Criteria for longitudinal data model selection based on Kullback's symmetric divergence

Bezza Hafidi, Nourddine Azzaoui

Criteria for longitudinal data model selection based on Kullback's symmetric divergence

Bezza Hafidi* — Nourddine Azzaoui**

* Faculty of science Agadir
University Ibn Zohr
Agadir, Morocco
Email: b.hafidi@ucam.ac.ma

** Laboratoire de Mathématiques, UMR CNRS 6620
Campus des Céseaux, BP 80026,
63171, Aubière Cedex, France
Email: nourddine.azzaoui@math.univ-bpclermont.fr

ABSTRACT. Recently, Azari et al (2006) showed that (AIC) criterion and its corrected versions cannot be directly applied to model selection for longitudinal data with correlated errors. They proposed two model selection criteria, AICc and RICc, by applying likelihood and residual likelihood approaches. These two criteria are estimators of the Kullback-Leibler's divergence distance which is asymmetric. In this work, we apply the likelihood and residual likelihood approaches to propose two new criteria, suitable for small samples longitudinal data, based on the Kullback's symmetric divergence. Their performance relative to others criteria is examined in a large simulation study.

RÉSUMÉ. Récemment, Azari et al. (2006) ont montré que le critère (AIC) ainsi que ses versions corrigées ne peuvent pas être directement appliqués aux données longitudinales avec des erreurs corrélées. Ils ont proposé deux critères, AICc et RICc, en utilisant la notion de vraisemblance et de vraisemblance résiduelle. Leurs critères sont des estimations de la distance asymétrique de divergence de Kullback-Leibler. Dans ce travail, nous proposons deux nouveaux critères adaptés aux échantillons de petites tailles de données longitudinales en se basant sur la divergence symétrique de Kulback et les approches de maximum de vraisemblance et de vraisemblance résiduelle. Les performances de ces critères sont examinées dans une étude de simulations.

KEYWORDS : Model selection, AIC, KIC, Longitudinal data analysis, Kullback's symmetric divergence.

MOTS-CLÉS : Sélection de modèles, AIC, KIC, Analyse des données longitudinales, divergence symétrique de Kullback.

Received, July 16, 2009
Revised, January 28, 2010,
Accepted, April 7, 2010.

ARIMA Journal, vol. 15 (2012), pp. 83-99

1. Introduction

Model selection criteria play an important role in applied statistical data analysis especially in parametric and non parametric regression, mixture models, time series analysis... The most known is the Akaike Information Criterion (AIC) which is applicable in different arrays of modeling, see Akaike [1]. The AIC criterion was designed as an unbiased estimator of a variant of the Kullback's directed divergence between the model which presumably gave rise to the data and a fitted approximating one.

The directed divergence, also known as the Kullback-Leibler information, is a measure of separation between two statistical models. It is asymmetric, meaning that an alternate directed divergence can be obtained by reversing the roles of the two models in the definition of the measure. The sum of the two directed divergences is the Kullback's symmetric divergence, which combines the information in both measures. Therefore, it functions as gauge of model disparity which is arguably more balanced than either of individual directed divergences. Model selection criteria based on the symmetric measure was first investigated by Cavanaugh [5]. He proposed an Akaike type criterion, named KIC, as an asymptotically unbiased estimator of the symmetric divergence in the case of large sample data.

When the sample size is small or the number of fitted parameters is large to moderate fraction of the sample size, both AIC and KIC criteria suffer from a negative bias and result in a serious overfitting. In order to overcome this problem, many corrected versions of AIC was proposed in different special situations. For instance, Hurvich and Tsai [16] proposed a corrected AIC (AICc) for linear and non-linear regression and autoregressive modeling. Their work was extended to the case of autoregressive moving average by Hurvich [14], to vector autoregressive modeling by Hurvich and Tsai [15] and multivariate regression by Bedrick et al. [3], Fujikoshi and Satoh [9].... On the other hand, using the Kullback's symmetric divergence Cavanaugh [4] proposed a corrected Kullback's information criterion (KICc) for linear models. Hafidi and Mkhadri [11] generalized the corrected KICc for multiple to multivariate regression and for univariate or multivariate autoregressive modeling.

Longitudinal data analysis has been a great deal of interest in the fields of clinical trials, epidemiology, agriculture and medicine over the last decade. These data arise when repeated measurements are obtained for an individual or more outcome variables at successive points in time. The successive measurements for each individual tend to be correlated. Therefore, it is necessary to take into account this correlation in order to produce proper analysis. The comprehensive synthesis of both theoretical, applied aspects, model structure details and parameter estimation about longitudinal analysis is given in Diggle [7], Hedeker and Gibbons [13], Fitzmaurice et al. [8], Hand and Crowder [12] and Jones [17].

Recently Azari et al. [2] showed that the classical corrected AICc criteria can not be di-

rectly applied to model selection for longitudinal data with correlated errors. They derived two model selection criteria: the first one is obtained by applying the maximum likelihood approach and the second is RICc derived by using the residual (restricted) likelihood approach. Both these two criteria are estimator of the Kullback-Leibler's divergence distance which is asymmetric. Cavanaugh [5, 4] suggested that the Kullback's symmetric distance is a preferable tool for model selection than the asymmetric one. In this paper, we use the kullback's symmetric divergence and we apply the likelihood and residual likelihood approaches to derive corrected version suitable for small sample longitudinal data. A large simulation investigations for these criteria show their appropriateness for model selection especially when the sample size and the signal to noise ration are small. In this case, they can be used as an alternative to classical criteria.
The paper is organized as follows. Section 2 is devoted to the description of model structure of longitudinal data. Derivation of the corrected KIC$_c$, is presented in Section 3. In Section 4, we presented the criteria obtained by using the residual likelihood. We end the paper with simulation results and with a small conclusion.

2. Preliminaries and notations

Let y_{ij} represents the measurement for the i^{th} subject at the time points t_{ij}. we denote by $Y_i = (y_{i1}, \ldots, y_{in_i})^t$ a $(n_i \times 1)$ vector of all repeated observations on the i^{th} subject. The sign $'^t'$ stands for the transpose. For simplicity, we assume that $n_i = n$ for all i. Suppose that the true model is given by

$$Y_i = X_{i0}\beta_0 + \varepsilon_i \qquad i = 1, \ldots, m, \qquad [1]$$

where $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in})$ is a $(n \times 1)$ vector at time n. It is assumed to follow a multivariate normal distribution with mean zero and covariance matrix $\sigma_0^2\Sigma_0$. β_0 is a $(p_0 \times 1)$ vector of unknown parameters. Finally X_{i0} is an $(n \times p_0)$ matrix with the j^{th} row $(x^0_{ij_1}, \ldots, x^0_{ij_{p_0}})$. Define the observation $(N \times 1)$ vector Y as $Y = (Y_1^t, \ldots, Y_m^t)^t$ where $N = nm$. Let $X_0 = (X_{10}^t, \ldots, X_{m0}^t)^t$ be a $(N \times p_0)$ matrix of explanatory variables and $\varepsilon_0 = (\varepsilon_1^t, \ldots, \varepsilon_m^t)^t$ the error vector. With this notation, we can form a general regression model

$$Y = X_0\beta_0 + \varepsilon_0,$$

where $\varepsilon_0 \sim \mathcal{N}(0, \sigma_0^2 V_0)$ and V_0 is an $(N \times N)$ block diagonal matrix with $(n \times n)$ blocks Σ_0.
In practice, we do not know the true model described below, thus we fit the data to a candidate one

$$Y_i = X_i\beta + \varepsilon_i \qquad i = 1, \ldots, m, \qquad [2]$$

where β is a $(p \times 1)$ vector and X_i is an explanatory $(n \times p)$ matrix with j^{th} row $(x_{ij_1}, \ldots, x_{ij_p})$ and $\varepsilon = (\varepsilon_1^t, \ldots, \varepsilon_m^t)^t$ is the vector error with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2\Sigma)$. As previous, the model (2) can be written as

$$Y = X\beta + \varepsilon,$$

where Y is a dependent $(N \times 1)$ vector, X is an $(N \times p)$ matrix of explanatory and $\varepsilon \sim \mathcal{N}(0, \sigma^2 V)$ and V is an $(N \times N)$ block diagonal matrix with $(n \times n)$ blocks Σ. With these previous considerations, we denote $\theta = (\beta, \sigma^2, \Sigma, p)$ and $\theta_0 = (\beta_0, \sigma_0^2, \Sigma_0, p_0)$ the parameter vectors of the candidate and the true models respectively. Let $f_{\theta_0}(Y)$ and $f_\theta(Y)$ represent the generating and candidate densities for data respectively. In this case, ignoring the constant term, the log-likelihood for the candidate model is given by

$$\log f_\theta(Y) = -\frac{1}{2}\left\{N \log \sigma^2 + m \log |\Sigma| + \frac{(Y - X\beta)^t V^{-1}(Y - X\beta)}{\sigma^2}\right\}. \qquad [3]$$

Similarly the log likelihood function of the true model is obtained by replacing θ, σ, β, Σ and V in equation (3) by θ_0, σ_0, β_0, Σ_0 and V_0 respectively.

REMARK. — It should be noted that, in longitudinal data analysis, the covariance matrix Σ is often assumed to have a special morphology. For example:

– **The uniform structure:** where $\Sigma = (\sigma_{jl})_{j,l=1\ldots n}$ is defined by

$$\sigma_{jl} = \begin{cases} \rho & \text{for} \quad t_{ij} \neq t_{il} \\ 1 & \text{for} \quad t_{ij} = t_{il} \end{cases}$$

and ρ is a positive correlation coefficient between two measurements on the same subject.

– **the exponential correlation structure:** in this case $\sigma_{jl} = \exp(-\gamma|t_{ij} - t_{il}|)$ where γ is the correlation decay rate between two measurements on the same unit.

– **The autoregressive correlation structure:** it is a particular case of the latter where the observation times are equally spaced for all j. In this setting σ_{jl} can be expressed as $\sigma_{jl} = \rho^{|j-l|}$ or $\sigma_{jl} = \frac{\rho^{|j-l|}}{1-\rho^2}$. The parameter ρ is the correlation between successive observations on the same subject.

In this paper we assume that the correlation structure of Σ and Σ_0 is expressed as a function of an unknown parameter vector ϕ and ϕ_0 respectively: meaning that $\Sigma = \Sigma(\phi)$ and $\Sigma_0 = \Sigma_0(\phi_O)$. As Azari et al. [2], we assume the consistency of the maximum likelihood and the restricted maximum likelihood estimators of Σ. In all the rest of the paper we replace the precedent notations of θ and θ_0 by $\theta = (\beta, \sigma^2, \phi)$ and $\theta_0 = (\beta_0, \sigma_0^2, \phi_0)$.

3. Derivation of the corrected (KICc) based on the Kulback's symmetric divergence

A measure of separation between the generating and a candidate model is given by the symmetric divergence (Kullback [18]). It is defined by

$$J(\theta_0,\theta)=\{d(\theta_0,\theta)-d(\theta_0,\theta_0)\}+\{d(\theta,\theta_0)-d(\theta,\theta)\}, \quad [4]$$

where $d(\theta_0,\theta)=I\!E_{\theta_0}\{-2\log f_\theta(Y)\}$ and $I\!E_{\theta_0}$ denotes the expectation with respect to $f_{\theta_0}(Y)$. Ignoring $d(\theta_0,\theta_0)$ which does not depend on θ, for the purpose of discriminating among various models depending on θ we propose a substitution of $J(\theta_0,\theta)$ as,

$$K(\theta_0,\theta)=d(\theta_0,\theta)+\{d(\theta,\theta_0)-d(\theta,\theta)\}. \quad [5]$$

In practice, the exact computation of this quantity is not easily accessible. To overcome this problem, Cavanaugh [5] proposed an asymptotically unbiased estimator of

$$\Omega(\theta_0)=I\!E_{\theta_0}\{K(\theta_0,\hat{\theta})\}. \quad [6]$$

in the case of large sample data, where $\hat{\theta}$ is the maximum likelihood estimator of θ, $K(\theta_0,\hat{\theta})$ has the same expression as in (5) by replacing θ by $\hat{\theta}$. This estimator is given by

$$KIC=N\log\hat{\sigma}^2+m\log|\hat{\Sigma}|+3(p+1). \quad [7]$$

The purpose of this section is to derive a corrected version of this criterion. For this, we assume that the candidate class of models includes the true model. This assumption is also used in the derivation of KIC [5] and its other versions see for instance [2, 4, 10, 11]. In this setting the columns of X can be rearranged so that $X_0\beta_0=X\beta^*$, where $\beta^*=(\beta_0^t,\beta_1^t)^t$ and β_1 is a $(p-p_0)\times 1$ vector of zeros. Under this assumption, we show the forthcoming proposition.

Proposition 1 *The criterion defined by*

$$KICc=N\log\hat{\sigma}^2+m\log|\hat{\Sigma}|+\frac{(p+1)(3N-p-2)}{N-p-2} \quad [8]$$

is an approximate unbiased estimator of $\Omega(\theta_0)$.

Proof: Ignoring the constant term, the log likelihood of the candidate model is given by

$$\log f_\theta(Y)=-\frac{1}{2}\left\{N\log\sigma^2+m\log|\Sigma|+\frac{(Y-X\beta)^tV^{-1}(Y-X\beta)}{\sigma^2}\right\}. \quad [9]$$

The maximum likelihood estimators of β,σ^2 and $V(\phi)$ are given by

$$\hat{\beta}=(X^t\hat{V}^{-1}X)^{-1}X^t\hat{V}^{-1}Y,\quad \hat{\sigma}^2=\frac{(Y-X\hat{\beta})^t\hat{V}^{-1}(Y-X\hat{\beta})}{N}, \quad [10]$$

where $\hat{V}^{-1}$ is $V^{-1}(\phi)$ evaluated at $\hat{\phi}$. This latter is obtained by maximizing the profile log-likelihood: $\ell(\phi) = -\frac{1}{2}(N\hat{\sigma}^2(\phi) + m\log|\Sigma(\phi)|)$, see for instance Azari [2] and for more details see Diggle ([7] p. 64-65).
From (5), the expectation of $K(\theta_0, \hat{\theta})$ has the form

$$\Omega(\theta_0) = I\!E_{\theta_0}\{d(\theta_0, \hat{\theta}) + d(\hat{\theta}, \theta_0) - d(\hat{\theta}, \hat{\theta})\}. \qquad [11]$$

For clarity, we emphasize that

$$d(\theta_0, \hat{\theta}) = I\!E_{\theta_0}\{-2\log f_\theta(Y)\}\Big|_{\theta=\hat{\theta}} \quad \text{and} \quad d(\hat{\theta}, \theta_0) = I\!E_{\theta}\{-2\log f_{\theta_0}(Y)\}\Big|_{\theta=\hat{\theta}}.$$

The notation $\Big|_{\theta=\hat{\theta}}$ means that we calculate the expectation and then we replace θ by $\hat{\theta}$ in the resulting formula. Now, we will compute each term of (11). By adding and subtracting $X\beta^*$, we have

$$\begin{aligned} d(\theta_0, \hat{\theta}) &= I\!E_{\theta_0}\{-2\log f_\theta(Y)\}\Big|_{\theta=\hat{\theta}} \\ &= I\!E_{\theta_0}\left\{N\log\sigma^2 + m\log|\Sigma| + \frac{(Y - X\beta)^t V^{-1}(Y - X\beta)}{\sigma^2}\right\}\Big|_{\theta=\hat{\theta}} \\ &= I\!E_{\theta_0}\left\{N\log\sigma^2 + m\log|\Sigma| + \frac{(Y - X\beta^*)^t V^{-1}(Y - X\beta^*)}{\sigma^2}\right. \\ &\qquad \left. + \frac{(\beta - \beta^*)^t X^t V^{-1} X(\beta - \beta^*)}{\sigma^2}\right\}\Big|_{\theta=\hat{\theta}}. \end{aligned}$$

It is easy to see that the expectation $I\!E_{\theta_0}\{N\log\sigma^2\}\Big|_{\theta=\hat{\theta}} = N\log\hat{\sigma}^2$ and similarly we have

$$I\!E_{\theta_0}\{m\log|\Sigma|\}\Big|_{\theta=\hat{\theta}} = m\log|\hat{\Sigma}|.$$

By using the same reasoning we also have,

$$I\!E_{\theta_0}\left\{\frac{(\beta - \beta^*)^t X^t V^{-1} X(\beta - \beta^*)}{\sigma^2}\right\}\Big|_{\theta=\hat{\theta}} = \frac{(\hat{\beta} - \beta^*)^t X^t \hat{V}^{-1} X(\hat{\beta} - \beta^*)}{\hat{\sigma}^2}.$$

On the other hand, since $(Y - X\beta^*)^t V^{-1}(Y - X\beta^*) = \varepsilon_0^t V^{-1}\varepsilon_0$, consequently we deduce that

$$I\!E_{\theta_0}\left\{\frac{(Y - X\beta^*)^t V^{-1}(Y - X\beta^*)}{\sigma^2}\right\}\Big|_{\theta=\hat{\theta}} = \frac{\sigma_0^2}{\hat{\sigma}^2} tr(\hat{V}^{-1}V_0),$$

where "tr" stands for trace. Finally,

$$d(\theta_0,\hat{\theta}) = N\log\hat{\sigma}^2 + m\log|\hat{\Sigma}| + \frac{\sigma_0^2}{\hat{\sigma}^2}tr(\hat{V}^{-1}V_0) + \frac{(\hat{\beta}-\beta^*)^t X^t\hat{V}^{-1}X(\hat{\beta}-\beta^*)}{\hat{\sigma}^2}. \quad [12]$$

Similarly, we have

$$\begin{aligned} d(\hat{\theta},\theta_0) &= I\!E_\theta\{-2\log f_{\theta_0}(Y)\}\Big|_{\theta=\hat{\theta}} \\ &= I\!E_\theta\left\{N\log\sigma_0^2 + m\log|\Sigma_0| + \frac{(Y-X\beta^*)^tV_0^{-1}(Y-X\beta^*)}{\sigma_0^2}\right\}\Big|_{\theta=\hat{\theta}} \\ &= I\!E_\theta\left\{N\log\sigma_0^2 + m\log|\Sigma_0| + \frac{(Y-X\beta)^tV_0^{-1}(Y-X\beta)}{\sigma_0^2}\right. \\ &\qquad \left. + \frac{(\beta-\beta^*)^tX^tV_0^{-1}X(\beta-\beta^*)}{\sigma_0^2}\right\}\Big|_{\theta=\hat{\theta}}. \end{aligned} \quad [13]$$

With the same technique as in (12), the expectation calculation leads to the equality

$$I\!E_\theta\left\{\frac{(Y-X\beta)^tV_0^{-1}(Y-X\beta)}{\sigma_0^2}\right\}\Big|_{\theta=\hat{\theta}} = I\!E_\theta\left\{\frac{\varepsilon^tV_0^{-1}\varepsilon}{\sigma_0^2}\right\}\Big|_{\theta=\hat{\theta}} = \frac{\hat{\sigma^2}}{\sigma_0^2}tr(\hat{V}V_0^{-1}).$$

This implies that

$$d(\hat{\theta},\theta_0) = N\log\sigma_0^2 + m\log|\Sigma_0| + \frac{\hat{\sigma}^2}{\sigma_0^2}tr(\hat{V}V_0^{-1}) + \frac{(\hat{\beta}-\beta^*)^tX^tV_0^{-1}X(\hat{\beta}-\beta^*)}{\sigma_0^2}. \quad [14]$$

The same reasoning leads to

$$\begin{aligned} d(\hat{\theta},\hat{\theta}) &= I\!E_\theta\{-2\log f_\theta(Y)\}\Big|_{\theta=\hat{\theta}} \\ &= I\!E_\theta\left\{N\log\sigma^2 + m\log|\Sigma| + \frac{(Y-X\beta)^tV^{-1}(Y-X\beta)}{\sigma^2}\right\}\Big|_{\theta=\hat{\theta}} \\ &= I\!E_\theta\left\{N\log\sigma^2 + m\log|\Sigma| + \frac{\varepsilon^tV^{-1}\varepsilon}{\sigma^2}\right\}\Big|_{\theta=\hat{\theta}} \\ &= N\log\hat{\sigma}^2 + m\log|\hat{\Sigma}| + N. \end{aligned} \quad [15]$$

Substituting (12), (14) and (15) into (11) we obtain

$$\Omega(\theta_0) = I\!E_{\theta_0}\left\{N\log\hat{\sigma}^2 + m\log|\hat{\Sigma}| + \frac{\sigma_0^2}{\hat{\sigma}^2}tr(\hat{V}^{-1}V_0)\right.$$

$$+\frac{(\hat{\beta}-\beta^*)^t X^t \hat{V}^{-1} X(\hat{\beta}-\beta^*)}{\hat{\sigma}^2}\Bigg\} \quad [16]$$

$$+I\!E_\theta\Bigg\{N\log\frac{\sigma_0^2}{\hat{\sigma}^2}+m\log\frac{|\Sigma_0|}{|\hat{\Sigma}|}+\frac{\hat{\sigma}^2}{\sigma_0^2}tr(\hat{V}V_0^{-1}) + \frac{(\hat{\beta}-\beta^*)^t X^t V_0^{-1} X(\hat{\beta}-\beta^*)}{\sigma_0^2}-N\Bigg\}. \quad [17]$$

REMARK. — Under the assumption that $\hat{\phi}$ is a consistent estimator of ϕ_0, Azari et al. [2] approximate $\hat{V}$ by V_0; i.e. $\hat{V} = V_0 + o_p(1)$. They have given a corrected version of the Akaike criterion and they showed that

$$AICc = N\log\hat{\sigma}^2 + m\log|\hat{\Sigma}| + 2\frac{N(p+1)}{N-p-2}, \quad [18]$$

is an approximate unbiased estimator of the first term in $\Omega(\theta_0)$ which is in our case the right hand side of equation (16).

Let us remark that

$$\begin{aligned}\hat{\beta} &= (X^t\hat{V}^{-1}X)^{-1}X^t\hat{V}^{-1}(X_0\beta_0+\varepsilon_0)\\ &= (X^t\hat{V}^{-1}X)^{-1}X^t\hat{V}^{-1}X\beta^* + (X^t\hat{V}^{-1}X)^{-1}X^t\hat{V}^{-1}\varepsilon_0\\ &= \beta^* + (X^t\hat{V}^{-1}X)^{-1}X^t\hat{V}^{-1}\varepsilon_0. \quad [19]\end{aligned}$$

Using the assumption that $\hat{\phi}$ is a consistent estimator of ϕ_0 and the fact that $\hat{\beta}-\beta^* = (X^t\hat{V}^{-1}X)^{-1}X^t\hat{V}^{-1}\epsilon_0$, we compute the error term given in equation (17). In the first term, we have the approximation

$$I\!E_{\theta_0}\Bigg\{m\log\frac{|\Sigma_0|}{|\hat{\Sigma}|}\Bigg\} \approx 0. \quad [20]$$

On the other hand, since $\hat{V} = V_0 + o_p(1)$ then the third term in the right-hand side of (17) becomes

$$\begin{aligned}I\!E_{\theta_0}\Bigg\{\frac{\hat{\sigma}^2}{\sigma_0^2}tr(\hat{V}V_0^{-1})\Bigg\} &\approx I\!E_{\theta_0}\Bigg\{\frac{\epsilon_0^t(\hat{V}^{-1}-\hat{V}^{-1}X(X^t\hat{V}^{-1}X)^{-1}X^t\hat{V}^{-1})\epsilon_0}{\sigma_0^2}\Bigg\}\\ &= I\!E_{\theta_0}\Bigg\{\frac{\epsilon_0^t(V_0^{-1}-V_0^{-1}X(X^tV_0^{-1}X)^{-1}X^tV_0^{-1})\epsilon_0}{\sigma_0^2}\Bigg\}\\ &= N-p. \quad [21]\end{aligned}$$

Furthermore, using the expression (19) and by a simple matrix calculus, the fourth term on the right-hand side of (17) becomes,

$$\begin{aligned} I\!E_{\theta_0}\left\{\frac{(\hat{\beta}-\beta^*)^t X^t V_0^{-1} X(\hat{\beta}-\beta^*)}{\sigma_0^2}\right\} &= I\!E_{\theta_0}\left\{\frac{\epsilon_0^t V_0^{-1} X(X^t V_0^{-1} X)^{-1} X^t V_0^{-1}\epsilon_0}{\sigma_0^2}\right\} \\ &= tr(X(X^t V_0^{-1} X)^{-1} X^t V_0^{-1}), \\ &= p. \end{aligned} \quad [22]$$

Now it remains to compute $I\!E_{\theta_0}\{N\log\frac{\sigma_0^2}{\hat{\sigma}^2}\}$. It is known that the distribution of $N(\frac{\hat{\sigma}^2}{\sigma_0^2})$ is a chi-square with $(N-p)$ degrees of freedom. From lemma (3) in [6], we have

$$I\!E_{\theta_0}\{N\log\frac{\sigma_0^2}{\hat{\sigma}^2}\} = (p+1) + o(1). \quad [23]$$

According to (18) and substituting (20), (21), (22) and (23) in the right side of (17), we obtain the estimator of $\Omega(\theta_0)$ given in the proposition.

REMARK. — It is important to note that if p held fixed and N tends to infinity, the second term in the expression (8) of KICc tends to $3(p+1)$. This yields the criterion KIC given in (7).

4. Derivation of RIC$_{sd}$ criterion based on residual likelihood and symmetric divergence

The method of restricted (or residual) maximum likelihood is a way of estimating variance components in a general linear model. It was first introduced by Patterson and Thompson [19]. By adopting the results of Diggle [7] or Verbyla [21], and omitting irrelevant terms, the restricted likelihood for the candidate model (2) is defined by

$$\log f_\theta^{(r)}(Y) = -\frac{1}{2}\left\{(N-p)\log\sigma^2 + m\log|\Sigma| + \log|X^t V^{-1} X| + \frac{Y^t A Y}{\sigma^2}\right\} \quad [24]$$

where $A = V^{-1} - V^{-1}X(X^tV^{-1}X)^{-1}X^tV^{-1}$.
The residual maximum likelihood estimator of (β, σ^2, ϕ) are given by the equalities

$$\tilde{\beta} = (X^t\tilde{V}^{-1}X)^{-1}X^t\tilde{V}^{-1}Y$$

and $\tilde{\sigma}^2 = (Y-X\beta)^t\tilde{V}^{-1}(Y-X\beta)/(N-p)$. Following Azari et al. [2], $\tilde{V}^{-1}$ is V evaluated at $\tilde{\phi}$ which is obtained in this setting by maximizing the residual likelihood given

by $\ell_r(\phi) = \ell(\phi) - \frac{1}{2}\log|X^tV^{-1}(\phi)X|$.
Similarly, the restricted likelihood function for the true model (1), can be obtained by replacing σ, β, Σ, X, p and V in equation (24) with σ_0, β_0, Σ_0, X_0, p_0 and V_0 respectively. A useful measure of the discrepancy between the residual log-likelihood function of candidate and true models is the Kullback's symmetric divergence defined by

$$\Omega_r(\theta_0) = I\!E_{\theta_0}\{d_r(\theta_0,\tilde{\theta}) + d_r(\tilde{\theta},\theta_0) - d_r(\tilde{\theta},\tilde{\theta})\}. \quad [25]$$

where $d_r(\theta_0,\theta) = I\!E_{\theta_0}\{-2\log f_\theta^{(r)}(Y)\}$. Under the assumption that the true model is include in the family of candidate models, we show the following proposition:

Proposition 2 *The criterion defined by*

$$\begin{aligned} RIC_{sd} &= (N-p)\log\tilde{\sigma}^2 + m\log|\tilde{\Sigma}| + p\log(N) + \frac{(N-p)^2}{N-p-2} \\ &\quad +(N-p)\left(\log(\frac{N-p}{2}) - \Psi(\frac{N-p}{2})\right) \end{aligned} \quad [26]$$

is an approximate unbiased estimator of the discrepancy $\Omega_r(\theta_0)$, where Ψ denote the digamma function.

Proof: We have

$$\Omega_r(\theta_0) = I\!E_{\theta_0}\{d_r(\theta_0,\tilde{\theta}) + d_r(\tilde{\theta},\theta_0) - d_r(\tilde{\theta},\tilde{\theta})\}. \quad [27]$$

From the model consideration, it is easy to see that $X^tA = 0$. This implies the following:

$$\begin{aligned} d_r(\theta_0,\tilde{\theta}) &= I\!E_{\theta_0}\{-2\log f_\theta^{(r)}(Y)\}\Big|_{\theta=\tilde{\theta}} \\ &= I\!E_{\theta_0}\Big\{(N-p)\log\sigma^2 + m\log|\Sigma| + \log|X^tV^{-1}X| \\ &\qquad + \frac{(Y-X\beta^*)^tA(Y-X\beta^*)}{\sigma^2}\Big\}\Big|_{\theta=\tilde{\theta}} \\ &= I\!E_{\theta_0}\Big\{(N-p)\log\sigma^2 + m\log|\Sigma| + \log|X^tV^{-1}X| \\ &\qquad + \frac{(Y-X_0\beta_0)^tA(Y-X_0\beta_0)}{\sigma^2}\Big\}\Big|_{\theta=\tilde{\theta}} \\ &= I\!E_{\theta_0}\Big\{(N-p)\log\sigma^2 + m\log|\Sigma| + \log|X^tV^{-1}X| + \frac{\varepsilon_0^tA\varepsilon_0}{\sigma^2}\Big\}\Big|_{\theta=\tilde{\theta}} \\ &= (N-p)\log\tilde{\sigma}^2 + m\log|\tilde{\Sigma}| + \log|X^t\tilde{V}^{-1}X| + \frac{\sigma_0^2}{\tilde{\sigma}^2}tr(\tilde{A}^{-1}V_0), \end{aligned} \quad [28]$$

where $\tilde{A}^{-1}$ is A evaluated at $\tilde{\phi}$. Similarly, we have

$$\begin{aligned} d_r(\tilde{\theta},\theta_0) &= I\!E_\theta\{-2\log f^{(r)}_{\theta_0}(Y)\}\Big|_{\theta=\tilde{\theta}} \\ &= I\!E_\theta\left\{(N-p_0)\log\sigma_0^2 + m\log|\Sigma_0| + \log|X^tV_0^{-1}X| + \frac{Y^tA_0Y}{\sigma_0^2}\right\}\Big|_{\theta=\tilde{\theta}}. \end{aligned}$$

Furthermore, we derive the following simple expectation equalities

$$I\!E_\theta\left\{\frac{Y^tA_0Y}{\sigma_0^2}\right\}\Big|_{\theta=\tilde{\theta}} = I\!E_\theta\left\{\frac{(Y-X\beta^*)^tA_0(Y-X\beta^*)}{\sigma_0^2}\right\}\Big|_{\theta=\tilde{\theta}} = \frac{\tilde{\sigma}^2}{\sigma_0^2}tr(\tilde{V}A_0).$$

With the same technique as in (13), we obtain

$$\begin{aligned} d_r(\tilde{\theta},\theta_0) &= (N-p)\log\sigma_0^2 + m\log|\Sigma_0| + \log|X^tV_0^{-1}X| + \frac{\tilde{\sigma}^2}{\sigma_0^2}tr(\tilde{V}A_0) \\ &\quad + \frac{(\tilde{\beta}-\beta^*)^tX^tA_0X(\tilde{\beta}-\beta^*)}{\sigma_0^2}. \end{aligned} \qquad [29]$$

As it was argued bellow, since $X^tA = 0$ then we have the following

$$I\!E_\theta\left\{\frac{Y^tAY}{\sigma^2}\right\}\Big|_{\theta=\tilde{\theta}} = I\!E_\theta\left\{\frac{(Y-X\beta)^tA(Y-X\beta)}{\sigma^2}\right\}\Big|_{\theta=\tilde{\theta}} = I\!E_\theta\left\{\frac{\varepsilon^tA\varepsilon}{\sigma^2}\right\}\Big|_{\theta=\tilde{\theta}} = tr(\tilde{V}A).$$

The third term of (27) is then given by:

$$\begin{aligned} d_r(\tilde{\theta},\tilde{\theta}) &= I\!E_\theta\{-2\log f^{(r)}_{\theta}(Y)\}\Big|_{\theta=\tilde{\theta}}, \\ &= I\!E_\theta\left\{(N-p)\log\sigma^2 + m\log|\Sigma| + \log|X^tV^{-1}X| + Y^tAY/\sigma^2\right\}\Big|_{\theta=\tilde{\theta}} \\ &= (N-p)\log\tilde{\sigma}^2 + m\log|\tilde{\Sigma}| + \log|X^t\tilde{V}^{-1}X| + tr(\tilde{V}\tilde{A}). \end{aligned} \qquad [30]$$

From (28), (29) and (30), we obtain,

$$\begin{aligned} \Omega_r(\theta_0) &= I\!E_{\theta_0}\left\{(N-p)\log\tilde{\sigma}^2 + m\log|\tilde{\Sigma}| + \log|X^t\tilde{V}^{-1}X| + \frac{\sigma_0^2}{\sigma^2}tr(\tilde{A}^{-1}V_0)\right\} \quad [31] \\ &\quad + I\!E_{\theta_0}\left\{(N-p)\log\frac{\sigma_0^2}{\tilde{\sigma}^2} + m\log\frac{|\Sigma_0|}{|\tilde{\Sigma}|} + \log\frac{|X^tV_0^{-1}X|}{|X^t\tilde{V}^{-1}X|} - tr(\tilde{V}\tilde{A})\right. \\ &\quad \left. + \frac{\tilde{\sigma}^2}{\sigma_0^2}tr(\tilde{V}A_0) + \frac{(\tilde{\beta}-\beta^*)^tX^tA_0X(\tilde{\beta}-\beta^*)}{\sigma_0^2} - p_0\log\sigma_0^2\right\}. \qquad [32] \end{aligned}$$

By adopting the same reasoning as Azari et al. [2], we can approximate equation (31) by

$$N \log \tilde{\sigma}^2 + m \log |\tilde{\Sigma}| + p \log(N) + \frac{(N-p)^2}{N-p-2}. \qquad [33]$$

By using the equality (19) and simple matrix simplifications, we have

$$\begin{aligned}
\Delta &\triangleq I\!E_{\theta_0}\left\{ \frac{(\tilde{\beta}-\beta^*)^t X^t A_0 X (\tilde{\beta}-\beta^*)}{\sigma_0^2} \right\} \\
&= I\!E_{\theta_0}\left\{ \frac{\varepsilon_0^t V_0^{-1} X^t (X^t V_0^{-1} X)^{-1} X^t A_0 X (X^t V_0^{-1} X)^{-1} X^t V_0^{-1} \varepsilon_0}{\sigma_0^2} \right\} \\
&= tr(X (X^t V_0^{-1} X)^{-1} X^t A_0) \\
&= tr(X (X^t V_0^{-1} X)^{-1} X^t V_0^{-1}) \\
&\qquad -tr(X (X^t V_0^{-1} X)^{-1} X^t V_0^{-1} X_0 (X_0^t V_0^{-1} X_0)^{-1} X_0^t V_0^{-1}). \\
&= p - p_0. \qquad [34]
\end{aligned}$$

Under the same assumption used in the previous section, $\tilde{\phi}$ is a consistent estimator of ϕ_0, we can approximate (in probability) $\tilde{V}$ by V_0 and $\tilde{\Sigma}$ by Σ_0. Hence, we have

$$I\!E_{\theta_0}\left\{ m \log \frac{|\Sigma_0|}{|\tilde{\Sigma}|} \right\} \approx 0 \quad \text{and} \quad I\!E_{\theta_0}\left\{ \log |X^t V_0^{-1} X| - \log |X^t \tilde{V}^{-1} X| \right\} \approx 0.$$

On the other hand,

$$\begin{aligned}
&I\!E_{\theta_0}\left\{ \frac{\tilde{\sigma}^2}{\sigma_0^2} tr(\tilde{V} A_0) \right\} \\
&= I\!E_{\theta_0}\left\{ \frac{\epsilon_0^t (\tilde{V}^{-1} - \tilde{V}^{-1} X (X^t \tilde{V}^{-1} X)^{-1} X^t \tilde{V}^{-1}) \epsilon_0}{(N-p)\sigma_0^2} \right. \\
&\qquad\qquad \left. \times \quad tr\{ (V_0^{-1} - V_0^{-1} X (X^t V_0^{-1} X)^{-1} X^t V_0^{-1}) \tilde{V} \} \right\} \\
&\approx I\!E_{\theta_0}\left\{ \frac{\epsilon_0^t (V_0^{-1} - V_0^{-1} X (X^t V_0^{-1} X)^{-1} X^t V_0^{-1}) \epsilon_0}{(N-p)\sigma_0^2} \right. \\
&\qquad\qquad \left. \times \quad tr\{ (V_0^{-1} - V_0^{-1} X (X^t V_0^{-1} X)^{-1} X^t V_0^{-1}) V_0 \} \right\} \\
&= I\!E_{\theta_0}\left\{ \frac{\epsilon_0^t (V_0^{-1} - V_0^{-1} X (X^t V_0^{-1} X)^{-1} X^t V_0^{-1}) \epsilon_0}{\sigma_0^2} \right\}. \\
&= N - p. \qquad [35]
\end{aligned}$$

Furthermore, we have

$$\begin{aligned} I\!E_{\theta_0}\{tr(\tilde{A}\tilde{V})\} &\approx I\!E_{\theta_0}\{tr(V_0^{-1} - V_0^{-1}X(X^tV_0^{-1}X)^{-1}X^tV_0^{-1})V_0)\} \\ &= N - p. \end{aligned} \quad [36]$$

Now it remains to compute $I\!E_{\theta_0}\{(N-p)\log\frac{\sigma_0^2}{\tilde{\sigma}^2}\}$. Since, the distribution of $(N-p)(\frac{\tilde{\sigma}^2}{\sigma_0^2})$ is a chi-square with $(N-p)$ degrees of freedom, then a simple calculus shows that:

$$I\!E_{\theta_0}\{(N-p)\log\frac{\tilde{\sigma}^2}{\sigma_0^2}\} = (N-p)\{\log\frac{2}{N-p} + \Psi(\frac{N-p}{2})\} \quad [37]$$

ignoring the constant term in (32) ($p_0\log\sigma_0^2$) and according to (33) and substituting (37), (35) and (36) in the right side of (32), we obtain the estimator of $\Omega_r(\theta_0)$ given in the proposition.

5. Simulation and conclusions

The purpose of this section is to study and compare the performance of both criteria KICc and RIC$_{sd}$, introduced in this paper, to those studied in Azari et al. [2], namely RIC and AICc, and the well known criteria AIC, BIC (Akaike [1], Schwarz [20]). The performances are evaluated as a function of the number of subjects, the repeated measurements, the SNR and the correlation structure of longitudinal models.
We generate 1000 realizations from model (1), with $p_0 = 3$, $\beta_0 = (1,2,3)^t$ and variables X_0 is a three column explanatory matrix. The explanatory variables of the candidate model were stored in a $p = 7$ column matrix X; where the first three columns of X are those of X_0. The explanatory variables were randomly generated from the standard normal distribution. The number of repeated measurement was fixed at $n = 10$ and the number of subjects m have been changed to take values $m = 1, m = 5, m = 10$ and $m = 30$. The SNR ratio was also controlled at 1, 5, and 10, where SNR $= \frac{\text{var}(x_{ij}^{0t}\beta_0)}{\text{var}(\epsilon_{ij})}$. We have considered the uniform correlation structure of longitudinal data with ρ tacking 0.5 and 0.9. The results are summarized in Table 1 in which we give the percentages of the correct model order selected by each criterion considered here. We recall that:

$$\begin{aligned} \text{BIC} &= N\log\hat{\sigma}^2 + m\log|\hat{\Sigma}| + p\log(N) \\ \text{AIC} &= N\log\hat{\sigma}^2 + m\log|\hat{\Sigma}| + 2(p+1) \\ \text{AICc} &= N\log\hat{\sigma}^2 + m\log|\hat{\Sigma}| + 2\frac{N(p+1)}{N-p-2} \\ \text{RIC} &= N\log\tilde{\sigma}^2 + m\log|\tilde{\Sigma}| + p\log(N) + \frac{(N-p)^2}{N-p-2}. \end{aligned}$$

From Table 1, we see that when SNR $= 1$ and $m = 1$ the KICc criterion performs better than the others criteria followed by AICc. However, RIC$_{sd}$ is superior than RIC, BIC, AIC and KIC. If the number of the subjects increases, RIC$_{sd}$ outperforms all criteria, except when SNR $= 1$ and $\rho = 0.5$ where BIC is the best. We note also, that in all setting KICc is better than AICc and when m is large ($m = 30$), BIC and RIC give approximatively the same results. This latter remark is mentioned also by Azari et al. [2]. Moreover, when SNR increases both RIC and RIC$_{sd}$ are improved and the performance of these criteria is better as ρ increases. On the other hand the performance of KICc and AICc decays as m increases.
Other examples of simulations not reported here, with the autoregressive correlation structure of longitudinal data, give the same results as in the preceding example.
In conclusion, when both the number of subjects and the repeated observations are small, it is preferable to use the KICc criterion. However, when m or n is large one should use RIC$_{sd}$ or BIC. This latter is favored for a small SNR ratio and RIC$_{sd}$ for a moderate to large SNR ratio.

6. Conclusion

In this paper, we have derived two model selection criteria, KICc and RIC$_{sd}$, for application in longitudinal data analysis. Our criteria are based on the Kullbak's symmetric divergence. A small simulation study is undertaken to compare the performance of our criteria to other well known criteria. Moreover, our simulation studies show that the KICc criterion outperforms all other criteria when the sample size and the SNR ration are small. RIC$_{sd}$ is favored when the sample size and the SNR ration are moderate to large. Furthermore, KICc and RIC$_{sd}$ are superior to AIC and RIC, respectively, which are based on the asymmetric divergence.

Acknowledgements

This article was partially written while the first author was in post-doctoral at Mathematical Institute of Burgundy. He wishes to thank Veronique Maume-Deschamps for her helpful discussions. He also wants to thank the regional council of Burgundy for his research grant. The authors would like to thank the Associate Editor and the anonymous referee for their comments which helped improving the presentation of this paper.

Table 1: Percentages of correct model order selection for the uniform correlation structure

		$m=1$ $n=10$	$m=5$ $n=10$	$m=10$ $n=10$	$m=30$ $n=10$
	AIC	20.9	66.7	69.7	70.5
	AICc	95	79.8	74.4	71.8
$\rho=0.5$	KIC	34.2	85.2	86.2	86.8
$SNR=1$	KICc	96.4	89.7	88.1	87.2
	BIC	32.4	93.4	96.1	98.4
	RIC	35.4	64.3	82.8	92.8
	RICsd	45.8	83.8	92.6	96.1
	AIC	21.1	67.3	69.7	72.7
	AICc	95.1	77.7	76.3	74.3
$\rho=0.5$	KIC	33.9	84.8	87.3	88.2
$SNR=5$	KICc	96.6	90.4	89.4	88.9
	BIC	32.1	89.2	90.0	94.9
	RIC	65.1	92.6	96.5	98
	RICsd	77.9	95.2	98.1	99
	AIC	20.0	64.8	69.9	74.2
	AICc	95.7	76.2	75.8	76.1
$\rho=0.5$	KIC	33.2	82.9	85.2	87.8
$SNR=10$	KICc	97.3	89.7	88.1	88.0
	BIC	32.0	92.4	96.2	99.0
	RIC	82	95.8	96.0	97.1
	RICsd	91	97.8	98.5	98.9
	AIC	24.9	66.4	69	72
	AICc	96.2	75.8	75.4	74.6
$\rho=0.9$	KIC	39.6	83.4	87.1	88.5
$SNR=1$	KICc	97.1	89.2	88.2	87.1
	BIC	34.2	92.6	95.6	98.7
	RIC	50.0	85.3	91.7	98.9
	RICsd	60.6	94.4	95.7	99.7
	AIC	21.2	67.7	70.7	75.5
	AICc	95.7	78	75.7	73.3
$\rho=0.9$	KIC	39.2	84.4	86.7	87.5
$SNR=5$	KICc	97.0	89.	88.7	88.1
	BIC	33.4	93.2	95.8	99.0
	RIC	89.4	95.7	98.5	99.0
	RICsd	94.6	98.1	99.1	99.9
	AIC	22.9	65.8	74.4	74.4
	AICc	97.2	77.5	79.3	75.2
$\rho=0.9$	KIC	39.7	83.9	87.3	88.2
$SNR=10$	KICc	97.8	89.6	90.8	88.7
	BIC	32.9	93.5	97.5	99.2
	RIC	95.1	98	99.1	99.8
	RICsd	97.6	98.6	99.9	100

7. References

[1] Akaike, H (1973). Information theory and an extension of the maximum likelihood principle.*International Symposium on Information Theory, 2 nd, Tsahkadsor, Armenian SSR*, 267–281.

[2] Azari, R. and Li, L. and Tsai, C.L. Longitudinal data model selection. *Computational Statistics and Data Analysis*, 2006, **50**, 3053–3066,

[3] Bedrick, E. J. and Tsai, C. L. (1994). Model selection for multivariate regression in small samples. *Biometrics*, **50**, 226-231.

[4] Cavanaugh, J.E.5 (2004). Criteria for Linear Model Selection Based on Kullback's Symmetric Divergence.*Australian & New Zealand Journal of Statistics*, **46**, 257–274.

[5] Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics and Probability Letters*, **44**, 333-344.

[6] Cavanaugh, J. E. (1997). Unifying the derivations of the Akaike and corrected Akaike information criteria. *Statistics and Probability Letters*, **33**, 201-208.

[7] Diggle,P (2002). *Analysis of Longitudinal Data*, Oxford University Press.

[8] Fitzmaurice, G.M. and Laird, N.M. and Ware, J.H. (2004). *Applied Longitudinal Analysis*, Wiley-IEEE.

[9] Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and Cp in multivariate linear regression. *Biometrika*, **84**, 707-716.

[10] Hafidi, B. (2006). A small-sample criterion based on Kullback's symmetric divergence for vector autoregressive modeling. *Statistics and Probability Letters*, **76**, 1647–1654.

[11] Hafidi, B. and Mkhadri, A. (2006). A corrected Akaike criterion based on Kullback's symmetric divergence: applications in time series, multiple and multivariate regression. *Computational Statistics and Data Analysis*, **50**, 1524–1550.

[12] Hand, D.J. and Crowder, M.J. (1995). *Practical Longitudinal Data Analysis*, Chapman & Hall/CRC.

[13] Hedeker, D. and Gibbons, R.D. (2005). *Applied Longitudinal Data Analysis*, Wiley; John Wiley distributor.

[14] Hurvich, C. M., Shumway, R. H. and Tsai, C.L. (1990). Improved estimators of Kulback-Leibler information for autoregressive model selection in small samples. *Biometrika* **77**, 709-719.

[15] Hurvich, C. M. and Tsai, C.L. (1993). A corrected Akaike information criterion for vector autoregressive model selection. *Journal of Time Series*, **14**, 271-279.

[16] Hurvich, C. M. and Tsai, C.L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297-307.

[17] Jones, R.H. (1993). *Longitudinal Data With Serial Correlation: a state-space approach*, Chapman & Hall/CRC.

[18] Kullback, S. (1968). *Information theory and statistics*. (Dover, New York).

[19] Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545.

[20] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*. **6**, 461-464.

[21] Verbyla, AP. (1990). A conditional derivation of residual maximum likelihood. *Aust. J. Stat.*, **32**, 227–230.