Web Working Papers
by

The Italian Group of Environmental Statistics

Gruppo di Ricerca per le Applicazione della Statistica
ai Problemi Ambientali

www.graspa.org

Composite likelihood model selection

Cristiano Varin e Paolo Vidoni

GRASPA Working paper n.18, Novembre 2003

Composite likelihood model selection

By Cristiano Varin

Department of Statistics, University of Padova

via C. Battisti 241/243, I-35121 Padova, Italy. E-mail: sammy@stat.unipd.it

and Paolo Vidoni

Department of Statistics, University of Udine

via Treppo 18, I-33100 Udine, Italy. E-mail: vidoni@dss.uniud.it

Summary

The Akaike information criterion has been derived under the assumptions that the model is "true", or it is a good approximation to the truth, and that parameter estimation is obtained by using likelihood-based methods. In this paper we relax these two assumptions, by allowing inference to be drawn through a very flexible class of pseudolikelihoods called composite likelihood. The merit of composite likelihood is to reduce the computational complexity so that is possible to deal with large datasets and very complex models, even when the use of standard likelihood or Bayesian methods is not feasible. In particular, we introduce a new class of model selection criteria based on composite likelihood. An application to the well-known Old Faithful geyser dataset is also given.

Some key words: AIC; hidden Markov model; Old Faithful geyser data; pairwise likelihood; pseudolikelihood; tripletwise likelihood.

1 Introduction

A popular approach to model selection in statistics is the AIC, namely the Akaike's information criterion (Akaike 1973). It is well known that the AIC has been derived under the assumptions that the model is "true", or it is a good approximation to the truth, and that parameter estimation is obtained by using likelihood-based methods. In this paper we relax these two assumptions, by allowing inference to be drawn through a very flexible class of pseudolikelihoods. In fact, in a number of applications, large correlated datasets make unfeasible the use of the likelihood function, since too computationally demanding. One possibility is to avoid full likelihood methods, or Bayesian strategies, and to adopt simpler pseudolikelihoods, like those belonging to the composite likelihood class (Lindsay 1988). A composite likelihood consists in a combination of valid likelihood objects, usually small subsets of data. It has good theoretical properties and it behaves well in many complex applications (Besag 1974, Azzalini 1983, Hjort & Omre 1994, Heagerty & Lele 1998, Nott & Rydén 1999, Parner 2001, Renard 2002, Henderson & Shimakura 2003, Varin, Høst & Skare 2003). We aim to generalize the AIC for dealing with this class of pseudolikelihoods, without assuming that the true model belongs to the working family of distributions. The paper is organized as follow. In Section 2, we restore the concept of composite likelihood, give some examples and generalize the Kullback-Leibler divergence to deal with model selection based on composite likelihood. In Section 3, we derive a first-order unbiased composite likelihood selection statistics. Finally, in Section 4, we show the potential usefulness of our methodology analysing the well-known Old Faithfull geyser dataset (Azzalini & Bowman 1990).

2 Model selection using composite likelihood

The term composite likelihood (Lindsay 1988) denotes a rich class of pseudolikelihoods based on likelihood-type objects. We start by restoring its definition.

Definition 1. Let $\{f(y;\theta), y \in \mathcal{Y}, \theta \in \Theta\}$ be a parametric statistical model, with $\mathcal{Y} \subseteq \mathbb{R}^n$,

$\Theta \subseteq \mathbb{R}^d$, $n \geq 1$, $d \geq 1$. Consider a set of events $\{\mathcal{A}_i : \mathcal{A}_i \subseteq \mathcal{F}, i \in I\}$, where $I \subseteq \mathbb{N}$ and $\mathcal{F}$ is some sigma algebra on $\mathcal{Y}$. Then, a *composite likelihood* is a function of θ defined as

$$\mathrm{CL}_f(\theta; y) = \prod_{i \in I} f(y \in \mathcal{A}_i; \theta)^{w_i},$$

where $f(y \in \mathcal{A}_i; \theta) = f\big(\{y_j \in y : y_j \in \mathcal{A}_i\}; \theta\big)$, with $y = (y_1, \ldots, y_n)$, while $\{w_i, i \in I\}$ is a set of suitable weights. The associated *composite loglikelihood* is $\log \mathrm{CL}_f(\theta; y)$.

Example We present three important examples of composite loglikelihoods.

(i) The *"full" loglikelihood*, given by $\log \mathrm{L}(\theta; y) = \log f(y; \theta)$.

(ii) The *pairwise loglikelihood*, defined as $\log \mathrm{PL}(\theta; y) = \sum_{j<k} \log f(y_j, y_k; \theta) w_{(j,k)}$, where the summation is over all the pairs (y_j, y_k), $j, k = 1, \ldots, n$, of observations. With a slight abuse of notation we denote with $w_{(j,k)}$ the weight associated to (y_j, y_k). Analogously, we may define the *tripletwise loglikelihood*, where triplets of observations are taken into account, and so on.

(iii) The *Besag's pseudologlikelihood*, defined as $\log \mathrm{BPL}(\theta; y) = \sum_{j=1}^{n} \log f(y_j | y_{(-j)}; \theta) w_j$, where the summation is over all the conditional events $\{y_j | y_{(-j)}\}$, with $y_{(-j)}$ the subset of the components of vector y without the j-th element y_j.

$\diamond$

The usefulness of the composite likelihood ideas naturally arises in an estimating function framework (Heyde 1997). Indeed, given the set of realized events $\{\mathcal{A}_i : \mathcal{A}_i \subseteq \mathcal{F}, i \in I\}$, the *maximum composite likelihood estimator* is usually defined as a solution of the *composite likelihood equation*

$$\nabla \log \mathrm{CL}_f(\theta; y) = 0, \tag{1}$$

where $\nabla \log \mathrm{CL}_f(\theta; y) = \sum_{i \in I} \nabla \log f(y \in \mathcal{A}_i; \theta) w_i$ is the *composite score function*. Hereafter, we use the notation $\nabla h(\theta)$ for the column vector of the first partial derivatives of

function $h(\theta)$, while $\nabla^2 h(\theta)$ is the symmetric matrix of second derivatives. Since the composite score function is a linear combination of unbiased estimating functions, then, under suitable regularity conditions (Lindsay 1988, Heyde 1997, Heagerty & Lele 1998, Nott & Rydén 1999), the maximum composite likelihood estimator is consistent and asymptotically normal distributed. In this paper, the composite likelihood is considered in order to define model selection procedures. In particular, we shall introduce a new selection criterion which may be viewed as a generalization of the AIC.

Let us consider a random sample $Y = (Y_1, \ldots, Y_n)$ from an unknown distribution with joint probability density function $g(y)$, with respect to a suitable dominating measure. In a realization of Y corresponds to a set of realized events such as $\{\mathcal{A}_i : \mathcal{A}_i \subseteq \mathcal{F}, i \in I\}$. Alternative parametric statistical models can be defined as plausible description for the observed data y. These models, viewed as parametric families of joint density functions, with respect to a suitable dominating measure, may or may not contain the true $g(y)$. Consider also a future random sample $Z = (Z_1, \ldots, Z_n)$, with the same distribution as Y; Y and Z are supposed to be independent. We are interested in the choice of the "best" model for forecasting Z, given a realization of Y, using composite likelihood methods.

If we adopt for Y and Z a parametric statistical model such as $\big\{f(y;\theta), y \in \mathcal{Y}, \theta \in \theta\big\}$, prediction statements, concerning the future random sample Z, may be conveniently based on the estimated density function $\widehat{f}(z) = f(z; \widehat{\theta}_{\mathrm{MCL}}(Y))$, where $\widehat{\theta}_{\mathrm{MCL}}(Y)$ is the maximum composite likelihood estimator for θ based on Y. Estimation is done within the assumed parametric statistical model. Thus, in this general framework, it is possible to specify a predictive model selection procedure based on the following generalization of the Kullback-Leibler divergence.

Definition 2. Given two density functions $g(z)$ and $h(z)$ for Z, the associated *composite Kullback-Leibler information* is defined by the non-negative quantity

$$I_c(g,h) = E_{g(z)}\{\log(\mathrm{CL}_g(Z)/\,\mathrm{CL}_h(Z))\} = \sum_{i\in I} E_{g(z)}\big\{\log g(Z \in \mathcal{A}_i) - \log h(Z \in \mathcal{A}_i)\big\} w_i, \tag{2}$$

where the expectation is with respect to $g(z)$, $\log \mathrm{CL}_g(Z) = \sum_{i\in I} w_i \log g(Z \in \mathcal{A}_i)$ and

$\log \mathrm{CL}_h(Z) = \sum_{i \in I} w_i \log h(Z \in \mathcal{A}_i)$.

Then, model selection can be approached on the basis of the expected composite Kullback-Leibler information between the true density $g(z)$ and the estimated density $\widehat{f}(z)$, under the assumed statistical model. Namely, we define the following theoretical criterion.

Definition 3. Let us consider the random samples Y and Z, as previously defined. The *expected composite likelihood information criterion* selects the model minimizing

$$E_{g(y)}\{I_c(g, \widehat{f})\} = \sum_{i \in I} E_{g(y)}\big[E_{g(z)}\big\{\log g(Z \in \mathcal{A}_i)\big\} - E_{g(z)}\big\{\log f(Z \in \mathcal{A}_i; \widehat{\theta}_{\mathrm{MCL}}(Y))\big\}\big] w_i, \quad (3)$$

where the expectations are with respect to the true distribution of Y and Z.

The composite Kullback-Leibler information $I_c(g, \widehat{f})$, considered in relation (3), is obtained from (2) with $h(z) = \widehat{f}(z)$, so that

$$\mathrm{CL}_h(Z) = \mathrm{CL}_f(\widehat{\theta}_{\mathrm{MCL}}(Y); Z) = \prod_{i \in I} f(Z \in \mathcal{A}_i; \widehat{\theta}_{\mathrm{MCL}}(Y))^{w_i}.$$

Indeed, with a slight abuse of notation, $\mathrm{CL}_g(Z)$ may be viewed as a constant function of θ. Note that the above likelihood terms are strictly defined and do not allow the presence of a multiplicative constant. In the particular case when the composite likelihood is in fact the likelihood function, the composite Kullback-Leibler divergence $I_c(g, \widehat{f})$ equals the usual Kullback-Leibler divergence given by

$$I(g, \widehat{f}) = E_{g(z)}\big\{\log g(Z)\big\} - E_{g(z)}\big\{\log f(Z; \widehat{\theta}_{\mathrm{ML}}(Y))\big\}, \quad (4)$$

where $\theta_{\mathrm{ML}}(Y)$ is the maximum likelihood estimator for θ based on Y. Thus, the generalization (2) is useful whenever the complete computation of the exact density $f(z; \theta)$, and of the associated likelihood function, is too demanding, and then not convenient or even possible. Therefore, this general approach may be fruitfully considered for modeling large collections of correlated data. Indeed, Y and Z are here defined as suitable n-dimensional random vector, with components not necessarily independent, identically distributed. If the components of vector $Y = (Y_1, \ldots, Y_n)$ are independent, identically distributed, the

Kullback-Leibler divergence (4) corresponds to that obtained for a one dimensional future random variable, with the same distribution as components of vector Y, multiplied by n. This particular case corresponds to the underlying assumptions adopted by Konishi & Kitagawa (1996), introducing an extended information criterion for model selection.

The model selection criterion (3), which points to the model minimizing the expected composite Kullback-Leibler information between $g(z)$ and $\widehat{f}(z)$, is equivalent to that one selecting the model maximizing the *expected predictive composite loglikelihood*

$$\begin{aligned} \varphi(g,f) &= E_{g(y)}\big[E_{g(z)}\big\{\log \mathrm{CL}_f(\widehat{\theta}_{\mathrm{MCL}}(Y);Z)\big\}\big] \\ &= \sum_{i\in I} E_{g(y)}\big[E_{g(z)}\big\{\log f(Z\in\mathcal{A}_i;\widehat{\theta}_{\mathrm{MCL}}(Y))\big\}\big]w_i. \end{aligned} \tag{5}$$

Note that equation (5) corresponds to the second term in the right hand side of (3). The selection statistic (5) can be considered as a theoretical criterion for (predictive) model selection, using composite likelihood. However, since it requires the knowledge of the true density $g(z)$, it is in fact unfeasible. Thus, model selection may be approached by maximizing a selection statistic $\Psi(Y;f)$, defined as a suitable estimator for $\varphi(g,f)$. In particular, we look for unbiased estimators, exactly or to the relevant order of approximation. The simplest way to estimate $\varphi(g,f)$, using the available random sample Y, is to consider

$$\Psi(Y;f) = \log \mathrm{CL}_f(\widehat{\theta}_{\mathrm{MCL}}(Y);Y) = \sum_{i\in I} \log f(Y\in\mathcal{A}_i;\widehat{\theta}_{\mathrm{MCL}}(Y))w_i,$$

which is obtained by simply substituting Z with Y, in the argument of the expectation in (5). In the following section we shall investigate the asymptotic properties of this naive model selection statistic and we shall present a first-order unbiased modification, which can be viewed as a generalization of the AIC, based on composite likelihood.

Example (continued) Write $\widehat{\theta}_{\mathrm{ML}}$, $\widehat{\theta}_{\mathrm{MPL}}$ and $\widehat{\theta}_{\mathrm{MBPL}}$ for the maximum likelihood, maximum pairwise likelihood and maximum Besag's pseudolikelihood estimators, respectively. Then, the expected predictive composite loglikelihood and its naive estimator $\Psi(Y;f)$ are as follows:

(i) for the "full" likelihood,

$$\varphi(g,f) = E_{g(y)}\big[E_{g(z)}\big\{\log f(Z;\widehat{\theta}_{\mathrm{ML}}(Y))\big\}\big], \quad \Psi(Y;f) = \log f(Y;\widehat{\theta}_{\mathrm{ML}}(Y));$$

(ii) for the pairwise likelihood,

$$\begin{aligned}\varphi(g,f) &= \sum_{j<k} E_{g(y)}\big[E_{g(z)}\big\{\log f(Z_j,Z_k;\widehat{\theta}_{\mathrm{MPL}}(Y))\big\}\big]w_{(j,k)},\\ \Psi(Y;f) &= \sum_{j<k} \log f(Y_j,Y_k;\widehat{\theta}_{\mathrm{MPL}}(Y))w_{(j,k)};\end{aligned}$$

(iii) for the Besag's pseudolikelihood,

$$\begin{aligned}\varphi(g,f) &= \sum_{j=1}^{n} E_{g(y)}\big[E_{g(z)}\big\{\log f(Z_j|Z_{(-j)};\widehat{\theta}_{\mathrm{MBPL}}(Y))\big\}\big]w_j,\\ \Psi(Y;f) &= \sum_{j=1}^{n} \log f(Y_j|Y_{(-j)};\widehat{\theta}_{\mathrm{MBPL}}(Y))w_j.\end{aligned}$$

◇

3 A first-order unbiased selection statistic

In the following lines, we study the asymptotic properties of the selection statistic $\Psi(Y;f)$ and, in particular, we prove that it is biased and it usually provides an overestimate of the target expectation $\varphi(g,f)$. Moreover, we introduce a new general selection statistic, which is defined as a simple AIC-type modification of $\Psi(Y;f)$ and turns out to be first-order bias corrected.

We shall consider suitable assumptions, which correspond to the requirements that the joint densities, defining the statistical model, are smooth and that the maximum composite likelihood estimator is consistent and asymptotically normal distributed, under a possibly misspecified model for Y.

Assumptions Recalling the notation and the definitions introduced in the previous section, we assume that the following conditions hold.

A. The parametric space Θ is a compact subset of $\mathbb{R}^d$, $d \geq 1$, and, for every fixed $y \in \mathcal{Y}$, the composite likelihood function is two times differentiable with continuity, with respect to θ.

B. The maximum composite likelihood estimator $\widehat{\theta}_{\mathrm{MCL}}(Y)$ is defined as a solution to the composite likelihood equation (1) and there exist a vector $\theta_* \in int(\Theta)$ such that, exactly or with an error term negligible as $n \to +\infty$,

$$E_{g(y)}\{\nabla \log \mathrm{CL}_f(\theta_*; Y)\} = 0.$$

C. The maximum composite likelihood estimator $\widehat{\theta}_{\mathrm{MCL}}(Y)$ is consistent, that is $\widehat{\theta}_{\mathrm{MCL}}(Y) = \theta_* + o_p(1)$, and asymptotically normal distributed, as $n \to +\infty$, with a suitable asymptotic covariance matrix.

Note that the first two assumptions correspond to the basic regularity conditions for the asymptotic properties of maximum likelihood, and in general maximum composite likelihood, estimators under a model which could be misspecified for Y (White 1994). The vector θ_* is a *pseudo-true parameter value*, defined as a value in $int(\Theta)$ such that the composite Kullback-Leibler divergence between $g(y)$ and $f(y;\theta)$ is minimal. If the true distribution belong to the working family of distributions, the model is correctly specified for Y, namely $g(y) = f(y;\theta_0)$, for some $\theta_0 \in int(\Theta)$. In this particular case, θ_0 the true parameter value. With regard to the third assumption, in order to prove the asymptotic normality of $\widehat{\theta}_{\mathrm{MCL}}(Y)$, we usually require that

$$\nabla^2 \log \mathrm{CL}(\theta_*; Y) = E_{g(y)}\{\nabla^2 \log \mathrm{CL}(\theta_*; Y)\} + o_p(n). \tag{6}$$

Now, we are ready to introduce the main result of this paper.

Theorem. *Under the assumptions A-C, the selection statistics $\Psi(Y; f)$, based on composite likelihood, is a biased estimator for $\varphi(g, f)$. More precisely, the first order bias term is*

$$E_{g(y)}\{\Psi(Y; f)\} - \varphi(g, f) = -\operatorname{tr}\{J(\theta_*)H(\theta_*)^{-1}\},$$

where

$$J(\theta) = \mathrm{var}_{g(y)}\{\nabla \log \mathrm{CL}(\theta; Y)\}, \quad H(\theta) = E_{g(\boldsymbol{y})}\{\nabla^2 \log \mathrm{CL}(\theta; Y)\}, \tag{7}$$

with $\theta \in \Theta$.

Proof. ¿From Lemma 1 and Lemma 2, given in the Appendix, we obtain the following asymptotic expansions for the expected predictive composite loglikelihood and the expectation of its naive estimator $\Psi(Y; f)$. That is,

$$\begin{aligned} \varphi(g, f) &= E_{g(y)}\{\log \mathrm{CL}(\theta_*; Y)\} + \frac{1}{2}\,\mathrm{tr}\{J(\theta_*)H(\theta_*)^{-1}\} + o(1), \\ E_{g(y)}\{\Psi(Y; f)\} &= E_{g(\boldsymbol{y})}\{\log \mathrm{CL}(\theta_*; Y)\} - \frac{1}{2}\,\mathrm{tr}\{J(\theta_*)H(\theta_*)^{-1}\} + o(1). \end{aligned}$$

Then, taking the difference of these two approximations, we state that $\Psi(Y; f)$ is a biased estimator for $\varphi(g, f)$ and the first order bias term is $-\,\mathrm{tr}\{J(\theta_*)H(\theta_*)^{-1}\}$. □

Using this result, it is immediate to define a new general criterion for model selection, which can be viewed as a generalization of the AIC, based on composite likelihood. Namely, we introduce the following model selection procedure.

Definition 4. Let us consider a random sample Y, as previously defined. The *composite likelihood information criterion* (CLIC) selects the model maximizing

$$\Psi^c(Y; f) = \Psi(Y; f) + \mathrm{tr}\{\widehat{J}(Y)\widehat{H}(Y)^{-1}\},$$

where $\widehat{J}(Y)$ and $\widehat{H}(Y)$ are suitable consistent, first order unbiased, estimators for $J(\theta_*)$ and $H(\theta_*)$, respectively, based on Y. Analogously, the CLIC selects the model minimizing $-\Psi^c(Y; f)$.

In the following, we shall consider the CLIC based on the selection statistic $-\Psi^c(Y; f)$, which is in accordance with the usual representation chosen for the AIC. It is immediate to see that, under assumptions A-C, $\Psi^c(Y; f)$ is a first order unbiased estimator for $\varphi(g, f)$. In order to apply the CLIC, we need to substitute $J(\theta_*)$ and $H(\theta_*)$ with some suitable estimators. In practice, this could be done by means of different strategies, which depend

on the particular selection problem taken into account and on the composite likelihood which is considered. In the following section, we shall present an application to Markov models and hidden Markov models, where a possible solution for this estimation problem is proposed.

Finally, a further important point regards the choice of the weights in the composite likelihood. Typically, the weights are chosen in order to cut off the pairs of not-neighboring observations, which should be less informative. The simpler weighting strategy is then to estimate the correlation range and put equal to zero all the pairs at a distance larger than such a range. A more accurate approach consists in choosing the pairs under some optimality criterion. For example, Nott & Rydén (1999) investigate the optimal weights for a pairwise likelihood applied to random set models, within an estimating function framework. However, this solution requires an extra amount of computations, which is not feasible for complex models, such as some spatial models.

Example (continued)

(i) The CLIC for the "full" likelihood is based on

$$\Psi^c(Y; f) = \log f(Y; \widehat{\theta}_{\mathrm{ML}}(Y)) + \mathrm{tr}\{\widehat{J}(Y)\widehat{H}(Y)^{-1}\}, \tag{8}$$

where $\widehat{J}(Y)$ and $\widehat{H}(Y)$ are convenient estimators for $J(\theta_*) = \mathrm{var}_{g(y)}\{\nabla \log f(Y; \theta_*)\}$ and $H(\theta_*) = E_{g(y)}\{\nabla^2 \log f(Y; \theta_*)\}$, respectively. In this case the CLIC corresponds to the Takeuchi's information criterion (Takeuchi 1976, Shibata 1989). Note that, if the components of Y are independent, identically distributed, the selection statistic (8) coincides with that obtained by Konishi & Kitagawa (1996), using the maximum likelihood estimator for θ. Moreover, if we (optimistically) assume that the model is correctly specified for Y, then $\theta_* = \theta_0$, $J(\theta_0) = -H(\theta_0)$ and (8) simplifies to the familiar AIC

$$\Psi^c(Y; f) = \log f(Y; \widehat{\theta}_{\mathrm{ML}}(Y)) - d.$$

(ii) The CLIC for the pairwise likelihood is based on

$$\Psi^c(Y;f) = \sum_{j<k} \log f(Y_j, Y_k; \widehat{\theta}_{\mathrm{MPL}}(Y)) w_{(j,k)} + \mathrm{tr}\{\widehat{J}(Y)\widehat{H}(Y)^{-1}\},$$

where $\widehat{J}(Y)$ and $\widehat{H}(Y)$ estimate, respectively,

$$J(\theta_*) = \sum_{j<k} \sum_{l<m} \mathrm{cov}_{g(y)}\{\nabla \log f(Y_j, Y_k; \theta_*), \nabla \log f(Y_l, Y_m; \theta_*)\} w_{(j,k)} w_{(l,m)},$$

and

$$H(\theta_*) = \sum_{j<k} E_{g(y)}\{\nabla^2 \log f(Y_j, Y_k; \theta_*)\} w_{(j,k)}.$$

(iii) The CLIC for the Besag's pseudolikelihood is based on

$$\Psi^c(Y;f) = \sum_{j=1}^{n} \log f(Y_j | Y_{(-j)}; \widehat{\theta}_{\mathrm{MBPL}}(Y)) w_j + \mathrm{tr}\{\widehat{J}(Y)\widehat{H}(Y)^{-1}\},$$

where $\widehat{J}(Y)$ and $\widehat{H}(Y)$ estimate, respectively,

$$J(\theta_*) = \sum_{j=1}^{n} \sum_{k=1}^{n} \mathrm{cov}_{g(y)}\{\nabla \log f(Y_j | Y_{(-j)}; \theta_*), \nabla \log f(Y_k | Y_{(-k)}; \theta_*)\} w_j w_k,$$

and

$$H(\theta_*) = \sum_{j=1}^{n} E_{g(y)}\{\nabla^2 \log f(Y_j | Y_{(-j)}; \theta_*)\} w_j.$$

◇

4 The Old Faithful geyser data

In this section we present an application of our model selection strategy to the Old Faithful geyser dataset discussed in Azzalini & Bowman (1990). The data consists in the time series of the duration of the successive eruptions at the Old Faithful geyser in the Yellowstone National Park in the period from 1 to 15 August 1985. Azzalini & Bowman (1990) and MacDonald & Zucchini (1997, §4.2) consider a binary version of this data obtained by thresholding the time series at 3 minutes. This discretization seems plausible since the

data can be described as short or long eruptions with very few values in between and there is a quite low variation within the two groups.

The main features of the data are summarized as follow. Let us label the short and the long eruptions with the states 0 and 1, respectively. The random variables N_r, $r = 0, 1$, indicate the corresponding number of observed eruptions. We have that $N_0 = 105$ and $N_1 = 194$; moreover, the one-step observed transition matrix is

$$\begin{pmatrix} N_{00} & N_{01} \\ N_{10} & N_{11} \end{pmatrix} = \begin{pmatrix} 0 & 104 \\ 105 & 89 \end{pmatrix},$$

where N_{rs}, $r, s = 0, 1$, in the number of one-step transition from state r to state s. Note that no transition from state 0 to itself is occurred. For the models discussed in the sequel, it is also relevant to consider the two-steps transitions. Since $N_{00} = 0$, only five triplets were observed. Being N_{rst}, $r, s, t = 0, 1$, the number of two-step transitions from state r to state s and then to state t, the non-null observations are: $N_{010} = 69$, $N_{110} = 35$, $N_{011} = 35$, $N_{101} = 104$ and $N_{111} = 54$.

In Azzalini & Bowman (1990), the time series is first analyzed by a first-order Markov chain model. Then, since this model does not fit very well the autocorrelation function, they move to a second-order Markov chain model, which seems more plausible. The same data are also analyzed by MacDonald & Zucchini (1997, §4.2). They consider some hidden Markow models based on the binomial distribution and compare them with the Markov chain models of Azzalini & Bowman (1990), using the AIC and the BIC, namely the Bayesian information criterion (Schwarz 1978). They conclude that both the AIC and the BIC indicate that the model for Old Faithful geyser data is the second-order Markov chain, even if the two-state binomial hidden Markov model is quite close in performance.

In the next lines, we discuss how the CLIC can be used for model selection in this dataset. We compare the three models highlighted by MacDonald & Zucchini (1997, §4.2) as the most effective, namely, the second-order Markov chain, the two-state hidden Markow model and the two-state second-order hidden Markow model. For completeness, also a simple two-state Markov chain has been included in the discussion. First at all, we have to

choose a useful composite likelihood for making inference in all the four models under competition. Since we have hidden Markow models, a composite likelihood based on marginal events can be conveniently considered, as we shall see in the following lines.

Let us start by recalling that the hidden Markow models constitute a rich and flexible class of statistical models for time series data, where the time evolution is determined by an unobservable latent process. A monograph on this topic is MacDonald & Zucchini (1997). A hidden Markow model is a double stochastic process $\{Y_i, X_i\}_{i\geq 1}$, where the observable random variables $\{Y_i\}_{i\geq 1}$ are assumed to be conditionally independent given a hidden Markov chain $\{X_i\}_{i\geq 1}$, describing the latent evolution of the system. Thereafter, we shall assume that this latent Markov chain is stationary and irreducible, with $m \in \mathbb{N}^+$ states and transition probability functions given by $f(x_i|x_{i-1};\theta)$, $i > 1$, with $\theta \in \Theta \subseteq \mathbb{R}^d$ an unknown parameter. Moreover, we denote by $f(y_i|x_i;\theta)$, $i \geq 1$, the conditional density function (or probability function) of Y_i given $X_i = x_i$, which does not depend on i. In this case, the bivariate process $\{Y_i, X_i\}_{i\geq 1}$ is stationary.

In this framework, the likelihood function for θ, based on the available observations $y = (y_1, \ldots, y_n)$, is

$$\mathrm{L}(\theta; y) = \sum_{x_1} \cdots \sum_{x_n} f(x_1) f(y_1|x_1;\theta) \prod_{i=2}^{n} f(x_i|x_{i-1};\theta) f(y_i|x_i;\theta), \tag{9}$$

where the summations are over the m states and the initial probability function $f(x_1)$ is not necessarily that related to the stationary distribution of the chain (Leroux 1992). Since the evaluation of (9) requires $O(m^n)$ computations, MacDonald & Zucchini (1997) rearrange the terms in the likelihood function in order to reduce significantly the computational burden. However, this rearrangement does not seem useful, if one desires to compute likelihood quantities such as the derivatives of the function $\log \mathrm{L}(\theta; y)$, with respect to θ, and their expectations with respect to the true unknown distribution.

An alternative to the full likelihood are the composite likelihoods based on small marginal events, which are much simpler to handle and can highly reduce the computational effort. The simpler useful composite likelihood is the pairwise likelihood based on

the pairs of subsequent observations

$$\mathrm{PL}(\theta; y) = \prod_{i=2}^{n} \sum_{x_{i-1}, x_i} f(x_{i-1}; \theta) f(x_i | x_{i-1}; \theta) f(y_{i-1} | x_{i-1}; \theta) f(y_i | x_i; \theta),$$

where the summation is over all the pairs of subsequent observations of the latent process. Note that this pairwise likelihood is obtained from the equation given in Section 2 by imposing a set of dummy weights, where the pairs corresponding to subsequent observations have weights equal to one.

However, the pairwise likelihood is not a good candidate for our model selection problem, since for second-order Markov chains the composite likelihood equation has an infinity number of solutions. Then, we have to move to the, slightly more complex, tripletwise likelihood. For a hidden Markow model, the tripletwise likelihood, based on the triplets of subsequent observations, is given by

$$\mathrm{TL}(\theta; y) = \prod_{i=3}^{n} \sum_{x_{i-2}, x_{i-1}, x_i} f(x_{i-2}, x_{i-1}, x_i; \theta) f(y_{i-2} | x_{i-2}; \theta) f(y_{i-1} | x_{i-1}; \theta) f(y_i | x_i; \theta),$$

where the summation is over all the triplets of subsequent observations of the latent process and $f(x_{i-2}, x_{i-1}, x_i; \theta)$ is the joint probability function of (X_{i-2}, X_{i-1}, X_i). When dealing with binary data, as with the dataset under discussion, if we assume stationarity, the tripletwise likelihood looks like

$$\mathrm{TL}(\theta; y) = \prod_{r,\, s,\, t \in \{0,1\}} \mathrm{pr}(Y_{i-2} = r, Y_{i-1} = s, Y_i = t)^{N_{rst}},$$

where N_{rst}, $r, s, t = 0, 1$, defined above, is the number of realized events $(Y_{i-2} = r, Y_{i-1} = s, Y_i = t)$, $i > 2$. Note that, in this case, $\mathrm{TL}(\theta; y)$ consists in eight terms. However, since in the Old Faithful geyser dataset there are no transitions from state 0 to itself, in fact only five terms enter in the function.

In the following, we shall present the four models under competition and we compute the corresponding tripletwise likelihoods. We implicitly assume that the assumptions A-C outlined in Section 3 are fulfilled. Indeed, the consistency and asymptotic normality of the tripletwise likelihood can be proved using the framework suggested in Renard (2002,

§3.2), whose assumptions are here satisfied because the Markov and hidden Markov models considered are indeed stationary.

The first model we consider is a two-states Markov chain with one-step transition probability matrix

$$\Gamma = \begin{pmatrix} 0 & 1 \\ a & 1-a \end{pmatrix},$$

with $a \in (0,1)$ an unknown parameter. Here, we assume that the probability of remaining in the state 0 is null, since in the dataset we observe $N_{00} = 0$. In order to compute the probabilities associated with the triplets (Y_{i-2}, Y_{i-1}, Y_i), $i > 2$, is convenient to consider the transition probability matrix whose entries are $\Delta_{(sr)(ts)} = \text{pr}(Y_{i-1} = s, Y_i = r | Y_{i-2} = t, Y_{i-1} = s) = \text{pr}(Y_i = r | Y_{i-1} = s, Y_{i-2} = t)$, $r, s, t = 0, 1$, $i > 2$,

$$\Delta^{MC} = \begin{pmatrix} 1-k & k & 0 & 0 \\ 0 & 0 & a & 1-a \\ 0 & 1 & 0 & 0 \\ 0 & 0 & a & 1-a \end{pmatrix},$$

where k is any real number in (0,1). The presence of this arbitrary value k is not relevant since, as noted by MacDonald & Zucchini (1997, §4.2), the pair $(0,0)$ can be disregarded without loss of information. The associated bivariate stationary distribution is

$$\pi^{MC} = \frac{1}{1+a} \begin{pmatrix} 0 & a \\ a & 1-a \end{pmatrix},$$

with entries $\pi_{rs} = \text{pr}(Y_{i-1} = r, Y_i = s)$, $r, s = 0, 1$, $i > 1$. Then, the non-null triplet

probabilities are, for $i > 2$,

$$\begin{aligned}
\mathrm{pr}(Y_{i-2} = 0, Y_{i-1} = 1, Y_i = 0) &= \Delta_{(10)(01)}\pi_{01} = \frac{a^2}{1+a}, \\
\mathrm{pr}(Y_{i-2} = 0, Y_{i-1} = 1, Y_i = 1) &= \Delta_{(11)(01)}\pi_{01} = \frac{a(1-a)}{1+a}, \\
\mathrm{pr}(Y_{i-2} = 1, Y_{i-1} = 0, Y_i = 1) &= \Delta_{(01)(10)}\pi_{11} = \frac{a}{1+a}, \\
\mathrm{pr}(Y_{i-2} = 1, Y_{i-1} = 1, Y_i = 0) &= \Delta_{(10)(11)}\pi_{11} = \frac{a(1-a)}{1+a}, \\
\mathrm{pr}(Y_{i-2} = 1, Y_{i-1} = 1, Y_i = 1) &= \Delta_{(11)(11)}\pi_{11} = \frac{(1-a)^2}{1+a}.
\end{aligned}$$

Note that $\mathrm{pr}(Y_{i-2} = 0, Y_{i-1} = 1, Y_i = 1) = \mathrm{pr}(Y_{i-2} = 1, Y_{i-1} = 1, Y_i = 0)$. This equivalence will also occur within the other models under competition. Here, $\theta = a$ and the tripletwise loglikelihood is given by

$$\begin{aligned}
\log \mathrm{TL}(a) = &-(N-2)\log(1+a) + N_{010}\log a^2 + (N_{011} + N_{110})\log((1-a)a) \\
&+ N_{101}\log a + N_{111}\log(1-a)^2.
\end{aligned}$$

The maximum tripletwise likelihood estimate is $\widehat{a}_{\mathrm{MTL}} = 0.5389$ and the corresponding function value is $\log \mathrm{TL}(\widehat{a}_{\mathrm{MTL}}) = -458.7104$. For comparison, the maximum likelihood estimate, as reported by MacDonald & Zucchini (1997, §4.2), is $\widehat{a}_{\mathrm{ML}} = 0.5412$.

The second model is a second-order two-states Markov chain. In this case the probabilities $\Delta_{(sr)(ts)}$, $r, s, t = 0, 1$, are such that

$$\Delta^{MC2} = \begin{pmatrix} 1-k & k & 0 & 0 \\ 0 & 0 & b & 1-b \\ 0 & 1 & 0 & 0 \\ 0 & 0 & c & 1-c \end{pmatrix},$$

with $b, c \in (0, 1)$ unknown parameters and k any real number in (0,1). The associated bivariate stationary distribution is

$$\pi^{MC2} = \frac{1}{2c + (1-b)} \begin{pmatrix} 0 & c \\ c & (1-b) \end{pmatrix}$$

and then the joint probabilities for the five relevant triplets are, for $i > 2$,

$$\begin{aligned}
\mathrm{pr}(Y_{i-2} = 0, Y_{i-1} = 1, Y_i = 0) &= \frac{cb}{2c + (1-b)}, \\
\mathrm{pr}(Y_{i-2} = 0, Y_{i-1} = 1, Y_i = 1) &= \frac{(1-b)c}{2c + (1-b)}, \\
\mathrm{pr}(Y_{i-2} = 1, Y_{i-1} = 0, Y_i = 1) &= \frac{c}{2c + (1-b)}, \\
\mathrm{pr}(Y_{i-2} = 1, Y_{i-1} = 1, Y_i = 1) &= \frac{(1-b)(1-c)}{2c + (1-b)}
\end{aligned}$$

and $\mathrm{pr}(Y_{i-2} = 1, Y_{i-1} = 1, Y_i = 0) = \mathrm{pr}(Y_{i-2} = 0, Y_{i-1} = 1, Y_i = 1)$. Since $\theta = (b, c)$, the tripletwise loglikelihood is

$$\begin{aligned}
\log \mathrm{TL}(b, c) = &-(N-2)\log(2c + 1 - b) + N_{010}\log(cb) + (N_{011} + N_{110})\log(c(1-b)) \\
&+ N_{101}\log c + N_{111}\log((1-b)(1-c)).
\end{aligned}$$

Here, the maximum tripletwise likelihood estimates are found to be $\widehat{b}_{MTL} = 0.6634$, $\widehat{c}_{MTL} = 0.3932$, which equals the maximum likelihood estimates. The corresponding value of the tripletwise loglikelihood is $\log \mathrm{TL}(\widehat{b}_{MTL}, \widehat{c}_{MTL}) = -451.5889$. Here, we note that the maximum tripletwise likelihood estimates allow the equality between the estimated and the observed probabilities for the five triplets, so we have

$$\widehat{\mathrm{pr}}(Y_{i-2} = 0, Y_{i-1} = 1, Y_i = 0) = \frac{\widehat{c}\widehat{b}}{2\widehat{c} + (1-\widehat{b})} = \frac{N_{010}}{n-2}$$

and similarly for the remaining triplets. Then, we can say that this model reaches the "best" fitting possible using the tripletwise likelihood.

The third model is a two-states hidden Markov model. The hidden process $\{X_i\}_{i \geq 1}$ is a Markov chain with the same one-step transition probabilities as in the Markov chain previously considered, that is

$$\Gamma = \begin{pmatrix} 0 & 1 \\ a & 1-a \end{pmatrix},$$

with $a \in (0, 1)$ unknown. The conditional probabilities for the observations given the latent

variables are

$$\begin{aligned}
&\mathrm{pr}(Y_i = y | X_i = 0) = \rho^y (1-\rho)^{1-y}, \quad y = 0, 1, \\
&\mathrm{pr}(Y_1 = 1 | X_i = 1) = 1,
\end{aligned}$$

for $i \geq 1$, with $\rho \in (0, 1)$ an unknown parameter. The relevant triplet probabilities are, for $i > 2$,

$$\begin{aligned}
\mathrm{pr}(Y_{i-2} = 0, Y_{i-1} = 1, Y_i = 0) &= \frac{(1-\rho)^2 a^2}{1+a}, \\
\mathrm{pr}(Y_{i-2} = 0, Y_{i-1} = 1, Y_i = 1) &= \frac{\rho(1-\rho)a^2 + (1-\rho)(1-a)a}{1+a}, \\
\mathrm{pr}(Y_{i-2} = 1, Y_{i-1} = 0, Y_i = 1) &= \frac{(1-\rho)a}{1+a}, \\
\mathrm{pr}(Y_{i-2} = 1, Y_{i-1} = 1, Y_i = 1) &= \frac{\rho^2 a^2 + 2\rho(1-a)a + \rho a + (1-a)^2}{1+a}
\end{aligned}$$

and $\mathrm{pr}(Y_{i-2} = 1, Y_{i-1} = 1, Y_i = 0) = \mathrm{pr}(Y_{i-2} = 0, Y_{i-1} = 1, Y_i = 1)$. Here $\theta = (a, \rho)$ and the maximum tripletwise likelihood estimates are $\widehat{a}_{MTL} = 0.8948, \widehat{\rho}_{MTL} = 0.2585$. The maximum likelihood estimates are again very similar, namely $\widehat{a}_{ML} = 0.827, \widehat{\rho}_{ML} = 0.225$. Moreover, we have $\log \mathrm{TL}(\widehat{a}_{MTL}, \widehat{\rho}_{MTL}) = -451.5889$, the same value obtained for the second-order Markov chain, again corresponding to the perfect equality between theoretical and observed triplets probabilities. Note that the tripletwise loglikelihood for the hidden Markow model is not a reparametrization of that for the second-order Markov chain.

The last model we take into consideration is a two-states second-order hidden Markov model. Here, the hidden process $\{X_i\}_{i \geq 1}$ is the same as the second-order Markov chain previously considered, while the conditional probabilities for the observations given the latent variables are as in the previous hidden Markow model. Then, the relevant triplet

probabilities are, for $i > 2$,

$$\begin{aligned}
\mathrm{pr}(Y_{i-2} = 0, Y_{i-1} = 1, Y_i = 0) &= \frac{(1-\rho)^2 cb}{2c + (1-b)}, \\
\mathrm{pr}(Y_{i-2} = 0, Y_{i-1} = 1, Y_i = 1) &= \frac{\rho(1-\rho)cb + (1-\rho)(1-b)c}{2c + (1-b)}, \\
\mathrm{pr}(Y_{i-2} = 1, Y_{i-1} = 0, Y_i = 1) &= \frac{(1-\rho)c}{2c + (1-b)}, \\
\mathrm{pr}(Y_{i-2} = 1, Y_{i-1} = 1, Y_i = 1) &= \frac{\rho^2 cb + 2\rho(1-b)c + \rho c + (1-b)(1-c)}{2c + (1-b)}
\end{aligned}$$

and $\mathrm{pr}(Y_{i-2} = 1, Y_{i-1} = 1, Y_i = 0) = \mathrm{pr}(Y_{i-2} = 0, Y_{i-1} = 1, Y_i = 1)$. Here $\theta = (b, c, \rho)$ and the maximum tripletwise likelihood estimates are $\widehat{b}_{MTL} = 0.8494$, $\widehat{c}_{MTL} = 0.6535$, $\widehat{\theta}_{MTL} = 0.2189$ and the maximized tripletwise loglikelihood is $\log \mathrm{TL}(\widehat{b}_{MTL}, \widehat{c}_{MTL}, \widehat{\rho}_{MTL}) = -451.5889$, that is the same value obtained in previous two models. In this case the tripletwise likelihood estimates differ considerably from the maximum likelihood estimates given by MacDonald & Zucchini (1997, §4.2), which correspond to $\widehat{b}_{\mathrm{ML}} = 0.717$, $\widehat{c}_{\mathrm{ML}} = 0.414$, $\widehat{\rho}_{\mathrm{ML}} = 0.072$). Such difference can be investigated by comparing the values of tripetwise and full likelihood at the two different estimates. We have $\log \mathrm{TL}(\widehat{b}_{\mathrm{ML}}, \widehat{c}_{\mathrm{ML}}, \widehat{\rho}_{\mathrm{ML}}) = -451.5941$ which is almost the same value of that obtained plugging the tripletwise estimates. On the other side, we found $\log \mathrm{L}(\widehat{b}_{MTL}, \widehat{c}_{MTL}, \widehat{\rho}_{\mathrm{MTL}}) = -128.0877$, but $\log \mathrm{L}(\widehat{b}_{\mathrm{ML}}, \widehat{c}_{ML}, \widehat{\rho}_{\mathrm{ML}}) = -126.8575$. However, the fact that the log tripletwise likelihood is almost the same if computed at the MTL or the maximum likelihood estimates, does not make difference for model selection conclusion under the CLIC.

In order to compute the CLIC, for the four alternative models, we have to estimate the mean, the variance and the covariance of the random variables N_{010}, N_{011}, N_{110}, N_{101} and N_{111}, with respect to the true unknown distribution. To obtain suitable estimates, we may consider a reuse sampling approach (Heagerty & Lumley 2000). The idea is to subdivide the time series in a set of overlapping subseries and use them as several samples from the true model. This is useful for our problem, if we assume that the unknown model for $\{Y_i\}_{i\geq 1}$ satisfies stationary. Let us assume that the data are split into $n-M+1$ overlapping subwindows of the same dimension M. Then, an estimator for p_{rst}, namely the probability, under the true unknown distribution, of the observed triplet $(Y_{i-2} = r, Y_{i-1} = s, Y_i = t)$,

$r, s, t = 0, 1$, $i > 2$, is given by

$$\widehat{p}_{rst} = \frac{1}{(n-M+1)M} \sum_{j=1}^{n-M+1} \sum_{i=j+2}^{j+M+1} I(y_{i-2} = r, y_{i-1} = s, y_i = t),$$

where $I(B)$ is the indicator function of the event B. Heagerty & Lumley (2000) show that the optimal dimension of the subwindows is $Cn^{1/3}$, where n is the number of observations, while C is a constant which depends on the strength of the dependence within the data. Simulation experiments suggest that a value of C between 4 and 8 should be good enough for most situations.

The estimated true probabilities, using 250 subwindows based on $M = 50$ observations, with $C \approx 8$, are $\widehat{p}_{010} = 0.226$, $\widehat{p}_{011} = 0.114$, $\widehat{p}_{110} = 0.340$, $\widehat{p}_{101} = 0.113$ and $\widehat{p}_{111} = 0.207$. We also tried with other values for C, but the results are almost unchanged. Finally, as reported in Table 1, the CLIC suggests that the hidden Markow model is slightly better than the second-order Markov chain, which is the opposite of the result obtained by MacDonald & Zucchini (1997, §4.2). In the table we report the values of the CLIC, the AIC and the BIC, for the four alternative models.

Insert Table 1.

5 Conclusions

We have presented a class of model selection criteria based on composite likelihood. Our methodology allows to make model selection even in computationally expensive models, without the condition that the assumed model is the true one or it is a good approximation to the truth. As an example, we analyzed the well-known Old Faithful geyser dataset by comparing binary Markov and hidden Markov models using the particular composite likelihood called tripletwise likelihood. This example has been chosen because we fell it instructive for describing the main features of our methodology. However, the usefulness of composite likelihood methods is evident if we consider more complex models. First at all, models with an hidden structure such as generalized linear mixed models (Breslow &

Clayton 1993) and state space models (Durbin & Koopman 2001). In both these classes of models, the computation of standard likelihood objects often requires the solution of untractable integrals whose dimension depends on the hidden part of the model. In many applications, the dimension of the integrals to be solved is so large that is very difficult, or even unfeasible, to make inference and also model selection. Then, the class of composite likelihood gives interesting alternatives since only a set of small dimensional integrals are considered in the computations. Moreover, the use of this methodology can be also considered for analyzing large datasets, such as those arising in many application of spatial statistics (Hjort & Omre 1994, Heagerty & Lele 1998, Nott & Rydén 1999). Conditional models like point processes models and autoregressive models used in image analysis (Besag 1986) are other strong potential area of application.

Acknowledgement

The authors would like to thank Professor A. Azzalini for helpful comments.

A Appendix

Derivation of Lemma 1 and Lemma 2

In this appendix, we present the two lemmas involved in the proof of the theorem given in Section 3.

Lemma 1. *Under the assumptions A-C, we have that*

$$\varphi(g, f) = E_{g(y)}\{\log \mathrm{CL}(\theta_*; Y)\} + \frac{1}{2}\,\mathrm{tr}\{J(\theta_*)H(\theta_*)^{-1}\} + o(1),$$

with $J(\theta_*)$ and $H(\theta_*)$ given by (7).

Proof. Let us consider the following stochastic Taylor expansion for $\log \mathrm{CL}(\widehat{\theta}_{\mathrm{MCL}}(Y); Z)$, around $\widehat{\theta}_{\mathrm{MCL}}(Y) = \theta_*$,

$$\begin{aligned}\log \mathrm{CL}(\widehat{\theta}_{\mathrm{MCL}}(Y); Z) &= \log \mathrm{CL}(\theta_*; Z) + (\widehat{\theta}_{\mathrm{MCL}}(Y) - \theta_*)^{\mathrm{T}} \nabla \log \mathrm{CL}(\theta_*; Z) \\ &\quad + \frac{1}{2}(\widehat{\theta}_{\mathrm{MCL}}(Y) - \theta_*)^{\mathrm{T}} \nabla^2 \log \mathrm{CL}(\theta_*; Z)](\widehat{\theta}_{\mathrm{MCL}}(Y) - \theta_*) + o_p(1).\end{aligned}$$

Since the random vectors Y and Z are independent, identically distributed, we state that $E_{g(z)}\{\log \mathrm{CL}(\theta_*; Z)\} = E_{g(y)}\{\log \mathrm{CL}(\theta_*; Y)\}$ and, as a consequence of assumption B, $E_{g(z)}\{\nabla \log \mathrm{CL}(\theta_*; Z)\} = 0$, exactly or to the relevant order of approximation. Thus, taking expectations term by term, with respect to the true distribution of Z, we have

$$\begin{aligned}E_{g(z)}[\log \mathrm{CL}(\widehat{\theta}_{\mathrm{MCL}}(Y); Z)] &= E_{g(y)}\{\log \mathrm{CL}(\theta_*; Y)\} \\ &\quad + \frac{1}{2}(\widehat{\theta}_{\mathrm{MCL}}(Y) - \theta_*)^{\mathrm{T}} E_{g(z)}[\nabla^2 \log \mathrm{CL}(\theta_*; Z)](\widehat{\theta}_{\mathrm{MCL}}(Y) - \theta_*) + o_p(1).\end{aligned}$$

Moreover, the mean value of the above expansion, with respect to the true distribution of Y, gives

$$\varphi(g, f) = E_{g(y)}\{\log \mathrm{CL}(\theta_*; Y)\} + \frac{1}{2} \mathrm{tr}\{H(\theta_*)V(\theta_*)\} + o(1), \tag{10}$$

where $V(\theta_*) = E_{g(y)}\{(\widehat{\theta}_{\mathrm{MCL}}(Y) - \theta_*)(\widehat{\theta}_{\mathrm{MCL}}(Y) - \theta_*)^{\mathrm{T}}\}$. The final step is to get an approximation for $V(\theta_*)$. Note that, by means of standard asymptotic arguments, $(\widehat{\theta}_{\mathrm{MCL}}(Y) - \theta_*)$ may approximated by

$$-\nabla^2 \log \mathrm{CL}(\widehat{\theta}_{\mathrm{MCL}}(Y); Y)^{-1} \nabla \log \mathrm{CL}(\theta_*; Y).$$

Then, since $\nabla^2 \log \mathrm{CL}(\widehat{\theta}_{\mathrm{MCL}}(Y); Y)$ can be further approximated by $\nabla^2 \log \mathrm{CL}(\theta_*; Y)$ we obtain that, using relation (6),

$$V(\theta_*) = H(\theta_*)^{-1} J(\theta_*) H(\theta_*)^{-1} + o(n). \tag{11}$$

Plugging (11) in (10) completes the proof. □

Lemma 2. *Under the assumptions A-C, we have that*

$$E_{g(y)}\{\Psi(Y; f)\} = E_{g(y)}\{\log \mathrm{CL}(\theta_*; Y)\} - \frac{1}{2} \mathrm{tr}\{J(\theta_*)H(\theta_*)^{-1}\} + o(1),$$

with $J(\theta_*)$ and $H(\theta_*)$ given by (7).

Proof. Let us consider the following stochastic Taylor expansion for the selection statistic $\Psi(Y; f)$, around $\widehat{\theta}_{\mathrm{MCL}}(Y) = \theta_*$,

$$\begin{aligned}\Psi(Y; f) = &\log \mathrm{CL}(\theta_*; Y) + (\widehat{\theta}_{\mathrm{MCL}}(Y) - \theta_*)^{\mathrm{T}} \nabla \log \mathrm{CL}(\theta_*; Y) \\ &+ \frac{1}{2}(\widehat{\theta}_{\mathrm{MCL}}(Y) - \theta_*)^{\mathrm{T}} \nabla^2 \log \mathrm{CL}(\theta_*; Y)(\widehat{\theta}_{\mathrm{MCL}}(Y) - \theta_*) + o_p(1).\end{aligned}$$

Since, by means of standard asymptotic arguments, $\nabla \log \mathrm{CL}(\theta_*; Y)$ may approximated by

$$-(\widehat{\theta}_{\mathrm{MCL}}(Y) - \theta_*)^{\mathrm{T}} \nabla^2 \log \mathrm{CL}(\theta_*; Y),$$

we obtain

$$\Psi(Y; f) = \log \mathrm{CL}(\theta_*; Y) - \frac{1}{2}(\widehat{\theta}_{\mathrm{MCL}}(Y) - \theta_*)^{\mathrm{T}} \nabla^2 \log \mathrm{CL}(\theta_*; Y)(\widehat{\theta}_{\mathrm{MCL}}(Y) - \theta_*) + o_p(1).$$

Then, taking expectations, with respect to the true distribution of Y, and using relations (6) and (11), complete the proof. □

References

Akaike, H. (1973), Information theory and extension of the maximum likelihood principle, *in* N. Petron & F. Caski, eds, 'Second Symposium on Information Theory', Budapest: Akademiai Kiado, pp. 267–281.

Azzalini, A. (1983), 'Maximum likelihood estimation of order m for stationary stochastic processes', *Biometrika* **70**, 381–387.

Azzalini, A. & Bowman, A. W. (1990), 'A look at some data on the old faithful geyser', *Applied Statistics* **39**, 357–365.

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

Besag, J. E. (1986), 'On the statistical analysis of dirty pictures', *Journal of the Royal Statistical Society, Series B* **36**, 259–279.

Breslow, N. E. & Clayton, D. G. (1993), 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association* **88**(421), 9 – 25.

Durbin, J. & Koopman, S. J. (2001), *Time series analysis by state space methods*, Oxford University Press.

Heagerty, P. J. & Lele, S. R. (1998), 'A composite likelihood approach to binary spatial data', *Journal of the American Statistical Association* **93**, 1099–1111.

Heagerty, P. J. & Lumley, T. (2000), 'Window subsampling of estimating functions with application to regression models', *Journal of the American Statistical Association* **95**, 197–211.

Henderson, R. & Shimakura, S. (2003), 'A serially correlated gamma frailty model for longitudinal count data', *Biometrika* **90**(2), 355–366.

Heyde, C. C. (1997), *Quasi-likelihood and its Application*, Springer Verlag, New York.

Hjort, N. L. & Omre, H. (1994), 'Topics in spatial statistics', *The Scandinavian Journal of Statistics* **21**, 289–357.

Konishi, S. & Kitagawa, G. (1996), 'Generalized information criteria in model selection', *Biometrika* **4**(83), 875–890.

Leroux, B. G. (1992), 'Maximum-likelihood estimation for hidden Markov models', *Stochastic Processes and their Applications* **40**, 127–143.

Lindsay, B. (1988), Composite likelihood methods, *in* N. U. Prabhu, ed., 'Statistical Inference from Stochastic Processes', Providence RI: American Mathematical Society.

MacDonald, I. L. & Zucchini, W. (1997), *Hidden Markov and Other Models for Discrete-valued Time Series*, Chapman & Hall.

Nott, D. J. & Rydén, T. (1999), 'Pairwise likelihood methods for inference in image models', *Biometrika* **86**(3), 661–676.

Parner, E. T. (2001), 'A composite likelihood approach to multivariate survival data', *The Scandinavian Journal of Statistics* **28**, 295–302.

Renard, D. (2002), Topics in Modelling Multilevel and Longitudinal Data, PhD thesis, Limburgs Universitair Centrum, Holland.

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

Shibata, R. (1989), Statistical aspects of model selection, *in* J. Willems, ed., 'From Data to Model', New York: Springer-Verlag, pp. 215–240.

Takeuchi, K. (1976), 'Distribution of information statistics and criteria for adequacy of models (in japanese)', *Mathematical Science* **153**, 12–18.

Varin, C., Høst, G. & Skare, Ø. (2003), Pairwise likelihood inference in spatial generalized linear mixed models, Technical report, Department of Statistics, University of Padova, Italy.

White, H. (1994), *Estimation, Inference and Specification Analysis*, Cambridge University Press, New York.

Model	AIC	BIC	CLIC
MC	272.48	279.88	460.01
MC2	**262.24**	**277.04**	452.65
HMM	262.62	277.42	**452.32**
HMM2	265.80	288.00	458.54

Table 1: Old Faithful geyser dataset. Values for the AIC, the BIC and the CLIC for the Markov chain (MC), the second-order Markov chain (MC2), the hidden Markow model (HMM) and the second-order hidden Markow model (HMM2). The values for the AIC and the BIC are those obtained by MacDonald & Zucchini (1997, §4.2).