

# Parameter Estimators for Gaussian Models with Censored Time Series and Spatio-temporal Data

C.A. Glasbey<sup>1</sup>, I.M. Nevison<sup>1</sup> & A.G.M. Hunter<sup>2</sup>

<sup>1</sup> Biomathematics and Statistics Scotland, JCMB, King's Buildings, Edinburgh EH9 3JZ, Scotland

<sup>2</sup> Environmental Division, SAC, Bush Estate, Penicuik EH26 0PH, Scotland

**Abstract.** Computationally-fast algorithms are considered for estimating parameters in Gaussian time series and spatio-temporal models from censored and/or missing data. The problem arises in fitting models involving Gaussian latent variables to environmental data. Spectral estimators and least-squares fits of auto- and cross-covariances are found to be of similar efficiency for fitting models to rainfall and solar radiation data.

**Keywords.** Fourier transform, latent variable, multivariate time series, rainfall, solar radiation

## 1 Introduction

Environmental variables such as temperature can be modelled as Gaussian processes, whereas others, such as rainfall and solar radiation, are far from Gaussian but can possibly be transformed to normality. See Jones & Phelps (1996) for a review of weather models. In particular, many models have been proposed for rainfall, based either on point processes (Rodriguez-Iturbe *et al.*, 1988), or constructed in two stages: first a binary rain/no-rain process and then a rainfall distribution applied to the wet periods (Katz & Parlange, 1995). However, such models are far more difficult than Gaussian ones to study analytically, to combine with models of other environmental variables, or to make use of in forecasting. Glasbey & Nevison (1997) developed an alternative approach: they applied a monotonic transformation to rainfall data to achieve marginal normality. This defines a latent Gaussian variable, with zero rainfall corresponding to censored values below a threshold. A similar approach has been taken with solar radiation (Graham *et al.*, 1996): if observed values are divided by the elevation of the sun at that position in space and time (Page, 1986), the resulting variable is approximately a stationary Gaussian process. However, data sets are incomplete because the latent variable is unobservable during the night.

Missing and/or censored data are problematic in both time series and spatio-temporal modelling. Kleiner *et al.* (1979) estimated spectra from time series containing outliers; Jones (1980) used Kalman filters to fit autoregressive-moving average models to data with missing values; Kedem (1980), for computational speed, considered parameter estimation from binary time series obtained by a hard-limiting transformation. Glasbey & Nevison (1997) and Graham *et al.* (1996) both used an ad hoc procedure to estimate parameters in the latent Gaussian process, by minimising the sum of squares

$$L_C = \sum_{j=1}^m \sum_{k=1}^m \sum_{t=0}^{n'/2} \left( \hat{C}_{jkt} - C_{jkt} \right)^2 .$$

Here  $C_{jkt}$  and  $\hat{C}_{jkt}$  are, respectively, the expected and sample cross-covariances between series  $j$  and  $k$  at time lag  $t$ , and there are  $m$  series of length  $n \geq n'$ . In this paper, for  $n \gg m$  we consider alternative, computationally-fast estimators and the optimal choice of  $n'$ . A spectral approach is developed for multivariate, stationary processes in Section 2, and applied to rainfall data in Section 3 and solar radiation data in Section 4. Finally, conclusions are drawn in Section 5.

## 2 Spectral likelihood

The negative log-likelihood of a multivariate, stationary, Gaussian time series can be approximated by its spectral representation as a set of independent complex Wishart distributions,

$$L_S = \frac{1}{2} \sum_{l=-n/2}^{n/2-1} \left\{ \log |S_l| + \text{trace} \left( S_l^{-1} \hat{S}_l \right) \right\}$$

(Brillinger, 1974, p 238). Here  $S_l$  and  $\hat{S}_l$  are, respectively, the  $m \times m$  complex matrices of cross-spectral and cross-periodogram coefficients at frequency  $l/n$ , so that

$$\hat{S}_{jkl} = \frac{1}{n} \left( \sum_{t=1}^n y_{jt} e^{-2\pi i l t/n} \right) \left( \sum_{t=1}^n y_{kt} e^{+2\pi i l t/n} \right) = \sum_{t=-n/2}^{n/2-1} \hat{C}_{jkt} e^{-2\pi i l t/n},$$

where  $y_j$  and  $y_k$  are the  $j$ th and  $k$ th time series, and  $n$  is assumed to be even. The approximation is exact if covariances are circulant, i.e.  $C_{jkt} = C_{jk(t-n)}$ , and otherwise applies asymptotically as  $n \rightarrow \infty$ . In particular, for a bivariate series, at each frequency

$$S = \begin{bmatrix} S_1 & S_c + iS_q \\ S_c - iS_q & S_2 \end{bmatrix},$$

where  $S_1$  and  $S_2$  are the spectra of the two series,  $S_c$  is the co-spectrum and  $S_q$  is the quotient spectrum, and  $L_S =$

$$\frac{1}{2} \sum_{l=-n/2}^{n/2-1} \left\{ \log (S_{1l} S_{2l} - S_{cl}^2 - S_{ql}^2) + \frac{S_{1l} \hat{S}_{2l} + S_{2l} \hat{S}_{1l} - 2S_{cl} \hat{S}_{cl} - 2S_{ql} \hat{S}_{ql}}{S_{1l} S_{2l} - S_{cl}^2 - S_{ql}^2} \right\}.$$

To illustrate the use of  $L_S$  and  $L_C$ , 100 independent series of length 1000 were simulated from AR(1) and ARMA(1,1) processes with  $\phi = 0.8$  and  $(\phi, \theta) = (0.8, 0.5)$ , respectively. For estimation using  $L_S$ , a value of  $n' < n$  will suffice, with  $\hat{S}$  replaced by

$$\hat{S}'_{jkl} = \sum_{t=-n'/2}^{n'/2-1} \hat{C}_{jkt} e^{-2\pi i l t/n'},$$

and similarly for  $S$ . It is only necessary for  $n'$  to be large enough so that autocorrelation coefficients  $C_t \approx 0$  for  $t > n'/2$ . Table 1 shows the root-mean-square errors of parameter estimators obtained by minimising  $L_S$  and  $L_C$  for a range of values of  $n'$ , using NAG routine E04JAF (NAG, 1993), a quasi-Newton algorithm which permits bounds on the parameters. The smallest values in each column are displayed in bold. For both models, with these parameter values  $n' \geq 100$  is sufficient for estimators based on  $L_S$  to be fully efficient, because  $\phi^{50} \approx 10^{-5}$ , and therefore several root-mean-square

Table 1.  $1000 \times$  root-mean-square errors of parameter estimators

model =	AR(1)		ARMA(1,1)			
parameter =	$\phi$		$\phi$		$\theta$	
criterion =	$L_S$	$L_C$	$L_S$	$L_C$	$L_S$	$L_C$
$n' = 2$	<b>20</b>	<b>20</b>				
4	29	20	72	72	115	115
6	23	21	50	47	74	72
10	22	23	42	<b>40</b>	60	<b>58</b>
20	21	26	41	44	58	66
50	20	29	39	52	56	87
100	<b>20</b>	30	<b>39</b>	55	<b>56</b>	96
200	<b>20</b>	30	<b>39</b>	56	<b>56</b>	98
500	<b>20</b>	30	<b>39</b>	56	<b>56</b>	98
1000	<b>20</b>	30	<b>39</b>	56	<b>56</b>	98

errors are displayed in bold. As is well known,  $\hat{C}_1/\hat{C}_0$  is an efficient estimator of  $\phi$  in an AR(1) process, so  $n' = 2$  is the optimal choice in  $L_C$ . For the ARMA(1,1) process, no choice of  $n'$  leads to fully efficient estimator using  $L_C$ , but  $n' = 10$  is almost efficient for these values of the parameters.

### 3 Rainfall application

The data analysed by Glasbey & Nevison (1997) were a univariate time series of ten years of hourly rainfall data ( $n = 87600$ ) at Turnhouse, Edinburgh. A monotonic transformation converted them to zero mean, unit variance, Gaussian variables, except that zero rainfall corresponded to censored values below a threshold. It is, therefore, not possible to compute  $\hat{C}$  directly. We have considered two alternatives.

1. The faster method is to compute the sample autocorrelations of the observed data using Fourier methods, then apply a transformation which relates expected correlations of the data to expected correlations of the latent variable.
2. Alternatively, for each time lag  $t$ , we use the EM algorithm to obtain a maximum likelihood estimate for  $\hat{C}_t$ , by alternating between computing the *expected* correlation, conditional on the censored data, using standard bivariate Gaussian distributional theory (Johnson & Kotz, 1972), and *maximising* the likelihood by equating the correlation coefficient with its sample value.

In both cases,  $S'$  is then obtained by Fourier transforming  $\hat{C}$ . Note,  $L_S$  is a pseudo-likelihood rather than a log-likelihood, because  $\hat{C}$  is not a set of sample correlation coefficients. Also, because the variance is known, we are using correlations rather than covariances, but the methodology in Section 2 applies equally to this situation. A third option would have been to use Markov chain Monte Carlo methods (Gilks *et al.*, 1996), by alternating between using a Gibbs sampler to simulate censored values and sampling parameter values from  $L_S$ . However, this would have been very computationally intensive and in this paper we are restricting ourselves to fast methods.

The EM-algorithm, in conjunction with minimising  $L_C$  for  $n' = 960$ , was used to fit to the data an ARMA(2,1) model parametrised as

$$C_t = \alpha \lambda_1^{|t|} + (1 - \alpha) \lambda_2^{|t|},$$

with  $1 \geq \alpha, \lambda_1, \lambda_2 \geq 0$ . Values obtained using NAG routine E04JAF were  $\hat{\alpha} = 0.83, \hat{\lambda} = (0.787, 0.979)$ . The efficiency of this and alternative estimators,

**Table 2.** 1000 × root-mean-square errors of parameter estimators in rainfall model

parameter =	$\alpha$				$\lambda_1$				$\lambda_2$			
$\hat{C}$ =	transform		EM		transform		EM		transform		EM	
criterion =	$L_S$	$L_C$	$L_S$	$L_C$	$L_S$	$L_C$	$L_S$	$L_C$	$L_S$	$L_C$	$L_S$	$L_C$
$n' = 6$	526	486	474	411	362	354	334	296	107	108	93	84
12	453	465	309	340	249	239	84	97	92	90	62	67
24	293	270	165	153	123	99	29	33	62	52	41	35
48	215	121	65	70	101	33	12	17	47	26	19	18
96	158	71	45	37	50	24	9	<b>12</b>	29	15	10	8
240	127	55	32	<b>29</b>	46	21	7	12	20	<b>11</b>	6	<b>5</b>
480	<b>126</b>	<b>54</b>	<b>31</b>	31	<b>46</b>	<b>21</b>	<b>7</b>	13	<b>20</b>	11	<b>6</b>	5
960	<b>126</b>	60	<b>31</b>	32	<b>46</b>	24	<b>7</b>	13	<b>20</b>	12	<b>6</b>	6

and of different values of  $n'$ , were compared by simulating 100 independent series with these values of the parameters and the same level of censoring, and then re-estimating the parameters. Results are summarised in Table 2, again by root-mean-square errors and with the smallest values in each column displayed in bold. We see that it is better to obtain  $\hat{C}$  using the EM-algorithm than by transformation, in which case there is little to choose between  $L_S$  and  $L_C$  as criteria, provided we know the appropriate value for  $n'$ . For  $L_C$  a value of  $n'$  around 240 appears to be best, while for  $L_S$  it is sufficient for  $n' \geq 240$ .

#### 4 Solar radiation application

Graham *et al.* (1996) analysed solar radiation data which had been recorded every 30 seconds between 8am and 4pm for 27 months at pairs of sites in Edinburgh. The sites were changed each month, and 12 different sites were used in total. It was found that dividing each observed radiation value by the elevation of the sun at that position in space and time was effective in removing temporal trends in both the mean and variance of solar radiation, provided times were restricted to those for which the solar angle exceeded 0.05 radians. Covariances were found to be well modelled by

$$C_{jkt} = \sigma^2 \phi \sqrt{D_{jk}^2 + \delta^2(t + \kappa E_{jk})^2},$$

where  $D_{jk}$  is the distance between sites  $j$  and  $k$  and  $E_{jk}$  is the distance site  $j$  is to the east of site  $k$ . Therefore, correlations between observations decay exponentially with increasing temporal and/or spatial separation, and in addition there is a time delay with more easterly sites experiencing fluctuations in radiation later.

The model was fitted separately to each of the 27 months of data, by minimising  $L_C$  with  $n' = 48$  and  $\hat{C}$  obtained by computing the sample autocovariance separately for each day and then averaging over the month. For this problem, full maximum likelihood estimation would have been possible, for example by approximating the series by bivariate autoregressive processes of high order (Jones & Vecchia, 1993), but this would have been computationally expensive. Average values obtained for parameters were  $\hat{\phi} = 0.95$ ,  $\hat{\delta} = 0.36$ ,  $\hat{\kappa} = 0.95$  and  $\hat{\sigma}^2 = 0.37$ . Again, efficiencies of alternative estimators, and of different values of  $n'$ , were compared by simulating 100 independent series with these values of the parameters and then re-estimating the parameters, subject to the bounds:

$$1 \geq \phi \geq 0, \quad 10 \geq \delta \geq 0, \quad 10 \geq \kappa \geq -10, \quad 100 \geq \sigma^2 \geq 0.$$

**Table 3.** Root-mean-square errors of parameter estimators in solar radiation model

par. =	$\phi \times 10^4$				$\delta \times 10^3$				$\kappa \times 10^2$				$\sigma^2 \times 10^3$			
series =	complete		8 h/day		complete		8 h/day		complete		8 h/day		complete		8 h/day	
crit. =	$L_S$	$L_C$	$L_S$	$L_C$	$L_S$	$L_C$	$L_S$	$L_C$	$L_S$	$L_C$	$L_S$	$L_C$	$L_S$	$L_C$	$L_S$	$L_C$
$n' = 6$	1229	<b>33</b>	1158	<b>49</b>	722	10	699	18	77	17	80	32	337	21	337	31
12	858	33	886	49	628	10	669	<b>18</b>	89	17	88	31	353	21	351	31
24	1082	33	1080	49	726	<b>9</b>	732	18	78	16	77	30	358	21	357	<b>31</b>
48	1060	34	1073	51	743	10	743	19	53	<b>15</b>	53	<b>28</b>	355	21	358	31
96	977	36	953	53	698	11	716	23	134	16	156	31	364	<b>21</b>	354	31
240	317	37	578	56	507	14	475	32	345	23	470	43	307	21	302	33
480	1579	41	2130	62	139	18	224	38	517	27	622	52	243	22	258	36
960	564	54	630	93	137	19	163	76	484	28	597	56	104	23	156	44
2000	211	59	257	140	92	19	105	1777	314	28	397	56	187	23	409	49
4000	<b>31</b>	59	<b>46</b>	140	<b>8</b>	19	<b>15</b>	1777	<b>14</b>	28	<b>26</b>	56	<b>21</b>	23	<b>31</b>	49
10000	<b>31</b>	59	<b>46</b>	140	<b>8</b>	19	<b>15</b>	1780	<b>14</b>	28	<b>26</b>	56	<b>21</b>	23	<b>31</b>	49

A single month was simulated ( $n = 86400$ ), with  $D$  and  $E$  set to typical values of 6 km and 4 km respectively, using a high-order bivariate autoregressive approximation. For larger values of  $n'$  a problem was encountered in that little or no data were available for  $\hat{C}_t$  when  $2160 \geq t \geq 720$  or  $5040 \geq t \geq 3600$  (in units of 30 seconds), so these terms were omitted from  $L_C$ . It is not so straightforward with  $L_S$ , and three approaches were tried:

1. Shorten  $\hat{C}$  by omitting the missing terms before applying the Fourier transform to obtain  $\hat{S}'$ , and do the same to  $C$  before obtaining  $S'$ ;
2. Set the missing terms in  $\hat{C}$  to zero before obtaining  $\hat{S}'$ , and do the same to  $C$  before obtaining  $S'$ ;
3. Set the missing terms in  $\hat{C}$  to the corresponding terms in  $C$  for current values of the model parameters, and then apply the Fourier transform to obtain  $\hat{S}'$ .

The final approach produced by far the best results, and these are the ones given in Table 3. For comparison, results are also given for the hypothetical case where the complete time series is observed. In both cases, for  $L_C$  a value of  $n' \leq 48$  was found to be satisfactory, but  $L_S$  was marginally better provided that  $n' \geq 4000$ . However, for smaller values of  $n'$ ,  $L_S$  performed very poorly and many instances occurred where parameter estimates were at the limits of their ranges.

## 5 Discussion

Computationally-fast algorithms have been considered for estimating parameters in Gaussian time series and spatio-temporal models from censored data. Spectral estimators and least-squares fits of auto- and cross-covariances have been found to be of similar efficiency for fitting models to rainfall and solar radiation data. The advantages of the spectral approach are that it is slightly more efficient, it has better theoretical properties, such as being known to be fully efficient if data are not missing, the variance of the process is automatically constrained to be positive definite, and there is no problem in choosing an appropriate value for  $n'$ , it simply has to be large. On the other hand, the least-squares approach is less sensitive to choices to be made between alternative ways of obtaining  $\hat{C}$ , is computationally faster because small values of  $n'$  are usually adequate, and is easier to generalise to larger numbers

of series and irregular sampling schemes. The spectral method only gains in computational efficiency if spatial data are collected on a rectangular grid. Finally, the least-squares criterion is possibly more robust to distributional assumptions, as with variograms (Cressie, 1991, pp 90-99), and its efficiency can be improved by extending to weighted least squares and generalised least squares criteria.

### Acknowledgements

The work was supported by funds from the Scottish Office Agriculture, Environment and Fisheries Department. The solar radiation data were collected under CEC Contract No. JOU2-CT92-0018.

### References

- Brillinger, D.R. (1974). *Time series : Data Analysis and Theory*. Holt, Rinehart and Winston: New York.
- Cressie, N.A.C. (1991). *Statistics for Spatial Data*. Wiley: New York.
- (ed.) Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall: London.
- Glasbey, C.A. & Nevison, I.M. (1997). Rainfall modelling using a latent Gaussian variable. In *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions* (T.G. Gregoire *et al.*, eds.). Lecture Notes in Statistics **122**, Springer: New York, 233-242.
- Graham, R., Glasbey, C.A. & Hunter, A.G.M. (1996). Consequences of decentralised PV on local network management. Final report: variation of solar energy across a region: spatio-temporal models. *SAC Report*, Bush Estate, Penicuik EH26 0PH, Scotland.
- Johnson, N.L. & Kotz, S. (1972). *Distributions in Statistics : Continuous Multivariate Distributions*. Wiley: New York.
- Jones, J.E. & Phelps, K. (1996). A review of meteorological data and weather generators for practical use in agricultural and horticultural modelling. *Aspects of Applied Biology*, **46**, 5-12.
- Jones, R.H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics*, **22**, 389-395.
- Jones, R.H. & Vecchia, A.V. (1993). Fitting continuous ARMA models to unequally spaced spatial data. *Journal of the American Statistical Association*, **88**, 947-954.
- Katz, R.W. & Parlange, M.B. (1995). Generalizations of chain-dependent processes: applications to hourly precipitation. *Water Resources Research*, **31**, 1331-1341.
- Kedem, B. (1980). *Binary Time Series*. Dekker: New York.
- Kleiner, B. Martin, R.D. & Thomson, D.J. (1979). Robust estimation of power spectra (with discussion). *Journal of the Royal Statistical Society, Series B*, **41**, 313-351.
- Numerical Algorithms Group (1993). *Library Manual Mark 16*. NAG Central Office, 256 Banbury Road, Oxford OX2 7DE, UK.
- (ed.) Page, J.K. (1986). *Prediction of Solar Radiation on Inclined Surfaces*. Solar Energy R & D in the European Community: Series F, Volume 3 – Solar Radiation Data. D. Reidel Publishing Company: Dordrecht.
- Rodriguez-Iturbe, I., Cox, D.R. & Isham, V. (1988). A point process model for rainfall: further developments. *Proceedings of the Royal Society, London, Series A*, **417**, 283-298.