HHS Public Access
Author manuscript
Stat Interface. Author manuscript; available in PMC 2015 July 13.

Published in final edited form as:
Stat Interface. 2014 October 1; 7(4): 531–542. doi:10.4310/SII.2014.v7.n4.a9.

Bayesian Case-deletion Model Complexity and Information Criterion

Hongtu Zhu[*],
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, hzhu@bios.unc.edu

Joseph G. Ibrahim[*], and
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, ibrahim@bios.unc.edu

Qingxia Chen[†]
Department of Biostatistics, Vanderbilt University, Nashville, TN, 37232, cindy.chen@vanderbilt.edu

Abstract

We establish a connection between Bayesian case influence measures for assessing the influence of individual observations and Bayesian predictive methods for evaluating the predictive performance of a model and comparing different models fitted to the same dataset. Based on such a connection, we formally propose a new set of Bayesian case-deletion model complexity (BCMC) measures for quantifying the effective number of parameters in a given statistical model. Its properties in linear models are explored. Adding some functions of BCMC to a conditional deviance function leads to a Bayesian case-deletion information criterion (BCIC) for comparing models. We systematically investigate some properties of BCIC and its connection with other information criteria, such as the Deviance Information Criterion (DIC). We illustrate the proposed methodology on linear mixed models with simulations and a real data example.

Keywords and phrases

Bayesian; Case influence measures; Cross Validation; Information criterion; Markov chain Monte Carlo; Model complexity

1. INTRODUCTION

The aim of this paper is to establish a formal connection between Bayesian case influence measures for assessing the influence of individual observations on a model and Bayesian predictive methods for choosing an appropriate dimension of a model and selecting the best model for a given dataset. In Bayesian analysis, such statistical measures are very important

[*]Dr. Zhu and Dr. Ibrahim's work was partially supported by RR025747-01, GM70335, CA74015, P01CA142538-01, MH086633, and EB005149-01 from the National Institutes of Health.
[†]Dr. Chen's work was partially supported by 1R21HL097334 and UL1 RR024975-01 from the National Institutes of Health.

and highly relevant in any formal statistical analysis, but their formal connection has not been fully explored. We will systematically examine the properties of these measures and establish such connections.

Bayesian case influence measures are developed to assess the influence of individual observations (or generally, a set of observations), but they also provide the importance of each observation in the analysis for a better model fit [33, 9, 25, 5, 7, 12, 11]. See [42, 43] for a comprehensive review of various Bayesian case influence measures and their properties. Among them, single case influence measures have been widely used for various specific statistical models including generalized linear models, time series models, survival models, and statistical models with missing data [18, 29, 20, 14, 12, 28, 43]. The influence of individual observations are often assessed either on the posterior distributions or the predictive distributions through case deletion. The two most popular Bayesian case influence measures are the Kullback-Leibler (KL) divergence [11] and the conditional predictive ordinate (CPO) [14, 12].

Bayesian predictive methods are developed to evaluate the predictive performance of a given model and to select a single model with the best predictive performance from a set of candidate models. For instance, many researchers have been interested in Bayesian model assessment tools based on criterion-based methods, such as the L-measure [17, 23, 15, 16, 8]. See [39] and [4] for an overview of recent progress in cross-validation procedures and Bayesian predictive methods for model assessment, selection, and comparison. The main challenge is to estimate predictive model accuracy by correcting for the bias inherent in the double use of the data including both fitting and prediction. Cross-validation (CV) is a natural way of estimating out-of-sample prediction error [12, 41]. However, since cross-validation requires repeated model fits, it is computation intensive, and hence, information criteria are commonly sought as alternative measures. Such information criteria include the Akaiki Information Criterion (AIC) [1], the Takeuchi Information Criterion (TIC) [35, 22], the Bayesian Information Criterion (BIC) [31, 24, 21], the Deviance Information Criterion (DIC) [32], and the Bayesian Predictive Information Criterion (BPIC) [2], among many others. All these information criteria incorporate different complexity terms for model choice and can be viewed as approximations to different versions of cross-validation [34, 33].

Despite the extensive literature on Bayesian diagnostic measures and Bayesian predictive methods, very little has been done on systematically examining their connections in general parametric models. Based on such connections, we also develop Bayesian case-deletion model complexity (BCMC) measures for quantifying the effective number of parameters in a given statistical model and a Bayesian case-deletion information criterion (BCIC) for comparing different models. We calculate BCMC and BCIC in two theoretical examples including linear models and linear mixed models. We will show that BCMC can be regarded as a measure of model complexity, and show its asymptotic equivalence to the effective number of parameters in various information criteria. We systematically investigate the connection of BCIC with cross-validation methods and other information criteria, such as TIC and DIC. 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.

The rest of this paper is organized as follows. In Section 2, we review Bayesian case influence measures and Bayesian predictive methods. We propose BCMC for measuring model complexity and BCIC for comparing different models. We also systematically establish the connections between our two new measures including BCMC and BCIC and many existing model complexity measures and information criteria. In Section 3, we illustrate the proposed methodology on linear mixed models using both simulations and a real dataset involving the Yale infant growth data. We conclude the paper with some discussion in Section 4.

2. METHODS

2.1 Bayesian Case Influence Measures

We consider a probability function for an $N \times 1$ vector $\boldsymbol{Y}^T=(\boldsymbol{Y}_1^T, \cdots, \boldsymbol{Y}_n^T)$, denoted by $p(\boldsymbol{Y}|\theta)$, where $\theta = (\theta_1, \cdots, \theta_p)^T$ is a $p \times 1$ vector in an open subset Θ of R^p, $\boldsymbol{Y}_i = (y_{i1}, \cdots, y_{im_i})^T$, and $N=\sum_{i=1}^n m_i$. Letting $p(\theta)$ be the prior distribution of θ, the posterior distribution for the full data $\boldsymbol{Y}$ is given by $p(\theta|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\theta)p(\theta)$. Moreover, the dimension of $\boldsymbol{Y}_i$ (or m_i), such as the number of repeated measures in each cluster of longitudinal studies, may vary across all i.

Bayesian case influence measures are primarily used to assess the influence of deleting an $N_S \times 1$ vector of observations, denoted by S, on posterior inferences regarding θ. We use a subscript '[S]' to denote the relevant quantity with all observations in S deleted. For example, if $S = \{i\}$, then $\boldsymbol{Y}_{[S]}$ is the corresponding observed data with all of $\boldsymbol{Y}_i$ deleted, whereas for $S = \{i_1, i_2\}$, $\boldsymbol{Y}_{[S]}$ is the corresponding observed data with $\boldsymbol{Y}_{i_1}$ and $\boldsymbol{Y}_{i_2}$ deleted. Moreover, 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. We use $\boldsymbol{Y}_S$ and $\boldsymbol{Y}_{[S]}$ to represent a subsample of $\boldsymbol{Y}$ consisting of all the observations in S and a subsample of $\boldsymbol{Y}$ with all observations in S ($\boldsymbol{Y}_S$) deleted, respectively. We also calculate $p(\theta|\boldsymbol{Y}_{[S]}) \propto p(\boldsymbol{Y}_{[S]}|\theta)p(\theta)$ as the posterior distribution of θ given $\boldsymbol{Y}_{[S]}$, where $p(\boldsymbol{Y}_{[S]}|\theta) = p(\boldsymbol{Y}|\theta)/p(\boldsymbol{Y}_S|\theta)$.

Following [43], we briefly introduce three types of Bayesian case influence measures based on case deletion. First, we consider the φ-influence of $\boldsymbol{Y}_{[S]}$, denoted by $D_\varphi(S)$, as a measure of the distance (discrepancy) between $p(\theta|\boldsymbol{Y}_{[S]})$ and $p(\theta|\boldsymbol{Y})$. Letting $R_{[S]}(\theta) = p(\theta|\boldsymbol{Y}_{[S]})/p(\theta|\boldsymbol{Y})$, then $D_\varphi(S)$ is given by

$$\mathrm{D}_\phi(S)=\int \phi_\alpha(R_{[S]}(\boldsymbol{\theta}))p(\boldsymbol{\theta}|\boldsymbol{Y})d\boldsymbol{\theta}, \quad (1)$$

where $\varphi_\alpha(u)$ 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 $\alpha = -1$. The $\varphi_1(\cdot)$ and $\varphi_{-1}(\cdot)$ lead to the Kullback-Leibler divergence (K-L divergence), whereas $\varphi(u) = \varphi_1(u) + \varphi_{-1}(u)$ leads to the symmetric K-L divergence. The L_1−distance and the χ^2−divergence correspond to $\varphi(u) = 0.5|u - 1|$ and $\varphi(u) = (u - 1)^2$, respectively [20].

Second, we consider *Cook's posterior mode distance*, denoted by CP(S), for quantifying the discrepancy between the posterior mode of θ with and without the ith case [10]. We define the posterior modes of θ for the full sample $\boldsymbol{Y}$ and a subsample $\boldsymbol{Y}_{[S]}$ as $\hat{\theta} = \text{argmax}_\theta \log p(\theta|\boldsymbol{Y})$ and $\hat{\theta}_{[S]} = \text{argmax}_\theta \log p(\theta|\boldsymbol{Y}_{[S]})$, respectively. Then, CP(S) is given by

$$\text{CP}(S)=(\hat{\boldsymbol{\theta}}_{[S]}-\hat{\boldsymbol{\theta}})^T G_\theta(\hat{\boldsymbol{\theta}}_{[S]}-\hat{\boldsymbol{\theta}}), \quad (2)$$

where G_θ is chosen to be a positive definite matrix. For instance, G_θ can be $J_N(\boldsymbol{\theta})=-\partial^2_{\boldsymbol{\theta}}\log p(\boldsymbol{\theta}|\boldsymbol{Y})=-\partial^2_{\boldsymbol{\theta}}\log p(\boldsymbol{Y}|\boldsymbol{\theta})-\partial^2_{\boldsymbol{\theta}}\log p(\boldsymbol{\theta})$ evaluated at $\hat{\theta}$, where $\partial^2_{\boldsymbol{\theta}}$ represents the second-order derivative with respect to θ. If $\partial^2_{\boldsymbol{\theta}}\log p(\hat{\boldsymbol{\theta}})=o_p(-\partial^2_{\boldsymbol{\theta}}\log p(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}))$, then CP($S$) is close to the well-known Cook's distance for deleting a set of observations [10, 44]. A large value of CP(S) implies more influence of the set S on the posterior mode.

Third, we consider *Cook's posterior mean distance*, denoted by CM(S), for quantifying the distance between the posterior mean of θ with and without the observations in S. Let $\tilde{\theta} = \int \theta \cdot p(\theta|\boldsymbol{Y})d\theta$ and $\tilde{\theta}_{[S]} = \int \theta \cdot p(\theta|\boldsymbol{Y}_{[S]})d\theta$ be, respectively, the posterior mean of θ for $\boldsymbol{Y}$ and $\boldsymbol{Y}_{[S]}$. The CM(S) is given by

$$\text{CM}(S)=(\tilde{\boldsymbol{\theta}}_{[S]}-\tilde{\boldsymbol{\theta}})^T W_\theta(\tilde{\boldsymbol{\theta}}_{[S]}-\tilde{\boldsymbol{\theta}}), \quad (3)$$

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

Computationally, the proposed case influence measures can all be approximated using only MCMC samples from the full posterior distribution, $p(\theta|\boldsymbol{Y})$. For diagnostic purposes, it is desirable to derive computationally feasible approximations to these case influence measures. For completion, we include an important theoretical result regarding such approximations, whose proof can be found in [43], as follows.

Proposition 1—*Assume that Assumptions C1–C4 in the Appendix hold and N_S is bounded by a fixed constant. We have the following results*:

- **a.** $D_\varphi(S) = 0.5\,\ddot{\varphi}(1) \times CP(S) + O_p(N^{-2}) = 0.5\ddot{\varphi}(1) \times CM(S) + O_p(N^{-2})$.
- **b.** $\hat{\theta}_{[S]} = \hat{\theta} + O_p(N^{-1}) = \hat{\theta} - [J_N(\hat{\theta})]^{-1}\partial_\theta \log p_S(\hat{\theta})[1+ O_p(N^{-1})]$.
- **c.** $\tilde{\theta}_{[S]} = \tilde{\theta} - [J_N(\hat{\theta})]^{-1}\partial_\theta \log p_S(\hat{\theta})[1 + O_p(N^{-1})]$.
- **d.** $D_\varphi(S) = 0.5\,\ddot{\varphi}(1)[\partial_\theta \log p_S(\hat{\theta})]^T [J_N(\hat{\theta})]^{-1}[\partial\theta \log p_S(\hat{\theta})][1 + O_p(N^{-1})]$, *where* $\ddot{\phi}(1)=\partial^2_u\phi(u)|_{u=1}$ *and* $p_S(\theta) = p(\boldsymbol{Y}_S|\boldsymbol{Y}_{[S]}, \theta)$ *is the conditional distribution of* $\boldsymbol{Y}_S$ *given* $\boldsymbol{Y}_{[S]}$.

Proposition 1 establishes a direct connection between $D_\varphi(S)$, CP(S) and CM(S) for any φ(·) and the one-step approximation of $\hat{\theta}_{[S]}$ and $\tilde{\theta}_{[S]}$ within the Bayesian framework. Proposition 1 provides a theoretical and computational approximation of $D_\varphi(S)$, denoted by AD(S;$\tilde{\theta}$), as

$$\text{AD}(S;\tilde{\boldsymbol{\theta}})=[\partial_{\boldsymbol{\theta}}\log p_S(\tilde{\boldsymbol{\theta}})]^T[J_N(\tilde{\boldsymbol{\theta}})]^{-1}[\partial_{\boldsymbol{\theta}}\log p_S(\tilde{\boldsymbol{\theta}})]. \quad (4)$$

The $\tilde{\theta}$ and $J_N(\tilde{\theta})$ can be easily computed from the MCMC samples. Moreover, it is straightforward to compute $\partial_\theta \log p_S(\theta) = \partial_\theta \log p(\boldsymbol{Y}|\theta) - \partial_\theta \log p(\boldsymbol{Y}_{[S]}|\theta)$. As an illustration, we consider a normal linear model to illustrate the calculation of Bayesian case influence measures.

Example 1—We consider a normal linear model as $\boldsymbol{Y} = \mathbf{X}\beta + \varepsilon$ or $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$, where β and $\boldsymbol{x}_i$ are $p \times 1$ vectors, β is unknown, $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T \sim N_n(0, \tau^{-1}\mathrm{I})$, and $\tau = 1/\sigma^2$ is assumed known for simplicity. We consider a conjugate prior for β as $N_p(\boldsymbol{\mu}_0, \tau^{-1}\Sigma_0)$. For a given set S, $p(\beta|\boldsymbol{Y})$ and $p(\beta|\boldsymbol{Y}_{[S]})$ are, respectively, given by

$$\boldsymbol{\beta}|\boldsymbol{Y} \sim N_p(\tilde{\boldsymbol{\beta}}, \tau^{-1}(\mathbf{X}^T\mathbf{X} + \boldsymbol{\Sigma}_0^{-1})^{-1})$$

and

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

where

$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0), \tilde{\boldsymbol{\beta}}_{[S]} = (\boldsymbol{X}_{[S]}^T\boldsymbol{X}_{[S]} + \boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{X}_{[S]}^T\boldsymbol{Y}_{[S]} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0)$, $\boldsymbol{X}_{[S]}$ is $\boldsymbol{X}$ with all $\boldsymbol{x}_i$ deleted for $i \in S$, and $\boldsymbol{Y}_{[S]}$ is $\boldsymbol{Y}$ with all y_i deleted for all $i \in S$. Note that $\boldsymbol{X}_{[S]}^T\boldsymbol{X}_{[S]} = \boldsymbol{X}^T\boldsymbol{X} - \sum_{i \in S}\mathbf{x}_i\mathbf{x}_i^T$ and $\boldsymbol{X}_{[S]}^T\boldsymbol{Y}_{[S]} = \boldsymbol{X}^T\boldsymbol{Y} - \sum_{i \in S}\mathbf{x}_i y_i$.

Let $S = \{i_1, \cdots i_{N_S}\}$ and $E_S = [\boldsymbol{e}_{i1}, \cdots, \boldsymbol{e}_{i_{N_S}}]$ be an $N \times N_S$ matrix, where $\boldsymbol{e}_k$ is an $N \times 1$ vector with a 1 at the k-th element and 0 elsewhere for $k \in S$. With some algebraic calculations, we have

$$\tilde{\boldsymbol{\beta}}_{[S]} = \tilde{\boldsymbol{\beta}} - (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}\boldsymbol{X}_S^T(\boldsymbol{I}_{N_S} - P_S)^{-1}\hat{e}_S,$$

where $\boldsymbol{X}_S = E_S^T\boldsymbol{X}, P_S = E_S^T P_{X0} E_S$, in which $P_{X0} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}\boldsymbol{X}^T$, and $\hat{e}_S = E_S^T(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})$. For the KL divergence, we get

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

Note that the posterior mode and the posterior mean are the same in this example. If we set $W_\theta = G_\theta = \tau(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})$, we have

$$\mathrm{CM}(S) = \mathrm{CP}(S) = \tau(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{[S]})^T(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{[S]}).$$

Since log $\log p_S(\boldsymbol{\theta}) = -0.5\tau \sum_{i \in S} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$ and $J_N(\tilde{\boldsymbol{\theta}}) = \tau(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})$, we have

$$\mathrm{AD}(S;\tilde{\boldsymbol{\theta}}) = \hat{e}_S^T \boldsymbol{X}_S \tau(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1} \boldsymbol{X}_S^T \hat{e}_S.$$

2.2 Cross Validation and Model Complexity

Bayesian case influence measures (BCIM) and cross-validation (CV) methods share the same strategy of splitting the data into two subsamples, but they differ from each other in validation [33, 34, 13, 4]. BCIM divides the data into a target sample $\boldsymbol{Y}_S$ and a training sample $\boldsymbol{Y}_{[S]}$ and then estimates $\tilde{\theta}_{[S]}$ based on the training sample $\boldsymbol{Y}_{[S]}$. Note that all development below is valid for $\hat{\theta}_{[S]}$, but we focus on the posterior mean from here on for notational simplicity. BCIM for a given set S represents the influential level of S. In contrast, the CV method divides the data into two subsamples including a training sample $\boldsymbol{Y}_{[S]}$ for model fitting and a validation sample $\boldsymbol{Y}_S$ for assessing model fit. Compared to BCIM, CV usually uses the predictive distribution $p(\tilde{\boldsymbol{Y}}_S|\boldsymbol{Y}_{[S]})$ for model validation, where $\tilde{\boldsymbol{Y}}_S$ is an independent copy of $\boldsymbol{Y}_S$. One choice of the predictive distribution is to use $p(\tilde{\boldsymbol{Y}}_S|\boldsymbol{Y}_{[S]}, \tilde{\theta}_{[S]})$, where $\tilde{\theta}_{[S]}$ is estimated based on $\boldsymbol{Y}_{[S]}$. Let N_B be an integer and $S_1, \cdots, S_{N_B}$ is a sequence of non-empty proper subsets of $\{(1, 1), \cdots, (n, m_n)\}$. The CV estimator of the model $p(\theta|\boldsymbol{Y})$ based on $I_S = (S_k)_{1 \le k \le N_B}$ is defined by

$$\mathrm{CVE}(I_S) = N_B^{-1} \sum_{S \in I_S} \log p(\boldsymbol{Y}_S | \boldsymbol{Y}_{[S]}, \tilde{\boldsymbol{\theta}}_{[S]}) = N_B^{-1} \sum_{S \in I_S} \log p_S(\tilde{\boldsymbol{\theta}}_{[S]}).$$

A challenging issue associated with BCIM and CV is to calculate the $\tilde{\theta}_{[S]}$'s for all possible splits. Most BCIM and CV methods split the data with a fixed size of the training sample. There are two major categories of splitting schemes including exhaustive data splitting and partial data splitting. Exhaustive data splitting includes the leave-M-out CV for all $N \ge M \ge 1$. For each fixed M, $N_B = N!/(M!(N - M)!)$ and I_S is the set of all possible sets with a fixed size M. However, except for relatively small M, it can be computationally restrictive to calculate BCIM and CV for every possible subset of the M data. Alternatively, one may consider partial data splitting methods, such as V-fold CV [4, 41].

An interesting question is whether there is any other connection between BCIM and CV besides the strategy of splitting the data. We can establish a connection between BCIM and CV by extending the well-known result on the asymptotic equivalence between CV and AIC [34]. We obtain the following theorems, whose detailed proofs can be found in the the Appendix.

Theorem 1—*Let N_S be a fixed constant. Then we have the following results*:

i. *Under Assumptions C1–C4 in the Appendix, CVE(I_S) has an asymptotic expansion as*

$$CVE(I_S)=N_B^{-1}\sum_{S_k\in I_S}\log p_{S_k}(\tilde{\boldsymbol{\theta}})-MAD(I_S)[1+o_p(1)], \quad (5)$$

where $MAD(I_S)=N_B^{-1}\sum_{S_k\in I_S}AD(S_k;\tilde{\boldsymbol{\theta}})$ *is the mean of* $AD(S_k;\tilde{\theta})$*'s.*

ii. *Under Assumptions C1, C2, and C5 in the Appendix, we have*

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

where $J_N(\tilde{\boldsymbol{\theta}})=-\partial^2_{\boldsymbol{\theta}}\log p(\boldsymbol{\theta}|\boldsymbol{Y})|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$ *and* $J_*=\lim_{N\to\infty}N^{-1}E[J_N(\boldsymbol{\theta}_*)]$, *in which the expectation is taken with respect to the true data generator and* θ_* *denotes the pseudo-true parameter* [6]. *Moreover,* $K_N(I_S|\tilde{\boldsymbol{\theta}})=N_B^{-1}\sum_{S_k\in I_S}[\partial_{\boldsymbol{\theta}}\log p_S(\boldsymbol{\theta})]^{\otimes 2}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$ *and*

$$K_*(I_S)=\lim_{N\to\infty}(N_B)^{-1}\sum_{S_k\in I_S}E\{[\partial_{\boldsymbol{\theta}}\log p_{S_k}(\boldsymbol{\theta}_*)]^{\otimes 2}\},$$

where $\boldsymbol{a}^{\otimes 2}=\boldsymbol{a}\boldsymbol{a}^T$ *for any vector* $\boldsymbol{a}$.

Theorem 1 shows a direct connection between CVE(I_S) and MAD(I_S) and an indirect connection between CVE(IS) and BCIM. According to Proposition 1, we can use the average of BCIMs to approximate MAD(I_S) as follows:

$$\mathrm{MAD}(I_S)=N_B^{-1}\sum_{S_k\in I_S}\mathrm{CP}(S_k)+O_p(N^{-1}). \quad (6)$$

A similar approximation also holds for both CM(S) and $D_\varphi(S)$. Moreover, MAD(I_S) is always nonnegative. Throughout the paper, based on MAD(I_S) and their approximations, we define the Bayesian case-deletion model complexity (BCMC) measures as

$$\mathrm{BCMC}(I_S)=NN_S^{-1}\times\mathrm{MAD}(I_S)\approx N_S^{-1}\mathrm{tr}[J_*^{-1}K_*(I_S)]. \quad (7)$$

We will show below that our BCMC measures can be regarded as a generalization of many existing measures of model complexity. We first consider single cluster deletion (or the leave-one-out CV) for clustered data, in which the $\boldsymbol{Y}_i$'s are independent for different i, but the components in each $\boldsymbol{Y}_i$ may be correlated. For the leave-one-out CV, we denote $I_{LOO}=\{\{1\},\cdots,\{n\}\}$. In this case, we have $N_B=n$, $p_{\{i\}}(\theta)=p(\boldsymbol{Y}_i|\theta)$,

$$K_N(I_{LOO}|\tilde{\boldsymbol{\theta}})=n^{-1}\sum_{i=1}^{n}\{\partial_{\boldsymbol{\theta}}\log p(\boldsymbol{Y}_i|\boldsymbol{\theta})\}^{\otimes 2}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}\to^p K_*(I_{LOO})=\lim_{n\to\infty}E\{K_N(I_{LOO}|\boldsymbol{\theta}_*)\},$$

and

$$J_N(\tilde{\boldsymbol{\theta}}) = -n^{-1}[\sum_{i=1}^{n}\partial_{\boldsymbol{\theta}}^2 \log p(\boldsymbol{Y}_i|\boldsymbol{\theta}) + \partial_{\boldsymbol{\theta}}^2 \log p(\boldsymbol{\theta})]|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \to^p J_* = \lim_{N\to\infty} E\{J_N(\boldsymbol{\theta}_*)\},$$

where $\to^p$ denotes convergence in probability. Let $p_* = \mathrm{BCMC}(I_{LOO})$ in this case. Using a uniform improper prior for θ, that is, $\partial_{\boldsymbol{\theta}}^2 \log p(\boldsymbol{\theta})=0$, p_* is the measure of model complexity in TIC. Furthermore, if the model $p(\boldsymbol{Y}|\theta)$ is correctly specified, then p_* reduces to p, the number of parameters, and $\mathrm{MAD}(I_{LOO}) = p + o_p(1)$. 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) [27, 30]. Moreover, $\mathrm{MAD}(I_{LOO})$ is also associated with the effective number of parameters, denoted by p_D, in DIC, where $p_D = E_{\theta|Y}[-2 \log p(\boldsymbol{Y}|\theta)] + 2 \log[p(\boldsymbol{Y}|\tilde{\theta})]$. Under the two conditions of approximately normal likelihoods and a uniform improper prior for θ, it can be shown that $p_D = \mathrm{tr}\{J_N(\tilde{\theta})E[(\theta - \tilde{\theta})^{\otimes 2}]\} + o_p(1)$ [32]. Moreover, using the fact that $E[(\theta - \tilde{\theta})^{\otimes 2}] = J_N(\theta_*)^{-1}K_N(ILOO|\theta_*)J_N(\theta_*)^{-1}[1 + o_p(1)]$ [6], we can obtain the following connections between p_D and p_*: $p_D = p_* + o_p(1)$. Thus, $\mathrm{MAD}(I_{LOO})$ has many of the same properties as p_D [32]. We also note that $\mathrm{MAD}(I_{LOO})$ is always nonnegative, whereas p_D is not.

Second, we consider multiple cluster deletion (or the leave-M clusters-out CV) for clustered data. Specifically, we focus on deleting every possible subset of data from M clusters and using it for validation. Let I_{LMO} be the set of all $N_B = \binom{n}{M}$ subsets with M clusters. If we set $S_1 = \{\{i_1\}, \cdots, \{i_M\}\}$, then we have

$$E\{[\partial_{\boldsymbol{\theta}} \log p_{S_1}(\boldsymbol{\theta}_*)]^{\otimes 2}\} = \sum_{i_k, i_k'} E\{\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_{i_k}|\boldsymbol{\theta}_*)\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_{i_k'}|\boldsymbol{\theta}_*)^T\} = \sum_{k=1}^{M} E\{\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_{i_k}|\boldsymbol{\theta}_*)^{\otimes 2}\}.$$

Therefore, by doing exhaustive data splitting, we have

$$N_B^{-1}\sum_{S_k\in I_S} E\{[\partial_{\boldsymbol{\theta}} \log p_{S_k}(\boldsymbol{\theta}_*)]^{\otimes 2}\} = \frac{M}{n}\sum_{i=1}^{n} E\{\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_i|\boldsymbol{\theta}_*)^{\otimes 2}\}, \quad (8)$$

which yields that $\mathrm{MAD}(I_{LMO}) = M \times \mathrm{MAD}(I_{LOO})$. If $m_1 = \cdots = m_n$, then $\mathrm{BCMC}(I_{LMO}) = \mathrm{BCMC}(I_{LOO})$. Similar discussions also hold for V-fold CV [4, 41].

Third, we consider single observation deletion $I_{SO} = \{\{(1, 1)\}, \cdots, \{(n,m_n)\}\}$ and examine $\mathrm{MAD}(I_{SO})$ for clustered data. We have $N_B = N = \sum_{i=1}^{n} m_i$ and

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

where $\boldsymbol{Y}_{i,[(i,j)]}$ denotes $\boldsymbol{Y}_i$ with $y_{i,j}$ deleted. The $K_N(ISO|\tilde{\theta})$ is given by

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

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

Example 1 (continued)—In this case, we have

$$\text{CVE}(I_S) = -0.5\tau N_B^{-1} \sum_{S_k \in I_S} \sum_{i \in S_k} (y_i - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}_{[S]})^2,$$

$$N_B^{-1} \sum_{S_k \in I_S} \log p_{S_k}(\tilde{\boldsymbol{\theta}}) = -0.5\tau N_B^{-1} \sum_{S_k \in I_S} \sum_{i \in S_k} (y_i - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}})^2,$$

$$\text{MAD}(I_S) = N_B^{-1} \sum_{S_k \in I_S} \hat{e}_{S_k}^T \boldsymbol{X}_{S_k} \tau (\boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1} \boldsymbol{X}_{S_k}^T \hat{e}_{S_k}.$$

According to Theorem 1, we have

$$\text{MAD}(I_S) = N_B^{-1} \text{tr}((\boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1} \sum_{S_k \in I_S} \boldsymbol{X}_{S_k}^T \boldsymbol{X}_{S_k})[1 + o_p(1)].$$

For the leave-one-out CV, $\text{BCMC}(I_{LOO})$ can be approximated by $\sum_{i=1}^{n} p_{ii}/n$, where the p_{ii}'s are the diagonal elements of P_{X0}. As Σ_0^{-1} converges to zero, which corresponds to a non-informative prior, $\text{BCMC}(I_{LOO})$ converges to the number of parameters in β.

2.3 Bayesian Case-deletion Information Criterion

Based on the development of BCMC(I_S) and CVE(I_S), we develop a new model selection criterion, called the Bayesian case-deletion information criterion (BCIC), to select an 'optimal' model from a pool of candidate models $\{M_l : l = 1, \cdots, L\}$ for the same dataset. Specifically, for model M_l and the deletion set I_S, BCIC is defined as

$$\mathrm{BCIC}(I_S, M_l) = -2\sum_{S_k \in I_S} \log p_{S_k}(\tilde{\boldsymbol{\theta}}(M_l), M_l) + (N_B N_S/N) C_n(I_S, \tilde{\boldsymbol{\theta}}(M_l), M_l), \quad (9)$$

where $\tilde{\theta}(M_l)$ is an estimator of θ and $p_{S_k}(\theta; M_l)$ denotes $p(\mathbf{Y}_{S_k}|\mathbf{Y}_{[S_k]}, \theta)$ under model M_l and $C_n(I_S, \tilde{\theta}(M_l), M_l)$ is a penalty term, which is a function of the data, the deletion set I_S, and an estimator of $\theta(M_l)$. In (9), $\sum_{S_k \in I_S} \log p_{S_k}(\tilde{\theta}(M_l), M_l)$ can be regarded as the conditional deviance function evaluated at $\tilde{\theta}(M_l)$. We choose an 'optimal' model, denoted by M_{opt}, which minimizes BCIC(I_S,M_l), as follows:

$$M_{opt}(I_S) = \mathrm{argmin}_{M_l : 1 \le l \le L} \mathrm{BCIC}(I_S, M_l).$$

Different forms of the model penalty $C_n(I_S, \tilde{\theta}(M_l), M_l)$ lead to different criteria. Two popular choices of $C_n(I_S, \tilde{\theta}(M_l), M_l)$ are the AIC-type penalty and the BIC-type penalty. For the AIC-type penalty, $C_n(I_S, \tilde{\theta}(M_l), M_l) = C_0 \times \mathrm{BCMC}(I_S)$, where C_0 is a positive scalar. In practice, similar to AIC, DIC, and TIC [1, 35, 22, 32], it is common to set $C_0 = 2$. For the BIC-type penalty, $C_n(I_S, \tilde{\theta}(M_l), M_l) = C_{0,n} \times \mathrm{BCMC}(I_S)$ with $\lim_{n\to\infty} C_{0,n} = \infty$. Similar to BIC, $C_{0,n}$ is often set as $\log(N)$ or other functions of N. Therefore, BCIC can be regarded as a generalization of existing model selection criteria.

Different deletion sets lead to slightly different BCIC(I_S,M_l) for all l. For instance, if we consider the single cluster deletion I_{LOO} and the single observation deletion I_{SO}, then we obtain different BCIC measures. Thus, it is possible that $M_{opt}(I_S)$ may vary across I_S. However, when we consider the leave-M clusters-out deletion for clustered data, we are able to obtain an invariance property of $M_{opt}(I_S)$. We are led to the following theorem.

Theorem 2—*Assume that* $\mathbf{Y}_i$*'s are independent and* $C_n(I_S, \tilde{\theta}(M_l), M_l) = \tilde{C}_{0,n} \times \mathrm{BCMC}(I_S)$, *where* $\tilde{C}_{0,n}$ *is in-dependent of* I_S *and* M_l*, but it may depend on n, we have the following results.*

i. *For the leave-M clusters-out CV, we have*

$$BCIC(I_{LMO}, M_l) = \begin{pmatrix} n-1 \\ M-1 \end{pmatrix} BCIC(I_{LOO}, M_l)$$ *and* $M_{opt}(I_{LMO}) = M_{opt}(I_{LOO})$ *for any* $M \ge 1$.

ii. *If* $BCIC(I_{LOO}, M_{opt}(I_{LOO})) - BCIC(I_{LOO}, M_l) >> O_p(N_B N^{-3/2})$ *for all* $M_l \ne M_{opt}(I_{LOO})$, *Assumption C6 holds, and we use* $MAD(I_S)$ *to approximate* $BCMC(I_S)$, *then* $M_{opt}(I_{LMO}) = M_{opt}(I_{LOO})$ *in probability 1 for any* $M \ge 1$.

Theorem 2 shows that BCIC(I_S,M_l) and $M_{opt}(I_S)$ are invariant for clustered data under different exhaustive splitting schemes. Due to Theorem 2, the two partitions of primary interest are now single cluster deletion (I_{LOO}) and single observation deletion (I_{SO}). Under I_{LOO}, BCIC can be simplified as

$$\mathrm{BCIC}(I_{LOO}, M_l) = -2(n-1)\log p(\boldsymbol{Y}|\tilde{\boldsymbol{\theta}}(M_l), M_l) + nC_{0,n}MC(I_{LOO}),$$

and under I_{SO}, BCIC can be simplified as

$$\mathrm{BCIC}(I_{SO}, M_l) = -2\left[N\log p(\boldsymbol{Y}|\tilde{\boldsymbol{\theta}}(M_l), M_l) - \sum_{i=1}^{n}\sum_{j=1}^{m_n}\log p(\boldsymbol{Y}_{i,j}|\tilde{\boldsymbol{\theta}}(M_l), M_l)\right] + NC_{0,n}MC(I_{SO}),$$

where $MC(I_{LOO})$ and $MC(I_{SO})$ are shown in Section 2.2. Note that, unlike cross validation, there is no much additional computational cost associated with BCIC procedure except the programming efforts to calculate $MC(I_S)$.

3. SIMULATIONS AND REAL DATA ANALYSIS

3.1 Simulation Studies

In this section, several simulation studies were carried out to investigate the finite sample performance of BCIC and compare BCIC with three existing Bayesian model selection criteria, including AIC, BIC, and DIC in linear mixed models. Specifically, we set $AIC = -2 \log p(Y|\tilde{\theta}(M_l)) + 2p$, $BIC = -2 \log p(Y|\tilde{\theta}(M_l)) + \log(N) \times p$, and $DIC = -2 \log p(Y|\tilde{\theta}(M_l)) + 2p_D$, where p is the number of parameters in the model and p_D is the effective number of parameters estimated by the posterior mean of the deviance minus the deviance of the posterior means. We consider both the leave-one cluster-out CV and the leave-one observation-out CV, the AIC- and BIC- type penalties, and calculate their associated BCICs.

Simulated datasets were generated from a linear mixed model with a random intercept. Specifically, we consider the following true model, given by $y_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + b_i + \varepsilon_{ij}$ for $i = 1, \cdots, n$ and $j = 1, \cdots, m_i$, where $x_{ij1} \sim Exp(1)$, $x_{ij2} = j$, $b_i \sim N(0, \tau^{-1}\xi^{-1})$, and $\varepsilon_{ij} \sim N(0, \tau^{-1})$. An additional covariate x_{ij3} was simulated from a $N(1, 1)$ distribution. The true parameter values were taken to be $\beta_0 = 2$, $\beta_1 = \beta_2 = 1$, $\tau = 0.1$, and $\xi = 1$ or $\xi = 0.04$, for $n = 10$ or $n = 20$. The values of ξ being 1 or 0.04 represent a medium or high intracluster correlation coefficient (ICC). We chose the priors as follows: $\pi(\beta, \tau, D^{-1}) \propto |D|^{-1/2}\tau^{-1}$ and $b|\tau, D \sim N_{nq}(0, \tau^{-1}(I_n \otimes D))$, where $D^{-1} = \xi$ in this simulation.

We considered five candidate models as follows:

M1 (true model) : $y_{ij}|x_{ij1}, x_{ij2} \sim N(\beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + b_i, \tau^{-1})$, $b_i \sim N(0, \tau^{-1}\xi^{-1})$;

M2 : $y_{ij}|x_{ij1}, x_{ij2} \sim N(\beta_0 + \beta_1 x_{ij2} + b_i, \tau^{-1})$, $b_i \sim N(0, \tau^{-1}\xi^{-1})$;

M3 : $y_{ij}|x_{ij1}, x_{ij2}, x_{ij3} \sim N(\beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + b_i, \tau^{-1})$, $b_i \sim N(0, \tau^{-1}\xi^{-1})$;

M4 : $y_{ij}|x_{ij1}, x_{ij2}, x_{ij3} \sim N(\beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij2}x_{ij3} + b_i, \tau^{-1})$, $b_i \sim N(0, \tau^{-1}\xi^{-1})$;

M5 : $y_{ij}|x_{ij1}, x_{ij2}, x_{ij3} \sim N(\beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \beta_4 x_{ij2}x_{ij3} + b_i, \tau^{-1})$, $b_i \sim N(0, \tau^{-1}\xi^{-1})$.

We generated 1, 000 simulated datasets from M1 and then calculated AIC, BIC, DIC, and BCIC for the five candidate models M1–M5.

Tables 1 and 2 show the number of times out of 1000 simulations that each rank was achieved for the true model M1 for all model selection criteria. The columns correspond to the rankings of AIC, BIC, and DIC under different settings, and the rows corresponds to the proposed BCIC criteria for different choices of k and I_S. Table 1 provides the results for the setting with $n = 10$ and m_i varying between 3 and 10, representing deletion of moderate numbers of observations in an unbalanced design, whereas Table 2 shows the results for the setting with $n = 20$ and m_i varying between 3 and 15, a setup with deletion of a relatively large number of observations in an unbalanced design. In the simulation, 1,000 burn-in and 5,000 Gibbs samples were used in the calculation. The convergence of the Gibbs sampler was checked by trace plots, but was not included here.

With $n = 10$, m_i from [3, 10], and ICC = 0.5, M1 was ranked number one 556 (= 349 + 118 + 53 + 33 + 3) times by AIC, 548 times by BIC, 467 times by DIC, 390 times by BCIC(I_{LOO}) and 544 times by BCIC(I_{SO}) for $C_0 = 2$, and 462 times by BCIC(I_{LOO}) and 561 times by BCIC(I_{SO}) for $C_{0,n} = \log(N)$, respectively. With ICC increasing to 0.96, M1 was ranked number one 675 times by AIC, 887 times by BIC, 536 times by DIC, 452 times by BCIC(I_{LOO}) and 652 times by BCIC(I_{SO}) for $C_0 = 2$, and 582 times by BCIC(I_{LOO}) and 875 times by BCIC(I_{SO}) for $C_{0,n} = \log(N)$, respectively.

With $n = 20$, m_i from [3, 15], and ICC= 0.5, M1 was ranked number one 719 times by AIC, 847 times by BIC, 571 times by DIC, 614 times by BCIC(I_{LOO}) and 724 times by BCIC(I_{SO}) for $C_0 = 2$, and 737 times by BCIC(I_{LOO}) and 837 times by BCIC(I_{SO}) for $C_{0,n} = \log(N)$, respectively. With ICC increasing to 0.96, M1 was ranked number one 727 times by AIC, 966 times by BIC, 587 times by DIC, 610 times by BCIC(I_{LOO}) and 749 times by BCIC(I_{SO}) for $C_0 = 2$, 839 times by BCIC(I_{LOO}) and 970 times by BCIC(I_{SO}) for $C_{0,n} = \log(N)$, respectively.

These results indicate that there is no single model selection criterion can dominate the rest. Considering different BCIC approaches, BCIC(I_{SO})s outperforms BCIC(I_{LOO})s for both the AIC- and BIC-type penalty terms, the BIC-type penalty term outperforms the AIC-type penalty term, and $BCIC(I_{SO})$ with BIC-type penalty has the best performance within BCIC model selection criteria. Compared with other existing model selection criteria, BCIC(I_{SO}) with $C0_{,n} = \log(N)$ perform similar to BIC, while BCIC(I_{SO}) with $C_0 = 2$ perform similar to AIC. The performances of DIC and BCIC(I_{LOO}) with $C_{0,n} = 2$ or $C_{0,n} = \log(N)$ are among the worst in all scenarios.

3.2 Yale Infant Growth Data

We consider the Yale infant growth data, which studies whether cocaine exposure during pregnancy may lead to the maltreatment of infants after birth, such as physical and sexual abuse. There are a total of 298 children with 3176 records recruited from two exposure groups, the cocaine exposure group and the unexposed group. In this dataset, a unique feature is that different children had different numbers of visits, ranging from 2 to 30 (interquantile range: 7–13), as well as different patterns of visits during the study period. See

Merikangas et al. [26] for a detailed description of the study design and data collection. We apply the proposed BCIC method and compare it to existing model selection criteria for these data to illustrate the application of BCIC.

The multivariate adaptive splines for the analysis of longitudinal data (MASAL) were used to analyze the Yale infant growth data in Zhang [40]. [40] selected the MASAL model

$$y_{ij}=\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\varepsilon_{ij},$$

where the $\boldsymbol{x}_{ij}$ are the potential fixed effects covariates, given by

$$\begin{aligned}\boldsymbol{x}_{ij}=&(1,d,(d-120)^+,(d-200)^+,(g_a-28)^+,\\ &d(g_a-28)^+,(d-60)^+(g_a-28)^+,\\ &(d-490)^+(g_a-28)^+,sd,s(d-120)^+)^T,\end{aligned} \quad (10)$$

in which d and g_a are the age at visit and gestation age, respectively, and s is the indicator for gender with 1 indicating a girl and 0 indicating a boy. In addition, we assume that $\varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{im_i})^T \sim N(\mathbf{0}, \Sigma_i(\tau, \xi))$ and $\Sigma_i(\tau,\xi)$ is determined by the dispersion parameter τ and additional parameters ξ. During this reanalysis, we considered two covariance structures for $\Sigma_i(\tau,\xi)$, including the AR(1) and compound symmetry (CS) structures, along with four sets of fixed effect covariates: (a) $\boldsymbol{x}_{ij}$; (b) $(\boldsymbol{x}_{ij}^T,(d-120)^+(g_a-28)^+)$; (c) $(\boldsymbol{x}_{ij}^T,(d-200)^+(g_a-28)^+)$; (d) $(\boldsymbol{x}_{ij}^T,s(d-200)^+)$. The combinations of different covariance structures and fixed effects lead to a total of eight candidate models. The same priors of Section 3.1 were used in the real data analysis. The additional correlation coefficient parameters in the AR(1) and CS had independent Unif(−1, 1) priors.

Table 3 shows the values of AIC, BIC, DIC, and four BCIC measures normalized by N_B as well as the ranks of all eight candidate models for each criterion. The best model selected by the different criteria are slightly different − AIC, BIC, and DIC ranked the mixed model with the fixed effects of $\boldsymbol{x}_{ij}^T\boldsymbol{\beta}$ and the AR(1) covariance structure as the best model, and the four BCIC measurements ranked the model with the fixed effects $(\boldsymbol{x}_{ij}^T,(d-200)^+(g_a-28)^+)^T\boldsymbol{\beta}$ and the AR(1) covariance structure of AR(1) as the best model. However, the numerical values of the measurements for the models ranked from 1–4 (all the models with AR(1) covariance structure) and for the models ranked from 5–8 (all the models with CS covariance structure) are almost indistinguishable, implying great uncertainty of the ranking decision. Furthermore, the finding that models with the AR(1) covariance structure always provides a better fit to these data than the models with the CS covariance structure, is consistent with the longitudinal nature of this dataset.

4. DISCUSSION

We have systematically examined the connection between Bayesian case influence measures and Bayesian predictive methods. Based on these connections, we have developed a BCMC measure for quantifying the effective number of parameters in a given statistical model and a BCIC measure for comparing models. We have systematically investigated some properties of BCIC and BCMC and their connections with cross-validation and other existing information criteria. We have shown that BCIC is a valuable tool for Bayesian model assessment.

APPENDIX: ASSUMPTIONS AND PROOFS

We need to introduce some notation. Let $F_N(\theta) = \partial_\theta \log p(\theta|\boldsymbol{Y})$ and $F_{N,[S]}(\theta) = \partial\theta \log p(\theta|\boldsymbol{Y}_{[S]})$. Under certain conditions [6], the posterior mode $\hat{\theta}$ converges to the θ_{n*} that minimizes $E\{-\log p(\theta|\boldsymbol{Y})\}$, where the expectation is taken with respect to the true distribution of $\boldsymbol{Y}$. For simplicity, we further assume that $\theta_{n*} = \theta_*$ 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, such as $O_p(1)$, $o_p(1)$, 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

$\hat{\theta}$ and $\hat{\theta}_{[S]}$ for all S are consistent estimates of $\theta_* \in \Theta^o$.

Assumption C2

Let $\Delta(\theta) = \theta - \theta_*$ and suppose

$$\log p(\boldsymbol{\theta}|\boldsymbol{Y}) = \log p(\boldsymbol{\theta}_*|\boldsymbol{Y}) + \Delta(\boldsymbol{\theta})^T F_N(\boldsymbol{\theta}_*) - 0.5\Delta(\boldsymbol{\theta})^T J_N(\boldsymbol{\theta}_*)\Delta(\boldsymbol{\theta})[1+o_p(1)]$$

and

$$\log p(\boldsymbol{\theta}|\boldsymbol{Y}_{[S]}) = \log p(\boldsymbol{\theta}_*|\boldsymbol{Y}_{[S]}) + \Delta(\boldsymbol{\theta})^T F_{N,[S]}(\boldsymbol{\theta}_*) - 0.5\Delta(\boldsymbol{\theta})^T J_{N,[S]}(\boldsymbol{\theta}_*)\Delta(\boldsymbol{\theta})[1+o_p(1)]$$

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

$$0 < \inf_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*, \delta_0 N^{-1/2})} \lambda_{\min}(n^{-1}J_N(\boldsymbol{\theta})) \le \sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*, \delta_0 N^{-1/2}} \lambda_{\max}(N^{-1}J_N(\boldsymbol{\theta})) < \infty,$$

and

$$0 < \min_{S\in I_S} \inf_{\theta\in B(\boldsymbol{\theta}_*, \delta_0 N^{-1/2})} \lambda_{\min}(N^{-1}J_{N,[S]}(\boldsymbol{\theta})) \le \max_{S\in I_S} \sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*, \delta_0 N^{-1/2})} \lambda_{\max}(N^{-1}J_{N,[S]}(\boldsymbol{\theta})) < \infty.$$

Assumption C3

Assume that for small $\theta_0 > 0$, if $N_S \le N_0$, a fixed constant, then

$$\max_{S\in I_S}\sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*,\delta_0)}\|\partial_{\boldsymbol{\theta}}\log p_S(\boldsymbol{\theta})\|=O_p(1)$$

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 C4

$\log p(\theta|\boldsymbol{Y})$ and $\log p(\theta|\boldsymbol{Y}_{[S]})$ for all $S \in I_S$ are Laplace regular [19].

Assumption C5

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

$$\sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*,\delta_0)}\|K_N(I_S|\boldsymbol{\theta})-E[K_N(I_S|\boldsymbol{\theta})]\|=o_p(1)$$

and

$$\sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_*,\delta_0)}\|J_N(I_S|\boldsymbol{\theta})-E[J_N(I_S|\boldsymbol{\theta})]\|=o_p(1).$$

Assumption C6

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

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 [3]. Sufficient conditions of Assumptions C1 and C2 have been extensively discussed in the literature [3]. Assumption C3 is 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 Assumption C3. Assumption C4 is needed to use the Laplace approximation formula [19, 36]. Assumption C5 is ensured by the law of large numbers [38]. Assumption C6 is usually ensured by central limit theory. Recall that $p_S(\theta) = p(\boldsymbol{Y}_S|\boldsymbol{Y}_{[S]}, \theta)$. If $p_S(\theta)$ only depends on a few observations in $\boldsymbol{Y}_{[S]}$, then we can apply the theory of U-statistics to establish Assumption C6 [37].

Proof of Theorem 1

It follows from Assumptions C1–C3 that we can expand $\log p_{S_k}(\tilde{\theta}_{[S_k]})$ at $\tilde{\theta}$ for each S and obtain

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

where $\Delta S_k = \tilde{\theta}_{[S_k]} - \tilde{\theta}$. It follows from Proposition 1 (c) that

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

which yields Theorem 1 (i). Theorem 1 (ii) directly follows from Assumptions C1, C2, and C5.

Proof of Theorem 2

We consider the exhaustive splitting for the leave-M clusters-out CV. For any $S_k = \{\{i_1\}, \cdots, \{i_M\}\}$, we have

$$\log p_{S_k}(\tilde{\boldsymbol{\theta}}(M_l), M_l) = \sum_{l=1}^{M} \log p(\boldsymbol{Y}_{i_l} | \boldsymbol{\theta}, M_l),$$

$$\sum_{S_k \in I_S} \log p_{S_k}(\tilde{\boldsymbol{\theta}}(M_l), M_l) = \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} \sum_{i=1}^{n} \log p(\boldsymbol{Y}_i | \boldsymbol{\theta}, M_l),$$

and

$$\sum_{S_k \in I_S} E\{[\partial_{\boldsymbol{\theta}} \log p_{S_k}(\boldsymbol{\theta}_*)]^{\otimes 2}\} = \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} \sum_{i=1}^{n} E\{\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}_i | \boldsymbol{\theta}_*)^{\otimes 2}\}.$$

Therefore, we have

$$\mathrm{BCIC}(I_{LMO}, M_l) = \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} \mathrm{BCIC}(I_{LOO}, M_l),$$

which yields Theorem 2 (i). Theorem 2 (ii) directly follows from Assumption C6.

REFERENCES

1. Akaike H. A New Look at the Statistical Model Identification. IEEE Transactions On Automatic Control. 1974; 19:716–723.

2. Ando T. Bayesian Predictive Information Criterion for the Evaluation of Hierarchical Bayesian and Empirical Bayes Models. Biometrika. 2007; 94:443–458.
3. Andrews DWK. Estimation When a Parameter Is on a Boundary. Econometrica. 1999; 67:1341–1383.
4. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Statistics Surveys. 2010; 4:40–79.
5. Box GEP, Tiao GC. A Bayesian approach to some outlier problems. Biometrika. 1968; 55:119–129. [PubMed: 5661037]
6. Bunke O, Milhaud X. Asymptotic Behavior of Bayes Estimates under Possibly Incorrect Models. The Annals of Statistics. 1998; 26(2):617–644.
7. Chaloner K. Bayesian residual analysis in the presence of censoring. Biometrika. 1991; 78:637–644.
8. Chen M-H, Dey DK, Ibrahim JG. Bayesian Criterion Based Model Assessment for Categorical Data. Biometrika. 2004; 91(1):45–63.
9. Cook RD. Detection of Influential Observation in Linear Regression. Technometrics. 1977; 19:15–18.
10. Cook, RD.; Weisberg, S. Residuals and Inuence in Regression. Chapman & Hall Ltd.; 1982.
11. Csiszár I. Information-type Measures of Difference of Probability Distributions and Indirect Observations. Studia Scientiarum Mathematicarum Hungarica. 1967; 2:299–318.
12. Geisser, S. Predictive Inference: an Introduction. London: Chapman & Hall Ltd.; 1993.
13. Geisser S, Eddy WF. A Predictive Approach to Model Selection (Corr: V75 P765). Journal of the American Statistical Association. 1979; 74:153–160.
14. Gelfand, AE.; Dey, DK.; Chang, H. Model Determination Using Predictive Distributions, with Implementation Via Sampling-based Methods (Disc: P160–167). In: Bernardo, JM.; Berger, JO.; Dawid, AP.; Smith, AFM., editors. Bayesian Statistics 4. Oxford: Oxford University Press; 1992. p. 147-159.
15. Gelfand AE, Ghosh SK. Model Choice: A Minimum Posterior Predictive Loss Approach. Biometrika. 1998; 85:1–11.
16. Ibrahim JG, Chen M-H, Sinha D. Criterion-based Methods for Bayesian Model Assessment. Statistica Sinica. 2001; 11(2):419–443.
17. Ibrahim JG, Laud PW. A Predictive Approach to the Analysis of Designed Experiments. Journal of the American Statistical Association. 1994; 89:309–319.
18. Johnson W, Geisser S. A Predictive View of the Detection and Characterization of Influential Observations in Regression Analysis. Journal of the American Statistical Association. 1983; 78:137–144.
19. Kass, RE.; Kadane, JB.; Tierney, L. Asymptotic Evaluation of Integrals Arising in Bayesian Inference. In: Page, C.; LePage, R., editors. Computing Science and Statistics: Proceedings of the Symposium on the Interface. Springer-Verlag Inc; 1990. p. 38-42.
20. Kass RE, Tierney L, Kadane JB. Approximate Methods for Assessing Influence and Sensitivity in Bayesian Analysis. Biometrika. 1989; 76:663–674.
21. Konishi S, Ando T, Imoto S. Bayesian information criteria and smoothing parameter selection in radial basis function networks. Biometrika. 2004; 91:27–43.
22. Konishi S, Kitagawa G. Generalised Information Criteria in Model Selection. Biometrika. 1996; 83:875–890.
23. Laud P, Ibrahim J. Predictive Model Selection. Journal of the Royal Statistical Society, Series B. 1995; 57:247–262.
24. Lv J, Liu JS. Model selection principles in misspecified models. J. R. Statist. Soc. B. 2014
25. McCulloch RE. Local Model Influence. Journal of the American Statistical Association. 1989; 84:473–478.
26. Merikangas KR, Stevens DE, Fenton B, Stolar M, O'Malley S, Woods S, Risch N. Co-morbidity and familial aggregation of alcoholism and anxiety disorders. Psychological Medicine. 1998; 28:773–788. [PubMed: 9723135]

27. Murata N, Yoshizawa S, Amari S. Network Information Criterion Determining the Number of Hidden Units for an Artificial Neural Network Model. IEEE Transactions on Neural Networks. 1994; 5:865–872. [PubMed: 18267861]
28. Peng F, Dey DK. Bayesian Analysis of Outlier Problems Using Divergence Measures. The Canadian Journal of Statistics / La Revue Canadienne de Statistique. 1995; 23:199–213.
29. Pettit LI. Diagnostics in Bayesian Model Choice. The Statistician: Journal of the Institute of Statisticians. 1986; 35:183–190.
30. Ripley, BD. Pattern Recognition and Neural Networks. Cambridge University Press; 1996.
31. Schwarz G. Estimating The Dimension of a Model. The Annals of Statistics. 1978; 6:461–464.
32. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian Measures of Model Complexity and Fit (with Discussion). Journal of the Royal Statistical Society, Series B: Statistical Methodology. 2002; 64:583–639.
33. Stone M. Cross-validatory Choice and Assessment of Statistical Predictions (with Discussion) (Corr:76V38 P102). Journal of the Royal Statistical Society, Series B: Methodological. 1974; 36:111–147.
34. Stone M. An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. Journal of the Royal Statistical Society, Series B: Methodological. 1977; 39:44–47.
35. Takeuchi K. Distribution of information statistics and critera for adequacy of models (in Japanese). Mathematical Sciences. 1976; 153:12–18.
36. Tierney L, Kass RE, Kadane JB. Fully Exponential Laplace Approximations to Expectations and Variances of Non-positive Functions. Journal of the American Statistical Association. 1989; 84:710–716.
37. van der Vaart, AW. Asymptotic Statistics. Cambridge University Press; 1998.
38. van der Vaart, AW.; Wellner, JA. Weak Convergence and Empirical Processes. Springer-Verlag Inc.; 1996.
39. Vehtari A, Ojanen J. A survey of Bayesian predictive methods for model assessment, selection and comparison. Statistics Surveys. 2012; 6:142–228.
40. Zhang H. Analysis of Infant Growth Curves Using Multivariate Adaptive Splines. Biometrics. 1999; 55:452–459. [PubMed: 11318199]
41. Zhang P. Model Selection Via Multifold Cross Validation. The Annals of Statistics. 1993; 21:299–313.
42. Zhu, H.; Ibrahim, J.; Cho, N.; Tang, N. A general review of Bayesian influence analysis. In: Chen, MH.; Dey, DK.; Muller, P.; Sun, D.; Ye, K., editors. Frontier of statistical decision making and Bayesian analysis. New York: Springer-Verlag; 2010. p. 219-237.
43. Zhu H, Ibrahim J, Cho N, Tang NS. Bayesian case-deletion measures for statistical models with missing data. Journal of Computational and Graphical Statistics. 2012; 21:253–271. [PubMed: 23399928]
44. Zhu HT, Ibrahim JG, Cho HS. Perturbation and Scaled Cook's distance. Annals of Statistics. 2012; 40:785–811. [PubMed: 23152646]

Table 1

Ranks of the true model M1 for BCIC, AIC, BIC, and DIC in the mixed model. The number of clusters is n = 10 and the number of individuals within each cluster, mi, varying between 3 and 10. Two levels of intracluster correlation coefficient (ICC) are considered.

BCIC Rank	AIC 1	AIC 2	AIC 3	AIC 4	AIC 5	BIC 1	BIC 2	BIC 3	BIC 4	BIC 5	DIC 1	DIC 2	DIC 3	DIC 4	DIC 5
$\xi = 1$ (ICC=0.5)															
BCIC(I_{LOO}) with $C_0 = 2$															
1	349	33	7	1	0	285	104	0	1	0	249	78	45	17	1
2	118	96	15	5	0	121	111	2	0	0	109	82	22	19	2
3	53	48	55	17	0	84	72	14	3	0	62	48	45	16	2
4	33	40	38	36	4	57	69	14	8	3	35	55	36	20	5
5	3	16	12	11	9	1	38	7	3	2	12	10	15	9	5
BCIC(I_{SO}) with $C_0 = 2$															
1	504	33	5	2	0	393	151	0	0	0	361	112	47	23	1
2	43	164	27	1	0	95	140	0	0	0	69	98	43	22	3
3	7	28	78	16	0	45	66	16	2	0	30	46	39	12	2
4	2	8	14	50	2	15	32	17	9	3	6	14	32	20	4
5	0	0	3	1	11	0	5	4	4	2	1	3	2	4	5
BCIC(I_{LOO}) with $C_{0,n} = \log(N)$															
1	360	64	28	10	0	360	98	2	2	0	265	113	58	23	3
2	125	95	40	14	0	98	165	9	0	2	117	87	43	23	4
3	47	41	30	17	5	60	65	8	5	2	48	40	32	17	3
4	20	18	17	19	3	21	40	11	4	1	20	22	20	12	3
5	4	15	12	10	5	9	26	7	4	0	17	11	10	6	2
BCIC(I_{SO}) with $C_{0,n} = \log(N)$															
1	420	86	45	10	0	497	59	4	1	0	334	133	69	23	2
2	136	140	66	23	1	45	314	7	0	0	131	129	62	40	4

BCIC Rank	AIC					BIC					DIC				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
3	0	7	15	19	5	4	17	20	4	1	2	9	24	6	5
4	0	0	1	17	3	2	4	6	7	2	0	1	7	11	2
5	0	0	0	1	4	0	0	0	3	2	0	1	1	1	2
$\xi = 0.04$ (ICC=0.96)															
BCIC(I_{LOO}) with $C_0 = 2$															
1	431	16	5	0	0	451	1	0	0	0	306	98	45	3	0
2	137	50	11	2	0	191	9	0	0	0	119	48	29	4	0
3	69	37	41	11	0	132	21	3	2	0	60	44	42	12	0
4	38	28	41	76	0	113	34	28	8	0	51	41	56	35	0
5	0	0	1	5	0	0	0	4	2	0	0	1	2	3	0
BCIC(I_{SO}) with $C_0 = 2$															
1	619	26	7	0	0	650	2	0	0	0	449	144	52	7	0
2	51	78	17	4	0	139	11	0	0	0	61	48	34	7	0
3	5	24	58	10	0	73	19	5	0	0	18	27	41	11	0
4	0	3	17	80	0	25	33	30	12	0	8	13	47	32	0
BCIC(I_{LOO}) with $C_{0,n} = \log(N)$															
1	502	53	19	8	0	572	7	2	1	0	359	135	78	10	0
2	106	44	26	13	0	163	18	7	1	0	100	43	36	10	0
3	39	18	35	23	0	84	23	7	1	0	45	22	34	14	0
4	28	16	18	41	0	68	14	15	6	0	31	31	23	18	0
5	0	0	1	9	0	0	3	4	3	0	1	1	3	5	0
BCIC(I_{SO}) with $C_{0,n} = \log(N)$															
1	670	117	69	19	0	852	18	5	0	0	520	215	115	25	0
2	5	14	23	26	0	29	35	4	0	0	13	13	27	15	0
3	0	0	7	32	0	5	11	18	5	0	3	3	21	12	0
4	0	0	0	17	0	1	1	8	7	0	0	1	11	5	0

Table 2

Ranks of the true model M1 for BCIC, AIC, BIC, and DIC in the mixed model. The number of clusters is n = 20 and the number of individuals within each cluster, mi, varying between 3 and 15. Two levels of intracluster correlation coefficient (ICC) are considered.

BCIC Rank	AIC					BIC					DIC				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
ξ = 1 (ICC=0.5)															
BCIC(I_{LOO}) with C_0 = 2															
1	581	23	9	1	0	557	57	0	0	0	427	113	68	6	0
2	95	44	15	2	0	126	30	0	0	0	84	42	25	5	0
3	29	18	42	14	0	88	14	1	0	0	35	38	27	2	1
4	14	17	35	59	0	76	26	21	2	0	25	24	56	19	1
5	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0
BCIC(I_{SO}) with C_0 = 2															
1	690	28	6	0	0	662	62	0	0	0	503	139	76	6	0
2	24	61	21	4	0	82	28	0	0	0	41	41	23	5	0
3	5	12	66	10	0	79	13	1	0	0	24	30	37	2	0
4	0	1	8	62	1	24	25	21	2	0	3	7	40	20	2
BCIC(I_{LOO}) with $C_{0,n}$ = log(N)															
1	598	59	56	24	0	705	28	4	0	0	452	159	112	14	0
2	102	34	29	15	1	101	79	1	0	0	93	44	34	9	1
3	15	6	12	22	0	33	10	11	1	0	20	10	18	6	1
4	4	3	4	13	0	8	11	5	0	0	6	4	10	4	0
5	0	0	0	2	0	0	0	1	1	0	0	0	2	0	0
BCIC(I_{SO}) with $C_{0,n}$ = log(N)															
1	639	78	86	34	0	817	18	2	0	0	508	182	132	15	0
2	80	24	13	22	1	29	104	7	0	0	62	35	28	13	2
3	0	0	2	16	0	1	6	11	0	0	1	0	12	5	0
4	0	0	0	4	0	0	0	2	2	0	0	0	4	0	0

BCIC Rank	AIC 1	AIC 2	AIC 3	AIC 4	AIC 5	BIC 1	BIC 2	BIC 3	BIC 4	BIC 5	DIC 1	DIC 2	DIC 3	DIC 4	DIC 5
$\xi = 0.04$ (ICC=0.96)															
BCIC(I_{LOO}) with $C_0 = 2$															
1	584	16	10	0	0	610	0	0	0	0	443	107	57	3	0
2	85	39	20	2	0	145	1	0	0	0	78	39	26	3	0
3	47	31	57	9	0	140	2	2	0	0	52	43	42	7	0
4	11	12	31	45	0	71	19	8	1	0	14	31	45	9	0
BCIC(I_{SO}) with $C_0 = 2$															
1	714	27	8	0	0	749	0	0	0	0	537	136	73	3	0
2	12	59	19	0	0	89	1	0	0	0	28	35	26	1	0
3	1	11	85	10	0	106	0	1	0	0	19	38	40	10	0
4	0	1	6	46	0	22	21	9	1	0	3	11	31	8	0
BCIC(I_{LOO}) with $C_{0,n} = \log(N)$															
1	676	72	75	16	0	833	4	2	0	0	532	169	127	11	0
2	39	20	23	16	0	89	7	1	1	0	40	31	22	5	0
3	10	3	18	13	0	35	6	3	0	0	11	15	14	4	0
4	2	3	2	11	0	9	5	4	0	0	4	5	7	2	0
BCIC(I_{SO}) with $C_{0,n} = \log(N)$															
1	727	98	112	33	0	960	9	1	0	0	586	213	155	16	0
2	0	0	6	14	0	5	11	4	0	0	1	7	9	3	0
3	0	0	0	6	0	1	2	3	0	0	0	0	4	2	0
4	0	0	0	3	0	0	0	2	1	0	0	0	2	1	0

Table 3

Model selection results of Yale infant growth data. CS: compound symmetric; x_{ij} is defined in (11), $\boldsymbol{x}_{ij}^{*}=(\boldsymbol{x}_{ij}^{T},(d-120)^{+}(g_a-28)^{+})$, $\boldsymbol{x}_{ij}^{\dagger}=(\boldsymbol{x}_{ij}^{T},(d-200)^{+}(g_a-28)^{+})$, and $\boldsymbol{x}_{ij}^{\ddagger}=(\boldsymbol{x}_{ij}^{T},s(d-200)^{+})$

Candidate Models		AIC	BIC	DIC	BCIC & C_0 = 2		BCIC & $C_{0,n}$ = log(N)	
Fixed Effect	**Covariance Structure**	**(Rank)**	**(Rank)**	**(Rank)**	I_{LOO}/N_B **(Rank)**	I_{SO}/N_B **(Rank)**	I_{LOO}/N_B **(Rank)**	I_{SO}/N_B **(Rank)**
x_{ij}	AR(1)	5037.81 (1)	5110.58 (1)	5047.17 (1)	4997.26 (4)	4998.06 (4)	5011.07 (4)	5011.11 (4)
	CS	6426.16 (8)	6498.93 (5)	6426.75 (8)	6380.94 (8)	6381.72 (8)	6397.27 (8)	6397.30 (8)
$\boldsymbol{x}_{ij}^{*}$	AR(1)	5039.48 (4)	5118.31 (4)	5048.62 (3)	4996.93 (3)	4997.76 (3)	5010.74 (3)	5010.79 (3)
	CS	6425.82 (7)	6504.65 (8)	6426.01 (7)	6378.61 (7)	6379.42 (7)	6394.93 (7)	6394.96 (7)
$\boldsymbol{x}_{ij}^{\dagger}$	AR(1)	5038.17 (2)	5117.00 (2)	5047.75 (2)	4995.63 (1)	4996.46 (1)	5009.43 (1)	5009.48 (1)
	CS	6424.45 (5)	6503.28 (6)	6423.48 (5)	6377.25 (6)	6378.06 (6)	6393.56 (5)	6393.59 (5)
$\boldsymbol{x}_{ij}^{\ddagger}$	AR(1)	5038.83 (3)	5117.65 (3)	5048.73 (4)	4996.28 (2)	4997.11 (2)	5010.09 (2)	5010.13 (2)
	CS	6424.46 (6)	6503.28 (7)	6424.33 (6)	6377.25 (5)	6378.06 (5)	6393.57 (6)	6393.60 (6)