diseases of Children*, 147(12):1324–1328.

Wei, B.-C. (1998). *Exponential Family Nonlinear Models*. Springer: Singapore.

Weiss, R. (1996). An approach to Bayesian sensitivity analysis. *Journal of the Royal Statistical Society, Series B: Methodological*, 58:739–750.

Weiss, R. E. and Cho, M. (1998). Bayesian marginal influence assessment. *Journal of Statistical Planning and Inference*, 71:163–177.

Weiss, R. E. and Cook, R. D. (1992). A graphical case statistic for assessing posterior influence. *Biometrika*, 79:51–55.

Weissfeld, L. A. (1990). Influence diagnostics for the proportional hazards model. *Statistics & Probability Letters*, 10:411–417.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–26.

Zhang, H. (1999). Analysis of infant growth curves using multivariate adaptive splines. *Biometrics*, 55:452–459.

Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, 21:299–313.

Zhu, H., Lee, S. Y., Wei, B. C., and Zhou, J. (2001). Case deletion measures for models with incomplete data. *Biometrika*, 88:727–737.

Zhu, H. and Zhang, H. (2006). Asymptotics for estimation and testing procedures under loss of identifiability. *Journal of Multivariate Analysis*, 97(1):19–45.

Zhu, H.-T., Ibrahim, J. G., Lee, S.-Y., and Zhang, H. (2007). Perturbation selection and influence measures in local influence analysis. *The Annals of Statistics*, 35:2565–2588.

Zhu, H.-T. and Lee, S.-Y. (2001). Local influence for incomplete-data models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 63(1):111–126.