

Comparison of approaches to inference in stationary $AR(1)$ models

Confronto fra metodi di inferenza nei modelli $AR(1)$ stazionari

Valentina Mameli, Monica Musio and Philip Dawid

Abstract In this paper, we consider an inference approach based on Hyvärinen’s local homogeneous scoring rule within the stationary first order autoregressive framework for which both full and pairwise likelihood-based inference are available. Simulation studies were conducted to compare the estimator found by resorting to the Hyvärinen score to the full and pairwise maximum likelihood estimators.

Abstract *Nel presente lavoro consideriamo un metodo di inferenza basato sulla regola di punteggio locale e omogenea di Hyvärinen nel contesto dei modelli autoregressivi stazionari del primo ordine per i quali sia il metodo della verosimiglianza completa che quella a coppie sono disponibili. Sono stati effettuati degli studi empirici per confrontare lo stimatore ottenuto con il metodo di Hyvärinen con gli stimatori di massima verosimiglianza classico e a coppie.*

Key words: Homogeneous scoring rules, Hyvärinen score, stationary first order autoregressive model, pairwise likelihood.

1 Introduction

Composite likelihood methods have become extremely popular in recent years due to their computational advantages in estimating the parameters of very complex statistical models; see [15] for a review on the subject. Among the best known composite likelihood methods, we mention the pairwise likelihood introduced by [10]. A viable alternative to the composite likelihood approach to overcome possible dif-

Valentina Mameli and Monica Musio
University of Cagliari, via Ospedale 72, Cagliari, e-mail: vmameli@unica.it, musio@unica.it

Philip Dawid
University of Cambridge, Wilberforce Road, Cambridge, e-mail: A.P.Dawid@statslab.cam.ac.uk

difficulties related to the derivation of the normalizing constant of the full likelihood has been proposed in [11, 12]. Their methodology is based on a local homogeneous proper scoring rule, known as the Hyvärinen score ([8]), defined as

$$S(x, Q) = \Delta \ln q(x) + \frac{1}{2} \|\nabla \ln q(x)\|^2, \quad (1)$$

where Q is the quoted distribution with density function q of a random variable X , taking values x in the set $\mathcal{X} = \mathbb{R}^k$ equipped with the standard norm $\|\cdot\|$. The notation ∇ denotes the gradient vector, and Δ the Laplacian operator. The procedure seems appealing since homogeneous local scoring rules do not require the specification of the normalising constant of the model, and the resulting estimator is asymptotically normal. The aim of this work is to compare numerically the behaviour of the estimators obtained from the Hyvärinen score, and the pairwise and full likelihood-based methods in models for which the maximum likelihood estimator is available. In particular, we focus our attention on the inference of the autoregressive parameter in stationary first order autoregressive processes. Results concerning comparisons between pairwise and full likelihood-based methods within the first order autoregressive framework have been proposed in [9].

The paper unfolds as follows. Section 2 gives a brief introduction to the basic notions on scoring rules. Section 3 describes first order autoregressive processes resorting to full and the pairwise likelihood-based methods, and to the Hyvärinen scoring rule approach. Section 4 summarizes the results of the simulation study conducted to compare the minimum Hyvärinen score estimator to both the full and the pairwise maximum likelihood estimators.

2 Scoring rules

A scoring rule is a loss function designed to measure the quality of a given probability distribution Q for a random variable X , in view of the result x of the random variable X , taking values in the set $\mathcal{X}$. Specifically, if a forecaster quotes a predictive distribution Q for X and the event x realizes, then the loss will be $S(x, Q)$; see [4]. The function $S(x, Q)$ takes values in the interval $(-\infty, \infty]$ and its expected value under P is denoted by $S(P, Q)$.

Definition 1. The scoring rule S is proper relative to the class of distributions $\mathcal{P}$ if

$$S(P, Q) \geq S(P, P), \quad \text{for all } P, Q \in \mathcal{P}. \quad (2)$$

It is strictly proper relative to $\mathcal{P}$ if $S(P, Q) \geq S(P, P)$ with equality if and only if $Q = P$.

Examples of proper scoring rules are the log-score, i.e. $S(x, q) = -\log q(x)$ ([7]), and the Hyvärinen score in formula (1). Note that we are implicitly identifying a distribution Q over $\mathcal{X}$ by its probability density q ; so the two notations $S(x, q)$ and

$S(x, Q)$ are indistinguishable.

Scoring rules whose values are not affected by any positive scale factor of the predictive density q are termed homogeneous. For instance, formula (1) remains unchanged if q is multiplied by a positive constant.

Another important feature of scoring rules is locality. A scoring rule is called local if it depends on the quoted distribution Q only through the value of the predictive density q of Q at the point x . A weaker definition is the definition of m -locality. A scoring rule is called local of order m if it depends on Q only through the values of the density q and the derivatives of q up to the order m at the point x .

An example of local scoring rule is the log-score, which depends only on the values of q at x . Whereas an example of homogeneous local scoring rule of order 2 is the Hyvärinen score, which involves the values of q , q' and q'' at the point x ; see formula (1).

2.1 Estimation using scoring rules

Let $(x_1, \dots, x_n)$ be a sample of size n from P_θ , a parametric family of distributions for the random variable X indexed by the parameter $\theta \in \Theta$, where Θ is an open subset of $\mathbb{R}^k$. Given a proper scoring rule S , let $S(x, \theta)$ denote $S(x, P_\theta)$.

Inference for the parameter θ may be performed using the total empirical score, i.e. $\sum_{i=1}^n S(x_i, \theta)$.

The minimum score estimate $\hat{\theta}_S$, if it exists, is the value of θ which minimizes the total empirical score, namely, $\hat{\theta}_S = \arg \min_{\theta} \sum_{i=1}^n S(x_i, \theta)$.

Under broad regularity conditions on the model, see for example [1], $\hat{\theta}_S$ is a root of the score equation:

$$s(\theta) := \sum_{i=1}^n s(x_i, \theta) = 0, \quad (3)$$

where $s(x, \theta)$ denotes the gradient vector of $S(x, \theta)$ with respect to θ . The score equation (3) yields an unbiased estimating equation; see [5]. It is worth pointing out that if S is the log-score, then the score equation (3) reduces to the likelihood equation, and the minimum score estimate $\hat{\theta}_S$ coincides with the maximum likelihood estimate $\hat{\theta}$.

Under suitable conditions on the model, the minimum score estimate $\hat{\theta}_S$ is consistent and asymptotically normally distributed with asymptotic covariance matrix given by the inverse of the Godambe information matrix $G(\theta)$ (see [6, 12]), namely $\hat{\theta}_S \sim N(\theta, \{nG(\theta)\}^{-1})$.

The Godambe information matrix G is defined as $G(\theta) := K(\theta)J(\theta)^{-1}K(\theta)$, where $J(\theta) = E(s(X, \theta)s(X, \theta)^T)$, and $K(\theta) = E(\nabla s(X, \theta)^T)$.

In the special case that θ is a scalar parameter, the ratio of $G(\theta)$ to the expected Fisher information $I(\theta)$ determines the asymptotic relative efficiency of $\hat{\theta}_S$ relative to the maximum likelihood estimator $\hat{\theta}$. Whereas, if θ is a d -dimensional vector,

the asymptotic relative efficiency of $\hat{\theta}_S$ relative to $\hat{\theta}$ is defined as the d -th root of the ratio of the determinants of $G(\theta)$ and $I(\theta)$; see [3].

3 The first order autoregressive process

The stationary univariate autoregressive process of order 1, denoted by $AR(1)$, is defined by the equation

$$y_t - \mu = \phi(y_{t-1} - \mu) + \lambda_t, \quad \text{with } t = 2, \dots, d, \quad (4)$$

where ϕ is a constant known as the autoregressive parameter, such that $|\phi| < 1$. Moreover, λ_t denotes a normal variable with 0 mean and variance σ^2 . The random variable y_1 is a Gaussian random variable with mean μ and variance $\frac{\sigma^2}{1-\phi^2}$. Then y_t is a linear combination of normal variables and $y_1, \dots, y_d$ are jointly normal with mean vector $\mu \mathbf{1}_d$, where $\mathbf{1}_d$ is the d -dimensional unit vector, and covariance matrix given by

$$\Omega = \frac{\sigma^2}{1-\phi^2} \begin{pmatrix} 1 & \phi & \phi^2 & \dots & \phi^{d-1} \\ \phi & 1 & \phi & \dots & \phi^{d-2} \\ \phi^2 & \phi & 1 & \dots & \phi^{d-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{d-1} & \phi^{d-2} & \phi^{d-3} & \dots & 1 \end{pmatrix}, \quad (5)$$

whose inverse is given by the matrix (see for instance [3])

$$\Omega^{-1} = \sigma^{-2} \begin{pmatrix} 1 & -\phi & 0 & \dots & 0 & 0 \\ -\phi & 1+\phi^2 & -\phi & \dots & 0 & 0 \\ 0 & -\phi & 1+\phi^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1+\phi^2 & -\phi \\ 0 & 0 & 0 & \dots & -\phi & 1 \end{pmatrix}. \quad (6)$$

Full likelihood approach for a single series

The overall full log-likelihood function of a single series is (see for example [13])

$$l(\theta) = -\frac{1}{2\sigma^2} \left\{ \sum_{t=1}^d (y_t - \mu)^2 + \phi^2 \sum_{t=2}^{d-1} (y_t - \mu)^2 - 2\phi \sum_{t=2}^d (y_t - \mu)(y_{t-1} - \mu) \right\} + \\ -\frac{d}{2} \log \sigma^2 + \frac{1}{2} \log(1 - \phi^2), \quad (7)$$

with $\theta = (\mu, \sigma, \phi)$.

The likelihood equations are

$$\begin{aligned}
l_\mu(\theta) &= \frac{1}{\sigma^2} \left\{ \sum_{t=1}^d (y_t - \mu) + \phi^2 \sum_{t=2}^{d-1} (y_t - \mu) - \phi \sum_{t=2}^d (y_t - \mu) - \phi \sum_{t=2}^d (y_{t-1} - \mu) \right\}; \\
l_\sigma(\theta) &= \frac{1}{\sigma^3} \left\{ \sum_{t=1}^d (y_t - \mu)^2 + \phi^2 \sum_{t=2}^{d-1} (y_t - \mu)^2 - 2\phi \sum_{t=2}^d (y_t - \mu)(y_{t-1} - \mu) \right\} - \frac{d}{\sigma}; \\
l_\phi(\theta) &= -\frac{1}{2\sigma^2} \left\{ \sum_{t=1}^d (y_t - \mu)^2 + 2\phi \sum_{t=2}^{d-1} (y_t - \mu)^2 - 2 \sum_{t=2}^d (y_t - \mu)(y_{t-1} - \mu) \right\} - \frac{\phi}{(1 - \phi^2)}.
\end{aligned}$$

Pairwise likelihood approach for a single series

The pairwise likelihood for contiguous pairs of observations of a single series is (see [13])

$$\begin{aligned}
pl(\theta) &= -\frac{1}{2\sigma^2} \left\{ \sum_{i=2}^d (y_i - \mu)^2 + \sum_{i=2}^d (y_{i-1} - \mu)^2 - 2\phi \sum_{i=2}^d (y_i - \mu)(y_{i-1} - \mu) \right\} + \\
&\quad - (d-1) \log \sigma^2 + \frac{(d-1)}{2} \log(1 - \phi^2).
\end{aligned} \tag{8}$$

The pairwise likelihood equations are (see Appendix in [13])

$$\begin{aligned}
pl_\mu(\theta) &= \frac{1-\phi}{\sigma^2} \left\{ \sum_{t=2}^d (y_t - \mu) + \sum_{t=2}^d (y_{t-1} - \mu) \right\}; \\
pl_\sigma(\theta) &= \frac{1}{\sigma^3} \left\{ \sum_{t=2}^d (y_t - \mu)^2 + \sum_{t=2}^d (y_{t-1} - \mu)^2 - 2\phi \sum_{t=2}^d (y_t - \mu)(y_{t-1} - \mu) \right\} - \frac{2(d-1)}{\sigma}; \\
pl_\phi(\theta) &= \frac{1}{\sigma^2} \sum_{t=2}^d (y_t - \mu)(y_{t-1} - \mu) - \frac{(d-1)\phi}{(1 - \phi^2)}.
\end{aligned}$$

The pairwise likelihood estimator denoted by $\hat{\theta}_p$ is the solution of the previous equations. Note that we differentiate with respect to σ and not with respect to σ^2 as in [13].

Hyvärinen approach for a single series

By using basic differentiation rules, it is easy to find the Hyvärinen score for the model

$$\begin{aligned}
HS(y, \theta) &= \frac{1}{2\sigma^4} \sum_{t=2}^{d-1} [(1 + \phi^2)(y_t - \mu) - \phi(y_{t-1} + y_{t+1} - 2\mu)]^2 - \frac{[2 + (d-2)(1 + \phi^2)]}{\sigma^2} \\
&\quad + \frac{[y_d - \mu - \phi(y_{d-1} - \mu)]^2}{2\sigma^4} + \frac{[y_1 - \mu - \phi(y_2 - \mu)]^2}{2\sigma^4}.
\end{aligned} \tag{9}$$

The minimum score estimate of θ , denoted by $\hat{\theta}_S$, can be found by minimizing the Hyvärinen score in the above equation or by solving simultaneously the following equations:

$$\begin{aligned}
HS_\mu(y, \theta) &= \frac{2\phi - (1 + \phi^2)}{\sigma^4} \sum_{t=2}^{d-1} [(1 + \phi^2)(y_t - \mu) - \phi(y_{t-1} + y_{t+1} - 2\mu)] + \\
&\quad + \frac{\phi - 1}{\sigma^4} \{[y_d - \mu - \phi(y_{d-1} - \mu)] + [y_1 - \mu - \phi(y_2 - \mu)]\}; \\
HS_\sigma(y, \theta) &= -\frac{2}{\sigma^5} \sum_{t=2}^{d-1} [(1 + \phi^2)(y_t - \mu) - \phi(y_{t-1} + y_{t+1} - 2\mu)]^2 + \frac{2[2 + (d-2)(1 + \phi^2)]}{\sigma^3} + \\
&\quad - \frac{2}{\sigma^5} \{[y_1 - \mu - \phi(y_2 - \mu)]^2 + [y_d - \mu - \phi(y_{d-1} - \mu)]^2\}; \\
HS_\phi(y, \theta) &= \frac{1}{\sigma^4} \sum_{t=2}^{d-1} [(1 + \phi^2)(y_t - \mu) - \phi(y_{t-1} + y_{t+1} - 2\mu)] [2\phi(y_t - \mu) - (y_{t-1} + y_{t+1} - 2\mu)] + \\
&\quad - \frac{2(d-2)\phi}{\sigma^2} - \frac{1}{\sigma^4} \{[y_d - \mu - \phi(y_{d-1} - \mu)](y_{d-1} - \mu) + [y_1 - \mu - \phi(y_2 - \mu)](y_2 - \mu)\}.
\end{aligned}$$

3.1 v independent series

Consider v independent first order autoregressive processes with the same parameter ϕ , in the special case that $\mu = 0$ and $\sigma = 1$. We can define a $v \times d$ random matrix Y where each row represents a stationary first order autoregressive processes with the same parameter ϕ . Moreover, each row of Y has a d -variate normal distribution with mean 0 and variance covariance matrix Ω defined at equation (5) and is independent of the other rows.

Inference for the parameter ϕ can be performed by resorting to the sum of v individual Hyvärinen scores or v score equations or even v pairwise score equations. Here, we exploit the fact that the sum-of-squares-and-products matrix $S = Y^T Y$, a sufficient statistic for the multivariate normal model, is characterized by the Wishart distribution with parameters v and Ω . Assume $v \geq d$, taking into consideration all of the properties of the derivatives of traces and determinants, it can be shown that the Hyvärinen score relative to the Wishart model based on variables $(s_{ij} : 1 \leq i \leq j \leq d)$ is

$$HS(S, \Omega) = -\frac{(v-d-1)}{2} \sum_{i=1}^d (s^{ii})^2 + \frac{1}{2} \sum_{i,j=1}^d \left[\frac{(v-d-1)}{2} s^{ji} - \frac{1}{2} \omega^{ji} \right]^2, \quad (10)$$

where ω^{ij} , s^{ij} , s_{ij} are the elements of the matrix Ω^{-1} , S^{-1} and S , respectively.

Since Ω depends only on the parameter ϕ , the score equation turns into

$$HS_\phi(S, \Omega) = -\frac{1}{2} \sum_{i,j=1}^d \left[\frac{(v-d-1)}{2} s^{ji} - \frac{1}{2} \omega^{ji} \right] \frac{\partial \omega^{ji}}{\partial \phi} = 0, \quad (11)$$

which, after taking account of the elements of the matrix Ω^{-1} defined in equation (6), leads to the cubic equation

$$\frac{(d+1-v)}{2} \left[\phi \sum_{i=2}^{d-1} s^{ii} - \sum_{i=1}^{d-1} s^{i+1i} \right] + \frac{\phi}{2} [(d-2)(1 + \phi^2) + d - 1] = 0. \quad (12)$$

4 Numerical assessment

We carried out a simulation study to asses the performance of the minimum Hyvärinen score estimator for the case that $\mu = 0$ and $\sigma = 1$. We consider 1000 random samples for $v = 2000$ independent $AR(1)$ models of length $d = 50$, for some values of the autoregressive coefficient ϕ . For each simulated sample, we estimate the parameter ϕ by resorting to the full likelihood method ($\hat{\phi}$), the pairwise likelihood ($\hat{\phi}_p$), the sum of v Hyvärinen scores ($\hat{\phi}_S$), and the Hyvärinen score based on the Wishart model ($\hat{\phi}_{HS}$). Results of the simulation study are summarized in Table 1. All calculations are performed with the software R ([14]).

It should be pointed out that, for large v , we could empirically estimate the function J for the sum of v Hyvärinen score by employing the mean of the functions $HS_{\phi}(y_{:i}, \hat{\phi}_S)^2$, for $i = 1, \dots, v$, where $y_{:i} = (y_{1i}, \dots, y_{di})$. We refer to [15] for a discussion of the estimation of J in the composite likelihood approach.

Table 1 Estimates with standard deviation (sd), root mean square error ($RMSE$), and absolute value of bias ($|BIAS|$), for varying values of ϕ , $v = 2000$, $d = 50$.*

(a)

ϕ	$\hat{\phi}$	$ BIAS $	sd	$RMSE$	$\hat{\phi}_p$	$ BIAS $	sd	$RMSE$
-0.6	-0.5999	0.0001	0.0025	0.0024	-0.5939	0.0061	0.0028	0.0066
-0.2	-0.2000	0	0.0031	0.0031	-0.1961	0.0039	0.0032	0.0050
0.0	0.0001	0.0001	0.0032	0.0032	0.0001	0.0001	0.0032	0.0032
0.3	0.3001	0.0001	0.0030	0.0030	0.2948	0.0052	0.0032	0.0061
0.7	0.7000	0	0.0022	0.0022	0.6950	0.0050	0.0025	0.0057

(b)

ϕ	$\hat{\phi}_S$	$ BIAS $	sd	$SRMSE$	$\hat{\phi}_{HS}$	$ BIAS $	$RMSE$
-0.6	-0.6002	0.0002	0.0041	0.0042	-0.6003	0.0003	0.0057
-0.2	-0.2000	0	0.0033	0.0033	-0.2000	0	0.0038
0.0	0.0001	0.0001	0.0032	0.0032	0.0000	0	0.0032
0.3	0.3001	0.0001	0.0035	0.0034	0.3001	0.0001	0.0043
0.7	0.6998	0.0002	0.0044	0.0045	0.6997	0.0003	0.0059

*All values in the Table are empirical estimates, except for the standard deviations of $\hat{\phi}$ which is based on the asymptotic formula of Fisher information function and the standard deviations of $\hat{\phi}_p$ and $\hat{\phi}_S$ which are obtained by resorting to an empirical estimate of the Godambe information function obtained by compounding the numerical estimate of J and the exact formula of K .

5 Discussion

Results from the simulation study reveal that $\hat{\phi}_S$ and $\hat{\phi}_{HS}$ produce, on average, estimates of the autoregressive parameter very close to the true values; see Table 1. Unreported results indicate that the minimum Hyvärinen score based on the Wishart model performs unsatisfactorily in terms of mean square error compared with the other methods when d and v are of comparable order, and when $|\phi|$ approaches 1. Moreover, the relative asymptotic efficiency (*ARE*) of both $\hat{\phi}_S$ and $\hat{\phi}_{HS}$ with respect to the maximum likelihood estimator decreases when $|\phi|$ increases and approaches 1. Although, we lose efficiency with respect to the maximum likelihood estimator as long as $|\phi|$ approaches 1, $\hat{\phi}_S$ may represent an appealing alternative to the pairwise likelihood approach in situations where computation of the normalizing constant is infeasible, since it has the advantages to use the full likelihood, and, when d is very small, to perform better in terms of bias. It would be also of interest to analyse the performance of $\hat{\phi}_S$ in relation to the nuisance parameter σ or to consider a vector parameter of interest $\theta = (\phi, \sigma)$. Indeed, as is known (see [2]), an aspect which deserves particular attention under this set-up is the dependence on the correlation structure of the consistency of the estimator when v is rather small with respect to d .

Acknowledgements This research was partially supported by *Progetto di Ricerca Fondamentale o di Base, L.R. 7/2007 annualità 2012, Università di Cagliari: “Sviluppi matematici nella Statistica metodologica con applicazioni a problemi reali”*.

References

1. Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.
2. Cox, D. R. and Reid, N. (2004), A note on pseudo-likelihood constructed from marginal densities, *Biometrika*, **91**, 729–737.
3. Davison, A. C. (2003). *Statistical Models*. Cambridge University Press, Cambridge.
4. Dawid, A. P. (1986). Probability forecasting. In *Encyclopedia of Statistical Sciences*, (ed. S. Kotz, N. L. Johnson, and C. B. Read), 210–218. Wiley-Interscience.
5. Dawid, A. P. and Lauritzen, S. L. (2005). The geometry of decision theory. In *Proceedings of the Second International Symposium on Information Geometry and its Applications*, 22–28. University of Tokyo.
6. Dawid, A. P., Musio, M., and Ventura, L. (2014). Minimum scoring rule inference, <http://arxiv.org/pdf/1403.3920v1.pdf>.
7. Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B*, **14**, 107–114.
8. Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning*, **6**, 695–709.
9. Jin, Z. (2010). *Aspects of Composite Likelihood Inference*. Ph.D. Thesis, University of Toronto. <http://hdl.handle.net/1807/26460>.
10. Le Cessie, S., Van Houwelingen, J.C. (1994). Logistic regression for correlated binary data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **43**, 95–108.

11. Musio, M. and Dawid, A. P. (2013). Local scoring rules: A versatile tool for inference. In Proceedings of the 59th ISI World Statistics Congress, Hong Kong. <http://2013.isiproceedings.org/Files/STS019-P3-S.pdf>.
12. Dawid, A. P., Musio, M. (2014). Theory and Applications of Proper Scoring Rules. Metron to appear, DOI: 10.1007/s40300-014-0039-y, <http://arxiv.org/pdf/1401.0398.pdf>.
13. Pace, L., Salvan, A. and Sartori, N. (2011). Adjusting composite likelihood ratio statistics. *Statistica Sinica*, **21**, 129–148.
14. R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.
15. Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, **21**, 5–42.