Working Paper Series, N. 6, May 2010

Pairwise likelihood inference in state space models with unknown stationary distribution

Nadia Frigo
Department of Statistical Sciences
University of Padua
Italy

Christophe Andrieu
Department of Mathematics
University of Bristol
United Kingdom

Abstract: We consider stationary state space models for which the stationary distribution is not known analytically. We analyze the problem of static parameter estimation based on pairwise likelihood functions, motivated by the fact that for these general models the evaluation of the full likelihood function is often computationally infeasible. We quantify the bias in stationary models where the invariant distribution is unknown. For these models, an on line Expectation- Maximization algorithm to obtain the maximum pairwise likelihood estimate is developed. We illustrate the method for a linear gaussian model and we give an empirical evidence of our Bias theorem.

Keywords: Composite likelihood, Stationary distribution, Bias, Expectation Maximization algorithm.

Final version (2010-05-20)

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://www.homes.stat.unipd.it/frigo

Pairwise likelihood inference in state space models with unknown stationary distribution

Nadia Frigo
Department of Statistical Sciences
University of Padua
Italy

Christophe Andrieu
Department of Mathematics
University of Bristol
United Kingdom

Abstract: We consider stationary state space models for which the stationary distribution is not known analytically. We analyze the problem of static parameter estimation based on pairwise likelihood functions, motivated by the fact that for these general models the evaluation of the full likelihood function is often computationally infeasible. We quantify the bias in stationary models where the invariant distribution is unknown. For these models, an on line Expectation- Maximization algorithm to obtain the maximum pairwise likelihood estimate is developed. We illustrate the method for a linear gaussian model and we give an empirical evidence of our Bias theorem.

Keywords: Composite likelihood, Stationary distribution, Bias, Expectation Maximization algorithm.

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. If the parameter describing the model were known, the inferential problem would be focused on the latent process through the sequence of joint posterior distribution. 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. (e.g. Sequential Monte Carlo (SMC) methods (aka particle filters) as described in Doucet et al. [2001]). This methodology is now well developed and the theory supporting this approach is also well established [Del Moral, 2004].

More realistically, the parameter will be unknown and need 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. There have been many attempts to develop elaborate sequential algorithms, but all of them suffer from a common intrinsic problem, namely path degeneracy: 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].

Douc et al. [2004] have recently proved some theoretical results on the consistency and asymptotic normality of the maximum likelihood estimator in state space models. Anyway, in the cases when the latent process is continuous, evaluation of the full likelihood function is infeasible. Approximated solutions, based on Monte Carlo or numerical methods, have been considered, but none of them are completely satisfactory. It is possible to overcome this problem by replacing the likelihood function with another function, easier to determine. In this direction, composite likelihood functions have been suggested. They consist of 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 aim at analyzing the problem of static parameter estimation based on pairwise likelihood functions. We will focus on stationary state space models for which the stationary distribution underlying the process is unknown. The main interest of the work is the evaluation of the bias in the estimate when the invariant distribution is replaced by a generic distribution. The outline of the paper is as follows. Section 2 presents the model and justifies the inferential procedure we will focus on. Section 3 provides the main results of the paper, giving an expression for the bias of the estimate in the case where the invariant distribution is unknown. In addition, we suggest a possible way to choose a suitable approximation for the invariant distribution. Section 4 describes an on line Expectation- Maximization algorithm in order to obtain the maximum pairwise likelihood estimate in a general state space framework and in Section 5 we illustrate this method for a linear gaussian model, giving an empirical evidence of our Bias theorem. Section 6 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

$$p_\theta(x_{1:n}, y_{1:n}) = \pi_\theta(x_1) g_\theta(y_1|x_1) \prod_{i=2}^{n} f_\theta(x_i|x_{i-1}) g_\theta(y_i|x_i), \tag{3}$$

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 $\{p(\theta, Y_{1:n}); n \geq 1\}$.

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

$$L(\theta; y_{1:n}) = p_\theta(y_{1:n}) = \int_{\mathcal{X}^n} \pi_\theta(x_1) g_\theta(y_1|x_1) \prod_{i=2}^{n} f_\theta(x_i|x_{i-1}) g_\theta(y_i|x_i) dx_{1:n}, \tag{4}$$

which is simply obtained by taking into account the dependence structure characterizing the model. In general, finding the invariant distribution requires solving an integral equation. This is not a simple task even for a specific kind of model. Hence, in many situations π_θ, i.e. the stationary distribution, is not known analytically. We denote with $p_\theta(y_{1:n}|\mu)$ the joint distribution of the observations when $X_1 \sim \mu$, obtained by substituting μ for the true invariant distribution π_θ in (4). With this notation, $p_\theta(y_{1:n}) := p_\theta(y_{1:n}|\pi_\theta)$.

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). Their results allow one to consider the case where π_θ, and hence the true likelihood, is unknown. The technique relies primarily on the forgetting properties of the filter, uniformly in θ. 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* (PL) [Le Cessie and Van Houwelingen, 1994],

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

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.

From now on, we focus 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

$$L_P^{(L)}(\theta; y_{1:n}) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{min\{i+L,n\}} p_\theta(y_i, y_j). \tag{6}$$

Note that $L_P^{(n-1)}(\theta; y_{1:n})$ corresponds to

$$L_P(\theta; y_{1:n}) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} p_\theta(y_i, y_j), \tag{7}$$

obtained choosing $\omega_{ij} = 1, \quad \forall i = 1, \dots, n-1, j = i+1, \dots, n$ in (5). This function takes into account all the $n(n-1)/2$ pairs of observations. Given the dependence structure of the model (1, 2), for every $i = 1, \dots, n-1, j = i+1 \dots, n$

$$p_\theta(y_i, y_j) = \int_{\mathcal{X}^{j-i+1}} \pi_\theta(x_i) g_\theta(y_i|x_i) \left[\prod_{k=i+1}^{j} f_\theta(x_k|x_{k-1}) \right] g_\theta(y_j|x_j) dx_{i:j}. \tag{8}$$

As discussed in Frigo [2010], the use of (6) instead of (7) is justified by theoretical and practical motivations. In particular, Frigo [2010] points out the asymptotic behavior of the normalized log likelihood

$$l_P^{(L)}(\theta; y_{1:n}) = \frac{1}{n-1} \sum_{i=1}^{n-1} \left[\frac{1}{L} \sum_{j=i+1}^{min\{i+L,n\}} log[p_\theta(y_i, y_j)] \right] \tag{9}$$

as n goes to infinity. Under suitable ergodic assumptions

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

where $l_P^{(L)}(\theta)$ is defined by

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

In addition, by ergodic and stationary assumptions,

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

So, if L goes to infinity, all the information about the dependence structure of the model are lost. Moreover, in the case where the invariant distribution is unknown, all the inference is carried out from $p_\theta(y_1|\mu) = \int_{\mathcal{X}} \mu(x_1) g_\theta(y_1|x_1) dx_1$, that might be completely wrong.

The characterization of the bias of the estimate introduced when π_θ is unknown is hence of great importance. Therefore we need an approximation for the bivariate density (8).

3 Bias of the estimate when π_θ is unknown

In many situations (exceptions are, for example, linear gaussian models for the dynamic of $\{X_k\}$ and the discrete case), invariant distribution is unknown. Denoting by $p_\theta(y_i, y_j|\mu)$ the bivariate density of the observations y_i, y_j when the process is wrongly initialized by $X_1 \sim \mu(\cdot)$ we have that

$$p_\theta(y_i, y_j|\mu) = \int_{\mathcal{X}^j} \mu(x_1) \left[\prod_{k=2}^{j} f_\theta(x_k|x_{k-1}) \right] g_\theta(y_i|x_i) g_\theta(y_j|x_j) dx_{1:j}.$$

The definition above yields the following approximation of the likelihood defined in (6)

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

The following result quantifies the bias of the estimate introduced when the true invariant distribution π_θ is replaced with a generic distribution μ. We denote $\hat{\theta}_P(\mu)$

a generic maximum of the resulting approximate pairwise likelihood (10). Assumptions under which Theorem 1 holds are summarized in the Appendix A. Middle results can be found in the Appendix B.

Theorem 1 (Bias theorem)**.** *There exist* $C \in (0, +\infty)$ *and* $\rho \in [0, 1)$ *such that for any* $\mu \in \mathcal{P}(\mathcal{X})$

$$|\hat{\theta}_P(\mu) - \theta^*| \leq C \left|[\nabla^2 l_P(\theta^*)]^{-1}\right| \left[\frac{||\mu - \pi_{\theta^*}||}{1-\rho} + ||\nabla\mu - \nabla\pi_{\theta^*}||\right].$$

Proof. Let us consider the following Taylor expansion around θ^* and $\mu \in \mathcal{P}(\mathcal{X})$ such that $[\theta^*, \hat{\theta}_P(\mu)] \subset \overset{\circ}{\Theta}$,

$$\begin{aligned} \nabla l_P(\hat{\theta}_P(\mu)) &= \nabla l_P(\theta^*) + (\hat{\theta}_P(\mu) - \theta^*) \int_0^1 \nabla^2 l_P(\theta^* + t(\hat{\theta}_P(\mu) - \theta^*))dt \\ &= \nabla l_P(\theta^*) + (\hat{\theta}_P(\mu) - \theta^*)[R(\mu) + \nabla^2 l_P(\theta^*)], \end{aligned} \tag{11}$$

where

$$R(\mu) := \int_0^1 \nabla^2 l_P(\theta^* + t(\hat{\theta}_P(\mu) - \theta^*) - \nabla^2 l_P(\theta^*))dt.$$

Since the set of parameters maximizing $l_P(\theta)$ includes the true parameter, $\nabla l_P(\theta^*) = 0$. Moreover, by definition, $\nabla l_P(\hat{\theta}_P(\mu), \mu) = 0$. Hence (11) can be written as

$$\nabla l_P(\hat{\theta}_P(\mu)) = \nabla l_P(\hat{\theta}_P(\mu), \mu) + (\hat{\theta}_P(\mu) - \theta^*)[R(\mu) + \nabla^2 l_P(\theta^*)],$$

leading to

$$(\hat{\theta}_P(\mu) - \theta^*) = [R(\mu) + \nabla^2 l_P(\theta^*)]^{-1}[\nabla l_P(\hat{\theta}_P(\mu)) - \nabla l_P(\hat{\theta}_P(\mu), \mu)].$$

We have that $R(\mu)$ vanishes as $||\mu - \pi_\theta||$ goes to zero. This follows from the Theorem 4 and from the continuity in θ of the function $\nabla^2 l_P(\theta)$. Using the result in Theorem 3, we can easily conclude. □

In the theorem above, the constant ρ characterizes the forgetting properties of $\{X_k\}$ a priori and conditional upon $\{Y_k\}$. This result confirms the intuition that the bias introduced when using μ instead of π_{θ^*} in the pairwise likelihood depends on how close μ is to π_{θ^*} and on the ergodic properties of $\{X_k\}$.

The problem now is how to choose the distribution μ. In the cases in which the invariant distribution π_θ is unknown but transitions $f_\theta(\cdot|x)$ are simple, the idea is to approximate the invariant distribution π_θ sampling from the transition kernel $f_\theta(\cdot|x)$ and to take advantage of the geometric ergodicity of the process. More precisely, the idea is to take

$$\mu(x_{i-r:i}) = \mu(x_{i-r}) \prod_{k=i-r+1}^{i} f_\theta(x_k|x_{k-1}), \tag{12}$$

where, under geometric ergodicity, the marginal

$$\mu(x_i) \to \pi_\theta(x_i),$$

as r goes to $+\infty$.
In more complex situations, the choice of μ has to be carefully done, taking into account that this will affect the bias of the estimate.

4 Pairwise likelihood inference via EM algorithm

In this section we describe how to obtain estimates for the parameter θ describing a general state space model. We focus on an on line Expectation- Maximization (EM) technique to minimize, with respect to θ, the Kullback- Leibler divergence $K_P^{(L)}(\theta, \theta^*)$, or equivalently to minimize $l_P^{(L)}(\theta)$. The key advantage of the average log pairwise likelihood function compared to the full likelihood is that it only requires the estimation of expectations with respect to distributions defined on $\mathcal{X}^{L+1}$. More precisely, this technique allows us to find

$$\min_{\theta \in \Theta} K_P^{(L)}(\theta, \theta^*),$$

where

$$\begin{aligned} K_P^{(L)}(\theta, \theta^*) &= \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}_{\theta^*} \left[\log \frac{p_{\theta^*}(y_1, y_j)}{p_\theta(y_1, y_j)} \right] \\ &= \frac{1}{L} \sum_{j=2}^{L+1} \int_{\mathcal{Y}^2} \log \frac{p_{\theta^*}(y_1, y_j)}{p_\theta(y_1, y_j)} p_{\theta^*}(y_1, y_j) dy_1 dy_j. \end{aligned} \tag{13}$$

This is clearly equivalent to maximize $l_P^{(L)}(\theta)$, where

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

The EM algorithm is a general method to find the maximum likelihood estimate of the parameters of an underlying distribution from a given data set when the data are incomplete or have missing values. There are two main applications of the EM algorithm. The first occurs when the data indeed have missing values, due to problems with or limitations of the observation process. The second occurs when optimizing the likelihood function is analytically intractable but the likelihood function can be simplified by assuming the existence of values for additional but missing (or hidden) parameters. The latter application is more common in the computational pattern recognition community.

In summary, each iteration of the EM algorithm consists of two steps:

(**E-step**) In the expectation step (from now on E-step) the missing data are estimated given the observed data and current estimate of the model parameters. This is achieved using the conditional expectation, explaining the choice of terminology.

(**M-step**) In the maximization step (from now on M-step), the likelihood function is maximized under the assumption that the missing data are known. The estimate of the missing data from the E-step are used in lieu of the actual missing data.

These steps define an efficient iterative procedure to compute the maximum likelihood estimate and convergence is assured, since the algorithm is guaranteed to increase the likelihood at each iteration.

This general method can be modified in order to obtain the maximum pairwise likelihood estimate in a state space framework, provided that the algorithm increases the pairwise likelihood at each iteration.

Instead of the full likelihood, we want to minimize here, with respect to θ, the Kullback- Leibler divergence $K_P^{(L)}(\theta, \theta^*)$ as defined in (13). Given an estimate θ_k of θ^*, at iteration $k+1$ we update our estimate via

$$\theta_{k+1} = \arg\max_{\theta \in \Theta} Q(\theta, \theta_k),$$

where we define $Q(\theta, \theta_k)$ as

$$\begin{aligned} Q(\theta, \theta_k) &= \frac{1}{L} \sum_{j=2}^{L+1} \int_{\mathcal{X}^j \times \mathcal{Y}^2} \log[p_\theta(y_1, y_j, x_{1:j})] p_{\theta_k}(x_{1:j}|y_1, y_j) p_{\theta^*}(y_1, y_j) dx_{1:j} dy_1 dy_j \\ &= \frac{1}{L} \sum_{j=2}^{L+1} \int_{\mathcal{X}^j \times \mathcal{Y}^{L+1}} \log[p_\theta(y_1, y_j, x_{1:j})] p_{\theta_k}(x_{1:j}|y_1, y_j) p_{\theta^*}(y_{1:L+1}) dx_{1:j} dy_{1:L+1} \\ &= \frac{1}{L} \sum_{j=2}^{L+1} \int_{\mathcal{X}^j \times \mathcal{Y}^{L+1}} \log[p_\theta(y_1, y_j, x_{1:j})] p_{\theta_k}(x_{1:j}|y_1, y_j) p_{\theta^*}(\mathbf{y}_1) dx_{1:j} d\mathbf{y}_1, \end{aligned}$$

where $\mathbf{y}_s = y_{s:s+L}$ denote the s-th block of observations. For every $\theta \in \Theta$, we see that an iteration of this EM algorithm decreases the value of $K_P^{(L)}(\theta, \theta^*)$, and the stationary points correspond to the zeros of $K_P^{(L)}(\theta, \theta^*)$. In particular, we have that

$$0 \leq Q(\theta_{k+1}, \theta_k) - Q(\theta_k, \theta_k) \leq K_P^{(L)}(\theta_k, \theta^*) - K_P^{(L)}(\theta_{k+1}, \theta^*). \tag{14}$$

Note that the inequality in (14) does not depend on the initial distribution. It holds even if the initial invariant distribution is replaced with any initial distribution. In practice for the model we will consider, it is necessary to compute a set of sufficient statistics $\Phi(\theta_k, \theta^*)$ at time k in order to evaluate the function $Q(\theta, \theta_k)$. To do that we have to compute the expectation with respect to $p_{\theta_k}(x_{1:j}|y_1, y_j) \times p_{\theta^*}(\mathbf{y}_1)$. Even if it is possible to maximize $Q(\theta, \theta_k)$ analytically, in practice Q can not be computed as the expectation is with respect to a measure dependent on the unknown parameter value θ^*. However, thanks to the ergodicity and stationary assumptions, the observed process $\{Y_n\}$ provides us with a sample from $p_{\theta^*}(\mathbf{y}_1)$ which can be used for the purpose of Monte Carlo integration.

In what follows, we illustrate this method for a linear and gaussian model. It is a simple example, where the invariant distribution is known, as well as the conditional distribution of the latent states given the pairs of observations.

5 EM calculations for the linear gaussian model

Let us consider the linear gaussian model

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

The choice of the parameter vector $\theta = (\phi, \tau^2, \sigma^2) \in (-1, 1) \times \mathbb{R}^+ \times \mathbb{R}^+$ ensures stationarity. So

$$\begin{aligned} \pi_\theta(x) &= N\left(x; 0, \frac{\tau^2}{1-\phi^2}\right) \\ f_\theta(x'|x) &= N(x'; \phi x, \tau^2) \\ g_\theta(y|x) &= N(y; x, \sigma^2). \end{aligned}$$

We develop here an on line EM procedure, as suggested above. In order to compute the function $Q(\theta, \theta_k)$, we have to derive $\log[p_\theta(y_1, y_j, x_{1:j})]$, for every $j = 2, \ldots, L+1$. We have that

$$\begin{aligned} \log[p_\theta(y_1, y_j, x_{1:j})] &= \log[\pi_\theta(x_1)] + \log[g_\theta(y_1|x_1)] + \log[g_\theta(y_j|x_j)] + \\ &+ \sum_{k=2}^{j} \log[f_\theta(x_k|x_{k-1})]. \end{aligned}$$

From the definition of the model and by linearity of Q we have that

$$\begin{aligned} Q(\theta, \theta_k) &= \frac{1}{2}\log[1-\phi^2] - \frac{\log[\tau^2]}{2} - \frac{1}{2}\log[\tau^2]\left(\frac{1}{L}\frac{L(L+1)}{2}\right) - \log[\sigma^2] + \\ &- \frac{1}{2\tau^2}\frac{1}{L}\sum_{j=2}^{L+1}\left[\mathbb{E}^{(j)}_{\theta_k,\theta^*}[X_1^2 + X_j^2] + (1+\phi^2)\sum_{k=2}^{j-1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[X_k^2] - 2\phi\sum_{k=2}^{j}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[X_k X_{k-1}]\right] + \\ &- \frac{1}{2\sigma^2}\frac{1}{L}\sum_{j=2}^{L+1}\left[\mathbb{E}^{(j)}_{\theta_k,\theta^*}[Y_1^2 + X_1^2 - 2X_1Y_1 + Y_j^2 + X_j^2 - 2X_jY_j]\right]. \end{aligned}$$

In practice, for this model, it is necessary to compute a set of sufficient statistics $\Phi_i(\theta_k, \theta^*), i = 1, \ldots, 4$ at time k, where

$$\begin{aligned} \Phi_1(\theta_k, \theta^*) &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[Y_1^2 + X_1^2 - 2X_1Y_1 + Y_j^2 + X_j^2 - 2X_jY_j] \\ \Phi_2(\theta_k, \theta^*) &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}[X_1^2 + X_j^2] \\ \Phi_3(\theta_k, \theta^*) &= \frac{1}{L}\sum_{j=1}^{L+1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}\left[\sum_{k=2}^{j-1} X_k^2\right] \\ \Phi_4(\theta_k, \theta^*) &= \frac{1}{L}\sum_{j=2}^{L+1}\mathbb{E}^{(j)}_{\theta_k,\theta^*}\left[\sum_{k=2}^{j} X_k X_{k-1}\right]. \end{aligned}$$

With this definition

$$Q(\theta,\theta_k) = \frac{1}{2}\log[1-\phi^2] - \frac{\log[\tau^2]}{2} - \frac{(L+1)\log[\tau^2]}{4} - \log[\sigma^2] + \\ - \frac{1}{2\tau^2}\left(\Phi_2(\theta_k,\theta^*) + (1+\phi^2)\Phi_3(\theta_k,\theta^*) - 2\phi\Phi_4(\theta_k,\theta^*)\right) - \frac{1}{2\sigma^2}\Phi_1(\theta_k,\theta^*).$$

Now, dropping for simplicity the dependence on $\theta, \theta^*, \theta_k$,

$$\begin{aligned}
\frac{\partial Q}{\partial \phi} &= \frac{\phi}{1-\phi^2} + \frac{1}{\tau^2}(\phi\Phi_3 - \Phi_4) = 0 \\
\frac{\partial Q}{\partial \tau^2} &= 1 + \frac{L+1}{2} - \frac{1}{\tau^2}(\Phi_2 + (1+\phi^2)\Phi_3 - 2\phi\Phi_4) = 0 \\
\frac{\partial Q}{\partial \sigma^2} &= 1 - \frac{1}{2\sigma^2}\Phi_1,
\end{aligned}$$

so

$$\begin{aligned}
\sigma^2 &= \frac{\Phi_1}{2} \\
0 &= \phi\tau^2 + (1-\phi^2)(\phi\Phi_3 - \Phi_4) \\
\tau^2 &= \frac{2}{L+3}(\Phi_2 + (1+\phi^2)\Phi_3 - 2\phi\Phi_4).
\end{aligned}$$

One can solve the equations above analytically (discarding solutions for ϕ that fall outside the interval $[-1,1]$ and keep among the remaining values).

For every $i = 1,\dots,4$, we recursively approximate the sufficient statistics $\Phi_i(\theta_k,\theta^*)$ with the following update, given here at time k,

$$\hat{\Phi}_i^{(k)} = (1-\gamma_k)\hat{\Phi}_i^{(k-1)} + \gamma_k \left[\frac{1}{L}\sum_{j=2}^{L+1} \mathbb{E}_{\theta_k}^{(j)}[\Psi_i(X_{1:j}, Y_k, Y_{k+j-1})|\mathbf{Y}_k]\right], \qquad (15)$$

where, for every function $h(\cdot)$, $\mathbb{E}_{\theta_k}^{(j)}[h(X_{1:j})|\mathbf{Y}_k]$ denotes the expectation of h with respect to $p_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})$ and for $i = 1,\dots,4$ we have implicitly defined

$$\Phi_i(\theta_k,\theta^*) := \frac{1}{L}\sum_{j=2}^{L+1} \mathbb{E}_{\theta_k,\theta^*}^{(j)}[\Psi_i(X_{1:j}, Y_k, Y_{k+j-1})].$$

We then substitute $\hat{\Phi}_i^{(k)}$ for $\Phi_i(\theta_k,\theta^*)$ and obtain θ_k by maximizing the Q function. If θ_k was constant and $\gamma_k = k^{-1}$, then $\hat{\Phi}_i^{(k)}$ would simply compute the arithmetic average of $\{\mathbb{E}_{\theta_k}^{(j)}[\Psi_i(X_{1:j}, Y_k, Y_{k+j-1})|\mathbf{Y}_k]\}$ for every $j = 2,\dots,L+1$, and converge towards $\Phi(\theta_k,\theta^*)$ by ergodicity. In fact, under mild suitable conditions, convergence is in general ensured for any non- increasing positive sequence $\{\gamma_k\}$ such that $\sum\gamma_k < \infty$ and $\sum\gamma_k^2 < \infty$. We can select $\gamma_k = Mk^{-\alpha}$ where $M > 0$ and $\frac{1}{2} < \alpha \leq 1$ thanks to the theory of stochastic approximation [Benveniste et al., 1990].

In a linear gaussian setup, $\mathbb{E}_{\theta_k}^{(j)}[\Psi_i(X_{1:j}, Y_k, Y_{k+j-1})|\mathbf{Y}_k]$ is known for every $j = 2,\dots,L+1$, since $p_\theta(x_{1:j}|y_1,y_j)$ is normal $N_j(x_{1:j};\mu,\Sigma)$. In this case, we do not need to use a further Monte Carlo approximation.

Remark 1. *The on line algorithm described above takes a block of observations* $\mathbf{y}_k$ *for each iteration of the Expectation- Maximization steps. It can be modified in order to consider more blocks in each iteration or to run more than one iteration for a single block.*

Calculation of the posterior density $p_\theta(x_{1:j}|y_1, y_j)$, and in particular its first and second moments, can be achieved by a modification of the general Kalman filter and smoother. Unlike the standard Kalman filter equations, in this contest, the conditioning is on the observations y_1 and y_j and not on all the observations between 1 and j. Prediction and update steps need to be modified in order to obtain the right moments of the posterior distributions $p_\theta(x_{1:j}|y_1, y_j)$. Roughly speaking, we pretend to run a Kalman filter with all the observations $y_{1:j}$ setting an infinity variance for the missing observations from time 2 to time $j-1$.

In those steps, quantities that depend on the variance σ^2 disappear. The innovation at time k and its covariance do not need to be computed, and this allows us to avoid dealing with infinite quantities. Moreover, the meaning of the steps for $k = 2, \dots, j-1$ is quite sensible: if we do not take into account the observations $y_{2:j-1}$, the update step is missing and so the predicted and updated estimates coincide.

The parameters μ and Σ of the density $p_\theta(x_{1:j}|y_1, y_j)$ can be obtained from the smoothing recursions.

We implement the on line EM algorithm described above in order to estimate the parameter $\theta = (\phi, \tau, \sigma)$ of the linear gaussian model. We consider a simulated time series of length $n = 10000$ from the linear gaussian model, with $\phi^* = 0.7$, $\sigma^* = 1, \tau^* = 1$ as true parameter values. We fix the maximum lag distance between the observations as $L = 4$. In order to reduce the variance of the estimate, we used the Polyak- Ruppert averaging procedure. The algorithm was run with $\gamma_k = k^{-0.5}$ for $k \leq 2000$ and $\gamma_k = (k - 2000)^{-0.8}$ for $k > 2000$. The results of this method are displayed in Figure 1. The convergence to the true value is reached in few iteration steps.

This simple example, where the invariant distribution is known, as well as the conditional distribution of the latent states given the pairs of observations, allows us to apply the idea of approximating the stationary distribution by sampling from the transition kernel. We give an empirical evidence of our Bias theorem. In light of this, we develop the strategy suggested in Equation (12), since the invariant distribution is supposed to be unknown, but transitions $f_\theta(\cdot|x)$ are simple. In practice, we approximate the invariant distribution sampling from the transition kernel $f_\theta(\cdot|x)$ and we take advantage of the geometric ergodicity of the process.

5.1 Approximation of the invariant distribution

We take a generic initial distribution $\mu(\cdot)$ for x_{-z} and we simulate a sufficiently long Markov chain from the transition kernel. Under geometric ergodicity, the marginal distribution of x_1 converges to $\pi(x_1)$ as z goes to $+\infty$.

In order to take into account the state before time 0, we define the function

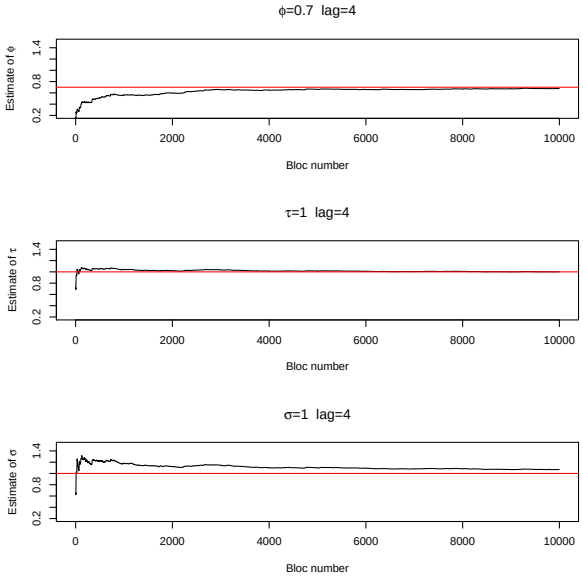

Figure 1: AR(1) model plus observation noise with $\theta^* = (0.7, 1, 1)$. Pairwise likelihood estimation using the on line EM algorithm with lag=4 denoting the maximum distance between the observations. Calculations based on a simulated series of length 10000. Initial value $\theta^{(0)} = (0.2, 0.5, 0.5)$.

$Q_z(\theta, \theta_k)$ as follows

$$Q_z(\theta, \theta_k) = \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta_k, \theta^*}[\log[p_\theta(y_1, y_j, x_{-z:j})]] =$$

$$= \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}^{(j)}_{\theta_k, \theta^*} \Bigg[\log[\mu_{x_{-z}}(x)] + \log[g_\theta(y_1|x_1)] + \log[g_\theta(y_j|x_j)] +$$

$$+ \sum_{k=-z+1}^{j} \log[f_\theta(x_k|x_{k-1})] \Bigg], \tag{16}$$

where $\mathbb{E}^{(j)}_{\theta_k, \theta^*}$ now denotes the expectation wrt $p_{\theta_k}(x_{-z:j}|y_1, y_j) p_{\theta^*}(y_1, y_j)$. If we choose $\mu_{x_{-z}}(\cdot)$ independent of θ, calculation of (16) and its maximization is derived in the same way as above.

We implement this idea for the AR(1) model, where stationary distribution is supposed to be unknown. We set as initial distribution $X_{-z} \sim \delta_6(x_{-z})$, where $\delta_a(x)$ denotes the Dirac delta mass density function at a and we take $z = 100$.

We report the distance between the estimate obtained taking δ_x, $x = 6$ as initial distribution and the estimate when the stationary distribution is known. In order to see how the idea suggested in (12) is useful, we compare it with the distance between the estimate obtained by approximate the invariant distribution by running a Markov chain of length $z = 100$ and the estimate when the stationary distribution is known. Reduction of the bias is displayed in Figure 2.

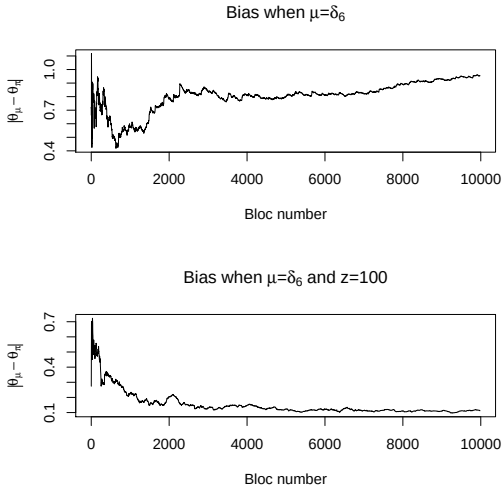

Figure 2: AR(1) model plus observation noise with $\theta^* = (0.7, 1, 1)$. Bias of the estimates when the invariant distribution is unknown and is approximated by taking as initial distribution $X_1 \sim \delta(x)$, where $x = 6$ (top) and $X_{-z} \sim \delta(x)$, where $x = 6$ and $z = 100$ (bottom).

Figure 3 reports the distance of the estimate with respect to the true parameter values.

6 Conclusion

In this paper we analyzed 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 pairwise likelihood functions.

Even if the models we considered are strictly stationary, in many situations invariant distribution is difficult (or even impossible) to compute. In this cases, it becomes important to quantify the bias in the estimate when stationary distribution is replaced with a generic approximation. When stationary distribution is unknown, objective functions need to be approximated and this leads to biased estimate of the parameters. We proved that the bias introduced when using a generic distribution instead of the stationary distribution in the pairwise likelihood function depends on how close the two distributions are, and on the ergodic properties of the latent process. To prove this result, we need uniform convergence of the pairwise likelihood function and of its gradient. In addition, we suggested a possible way to choose a suitable approximation for the invariant distribution. In the case in which the invariant distribution is unknown, but transitions for the latent process are simple, the idea is to approximate the invariant distribution sampling from this transition kernel and to take advantage of the geometric ergodicity of the process.

We focused on numerical methods to compute estimates of the parameter de-

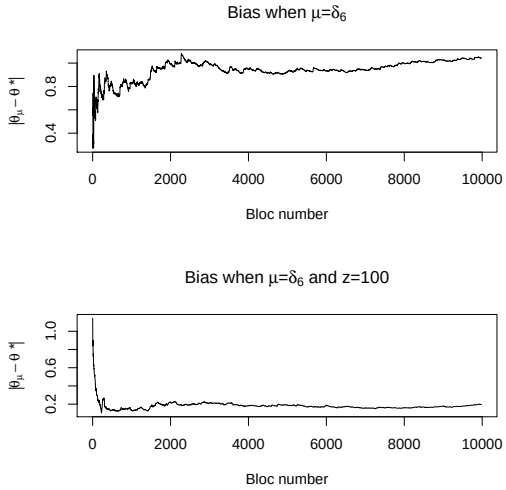

Figure 3: AR(1) model plus observation noise with $\theta^* = (0.7, 1, 1)$. Bias of the estimates with respect to the true parameter values when the invariant distribution is unknown and is approximated by taking as initial distribution $X_1 \sim \delta(x)$, where $x = 6$ (top) and $X_{-z} \sim \delta(x)$, where $x = 6$ and $z = 100$ (bottom).

scribing a general state space model. We presented an on line Expectation- Maximization algorithm to obtain the maximum pairwise likelihood estimate in a general state space framework. This algorithm is proved to increase the pairwise likelihood at each iteration step. We illustrated this method for a linear gaussian model, deriving the update equations in fairly explicit details. We modified standard Kalman filter recursions in order to take into account conditioning on pairs of observations instead of all observations. In this simple example, we gave an empirical evidence of our Bias theorem, i.e. starting from a generic distribution and sampling from the transition kernel reduces the bias in the estimates for each parameter in the model.

Further research will focus on numerical methods to compute estimate of the parameter in more general contexts.
In scenarios where $\mathbb{E}^{(j)}_{\theta_k}[\Psi_i(X_{1:j}, Y_k, Y_{k+j-1})|\mathbf{Y}_k]$, i.e. the expectation of Ψ with respect to $p_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})$ as defined in (15), does not admit an analytical expression, a further Monte Carlo approximation can be used. Assume that a good approximation $q_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})$ of $p_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})$ is available, and that it is easy to sample from $q_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})$. In this case, the expectation step will be altered as follows

- Sample $X^{(i)}_{1:j}$ from $q_{\theta_k}(\cdot|y_k, y_{k+j-1})$, for $i = 1, \ldots, N$
- Approximate $\Phi(\theta_k, \theta^*)$ as

$$\hat{\Phi}^{(k)} = (1 - \gamma_k)\hat{\Phi}^{(k-1)} + \gamma_k \left[\frac{1}{L} \sum_{j=2}^{L+1} \sum_{i=1}^{N} W^{(i)}_k \Psi(X^{(i)}_{1:j}, Y_k, Y_{k+j-1}) \right],$$

where

$$W_k^{(i)} \propto \frac{p_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})}{q_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})}, \quad \sum_{i=1}^{N} W_k^{(i)} = 1.$$

As N increases, the importance sampling approximation converges towards the true expectation. Note that if it is possible to sample from $p_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})$ exactly, then it is not necessary to have a large number N of samples and a single one might even be sufficient. Indeed it is only necessary to produce estimates of $\mathbb{E}_{\theta_k}^{(j)}[\Psi_i(X_{1:j}, Y_k, Y_{k+j-1})|\mathbf{Y}_k]$. We underline that the algorithm above leads to asymptotically biased estimates, but that this can be corrected by considering the following recursion for the estimation of the conditional expectation

$$\begin{aligned}
\hat{F}_k &= (1-\gamma_k)\hat{F}_{k-1} + \gamma_k \left[\frac{1}{L}\sum_{j=2}^{L+1} \frac{1}{N}\sum_{i=1}^{N} \frac{p_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})}{q_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})} \Psi(X_{1:j}^{(i)}, Y_k, Y_{k+j-1}) \right], \\
\hat{N}_k &= (1-\gamma_k)\hat{N}_{k-1} + \gamma_k \left[\frac{1}{L}\sum_{j=2}^{L+1} \frac{1}{N}\sum_{i=1}^{N} \frac{p_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})}{q_{\theta_k}(x_{1:j}|y_k, y_{k+j-1})} \right],
\end{aligned}$$

and let $\hat{\Phi}^{(k)} = \frac{\hat{F}_k}{\hat{N}_k}$. It is also possible to use rejection sampling or SMC techniques to approximate this expectation.

This idea may represent a starting point for subsequent extensions to more complex models.

A Assumptions

Our results hold under the following assumptions

(A1) 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.

(A2) We assume that f_θ and g_θ are twice continuously differentiable and that there exist $\underline{f_0}, \underline{g_0} > 0$ and $\overline{f_0}, \overline{g_0}, \overline{f_1}, \overline{g_1}, \overline{f_2}, \overline{g_2} < +\infty$ such that for all $x, x', y, \theta \in \mathcal{X}^2 \times \overline{\mathcal{Y}} \times \Theta$

$$\begin{aligned}
&\underline{f_0} \leq f_\theta(x'|x) \leq \overline{f_0}, \qquad \underline{g_0} \leq g_\theta(y|x) \leq \overline{g_0} \\
&|\nabla \log f_\theta(x'|x)| < \overline{f_1}, \qquad |\nabla \log g_\theta(y|x)| < \overline{g_1} \\
&|\nabla^2 \log f_\theta(x'|x)| < \overline{f_2} \quad \text{and} \quad |\nabla^2 \log g_\theta(y|x)| < \overline{g_2}.
\end{aligned}$$

In addition, we assume that $\nabla^2 \log f_\theta(x'|x)$ and $\nabla^2 \log g_\theta(y|x)$ are continuous in θ, uniformly in $x, x', y, \in \mathcal{X}^2 \times \mathcal{Y}$ and that $\sup_{\theta\in\Theta} |\nabla \log \mu| \leq \bar{\mu}$, with $\bar{\mu} \in (0, \infty)$, $\mu \in \mathcal{P}(\mathcal{X})$.

Assumptions (A2) implies that for all $x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})$,

$$P(x, A) := \int_A f_\theta(x'|x)dx' \geq \underline{f_0}\lambda(A),$$

where λ denotes the Lebesgue measure. This means that X has a unique invariant measure π_θ and is uniformly ergodic [Meyn and Tweedie, 1993].

B Technical and middle results

In this section we prove some uniform convergence results for $l_P^{(L)}(\theta, \mu)$ and its derivative. Hereafter, for simplicity, we drop the L index in $l_P^{(L)}(\cdot, \cdot) := l_P(\cdot, \cdot)$.
The first result states that $l_P(\theta, \mu)$ converges uniformly in θ to $l_P(\theta, \nu)$ as the total variation distance between μ and ν tends to zero (even if μ, ν can depend on θ, we omit the explicit dependence for notational convenience).

Theorem 2. *There exists a constant $C \in (0, +\infty)$ such that for any $\mu, \nu \in \mathcal{P}(\mathcal{X})$, $\theta \in \Theta$ and $L \geq 1$*

$$|l_P(\theta, \mu) - l_P(\theta, \nu)| \leq C||\mu - \nu||.$$

Proof. By definition

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

and using the following identity valid for any $x, y \in (0, +\infty)$,

$$|\log x - \log y| \leq \frac{|x - y|}{x \wedge y}, \tag{17}$$

we have

$$\begin{aligned} |l_P(\theta, \mu) - l_P(\theta, \nu)| &\leq \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}_{\theta^*} \left[\frac{|p_\theta(y_1, y_j|\mu) - p_\theta(y_1, y_j|\nu)|}{p_\theta(y_1, y_j|\mu) \wedge p_\theta(y_1, y_j|\nu)} \right] \\ &\leq \frac{1}{L} \sum_{j=2}^{L+1} C||\mu - \nu|| = C||\mu - \nu||. \end{aligned}$$

□

Now we look at the derivative of $l_P(\theta, \mu)$. For every $\mu, \nu \in \mathcal{P}(\mathcal{X})$ the difference of the gradient of two approximated pairwise likelihood of order L is

$$\nabla l_P(\theta, \mu) - \nabla l_P(\theta, \nu) = \frac{1}{L} \sum_{j=2}^{L+1} \mathbb{E}_{\theta^*} \left[\nabla \log p_\theta(y_1, y_j|\mu) - \nabla \log p_\theta(y_1, y_j|\nu) \right] \tag{18}$$

and

$$\begin{aligned} \nabla \log p_\theta(y_1, y_j, x_{1:j}|\mu) =& \nabla \log \mu(x_1) + \nabla \log g_\theta(y_1|x_1) + \\ &+ \sum_{k=2}^{j} \nabla \log f_\theta(x_k|x_{k-1}) + \nabla \log g_\theta(y_j|x_j). \end{aligned}$$

We prove the following result.

Theorem 3. *There exists a constant $C \in (0, +\infty)$ and $\rho \in [0, 1)$ such that for every $\mu, \nu \in \mathcal{P}(\mathcal{X})$, $\theta \in \Theta$, $L \geq 1$,*

$$|\nabla l_P(\theta, \mu) - \nabla l_P(\theta, \nu)| \leq C \left[\frac{||\mu - \nu||}{1 - \rho} + ||\nabla\mu - \nabla\nu|| \right].$$

Proof. Let us analyze the generic term of the sum (18). By an extension of the Fisher's identity

$$\begin{aligned}
&\nabla \log p_\theta(y_1, y_j|\mu) - \nabla \log p_\theta(y_1, y_j|\nu) \\
=&\mathbb{E}_{\theta^*}[\nabla \log p_\theta(y_1, y_j, x_{1:j}|\mu)|Y_1, Y_j, \mu] - \mathbb{E}_{\theta^*}[\nabla \log p_\theta(y_1, y_j, x_{1:j}|\nu)|Y_1, Y_j, \nu] \\
=&\int \nabla \log g_\theta(y_1|x_1) \left(p_\theta(x_{1:j}|y_1, y_j, \mu) - p_\theta(x_{1:j}|y_1, y_j, \nu)\right) dx_{1:j} \\
&+ \int \nabla \log g_\theta(y_j|x_j) \left(p_\theta(x_{1:j}|y_1, y_j, \mu) - p_\theta(x_{1:j}|y_1, y_j, \nu)\right) dx_{1:j} \\
&+ \sum_{k=2}^{j} \int \nabla \log f_\theta(x_k|x_{k-1}) \left(p_\theta(x_{1:j}|y_1, y_j, \mu) - p_\theta(x_{1:j}|y_1, y_j, \nu)\right) dx_{1:j} \\
&+ \left[\int \nabla \log \mu(x_1) p_\theta(x_{1:j}|y_1, y_j, \mu) dx_{1:j} - \int \nabla \log \nu(x_1) p_\theta(x_{1:j}|y_1, y_j, \nu) dx_{1:j} \right] \\
:=&T_1 + T_2 + T_3 + T_4.
\end{aligned}$$

We study the terms T_1, T_2, T_3, T_4 separately. Let us start with T_1.

$$\begin{aligned}
T_1 :=& \int \nabla \log g_\theta(y_1|x_1) \left(p_\theta(x_{1:j}|y_1, y_j, \mu) - p_\theta(x_{1:j}|y_1, y_j, \nu)\right) dx_{1:j} \\
=& \int \nabla \log g_\theta(y_1|x_1) \left[p_\theta(x_1|y_1, y_j, \mu) - p_\theta(x_1|y_1, y_j, \nu)\right] dx_1.
\end{aligned}$$

Under Assumptions $(A2)$,

$$\begin{aligned}
|T_1| &\leq \sup_{x_1} |\nabla \log g_\theta(y_1|x_1)| \cdot ||p_\theta(X_1 \in \cdot|y_1, y_j, \mu) - p_\theta(X_1 \in \cdot|y_1, y_j, \nu)|| \\
&\leq C||\mu - \nu||. \qquad (19)
\end{aligned}$$

Analogous calculations for T_2 yield to

$$|T_2| \leq C\rho^{j-1}||\mu - \nu||. \qquad (20)$$

Now, for every $k = 2, \ldots, j$

$$\begin{aligned}
&\int \nabla \log f_\theta(x_k|x_{k-1}) \left(p_\theta(x_{1:j}|y_1, y_j, \mu) - p_\theta(x_{1:j}|y_1, y_j, \nu)\right) dx_{1:j} \\
=&\int \left(p_\theta(x_{1:k-2}, x_{k-1:k}, x_{k+1:j}|y_1, y_j, \mu) - p_\theta(x_{1:k-2}, x_{k-1:k}, x_{k+1:j}|y_1, y_j, \nu)\right) \\
&\quad \cdot \nabla \log f_\theta(x_k|x_{k-1}) dx_{1:j} \\
=&\int \nabla \log f_\theta(x_k|x_{k-1}) \left[p_\theta(x_{k-1}, x_k|y_1, y_j, \mu) - p_\theta(x_{k-1}, x_k|y_1, y_j, \nu)\right] dx_{k-1:k}
\end{aligned}$$

$$
\begin{aligned}
&= \int \left[p_\theta(x_k|x_{k-1}, y_1, y_j, \mu)p_\theta(x_{k-1}|y_1, y_j, \mu) - p_\theta(x_k|x_{k-1}, y_1, y_j, \nu)p_\theta(x_{k-1}|y_1, y_j, \nu)\right] \\
&\qquad \cdot \nabla \log f_\theta(x_k|x_{k-1}) dx_{k-1:k} \\
&= \int \nabla \log f_\theta(x_k|x_{k-1}) p_\theta(x_k|x_{k-1}, y_1, y_j) \left[p_\theta(x_{k-1}|y_1, y_j, \mu) - p_\theta(x_{k-1}|y_1, y_j, \nu)\right] dx_{k-1:k} \\
&= \int \left[\int \nabla \log f_\theta(x_k|x_{k-1}) p_\theta(x_k|x_{k-1}, y_1, y_j) dx_k\right] \\
&\qquad \cdot \left[p_\theta(x_{k-1}|y_1, y_j, \mu) - p_\theta(x_{k-1}|y_1, y_j, \nu)\right] dx_{k-1} \\
&= \int \Psi(x_{k-1}) \left[p_\theta(x_{k-1}|y_1, y_j, \mu) - p_\theta(x_{k-1}|y_1, y_j, \nu)\right] dx_{k-1},
\end{aligned}
$$

where $\Psi(x_{k-1}) := \int \nabla \log f_\theta(x_k|x_{k-1}) p_\theta(x_k|x_{k-1}, y_1, y_j) dx_k$. Moreover,

$$
\sup_{x_{k-1}} |\Psi(x_{k-1})| \leq \overline{f_1} \int p_\theta(x_k|x_{k-1}, y_1, y_j) dx_k = \overline{f_1}. \tag{21}
$$

$$
|T_3| \leq \sum_{k=2}^{j} C\rho^{k-2} ||\mu - \nu|| \leq \frac{C||\mu - \nu||}{1-\rho}. \tag{22}
$$

The last term in the sum can be written as

$$
\begin{aligned}
&\int \nabla \log \mu(x_1) p_\theta(x_{1:j}|y_1, y_j, \mu) dx_{1:j} - \int \nabla \log \nu(x_1) p_\theta(x_{1:j}|y_1, y_j, \nu) dx_{1:j} \\
&= \int \nabla \log \mu(x_1) p_\theta(x_1, x_{2:j}|y_1, y_j, \mu) dx_{1:j} - \int \nabla \log \nu(x_1) p_\theta(x_1 x_{2:j}|y_1, y_j, \nu) dx_{1:j} \\
&= \int \nabla \log \mu(x_1) p_\theta(x_1|y_1, y_j, \mu) dx_1 - \int \nabla \log \nu(x_1) p_\theta(x_1|y_1, y_j, \nu) dx_1 \\
&= \int \frac{\nabla \mu(x_1)}{\mu(x_1)} \frac{p_\theta(x_1, y_1, y_j|\mu)}{p_\theta(y_1, y_j|\mu)} dx_1 - \int \frac{\nabla \nu(x_1)}{\nu(x_1)} \frac{p_\theta(x_1, y_1, y_j|\nu)}{p_\theta(y_1, y_j|\nu)} dx_1 \\
&= \int \nabla \mu(x_1) \frac{p_\theta(y_1, y_j|x_1)}{p_\theta(y_1, y_j|\mu)} dx_1 - \int \nabla \nu(x_1) \frac{p_\theta(y_1, y_j|x_1)}{p_\theta(y_1, y_j|\nu)} dx_1 \\
&= \int p_\theta(y_1, y_j|x_1) \left[\frac{\nabla \mu(x_1)}{p_\theta(y_1, y_j|\mu)} - \frac{\nabla \nu(x_1)}{p_\theta(y_1, y_j|\nu)}\right] dx_1.
\end{aligned}
$$

Since $p_\theta(y_1, y_j|x_1)$ is a bounded function of x_1 and the term in square brackets can be written as

$$
\begin{aligned}
\frac{\nabla \mu(x_1)}{p_\theta(y_1, y_j|\mu)} - \frac{\nabla \nu(x_1)}{p_\theta(y_1, y_j|\nu)} &= \frac{\nabla \mu(x_1) p_\theta(y_1, y_j|\nu) - \nabla \nu(x_1) p_\theta(y_1, y_j|\mu)}{p_\theta(y_1, y_j|\mu) p_\theta(y_1, y_j|\nu)} \\
&= \frac{(\nabla \mu(x_1) - \nabla \nu(x_1)) p_\theta(y_1, y_j|\nu) - \nabla \nu(x_1)(p_\theta(y_1, y_j|\mu) - p_\theta(y_1, y_j|\nu))}{p_\theta(y_1, y_j|\mu) p_\theta(y_1, y_j|\nu)}.
\end{aligned}
$$

Under Assumptions (A2) we have that

$$
|T_4| \leq C(||\nabla \mu - \nabla \nu|| + ||\mu - \nu||). \tag{23}
$$

From the results (19, 20, 22, 23), we conclude that

$$\begin{aligned} |\nabla l_P(\theta,\mu) - \nabla l_P(\theta,\nu)| &\leq \frac{1}{L}\sum_{j=2}^{L+1} \mathbb{E}_{\theta^*} |\nabla \log p_\theta(y_1, y_j|\mu) - \nabla \log p_\theta(y_1, y_j|\nu)| \\ &\leq C\left[\frac{||\mu-\nu||}{1-\rho} + ||\nabla\mu - \nabla\nu||\right]. \end{aligned}$$

□

Let us define the set

$$\hat{\theta}_P(\mu) := \arg\max_{\theta\in\Theta} l_P(\theta,\mu),$$

where, as usual

$$l_P(\theta,\mu) = \frac{1}{L}\sum_{j=2}^{L+1} \mathbb{E}_{\theta^*}[\log p_\theta(y_1, y_j|\mu)]$$

and $l_P(\theta,\pi_\theta) := l_P(\theta)$, being π_θ the unique stationary distribution. The set $\hat{\theta}_P(\mu)$ is not empty since Θ is compact and $l_P(\theta,\mu)$ is continuous from Assumptions (A2) whenever μ is continuous. For any $\epsilon \in (0,+\infty)$ and $\theta_0 \in \Theta$, let $B(\theta_0,\epsilon) = \{\theta \in \Theta : |\theta-\theta_0| \leq \epsilon\}$ and for any set $A \in \Theta$ let $d(\theta_0, A) = \inf\{|\theta-\theta_0| : \theta \in A\}$ the distance between θ_0 and the set A. Theorem 1 quantifies the bias when the (unknown) invariant distribution π_θ is replaced with a generic μ, that is the bias of the estimate introduced by maximizing $l_P(\theta,\mu)$ instead of $l_P(\theta)$. The result says that the bias depends on how close μ_{θ^*} is to π_{θ^*} and on the ergodicity properties of $\{X_k\}$, where θ^* denotes the true parameter. We prove the following statement.

Theorem 4. *Assume (A1). Then for any sequence of measures $\{\mu_k, k \geq 1\}$ with uniformly continuous (in θ) density such that $||\mu_k - \pi_\theta||$ goes to zero and for any $\epsilon > 0$ such that $B(\theta^*,\epsilon) \subset \overset{\circ}{\Theta}$, there exists $\underline{k}$ such that $\forall k \geq \underline{k}$, $l_P(\theta,\mu_k)$ has its maxima $\hat{\theta}(\mu_k)$ in $B(\theta^*,\epsilon)$ and*

$$\lim_{||\mu_k-\pi_\theta||\to 0} d(\theta^*, \hat{\theta}(\mu_k)) = 0. \tag{24}$$

Proof. Let ϵ be a strictly positive constant. We proceed by contradiction. Assume there exists a sequence of measures $\{\mu_k, k \geq 1\}$ with uniformly continuous (in θ) density such that $||\mu_k - \pi_\theta||$ goes to zero and $\hat{\theta}(\mu_k) \notin B(\theta^*,\epsilon)$. This means that the estimates obtained by maximizing $l_P(\theta,\mu_k)$ with respect to θ are far from the true parameter value. Hence $|\hat{\theta}(\mu_k) - \theta^*| > \epsilon \geq 0$. By definition of $\hat{\theta}(\mu_k)$, we have that

$$l_P(\theta^*,\mu_k) \leq l_P(\hat{\theta}(\mu_k),\mu_k).$$

Since $\{\hat{\theta}(\mu_k)\} \subset \Theta$, and Θ is bounded, it has at least an accumulation point $\tilde{\theta}^*$ corresponding to a subsequence of $\{\hat{\theta}(\mu_k)\}$. From Theorem 2, $l_P(\theta,\mu_k)$ converges uniformly to $l_P(\theta)$ as $||\mu_k - \pi_\theta||$ goes to zero and consequently

$$l_P(\tilde{\theta}^*) \geq l_P(\theta^*)$$

with $|\tilde{\theta}^* - \theta^*| > \epsilon \geq 0$. This contradicts the fact that θ^* is the unique strong maximum of $l_P(\theta)$. Equation (24) obviously holds. □

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.

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

A. Benveniste, M. Metivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximation*. Springer-Verlag, New York, 1990.

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.

N. Frigo. Composite likelihood function in state space models. Working Paper 5-2010, Department of Statistical Sciences, University of Padua, Italy, May 2010.

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.

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.

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Springer, London, 1993.

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.

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