Working Paper Series, N. 5, May 2010

Composite Likelihood Function in State Space Models

Nadia Frigo
Department of Statistical Sciences
University of Padua
Italy

Abstract: In general state space models, where the computational effort required in the evaluation of the full likelihood function is infeasible, we analyze the problem of static parameter estimation based on composite likelihood functions, in particular pairwise likelihood functions. We discuss consistency and efficiency properties of the estimators obtained by maximizing these functions in state space scenario, linking these properties to the characteristics of the model. We empirically compare the efficiency between maximum pairwise likelihood and maximum full likelihood estimators. We suggest the existence of a 'best' distance between pairs of observations, in terms of variance of the maximum pairwise likelihood estimator.

Keywords: Pairwise likelihood, Split data likelihood, Efficiency.

Final version (2010-05-18)

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Nadia Frigo
tel: +39 049 827 4151
nadia@stat.unipd.it
http://homes.stat.unipd.it/frigo

Composite Likelihood Function in State Space Models

Nadia Frigo
Department of Statistical Sciences
University of Padua
Italy

Abstract: In general state space models, where the computational effort required in the evaluation of the full likelihood function is infeasible, we analyze the problem of static parameter estimation based on composite likelihood functions, in particular pairwise likelihood functions. We discuss consistency and efficiency properties of the estimators obtained by maximizing these functions in state space scenario, linking these properties to the characteristics of the model. We empirically compare the efficiency between maximum pairwise likelihood and maximum full likelihood estimators. We suggest the existence of a 'best' distance between pairs of observations, in terms of variance of the maximum pairwise likelihood estimator.

Keywords: Pairwise likelihood, Split data likelihood, Efficiency.

1 Introduction

State space models are a general class of time series capable of modeling dependent observations in a natural and interpretable way. They consist of a Markov process (called hidden/latent state process) not observed directly, but only through another process. When the parameter describing the model is known, sequential inference on the latent process is typically based on the sequence of joint posterior distributions, where each summarizes all the information collected about the latent process up to the current time. Sequential estimation of these distributions is achieved by *optimal filtering* recursions. Such recursions rarely admit a closed form expression, but it is possible to resort to efficient numerical approximations. Sequential Monte Carlo (SMC) methods (aka particle filters) are a class of numerical algorithms available to approximate the sequence of joint posterior distributions sequentially in time [Doucet et al., 2001]. This methodology is now well developed and the theory supporting this approach is also well established [Del Moral, 2004].

In most real-world scenarios, the parameter is unknown and needs to be estimated. Although apparently simpler than optimal filtering, the static parameter estimation problem has proved to be much more difficult: no closed form solutions are, in general, available, even for linear gaussian and finite state space hidden Markov models. A possible way to address this problem is based on SMC methods. There have been many attempts to develop elaborate sequential algorithms, but all of them suffer from a common intrinsic problem, namely *path degeneracy*. This phenomenon

is the result of the resampling stage and has long been observed [Gordon et al., 1993]. It reflects a fundamental weakness of SMC methods: with limited resources, it is not possible to consistently estimate the sequence of posterior distributions at every instant time [Del Moral, 2004]. Direct application of SMC techniques is hence inappropriate for static parameter inference [Chopin, 2004, Kitagawa, 1998, Liu and West, 2001, Andrieu et al., 1999, Fernhead, 2002, Gilks and Berzuini, 2001, Storvik, 2002]. A different approach consists on developing an inferential procedure based on full likelihood function to compute point estimates from the data. Recently, some results on the consistency and asymptotic normality of the maximum likelihood estimator in state space models have been proved [Douc et al., 2004]. Anyway, when the latent process is continuous, the computational effort required in the evaluation of the full likelihood function is infeasible. Approximated solutions, based on Monte Carlo or numerical methods, have been considered, but none of the proposed solutions are completely satisfactory. A possible way to overcome this problem is to replace the likelihood function by another function, easier to determine. In this direction, composite likelihood approaches have been suggested. The term composite likelihood indicates a likelihood type object formed by taking the product of individual component likelihoods, each of which corresponds to a marginal or conditional event. This is useful when the joint density is difficult to evaluate but computing likelihoods for some subsets of the data is possible, as in general state space models framework. This idea dates back probably to Besag [1974] even though the term composite likelihood was stated by Lindsay [1988].

In this paper we analyze the problem of static parameter estimation based on composite likelihood functions, in particular pairwise likelihood functions. We study the asymptotic properties of the pairwise likelihood function and of the parameter estimators obtained by maximizing this function in state space scenario, in connection with stationary and ergodic properties of the processes involved. The paper is organized as follows. In Section 2 we present the model and we define two particular cases of composite likelihood function, i.e.pairwise likelihood (PL) and split data likelihood (SDL) functions. In Section 3 we discuss which kind of pairwise likelihood function is better to use among some possible choices for the weights. In Section 4 we study the asymptotic properties of the maximum PL estimator, related to the characteristic of the state space model. In particular, we prove the consistency of the maximum PL estimator of order L. Section 5 gives some comments about the loss of efficiency of maximum PL estimator wrt the maximum likelihood estimator and in Section 6 we empirically compare the efficiency between maximum PL and maximum full likelihood estimators as well as the efficiency between maximum SDL (when blocks of observations are allowed to overlap) and maximum PL estimators. Section 7 gives some concluding remarks.

2 The Framework

State space models can be defined in the following form. For any parameter $\theta \in \Theta$, the hidden/latent state process $\{X_k; k \geq 1\} \subset \mathcal{X}^{\mathbb{N}}$ is a Markov process, characterized

by its Markov transition probability distribution $f_\theta(x'|x)$, i.e. $X_1 \sim \nu$ and for $n \geq 1$,

$$X_{n+1}|(X_n = x) \sim f_\theta(\cdot|x). \tag{1}$$

The process $\{X_k; k \geq 1\}$ is observed, not directly, but through another process $\{Y_k; k \geq 1\} \subset \mathcal{Y}^{\mathbb{N}}$. The observations are assumed to be conditionally independent given $\{X_k; k \geq 1\}$, and their common marginal probability distribution is of the form $g_\theta(y|x)$, i.e. for $1 \leq n \leq m$,

$$Y_n|(X_1, \ldots, X_n = x, \ldots, X_m) \sim g_\theta(\cdot|x). \tag{2}$$

From now on, we will assume that the process $\{Z_k; k \geq 1\} = \{(X_k, Y_k); k \geq 1\}$ is stationary (in the strict sense) with joint distribution given by

$$p_\theta(x_{1:n}, y_{1:n}) = \pi_\theta(x_1) g_\theta(y_1|x_1) \prod_{i=2}^{n} f_\theta(x_i|x_{i-1}) g_\theta(y_i|x_i),$$

where we denote by π_θ the marginal for $\{X_k; k \geq 1\}$ of the invariant distribution. We assume that there is a 'true' parameter value θ^* generating the data $\{Y_k; k \geq 1\}$ and that this value is unknown. We focus here on point estimation methods developing an inferential procedure based on likelihood quantities to compute point estimates of θ^* from $\{Y_k; k \geq 1\}$ rather than a series of estimates of the posterior distributions $\{p(\theta, Y_{1:n}); n \geq 1\}$. As a result no particle method is required in the parameter space, and it should also be pointed out that SMC methods in the state space $\mathcal{X}$ are, in general, also not necessary.

The most natural approach of point estimate consists of maximizing the series of likelihoods $\{p_\theta(Y_{1:n}); n \geq 1\}$. With our notation, the likelihood for a sequence of observations $y_1, \ldots, y_n$ is

$$L(\theta; y_{1:n}) = p_\theta(y_{1:n}) = \int_{\mathcal{X}^n} \pi_\theta(x_1) g_\theta(y_1|x_1) \prod_{i=2}^{n} f_\theta(x_i|x_{i-1}) g_\theta(y_i|x_i) dx_{1:n},$$

which is simply obtained by taking into account the dependence structure characterizing the model.

Recently, some results on the consistency and asymptotic normality of the maximum likelihood estimator (MLE) can be found in Douc et al. [2004] (see also the references therein). Anyway, when $\{X_k; k \geq 1\}$ is continuous, evaluation of the full likelihood requires an integration over an n-dimensional space. This task is insurmountable for typical values of n and exact methods for computing and maximizing the likelihood function are usually not feasible. Approximated solutions, based on Monte Carlo or numerical methods, have been considered, but none of the proposed solutions are completely satisfactory. Markov Chain Monte Carlo (MCMC) methods are usually difficult to implement while Particle Filters (PF) are well suited but suffer from the well known degeneracy problem.

Even if the full likelihood approach is the most natural and leads to an efficient estimation of the parameter, the computational effort required in the evaluation and maximization of the function suggests to develop new procedures in order to reduce

the computational burden. In this way it is possible to fit highly structured statistical models, even when the use of standard likelihood methods is not practically possible. A possible way to overcome this problem is to replace the likelihood by another function, easier to determine. Any function which (asymptotically) has its maximum at the true parameter point is a potential candidate. In this direction composite likelihood approaches have been suggested. Given the observations $y_{1:n}$, a composite likelihood is defined by specifying a set of K marginal or conditional events $A_k(y_{1:n}), k = 1, \dots, K$, with likelihood given by $L_k(\theta; y_{1:n}) = L(\theta; A_k(y_{1:n}))$. Then, the composite likelihood is obtained by composing these likelihood objects and it corresponds to

$$L_C(\theta; y_{1:n}) = \prod_{k=1}^{K} L_k(\theta; y_{1:n})^{\omega_k},$$

with ω_k suitable non-negative weights. This class contains, and thus generalizes, the usual ordinary likelihood, as well as many other interesting alternatives. Examples include the Besag pseudolikelihood [Besag, 1974, 1977], the m-th order likelihood for stationary processes [Azzalini, 1983] and composite likelihoods constructed from marginal densities [Cox and Reid, 2004]. Typical attention is paid to compositions of low-dimensional marginals, since their computation involves usually lower dimensional integrals. This is the case of the *pairwise likelihood* [Le Cessie and Van Houwelingen, 1994],

$$L_{P,\omega}(\theta; y_{1:n}) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} p_\theta(y_i, y_j)^{\omega_{ij}}, \tag{3}$$

where $\omega_{ij}, i = 1, \dots, n-1, j = i+1, \dots, n$ are suitable non-negative weights, or of the *split data likelihood* (SDL) proposed by Ryden [1994] as an alternative to maximum likelihood for inference in hidden Markov models. This is a composite likelihood constructed by splitting the $n = mL$ observations into m groups of fixed size L and assuming these groups are independent

$$L_{SD}(\theta; y_{1:n}) = \prod_{i=1}^{m} p_\theta(y_{L(i-1)+1:iL}).$$

In the SDL framework, it is also possible to consider overlapping blocks of the form $(Y_{1:L}, Y_{2:L+1}, \dots, Y_{n-L+1:n})$. In this case we define

$$L_{SD}^{(ov)}(\theta; y_{1:n}) = \prod_{i=1}^{n-L+1} p_\theta(y_{i:L+i-1}). \tag{4}$$

3 Different choices for the weights

We consider the pairwise likelihood function and the asymptotic properties of the parameter estimator obtained by maximizing this function in state space scenario.

Starting from (3), a suitable choice for the weights allows one to consider different types of PL. We consider only $0-1$ weights and we shall concentrate on the PL

that takes into account all the $n(n-1)/2$ pairs (obtained choosing $\omega_{ij} = 1, \quad \forall i = 1, \dots, n-1, j = i+1, \dots, n$), that is

$$L_P(\theta; y_{1:n}) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} p_\theta(y_i, y_j) \tag{5}$$

and on the so called L-th order PL, which is based on all the pairs of observations with a lag distance not greater than $L \in \{1, \dots, n-1\}$, that is

$$L_P^{(L)}(\theta; y_{1:n}) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{min\{i+L,n\}} p_\theta(y_i, y_j).$$

Note that $L_P^{(n-1)}(\theta; y_{1:n})$ corresponds to (5). Given the dependence structure of the model (1, 2), for every $i = 1, \dots, n-1, j = i+1 \dots, n$

$$p_\theta(y_i, y_j) = \int_{\mathcal{X}^{j-i+1}} \pi_\theta(x_i) g_\theta(y_i|x_i) \left[\prod_{k=i+1}^{j} f_\theta(x_k|x_{k-1}) \right] g_\theta(y_j|x_j) dx_{i:j}. \tag{6}$$

The numerical computation of (6) involves in general a $(j-i+1)$- dimensional integral. If $j-i$ is bounded by a constant that does not depend on n, the computation is likely easier compared to the full likelihood approach. In the case of pairwise likelihood with all the pairs, the integral dimension increases with n, so its evaluation might be still infeasible, depending on the structure of $f_\theta(\cdot|x)$. This is one of the motivations why people usually do not work with pairwise likelihood with all the pairs but prefer using pairwise likelihood of order L, for some $L \geq 1$.
Moreover, even if the computation of (5) were feasible, the process has good properties and the invariant distribution is known, we expect that the normalized log pairwise likelihood

$$l_P(\theta; y_{1:n}) = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{1}{n-i} \sum_{j=i+1}^{n} log[p_\theta(y_i, y_j)] \tag{7}$$

will be well approximated by

$$\frac{1}{n-1} \sum_{i=1}^{n-1} \frac{1}{n-i} \sum_{j=m_n+i+1}^{n} log[p_\theta(y_i) p_\theta(y_j)] \tag{8}$$

for n large enough, where m_n is chosen in such a way that, for every i, $m_n/(n-i)$ and $\frac{m_n log n}{n}$ go to zero as n goes to infinity.

Roughly speaking, (8) tells us that if n grows there are more pairs that are far apart than pairs that are close and, if the process is ergodic, the pairs that are far away act as they were independent. In this case it is clear that all the information about the dependence structure of the model are lost, since only the marginal density is taken into account. More precisely, we prove the following theorem

Theorem 1. *Under the Assumptions* $(C1)$ *and* $(C2)$ *defined in Appendix A*

$$l_P(\theta; y_{1:n}) \approx \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{1}{n-i} \sum_{j=m_n+i+1}^{n} log[p_\theta(y_i) p_\theta(y_j)]$$

for n *large enough, where, for every* i, $m_n/(n-i)$ *and* $\frac{m_n log n}{n}$ *go to zero as* n *goes to infinity.*

Proof. By definition (7),

$$l_P(\theta; y_{1:n}) = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{1}{n-i} \left[\sum_{j=i+1}^{m_n+i} log[p_\theta(y_i, y_j)] + \sum_{j=m_n+i+1}^{n} log[p_\theta(y_i, y_j)] \right],$$

where for every i, $m_n/(n-i)$ goes to zero as n goes to infinity to ensure that m_n does not grow 'too much' compared to n and hence the second sum makes sense. We concentrate first in the term

$$L_1(n, m_n) := \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{1}{n-i} \sum_{j=i+1}^{m_n+i} log[p_\theta(y_i, y_j)].$$

We have that

$$\begin{aligned} |L_1(n, m_n)| &\leq \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{1}{n-i} \sum_{j=i+1}^{m_n+i} |log[p_\theta(y_i, y_j)]| \\ &\leq \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{C m_n}{n-i}, \end{aligned}$$

with $C \in (0, +\infty)$. The result above follows from Assumption (C1), which ensure that $p_\theta(y_i, y_j)$ is bounded away from zero for every i, j, and from the following identity, valid for any $x, y \in (0, +\infty)$,

$$|\log x - \log y| \leq \frac{|x-y|}{x \wedge y}. \tag{9}$$

Now

$$\begin{aligned} \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{C m_n}{n-i} &= \frac{C m_n}{n-1} \sum_{i=1}^{n-1} \frac{1}{i} \\ &\approx \frac{C m_n}{n-1} (log[n-1] + \gamma), \end{aligned}$$

where γ is the Euler constant. If $\frac{m_n log n}{n}$ goes to zero as n goes to infinity, then the term $L_1(n, m_n)$ goes to zero. Hence, the contribution to the log pairwise likelihood of the pairs with lag distance not greater than m_n vanishes as n goes to infinity.

Note that this condition holds, for example, when m_n is a constant.
We look now at

$$L_2(n, m_n) := \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{n-i}\sum_{j=m_n+i+1}^{n} log[p_\theta(y_i, y_j)].$$

We can rewrite $L_2(n, m_n)$ as

$$\begin{aligned} L_2(n, m_n) =& \frac{1}{n-1}\sum_{i=1}^{n-1}\Big[\frac{1}{n-i}\sum_{j=m_n+i+1}^{n}(log[p_\theta(y_i, y_j)] - log[p_\theta(y_i)p_\theta(y_j)]) + \\ &+\frac{1}{n-i}\sum_{j=m_n+i+1}^{n} log[p_\theta(y_i)p_\theta(y_j)]\Big]. \end{aligned}$$

By ergodic properties, there exist constants $\tilde{C} \in (0, +\infty)$ and $\rho \in [0, 1)$ such that, for every i, j

$$|p_\theta(y_i, y_j) - p_\theta(y_i)p_\theta(y_j)| \leq \tilde{C}\rho^{j-i}.$$

Using again identity (9), the absolute value of first term in $L_2(n, m_n)$ satisfies

$$\begin{aligned} &\left|\frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{n-i}\sum_{j=m_n+i+1}^{n}(log[p_\theta(y_i, y_j)] - log[p_\theta(y_i)p_\theta(y_j)])\right| \\ &\leq \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{n-i}\sum_{j=m_n+i+1}^{n} C\rho^{j-i} \leq \frac{C}{(1-\rho)(n-1)}\sum_{i=1}^{n-1}\frac{\rho^{m_n}-\rho^{n-i}}{n-i} \\ &= \frac{C\rho^{m_n}}{(1-\rho)(n-1)}\sum_{i=1}^{n-1}\frac{1}{i} - \frac{C}{(1-\rho)(n-1)}\sum_{i=1}^{n-1}\frac{\rho^i}{i}, \end{aligned}$$

for a suitable constant $C \in (0, +\infty)$. For n large enough

$$\frac{C\rho^{m_n}}{(1-\rho)(n-1)}\sum_{i=1}^{n-1}\frac{1}{i} \approx \frac{C\rho^{m_n}}{(1-\rho)(n-1)}(log[n-1]+\gamma) \overset{n\to+\infty}{\to} 0,$$

since ρ is a constant less than one. On the other hand

$$\begin{aligned} &\frac{C}{(1-\rho)(n-1)}\sum_{i=1}^{n-1}\frac{\rho^i}{i} \leq \frac{C}{(1-\rho)(n-1)}\sum_{i=1}^{n-1}\rho^i \\ &= \frac{C}{(1-\rho)(n-1)}\left(\frac{1-\rho^n}{1-\rho}-1\right) \overset{n\to+\infty}{\to} 0. \end{aligned}$$

We have that

$$L_2(n, m_n) \approx \frac{1}{n-i}\sum_{j=m_n+i+1}^{n} log[p_\theta(y_i)p_\theta(y_j)]$$

for n large enough and combining this with the result about $L_1(n, m_n)$, we are able to conclude that

$$l_P(\theta; y_{1:n}) \approx \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{n-i}\sum_{j=m_n+i+1}^{n} log[p_\theta(y_i)p_\theta(y_j)].$$

□

4 Strong consistency of the maximum PL estimator of order L

In this section we consider the PL function of order L, defined as

$$L_P^{(L)}(\theta; y_{1:n}) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{min\{i+L,n\}} p_\theta(y_i, y_j), \tag{10}$$

where $p_\theta(y_i, y_j)$ is defined by (6) and $L \geq 1$ is a fixed constant (under the hypothesis that π_θ is known). We study the properties of the maximum PL estimator, i. e. the estimator obtained by maximizing (10) with respect to the parameter θ. We denote by $\hat{\theta}_P^{(L)}$ any global maximum point of $L_P^{(L)}(\theta; y_{1:n})$. In order to study the properties of $\hat{\theta}_P^{(L)}$ we need to point out the asymptotic behavior of the normalized log likelihood

$$l_P^{(L)}(\theta; y_{1:n}) = \frac{1}{n-1} \sum_{i=1}^{n-1} \left[\frac{1}{L} \sum_{j=i+1}^{min\{i+L,n\}} log[p_\theta(y_i, y_j)] \right] \tag{11}$$

as n goes to infinity. Since $L^{-1} \sum_{j=i+1}^{min\{i+L,n\}} log[p_\theta(y_i, y_j)]$ is a function of the observations $(y_i, \ldots, y_{i+L})$ (let us denote this function as φ), under suitable ergodic assumptions

$$\begin{aligned} & \frac{1}{n-1} \sum_{i=1}^{n-1} \varphi(y_i, \ldots, y_{i+L}) \overset{n\to+\infty}{\to} \mathbb{E}_{\theta^*}[\varphi(Y_1, \ldots, Y_{L+1})] = \\ = & \int_{\mathcal{Y}^{L+1}} \varphi(y_1, \ldots, y_{L+1}) p_{\theta^*}(y_{1:L+1}) dy_{1:L+1} \\ = & \frac{1}{L} \sum_{j=2}^{L+1} \int_{\mathcal{Y}^2} log[p_\theta(y_1, y_j)] p_{\theta^*}(y_1, y_j) dy_1 dy_j, \end{aligned} \tag{12}$$

where $\mathbb{E}_{\theta^*}[\cdot]$ is the expectation associated to the stationary process $\{Z_k; k \geq 1\}$ generated by the model defined in (1) and (2) for $\theta = \theta^* \in \Theta$.
Hence

$$\lim_{n\to+\infty} l_P^{(L)}(\theta; y_{1:n}) = l_P^{(L)}(\theta),$$

where $l_P^{(L)}(\theta)$ is defined by (12). With appropriate conditions, it can be shown that the set of parameters maximizing $l_P^{(L)}(\theta)$ includes the true parameter and hence the L-th order PL is an objective function that, when maximized, leads to a reasonable estimator of the parameter. This follows from the fact that maximizing $l_P^{(L)}(\theta)$ is equivalent to minimizing the following Kullback-Leibler divergence

$$K_P^{(L)}(\theta, \theta^*) = l_P^{(L)}(\theta^*) - l_P^{(L)}(\theta) \geq 0.$$

Varin and Vidoni [2005] called $K_P^{(L)}(\theta, \theta^*)$ *composite Kullback-Leibler divergence* since it can be seen as the linear combination of the Kullback-Leibler divergences

associated with each component of the composite likelihood. In this case

$$K_P^{(L)}(\theta, \theta^*) = \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}_{\theta^*} \left[log \frac{p_{\theta^*}(y_1, y_j)}{p_\theta(y_1, y_j)} \right], \tag{13}$$

which preserves the non-negativity as soon as the ordinary Kullback-Leibler divergence does (see Appendix B).

Following the standard technique introduced by Wald [1949] and asking that the bivariate process $\{X_k, Y_k\}$ is uniformly ergodic and that the functions f_θ and g_θ are continuous in θ, the estimator obtained by maximizing the pairwise likelihood of order L is *strongly consistent*, i. e. it converges almost surely to the true parameter value as n goes to infinity.
More precisely, we prove the following theorem (middle results can be found in Appendix B)

Theorem 2 (Strong consistency)**.** *Assume that conditions* $(C1-C7)$ *in Appendix A hold and let* $\hat{\theta}_P^{(L)}$ *be the L-order pairwise likelihood estimator based on n observations. Then* $\hat{\theta}_P^{(L)} \to \theta^*$ P_{θ^*}*-almost surely as* $n \to \infty$*.*

Proof. Given an arbitrary $\epsilon > 0$, set $S_\epsilon = \{\theta \in \Theta; |\theta - \theta^*| < \epsilon\}$ and $C = \Theta \cap S_\epsilon^c$. Lemma 3 allows us to choose a positive number $\bar{b}$ such that, for every $j = 2, \ldots, L+1$

$$\mathbb{E}_{\theta^*} \left[\sup_{\theta:|\theta|>\bar{b}} \log p_\theta(y_1, y_j) \right] \leq \mathbb{E}_{\theta^*} \left[\log p_{\theta^*}(y_1, y_j) \right] - 1 \tag{14}$$

and let $C_1 = C \cap \{\theta \in \Theta; |\theta| \leq \bar{b}\}$. It follows from Lemma 1 and Lemma 2 that for each $\theta \in C_1$ there is a $\epsilon_\theta > 0$ and an open neighborhood G_θ of θ such that

$$\mathbb{E}_{\theta^*} \left[\sup_{\theta' \in G_\theta} \log p_{\theta'}(y_1, y_j) \right] \leq \mathbb{E}_{\theta^*} \left[\log p_\theta(y_1, y_j) \right] \leq \mathbb{E}_{\theta^*} \left[\log p_{\theta^*}(y_1, y_j) \right] - \epsilon_\theta. \tag{15}$$

Note that C_1 is a compact set (from Assumption C2) and thus there is a finite set $\{\theta_1, \ldots, \theta_d\} \subseteq \Theta$ such that $C_1 \subseteq \cup_{i=1}^d G_i$, where $G_i = G_{\theta_i}$ and define $G_0 = \{\theta \in \Theta; |\theta| > \bar{b}\}$. We have that

$$\sup_{\theta \in S_\epsilon^c} \left(\log L_P^{(L)}(\theta; y_{1:n}) - \log L_P^{(L)}(\theta^*; y_{1:n}) \right) =$$

$$= \max_{0 \leq i \leq d} \left(\sup_{\theta \in G_i} \log L_P^{(L)}(\theta; y_{1:n}) - \log L_P^{(L)}(\theta^*; y_{1:n}) \right).$$

From Assumption (C1), for every $i, 1 \leq i \leq d$

$$\sup_{\theta \in G_i} \left(l_P^{(L)}(\theta; y_{1:n}) - l_P^{(L)}(\theta^*; y_{1:n}) \right) \overset{n \to \infty}{\to}$$

$$\frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}_{\theta^*} \left[\sup_{\theta \in G_i} \log p_\theta(y_1, y_j) \right] - \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}_{\theta^*} \left[\log p_{\theta^*}(y_1, y_j) \right]$$

and by Equation (15) the right term above is less or equal to $-\epsilon_{\theta_i} < 0$.
Again by Assumption (C1)

$$\sup_{\theta\in G_0}\left(l_P^{(L)}(\theta;y_{1:n})-l_P^{(L)}(\theta^*;y_{1:n})\right)\overset{n\to\infty}{\to}$$
$$\frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}_{\theta^*}\left[\sup_{\theta\in G_0}\log p_\theta(y_1,y_j)\right]-\frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}_{\theta^*}\left[\log p_{\theta^*}(y_1,y_j)\right]$$

and by Equation (14) the right term above is less or equal to $-1 < 0$.
This proves that

$$\max_{0\le i\le d}\left(\sup_{\theta\in G_i}\log[L(n-1)l_P^{(L)}(\theta;y_{1:n})]-\log[L(n-1)l_P^{(L)}(\theta^*;y_{1:n})]\right)$$
$$\overset{n\to\infty}{\to}-\infty\quad\mathbb{P}_{\theta^*}-a.s.,$$

that is

$$\mathbb{P}_{\theta^*}\left\{\lim_{n\to\infty}\sup_{\theta\in S_\epsilon^c}\left(\log L_P^{(L)}(\theta;y_{1:n})-\log L_P^{(L)}(\theta^*;y_{1:n})\right)=-\infty\right\}=1. \tag{16}$$

Now, we use the result in (16) to prove the strong consistency of $\hat\theta_P^{(L)}$, i.e. that $\mathbb{P}_{\theta^*}\left\{\lim_{n\to\infty}\hat\theta_P^{(L)}=\theta^*\right\}=1$. Since $\hat\theta_P^{(L)}$ is a global maximum point of $L_P^{(L)}(\theta;y_{1:n})$, we have that

$$L_P^{(L)}(\hat\theta_P^{(L)};y_{1:n})\ge L_P^{(L)}(\theta^*;y_{1:n})$$

for all n. It is sufficient to prove that for any $\epsilon>0$ the probability that there exists a limit point $\hat\theta$ of the sequence $\{\hat\theta_P^{(L)}\}$ such that $|\hat\theta-\theta^*|>\epsilon$ is zero. If such a $\hat\theta$ exists than $\sup_{\theta\in S_\epsilon^c}L_P^{(L)}(\theta;y_{1:n})\ge L_P^{(L)}(\hat\theta_m^{(L)};y_{1:n})$ for infinitely many n. But then

$$\frac{\sup_{\theta\in S_\epsilon^c}L_P^{(L)}(\theta;y_{1:n})}{L_P^{(L)}(\theta^*;y_{1:n})}>0$$

for infinitely many n. Since, according to (16), this is an event with probability zero, we have shown that the probability that all limit points $\hat\theta$ of $\{\hat\theta_P^{(L)}\}$ satisfy the inequality $|\hat\theta-\theta^*|\le\epsilon$ is one. By the arbitrariness of ϵ, $\hat\theta_P^{(L)}$ is strongly consistent. □

Remark 1. *Starting from expression (12), for j large enough, by ergodicity, $p_\theta(y_1,y_j)$ is well approximated by $p_\theta(y_1)p_\theta(y_j)$, for every $\theta\in\Theta$. Hence, for j large enough*

$$\int_{\mathcal{Y}^2}log[p_\theta(y_1,y_j)]p_{\theta^*}(y_1,y_j)dy_1dy_j$$

is well approximated by

$$\int_{\mathcal{Y}^2}log[p_\theta(y_1)p_\theta(y_j)]p_{\theta^*}(y_1)p_{\theta^*}(y_j)dy_1dy_j=$$
$$=\int_{\mathcal{Y}}log[p_\theta(y_1)]p_{\theta^*}(y_1)dy_1+\int_{\mathcal{Y}}log[p_\theta(y_j)]p_{\theta^*}(y_j)dy_j.$$

By stationarity assumption, $p_\theta(y_1) = p_\theta(y_j)$ for every j and for every $\theta \in \Theta$ and using Cesaro sum we have that

$$\lim_{L\to+\infty} \frac{1}{L} \sum_{j=2}^{L+1} \int_{\mathcal{Y}^2} log[p_\theta(y_1, y_j)] p_{\theta^*}(y_1, y_j) dy_1 dy_j = 2 \int_{\mathcal{Y}} log[p_\theta(y_1)] p_{\theta^*}(y_1) dy_1.$$

Hence, if L is allowed to grow to infinity, all the information about the dependence structure of the model are lost, since only the marginal density is taken into account.

In the next section, we discuss the loss of asymptotic efficiency introduced by the use of $l_P^{(L)}(\theta)$ in place of the full likelihood, through the analysis of the asymptotic variance of the estimator $\hat{\theta}_P^{(L)}$.

5 Loss of efficiency

If L is fixed, the use of L-th order PL suggests that information about the parameter can be extracted from the dependence structure of the pairs of observations with a lag distance not greater than L. Usually, it happens that the maximum pairwise likelihood estimators tend to lose efficiency, with respect to those based on the full likelihood. Even if this behavior is obviously reliable, until now, no general results about the evaluation of this gap are available.

Instead of comparing the efficiency between the PL and the full likelihood, we would like to compare the efficiency of SDL and PL. This choice is justified by the fact that for general state space models the full likelihood function is unavailable, as we discussed before, and hence the estimator obtained by maximizing this function is not actually a real alternative. Moreover, for non overlapping version of split data likelihood estimator, some theoretical results about the behavior of its variance have already been proved. Anyway, maximum full likelihood estimator is the benchmark we have to refer to when we discuss efficiency of any estimator.

We would like to take into account the overlapping version of the SDL, as defined in (4) and consider the case where π_θ is known. We expect that, for a general model, there will be a loss of efficiency when we use PL of order L instead of overlapping SDL. The quantification of this loss can be achieved through the evaluation of the asymptotic variance of the estimator $\hat{\theta}_P^{(L)}$. Andrieu et al. [2007] characterize the asymptotic variance in the non overlapping version of SDL, called Σ_L, and quantify the loss of efficiency by comparing Σ_L to its counterpart associated to the full likelihood based criterion. More precisely, they state that there exists a $C \in (0, +\infty)$ and $\rho \in [0, 1)$ such that for any $L \geq 2$

$$|\Sigma - \Sigma_L| \leq C \left[\frac{\log(L)^2}{L \log(\rho)^2} + \frac{\rho}{L(1-\rho)} + \frac{\rho^{L+1}}{1-\rho^L} \right], \tag{17}$$

where Σ denotes the asymptotic variance of the full likelihood estimator. Since

$$\Sigma_L = H_L^{-1}(\theta^*) G_L(\theta^*) H_L^{-T}(\theta^*),$$

where

$$H_L(\theta^*) = \frac{1}{L}\mathbb{E}[\nabla logp_{\theta^*}(\mathbf{Y}_0)\nabla^T logp_{\theta^*}(\mathbf{Y}_0)],$$

$$G_L(\theta^*) = \frac{1}{L}\mathbb{E}[\nabla logp_{\theta^*}(\mathbf{Y}_0)\nabla^T logp_{\theta^*}(\mathbf{Y}_0)] + \frac{2}{L}\sum_{k=1}^{+\infty}\mathbb{E}[\nabla logp_{\theta^*}(\mathbf{Y}_0)\nabla^T logp_{\theta^*}(\mathbf{Y}_k)],$$

with $\mathbf{Y}_k = (Y_{kL+1}, \ldots, Y_{(k+1)L})$, the result in (17) comes from the fact that

$$\begin{aligned}
\frac{1}{L}|\mathbb{E}[\nabla logp_{\theta^*}(\mathbf{Y}_0)\nabla^T logp_{\theta^*}(\mathbf{Y}_0)] - \Sigma| &\leq \frac{\rho}{L(1-\rho)}, \\
\frac{1}{L}|\mathbb{E}[\nabla logp_{\theta^*}(\mathbf{Y}_0)\nabla^T logp_{\theta^*}(\mathbf{Y}_1)]| &\leq C\frac{\log(L)^2}{L\log(\rho)^2}, \\
\frac{1}{L}|\mathbb{E}[\nabla logp_{\theta^*}(\mathbf{Y}_0)\nabla^T logp_{\theta^*}(\mathbf{Y}_k)]| &\leq CL\rho^{(k-1)L+1} \qquad \forall k \geq 2,
\end{aligned}$$

for a suitable $C \in (0, +\infty)$ and $\rho \in [0, 1)$. Equation (17) proves that the loss of efficiency compared to the maximum likelihood estimator vanishes as L increases and depends on the mixing properties of the model. Extending their results to the overlapping version of the maximum split data likelihood estimator is far from being easy. The difficulties arise because the dependency structure between blocks is more complex when blocks are allowed to overlap instead of being disjoint. This translates to a more complicated calculation for the counterpart of $G_L(\theta^*)$, necessary to evaluate the asymptotic variance of the estimator.

For our purpose, we shall evaluate the asymptotic variances of the estimators obtained by maximizing (4) and (10). We refer to these quantities as $\Sigma_{SD}^{(ov)}$ and $\Sigma_P^{(L)}$ respectively. As we discussed above, evaluation of $\Sigma_{SD}^{(ov)}$ and a fortiori evaluation of $\Sigma_P^{(L)}$ is not easy to obtain, even for simple models. A deep theoretical analysis of the efficiency problem in pairwise and overlapping split data likelihood inferential procedures is beyond the scope of this paper. Anyway, while we suspect that $\Sigma_{SD}^{(ov)}$ still decreases if L grows, we do not expect that $\Sigma_P^{(L)}$ will do the same if L grows, unlike Σ_L does. This idea is consistent to Varin and Vidoni [2009]. In the next section, we give an empirical evidence of these behaviors and we suggest the existence of a 'best' lag L, in term of variance of the PL estimator. Anyway, how to determine L optimally is still a good open question.

6 Simulation study about efficiency

In this section, we empirically compare the efficiency between maximum pairwise likelihood and maximum full likelihood estimators, as well as the efficiency between maximum overlapping split data likelihood and maximum pairwise likelihood estimators. Even if we do not have theoretical results that state the behavior of their variances, our intuitions, suggested in Section 5, are confirmed in this simple example. We consider a linear gaussian state space model, where invariant distribution is known and the likelihood function is available in a closed form. Even if it is only an empirical study in a simple context, these preliminary results may be a useful guide

when we move to more complex settings where we can not compute the likelihood function in a closed form.

We illustrate here by means of simulation experiments, the performance of the maximum pairwise likelihood estimator of order L and we compare it with the maximum split data likelihood estimator, where the blocks defining the likelihood function are allowed to overlap.
We consider a state space model where the latent process follows an autoregressive dynamic and the marginal distributions of the observations are explicitly known

$$\begin{aligned} X_{n+1} &= \phi X_n + W_n, \quad W_n \sim N(0, \tau^2) \\ Y_n &= X_n + V_n, \quad V_n \sim N(0, \sigma^2). \end{aligned}$$

In this case

$$f_\theta(x'|x) = N(\phi x, \tau^2) \quad \text{and} \quad g_\theta(y|x) = N(x, \sigma^2).$$

We assume that the process is stationary, so $|\phi| < 1$ and $\pi_\theta \sim N\left(0, \frac{\tau^2}{1-\phi^2}\right)$. The unknown parameter is $\theta = (\phi, \tau, \sigma)$. The full likelihood function is available in a closed form and it can be efficiently computed by the Kalman filter recursions. Thus we can compare the performance of the maximum likelihood, the maximum pairwise likelihood and the maximum split data likelihood estimators. Moreover, we can empirically compare the variance of the maximum pairwise and the split data likelihood estimators in order to study their relationship in term of efficiency. Since we set the parameter space in such a way that the process is stationary, the bivariate distribution of the pairs $(Y_i, Y_j), i = 1, \ldots, n-1; j = i+1, \ldots, n$ is

$$\begin{pmatrix} Y_i \\ Y_j \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 + \frac{\tau^2}{1-\phi^2} & \phi^{j-i}\frac{\tau^2}{1-\phi^2} \\ \phi^{j-i}\frac{\tau^2}{1-\phi^2} & \sigma^2 + \frac{\tau^2}{1-\phi^2} \end{pmatrix} \right\},$$

and hence the pairwise likelihood of order L is easy to compute.
It is worthwhile to underline that the statistical model corresponding to the choice $L = 1$ is not identifiable. If $L = 1$ there exist at least two different sets of parameters values for θ which give the same value for the pairwise likelihood function. This problem can be easily overcome by adding pairs at lag distance greater than one.
On the other hand, under stationarity conditions, the marginal distribution of the blocks $(Y_i, \ldots, Y_{L+i-1}), i = 1, \ldots, n-L+1$ is

$$\begin{pmatrix} Y_i \\ \vdots \\ Y_{L+i-1} \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 + \frac{\tau^2}{1-\phi^2} & \phi\frac{\tau^2}{1-\phi^2} & \ldots & \phi^{L-1}\frac{\tau^2}{1-\phi^2} \\ \phi\frac{\tau^2}{1-\phi^2} & \ddots & \ddots & \phi^{L-2}\frac{\tau^2}{1-\phi^2} \\ \vdots & \ddots & \ddots & \vdots \\ \phi^{L-1}\frac{\tau^2}{1-\phi^2} & \ldots & \ldots & \sigma^2 + \frac{\tau^2}{1-\phi^2} \end{pmatrix} \right\},$$

and hence the split data likelihood with blocks of length L turns out to be easy to compute.

We compare the empirical properties of $\hat{\theta}_P^{(L)}$ and $\hat{\theta}_{SD}^{(L)}$, with $L = 2, \ldots, 29$. We consider 300 time series of length $n = 1000$ from the AR(1) model plus additive

Table 1: AR(1) model plus observation noise, with $\theta^* = (0.7, 1, 1)$. Sample mean and standard deviation (in brackets), for the maximum likelihood estimator $\hat{\theta}_{ML}$. Calculations based on 300 simulated time series of length 1000.

$\hat{\phi}_{ML}$	$\hat{\tau}_{ML}$	$\hat{\sigma}_{ML}$
0.6952	0.996	0.9932
(0.0473)	(0.0952)	(0.0798)

observation noise, with $\phi^* = 0.7$, $\sigma^* = 1, \tau^* = 1$ as true parameter values. Hereafter, in order to find the maximum point of the pairwise and split data likelihood functions, we adopt an optimization procedure based on the Nelder and Mead downhill simplex method, with a relative convergence tolerance of 10^{-8}. We repeat the optimization procedure starting from different values in the parameter space, finding similar results. The sample means and standard deviations for the maximum pairwise and split data likelihood estimators, for some L, are summarized in Table 2 as well as for the maximum full likelihood estimator (Table 1). The results presented here are obtained taking as starting values for the optimization procedure $\phi^0 = 0.9, \sigma^0 = 0.8, \tau^0 = 0.5$.

Table 2: AR(1) model plus observation noise, with $\theta^* = (0.7, 1, 1)$. Sample means and standard deviations (in brackets) for the maximum pairwise likelihood estimator $\hat{\theta}_P^{(L)}$ and split data likelihood estimator $\hat{\theta}_{SD}^{(L)}$ as L increases. Calculations based on 300 simulated time series of length 1000.

	Pairwise Likelihood			Split data Likelihood		
Lag	$\hat{\phi}_P^{(L)}$	$\hat{\tau}_P^{(L)}$	$\hat{\sigma}_P^{(L)}$	$\hat{\phi}_{SD}^{(L)}$	$\hat{\tau}_{SD}^{(L)}$	$\hat{\sigma}_{SD}^{(L)}$
2	0.6968	0.9928	0.992	0.6968	0.9929	0.9919
	(0.0561)	**(0.119)**	**(0.095)**	(0.0562)	(0.1191)	(0.0951)
3	0.6963	0.9936	0.9944	0.6956	0.9953	0.9923
	(0.0494)	**(0.1006)**	**(0.0827)**	(0.0507)	(0.1048)	(0.0866)
4	0.696	0.9939	0.9948	0.6952	0.9961	0.9924
	(0.0481)	**(0.0963)**	**(0.0803)**	(0.049)	(0.1003)	(0.0837)
5	0.6964	0.9932	0.9951	0.6951	0.9964	0.9924
	(0.0487)	(0.0983)	(0.0832)	(0.0484)	(0.0985)	(0.0825)
6	0.6954	0.9954	0.9925	0.6951	0.9965	0.9925
	(0.0505)	(0.1041)	(0.0882)	(0.048)	(0.0975)	(0.0817)
7	0.6945	0.9976	0.99	0.6951	0.9965	0.9926
	(0.052)	(0.1085)	(0.0922)	(0.0479)	(0.0971)	(0.0814)
8	0.694	0.9987	0.9885	0.6951	0.9965	0.9927
	(0.0534)	(0.113)	(0.0955)	(0.0478)	(0.0969)	(0.0811)

Table 2: Continued on next page

Table 2: continued from previous page

	Pairwise Likelihood			Split data Likelihood		
Lag	$\hat{\phi}_P^{(L)}$	$\hat{\tau}_P^{(L)}$	$\hat{\sigma}_P^{(L)}$	$\hat{\phi}_{SD}^{(L)}$	$\hat{\tau}_{SD}^{(L)}$	$\hat{\sigma}_{SD}^{(L)}$
9	0.6937	0.9995	0.9871	0.6951	0.9965	0.9928
	(0.055)	(0.1173)	(0.0995)	(0.0477)	(0.0968)	(0.0809)
10	0.6937	0.9995	0.9869	0.6951	0.9964	0.9928
	(0.0558)	(0.12)	(0.1008)	(0.0477)	(0.0967)	(0.0807)
11	0.6935	0.9998	0.9864	0.6952	0.9964	0.9929
	(0.0566)	(0.1225)	(0.1026)	(0.0477)	(0.0966)	(0.0806)
12	0.6937	0.9993	0.9865	0.6951	0.9964	0.9929
	(0.0573)	(0.1246)	(0.1042)	(0.0476)	(0.0965)	(0.0805)
13	0.6943	0.9977	0.9875	0.6952	0.9963	0.993
	(0.0584)	(0.1279)	(0.1065)	(0.0476)	(0.0964)	(0.0804)
14	0.6946	0.9971	0.9879	0.6952	0.9962	0.9931
	(0.0588)	(0.1294)	(0.1075)	(0.0476)	(0.0964)	(0.0802)
15	0.6949	0.9962	0.9886	0.6952	0.9962	0.9931
	(0.0593)	(0.1307)	(0.1083)	(0.0475)	(0.0963)	(0.0802)
16	0.695	0.9957	0.9889	0.6952	0.9962	0.9931
	(0.0598)	(0.1322)	(0.1092)	(0.0475)	(0.0963)	(0.0801)
17	0.6953	0.995	0.9894	0.6952	0.9962	0.9932
	(0.0601)	(0.1334)	(0.1095)	(0.0475)	(0.0963)	(0.08)
18	0.6955	0.9945	0.9897	0.6952	0.9962	0.9932
	(0.0605)	(0.1345)	(0.1105)	(0.0475)	(0.0963)	(0.08)
19	0.6958	0.9937	0.9904	0.6952	0.9962	0.9932
	(0.0607)	(0.1354)	(0.111)	(0.0475)	(0.0963)	(0.08)
20	0.6957	0.9938	0.9902	0.6952	0.9962	0.9932
	(0.061)	(0.1362)	(0.1113)	(0.0475)	(0.0963)	(0.08)
21	0.6958	0.9935	0.9906	0.6951	0.9962	0.9933
	(0.0611)	(0.1363)	(0.1108)	(0.0474)	(0.0963)	(0.0799)
22	0.696	0.9932	0.9907	0.6951	0.9962	0.9933
	(0.0613)	(0.1372)	(0.1117)	(0.0474)	(0.0963)	(0.0799)
23	0.6961	0.9928	0.9912	0.6951	0.9962	0.9933
	(0.0612)	(0.1371)	(0.1117)	(0.0474)	(0.0964)	(0.0799)
24	0.6959	0.9933	0.9906	0.6951	0.9962	0.9933
	(0.0614)	(0.1374)	(0.1121)	(0.0474)	(0.0963)	(0.0798)
25	0.6959	0.9932	0.9908	0.6951	0.9962	0.9934
	(0.0612)	(0.137)	(0.1116)	(0.0473)	(0.0963)	(0.0798)
26	0.696	0.9931	0.9908	0.6951	0.9961	0.9934
	(0.0612)	(0.1371)	(0.1118)	(0.0474)	(0.0963)	(0.0798)
27	0.6961	0.9928	0.9912	0.6951	0.9961	0.9934
	(0.0611)	(0.1366)	(0.111)	(0.0473)	(0.0963)	(0.0798)
28	0.6959	0.9932	0.9907	0.6951	0.9961	0.9934
	(0.0612)	(0.1371)	(0.1118)	(0.0473)	(0.0964)	(0.0798)

Table 2: Continued on next page

Table 2: continued from previous page

	Pairwise Likelihood			Split data Likelihood		
Lag	$\hat{\phi}_P^{(L)}$	$\hat{\tau}_P^{(L)}$	$\hat{\sigma}_P^{(L)}$	$\hat{\phi}_{SD}^{(L)}$	$\hat{\tau}_{SD}^{(L)}$	$\hat{\sigma}_{SD}^{(L)}$
29	0.6959	0.9932	0.9907	0.6951	0.9961	0.9935
	(0.0612)	(0.1369)	(0.1115)	(0.0474)	(0.0964)	(0.0798)

We clearly see that the behavior of the variance of the maximum pairwise likelihood estimator is not monotonic: this is consistent with the existence of a 'best' lag distance L^*, in terms of minimum variance. Table 2 reports also the estimates and the variances of the estimates referred to the maximum split data likelihood estimator. Our empirical study shows that the variance in this case decreases to the variance of the maximum full likelihood estimator as L grows. These results empirically prove that maximum SDL estimator goes to the maximum full likelihood estimator as the length of the blocks of observations goes to infinity [Andrieu et al., 2007].

Figure 1 displays the behavior of the variances of the two estimators (PL and SDL) compared to the variance of the maximum full likelihood estimator. We clearly identify L^* as $L^* = 4$ and the monotonic decreasing trend of the SDL variance.

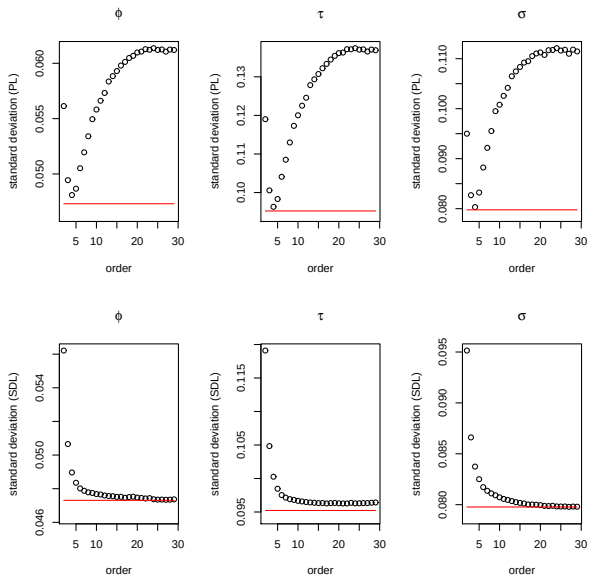

Figure 1: AR(1) model plus observation noise, with $\theta^* = (0.7, 1, 1)$. Standard deviations for the maximum likelihood estimator $\hat{\theta}_{ML}$ (solid line), the maximum pairwise likelihood estimator $\hat{\theta}_P^L$ (top) and the maximum split data likelihood estimator $\hat{\theta}_{SD}^{(L)}$ (bottom), with $L = 2, \ldots, 29$ denoting the maximum distance between the observations. Calculations based on 300 simulated time series of length 1000.

We repeat the simulation changing the values of θ^*. While σ^* and τ^* (and in particular the signal-to noise ratio τ^2/σ^2) do not seem to affect the optimal value

L^*, increasing the value of ϕ^* results in a bigger optimal value L^* (we recover the best order equal to six given in Varin and Vidoni [2009] when $\phi^* = 0.95$). This is probably connected to the weaker or stronger dependence structure of the pairs. Anyway, the fact that the optimal choice for the lag distance between the pairs depends on the unknown true parameter values makes its investigation ambiguous in real scenarios.

7 Conclusion

This paper dealt with the problem of static parameter estimation in general state space models. Given the difficulties arising in this framework, we have focused on inferential procedures based on composite likelihood functions, in particular *pairwise* and *split data likelihood* functions. Asymptotic properties of the parameter estimators obtained by maximizing these functions in general state space scenario were investigated. We proved that standard results, as strong consistency, depend on the properties of the processes involved, in particular stationarity and ergodicity that ensure forgetting behavior of the filter.

We also investigated efficiency problem in pairwise and split data likelihood framework as L, i.e. the lag distance between pairs or the length of a block, respectively, increases. We empirically proved that the loss of efficiency, with respect to maximum likelihood estimator, of the maximum split data likelihood estimator vanishes as L increases, while the variance of the maximum pairwise likelihood estimator decreases until a certain L^* and then it tends to increase. Anyway, until now, no general results about evaluation of this loss are available, even if this behavior is observed also in Varin and Vidoni [2009]. Moreover, we suggested the existence of a 'best lag' L^*, in terms of variance of the maximum pairwise likelihood estimator. However, we have not theoretically analyzed yet how to determine such value. In further research, we will investigate this topic through the evaluation of the asymptotic variance of the maximum pairwise likelihood estimator, in order to obtain an expression that depends on the lag distance L. We would like to follow the idea of Andrieu et al. [2007] in the non overlapping version of split data likelihood function. To do that, we need to quantify the loss of efficiency with respect to the full likelihood function. This requires a deep study of the dependence structures between pairs of observations, exploiting ergodic properties of the processes involved.

A Assumptions

Our results hold under the following assumptions

(C1) There exists $\underline{f_0}, \underline{g_0} > 0$ and $\overline{f}_0, \overline{g}_0 < \infty$ such that far all $x, x', y, \theta \in \mathcal{X}^2 \times \mathcal{Y} \times \Theta$

$$\underline{f_0} \leq f_\theta(x'|x) \leq \overline{f}_0, \quad \underline{g_0} \leq g_\theta(y|x) \leq \overline{g}_0$$

(C2) Θ is a compact set, θ^* is a unique global maximum of $l_P(\theta)$ and belongs to the

interior of Θ, denoted $\overset{\circ}{\Theta}$. Moreover $l_P(\theta)$ is twice continuously differentiable on $\overset{\circ}{\Theta}$ and $H_P(\theta^*) := \nabla^2 l_P(\theta^*)$ is positive definite.

(C3) f_θ and g_θ are continuous as functions of θ

There is an integer $L \geq 1$ such that, for every $j = 2, \ldots, L+1$

(C4) $p_\theta(y_1, y_j) = p_{\theta^*}(y_1, y_j)$ if and only if $\theta = \theta^*$

(C5) for the true parameter value θ^* we have $\mathbb{E}[|log[p_{\theta^*}(y_1, y_j)]|] < \infty$

(C6) for each θ there is a $\delta > 0$ (sufficiently small) such that

$$\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta':|\theta'-\theta|\leq\delta} \log p_{\theta'}(y_1, y_j)\right)^+\right] < \infty$$

and there is a $b > 0$ (sufficiently large) such that

$$\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta':|\theta'|>b} \log p_{\theta'}(y_1, y_j)\right)^+\right] < \infty,$$

where h^+ denotes the positive part of the function h

(C7) if $\lim_{i\to\infty} |\theta_i| = \infty$ then $\lim_{i\to\infty} p_{\theta_i}(y_1, y_j) = 0$.

Condition $(C1)$ implies that the process $\{X_k, Y_k\}$ is an uniformly ergodic Markov chain.

B Technical results about consistency

We prove here some middle results necessary to state that pairwise likelihood estimator is strongly consistent, as proved in Theorem 2. We start with a lemma concerning the L-dimensional Kullback-Leibler information. We recall that under the ergodicity assumption (C1), the log pairwise likelihood $l_P^{(L)}(\theta, y_{1:n})$ satisfies

$$\lim_{n\to\infty} l_P^{(L)}(\theta, y_{1:n}) = l_P^{(L)}(\theta).$$

Lemma 1. *Assume that Conditions $(C4 - C6)$ hold. Then $K_P^{(L)}(\theta, \theta^*) \geq 0$ with equality if and only if $\theta^* = \theta$.*

Proof. By Conditions $(C6)$, the expected values $l_P^{(L)}(\theta^*)$ and $l_P^{(L)}(\theta)$ exist. Because of the Assumption $(C5)$, we have that $l_P^{(L)}(\theta^*)$ is finite. If $l_P^{(L)}(\theta) = -\infty$, Lemma 1 obviously holds. Thus we shall consider the case when $l_P^{(L)}(\theta)$ is finite. Then

$K_P^{(L)}(\theta,\theta^*) = l_P^{(L)}(\theta^*) - l_P^{(L)}(\theta)$ exists finite. For every $j = 2,\dots,L+1$, by Jensen inequality we have that

$$\begin{aligned}
&\int_{\mathcal{Y}^2} \log \frac{p_\theta(y_1,y_j)}{p_{\theta^*}(y_1,y_j)} p_{\theta^*}(y_1,y_j)dy_1dy_j \\
&\leq \log\left[\int_{\mathcal{Y}^2} \frac{p_\theta(y_1,y_j)}{p_{\theta^*}(y_1,y_j)} p_{\theta^*}(y_1,y_j)dy_1dy_j\right] \\
&= \log\left[\int_{\mathcal{Y}^2} p_\theta(y_1,y_j)dy_1dy_j\right] = \log[1] = 0. \qquad (18)
\end{aligned}$$

Since

$$-K_P^{(L)}(\theta,\theta^*) = \frac{1}{L}\sum_{j=2}^{L+1}\int_{\mathcal{Y}^2} \log \frac{p_\theta(y_1,y_j)}{p_{\theta^*}(y_1,y_j)} p_{\theta^*}(y_1,y_j)dy_1dy_j$$

and given the result in (18), $K_P^{(L)}(\theta,\theta^*) \geq 0$ and this proves the first part of the lemma. The equality holds if and only if, for every $j = 2,\dots,L+1$, $p_\theta(y_1,y_j) = p_{\theta^*}(y_1,y_j)$ almost everywhere. By Condition $(C4)$, this is true if and only if $\theta = \theta^*$. □

Lemma 2. *Assume that Conditions* $(C3)$ *and* $(C6)$ *hold. Then for every* $\theta \in \Theta$ *and for every* $j = 2,\dots,L+1$

$$\lim_{\delta\to 0} \mathbb{E}_{\theta^*}\left[\sup_{\theta':|\theta'-\theta|\leq\delta} \log p_{\theta'}(y_1,y_j)\right] = \mathbb{E}_{\theta^*}\left[\log p_\theta(y_1,y_j)\right].$$

Proof. By Condition $(C3)$, $p_\theta(y_1,y_j)$ is continuous for all y_1, y_j, $j = 2,\dots,L+1$. Then

$$\lim_{\delta\to 0}\left(\sup_{\theta':|\theta'-\theta|\leq\delta} \log p_{\theta'}(y_1,y_j)\right)^+ = (\log p_\theta(y_1,y_j))^+,$$

except perhaps on a set whose probability measure is zero.
Since $\left(\sup_{\theta':|\theta'-\theta|\leq\delta} \log p_{\theta'}(y_1,y_j)\right)^+$ is an increasing function of δ, it follows from Assumption $(C6)$ that

$$\begin{aligned}
&\lim_{\delta\to 0} \mathbb{E}_{\theta^*}\left[\left(\sup_{\theta':|\theta'-\theta|\leq\delta} \log p_{\theta'}(y_1,y_j)\right)^+\right] \\
&\qquad = \mathbb{E}_{\theta^*}\left[\lim_{\delta\to 0}\left(\sup_{\theta':|\theta'-\theta|\leq\delta} \log p_{\theta'}(y_1,y_j)\right)^+\right] = \mathbb{E}_{\theta^*}\left[(\log p_\theta(y_1,y_j))^+\right]. \qquad (19)
\end{aligned}$$

Again by Condition $(C3)$

$$\lim_{\delta\to 0}\left(\sup_{\theta':|\theta'-\theta|\leq\delta} \log p_{\theta'}(y_1,y_j)\right)^- = (\log p_\theta(y_1,y_j))^-,$$

except perhaps on a set whose probability measure is zero, where h^- denotes the negative part of the function h. Then the relation

$$\lim_{\delta\to 0}\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta':|\theta'-\theta|\leq\delta}\log p_{\theta'}(y_1,y_j)\right)^-\right]=\mathbb{E}_{\theta^*}\left[(\log p_\theta(y_1,y_j))^-\right] \tag{20}$$

is clearly satisfied in both cases, when $\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta':|\theta'-\theta|\leq\delta}\log p_{\theta'}(y_1,y_j)\right)^-\right]$ is finite and when it is equal to $+\infty$. Lemma 2 is a consequence of (19) and (20). □

Lemma 3. *Assume that Conditions* $(C3, C6, C7)$ *hold. Then, for every* $j = 2,\ldots,L+1$

$$\lim_{b\to\infty}\mathbb{E}_{\theta^*}\left[\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j)\right]=-\infty.$$

Proof. From Assumptions $(C3)$ and $(C7)$

$$\lim_{b\to\infty}\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j)=\lim_{\theta\to\infty}\log p_\theta(y_1,y_j)=-\infty.$$

According to Assumption $(C6)$,

$$\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j)\right)^+\right]<\infty,$$

and since $\left(\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j)\right)^+$ is a decreasing function of b we have that

$$\lim_{b\to\infty}\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j)\right)^+\right]=0. \tag{21}$$

Since $\left(\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j)\right)^-$ is an increasing function of b, in the same way we have that

$$\lim_{b\to\infty}\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j)\right)^-\right]=+\infty \tag{22}$$

in both cases, when $\mathbb{E}_{\theta^*}\left[\left(\sup_{\theta:|\theta|>b}\log p_\theta(y_1,y_j)\right)^-\right]$ is finite and when it is equal to $+\infty$. Lemma 3 is a consequence of (21) and (22).

□

References

C. Andrieu, J. F. G. De Freitas, and A. Doucet. Sequential MCMC for bayesian model selection. In *Proceedings IEEE Workshop Higher Order Statistics*, 1999.

C. Andrieu, A. Doucet, and V. B. Tadic. On-line parameter estimation in general state-space models using pseudo-likelihood. 2007. URL `http://www.maths.bris.ac.uk/~maxca/preprints/andrieu\_doucet\_\\tadic\_2007`.

A. Azzalini. Maximum likelihood of order m for stationary stochastic processes. *Biometrika*, 70:367–81, 1983.

J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Stat. Soc. Ser. B*, 36:192–236, 1974.

J. Besag. Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika*, 64:616–18, 1977.

N. Chopin. Central limit theorem for sequential Monte Carlo methods and its application to bayesian inference. *Ann. Statist.*, 32:2385–411, 2004.

D. R. Cox and N. Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91:729–37, 2004.

P. Del Moral. *Feyman- Kac formulae. Genealogical and interacting particle approximations*. Probability and Applications. Springer, New York, 2004.

R. Douc, E. Moulines, and T. Ryden. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.*, 32:2254–304, 2004.

A. Doucet, J. F. G. de Freitas, and N. J. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer- Verlag, New York, 2001.

P. Fernhead. MCMC, sufficient statistics and particle filter. *J. Comput. Graph. Statist.*, 11:848–62, 2002.

W. R. Gilks and C. Berzuini. Following a moving target- Monte Carlo inference for dynamic bayesian models. *J. R. Stat. Soc. Ser. B*, 63:127–46, 2001.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proc. F*, 140:107–13, 1993.

G. Kitagawa. A self- organizing state- space model. *J. Amer. Statist. Assoc.*, 93: 1203–15, 1998.

S. Le Cessie and J. C. Van Houwelingen. Logistic regression for correlated binary data. *J. R. Stat. Soc. Ser. C*, 43:95–108, 1994.

B. G. Lindsay. Composite likelihood methods. *Contemp. Math.*, 80:221–39, 1988.

J. Liu and M. West. Combining parameter and state estimation in simulation- based filtering. In *Sequential Monte Carlo Methods in Practice*. 2001.

T. Ryden. Consistent and asymptotically normal parameter estimates for hidden Markov models. *Ann. Statist.*, 22:1841–95, 1994.

G. Storvik. Particle filters in state space models with the presence of unknown static parameters. *IEEE Trans. Signal Process.*, 50:281–89, 2002.

C. Varin and P. Vidoni. A note on composite likelihood inference and model selection. *Biometrika*, 92:519–28, 2005.

C. Varin and P. Vidoni. Pairwise likelihood inference for general state space models. *Econometric Rev.*, 28:170–85, 2009.

A. Wald. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statistics*, 29:595–601, 1949.

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it
Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Department of Statistical Sciences
University of Padua
Italy