JOURNAL OF COMPUTATIONAL PHYSICS 101, 224–226 (1992)

NOTE

Correlation Length of Time Series in Statistical Simulations

Molecular simulation calculations lead to strongly correlated data series. In these cases the calculated variance is too low. Since the temperature of a sample of molecules is proportional to the variance of their velocities this will affect the calculated temperature. For a correct estimation of the temperature the correlation of the series has to be estimated.

In a recent paper Morales *et al.*, [1] investigated a method to estimate the correlation length in data series. They modified a procedure originally given in Straatsma *et al.* [2]. These authors considered a correlated time series X_i, $i=1, ..., n$, with constant spacing and calculated the variance of the mean $\bar{X}=(1/n)\sum_{i=1}^{n} X_i$ by

$$\operatorname{var}(\bar{X})=\frac{c_0}{n}\left[1+2\sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)r_k\right], \tag{1}$$

where $c_0=E(X_i-E(X_i))^2$ denotes the variance of X_i and

$$r_k=E(X_i-E(X_i))(X_{i+k}-E(X_{i+k}))/c_0$$

denotes the correlation between X_i and X_{i+k}. $\tau:=\sum_{k=1}^{\infty}(1-k/n)\,r_k$ is defined as the "correlation length" of the series.

Under the assumption that there exists a maximum lag K for which r_k differs from zero the expression $(1-k/n)$ in (1) is neglected and the correlation length of the series is estimated by (Straatsma *et al.* [2])

$$\tau_S=\sum_{k=1}^{K}\frac{c'_k}{c'_0}, \tag{2}$$

where

$$c'_k=\frac{1}{n-k}\sum_{i=1}^{n-k}(X_i-\bar{X})(X_{i+k}-\bar{X}). \tag{3}$$

Morales *et al.* proposed the modification of this quantity

$$\tau_M=\sum_{k=1}^{K}\left(1-\frac{k}{n}\right)\frac{c'_k}{c'_0} \tag{4}$$

which does not neglect the term $(1-k/n)$.

We applied these methods to estimate the correlation length of a series of samples given by a "direct simulation Monte Carlo" calculation for a rarefied gas flow. The method of Straatsma *et al.* yields estimates rather randomly distributed. For example, for a sample size $n=1000$ we obtain values for τ_S between about -500 and $+500$. On the other hand, the estimate τ_M always yields the value $-\frac{1}{2}$.

The behavior of the estimates can be explained by the fact that we could not assume the existence of a maximum lag K (for which r_k differs from zero) because nothing was known about the correlation structure of our series. Therefore the summation in (2) and (4) was performed from 1 to $n-1$. In this case we have $K=n-1$ and the expression $(1-k/n)$ is not negligible for the estimate τ_S. For this reason one may expect that τ_M is a correct estimate of the correlation length. But unfortunately τ_M always reduces to $-\frac{1}{2}$ in the case $K=n-1$ which can be shown as follows:

$$\begin{aligned}
c'_0\tau_M &= \frac{1}{n}\sum_{k=1}^{n-1}\sum_{i=1}^{n-k}(X_i-\bar{X})(X_{i+k}-\bar{X})\\
&= \frac{1}{n}\sum_{i=1}^{n-1}\sum_{k=1}^{n-i}(X_i-\bar{X})(X_{i+k}-\bar{X})\\
&= \frac{1}{n}\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}(X_i-\bar{X})(X_k-\bar{X})\\
&= \frac{1}{n}\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}(X_iX_k-X_i\bar{X}-X_k\bar{X}+\bar{X}^2)\\
&= \frac{1}{n}\Bigg[\underbrace{\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}(X_iX_k)}_{\text{(I)}}\\
&\quad -\underbrace{\bar{X}\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}(X_i+X_k)}_{\text{(II)}}+\frac{n(n-1)}{2}\bar{X}^2\Bigg].
\end{aligned} \tag{5}$$

The first term reduces to

$$\begin{aligned}
2\times(\mathrm{I}) &= 2\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}(X_iX_k)=\sum_{i=1}^{n}\sum_{k\neq i}(X_iX_k)\\
&= \sum_{i=1}^{n}\sum_{k=1}^{n}(X_iX_k)-\sum_{i=1}^{n}X_i^2\\
&= n^2\bar{X}^2-\sum_{i=1}^{n}X_i^2
\end{aligned}$$

0021-9991/92 $5.00

and the second gives

$$(\mathrm{II}) = \sum_{i=1}^{n-1} \sum_{k=i+1}^{n} (X_i + X_k) = (n-1)\, n\bar{X}$$

which can be proved by induction. Thus from (5) we obtain

$$c_0' \tau_M = \frac{1}{n}\left[-\frac{1}{2} \sum_{i=1}^{n} X_i^2 + \frac{1}{2} \bar{X}^2 n \right]$$

$$= -\frac{1}{2}\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 = -\frac{1}{2} c_0',$$

and this yields $\tau_M = -\frac{1}{2}$.

Usually the summation in (2) and (4) is not performed from 1 to $n-1$ because, when K increases, the number of terms in the summation diminishes, and in the limiting case, $K = n-1$, one is comparing the correlation of the first and last terms with respect to the mean of all n points $\bar{X}$. Nevertheless, this extreme case indicates some difficulties when τ_S or τ_M is calculated with a too large maximum lag K. Since τ_S varies very much and τ_M is nearly constant for large K, values for these estimates should only be given together with the used K, which has to be chosen very carefully. In practical applications, if a maximum lag cannot be determined from the data, these estimates are not recommended. We will illustrate these phenomena in a simulation study given below.

From a statistical point of view, correct estimates of r_k are given by Kendall [3] and Jenkins and Watts [4]

$$\widehat{r_k} = \frac{\sum_{i=1}^{n-k} (X_i - \bar{X}_1)(X_{i+k} - \bar{X}_2)}{\sqrt{\sum_{i=1}^{n-k} (X_i - \bar{X}_1)^2}\, \sqrt{\sum_{i=1}^{n-k} (X_{i+k} - \bar{X}_2)^2}}, \qquad (6)$$

where $\bar{X}_1$ denotes the mean of the first $n-k$ observations $X_1, ..., X_{n-k}$ and $\bar{X}_2$ is the mean of the last $n-k$ observations $X_{k+1}, ..., X_n$ of the series. For most purposes this estimate can be modified so as to measure all the variables about the mean of the whole series, i.e.,

$$\widetilde{r_k} = \frac{\sum_{i=1}^{n-k} (X_i - \bar{X})(X_{i+k} - \bar{X})}{\sqrt{\sum_{i=1}^{n-k} (X_i - \bar{X})^2}\, \sqrt{\sum_{i=1}^{n-k} (X_{i+k} - \bar{X})^2}}. \qquad (7)$$

The corresponding expressions for the correlation length are given by

$$\hat{\tau} = \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) \widehat{r_k}, \qquad \tilde{\tau} = \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) \widetilde{r_k}, \qquad (8)$$

where $\hat{r}_{n-1}$ is set to zero.

To compare these four procedures we constructed the following time series: Let $X_1 = 0$ and define X_i for $i \geqslant 2$ by

$$X_i = \begin{cases} X_{i-1} + 1 & \text{if } U_i < P \\ U_i & \text{if } U_i \geqslant P, \end{cases}$$

where U_i is randomly distributed on [0, 1] and $P \in (0, 1)$ is a given probability. Note that the case $P = 1$ gives a completely dependent series while $P = 0$ yields an independent series of random numbers uniformly distributed on [0, 1]. For series of length $n = 1000$ the correlation length is estimated by τ_S, τ_M, $\hat{\tau}$, and $\tilde{\tau}$. This procedure is repeated 10,000 times and means and standard deviation of the estimates are given in Tables I and II. These calculations are very time consuming and were performed on a Cray Y-MP in about 9000 s.

The results show that there are great differences between all four procedures. For increasing P the correlation length

TABLE I

Mean Values of the Different Estimates for the "Correlation Length"

P	0.00	0.25	0.50	0.75	1.00
τ_S	−1.0016	−1.3133	−1.9546	−3.7732	−665.67
τ_M	−0.5000	−0.5000	−0.5000	−0.5000	−0.500
$\tilde{\tau}$	−0.5001	−0.4798	−0.4560	−0.3539	87.520
$\hat{\tau}$	−0.3034	−0.1840	0.0779	0.8029	499.499

TABLE II

Standard Deviations

P	0.00	0.25	0.50	0.75	1.00
τ_S	0.0575	0.1090	0.1697	0.2537	0.0000
τ_M	0.0000	0.0000	0.0000	0.0000	0.0000
$\tilde{\tau}$	0.0004	0.0017	0.0030	0.0100	0.0000
$\hat{\tau}$	0.0082	0.0146	0.0335	0.0692	0.0000

TABLE III

Calculation of τ's by Summation over r_k for $k = 1$, K and $P = 0.0$

K	10	20	30	50	100	250	500	999
τ_S	-0.0101	-0.0190	-0.0297	-0.0486	-0.0979	-0.2486	-0.5007	-1.0016
τ_M	-0.0100	-0.0188	-0.0292	-0.0473	-0.0929	-0.2172	-0.3738	-0.5000
$\tilde{\tau}$	-0.0100	-0.0188	-0.0293	-0.0474	-0.0930	-0.2174	-0.3741	-0.5001
$\hat{\tau}$	-0.0100	-0.0188	-0.0293	-0.0474	-0.0926	-0.2106	-0.3045	-0.3034

TABLE IV

Calculation of τ's by Summation over r_k for $k=1$, K and $P=0.5$

K	10	20	30	50	100	250	500	999
τ_S	0.9197	0.8912	0.8606	0.8071	0.6650	0.2340	−0.5028	−1.9546
τ_M	0.9180	0.8900	0.8601	0.8088	0.6774	0.3225	−0.1376	−0.5000
$\tilde{\tau}$	0.9180	0.8900	0.8601	0.8089	0.6777	0.3240	−0.1363	−0.4560
$\hat{\tau}$	0.9180	0.8900	0.8601	0.8090	0.6789	0.3441	−0.0739	−0.0779

of the series also increases. This behavior is best represented by the estimate $\hat{\tau}$ and with some limitations by $\tilde{\tau}$. A further advantage of the estimates $\hat{\tau}$ and $\tilde{\tau}$ over the others is their small standard deviation. Therefore we recommended the approximations $\hat{\tau}$ and $\tilde{\tau}$ for the correlation length of time series if there exists no maximum lag K for which r_k differs from zero.

In practical applications the decision if r_k differs from zero is quite relative and often depends on physical reasoning. This yields extreme difficulties in the interpretation of the results of the estimations. To give an example we consider the above series for $P=0.0$, 0.5, 1.0 and $n=1000$. The correlation length was calculated as the mean values of 10,000 independent series for $K=10$, 20, 30, 50, 100, 250, 500, 1000. The results are given in Tables III–V.

Obviously there are no great differences between all four estimates if the maximum lag is less than 20. The lag K for which the estimates differ depends essentially on the correlation structure of the series. The difference between the estimates increase with increasing K, which corresponds to the results given in Table I.

TABLE V

Calculation of τ's by Summation over r_k for $k=1$, K and $P=1.0$

K	10	20	30	50	100	250	500	999
τ_S	9.8892	19.5743	29.0511	47.3641	89.2233	176.707	165.916	−665.67
τ_M	9.8350	19.3701	28.6054	46.1782	84.9010	157.844	155.625	−0.500
$\tilde{\tau}$	9.9427	197725	29.4770	48.4586	92.8003	185.693	186.875	87.520
$\hat{\tau}$	9.9450	19.7900	29.5350	48.7250	94.9500	218.625	374.750	499.499

For these reasons we recommend a careful use of the correlation length τ and its corresponding estimates. If a value for the maximum lag K can be prescribed that is not too large, then all the methods can be applied. It must be emphasized that results based on τ should only be given together with the maximum lag K which has been used in the calculations.

Unfortunately, especially in a physical application, it is often difficult to give a maximum lag K in advance. This leads back to our original example of the Monte Carlo simulation. Nothing is known about the correlation structure. It is only known that the molecules are rarely affected by collisions and the resulting time series are correlated. In this type of problem we recommend the use of $\hat{\tau}$ for the correlation length since it is most reliable. The computational effort is negligible because this is only done during a test phase to check the degree of correlation.

REFERENCES

1. J. J. Morales, M. J. Nuevo, and L. F. Rull, *J. Comput. Phys.* **89**, 432 (1990).
2. T. P. Straatsma, H. J. C. Berendsen, and A. J. Stam, *Mol. Phys.* **57** (1) 89 (1986).
3. M. G. Kendall, *Time Series* (Griffin, London, 1973), p. 40.
4. G. M. Jenkins and D. G. Watts, *Spectral Analysis and its Applications* (Holden–Day, San Francisco, 1968), p. 182.

Received September 19, 1990; revised May 24, 1991

S. Dietrich

Institute for Theoretical Fluid Mechanics
DLR, Bunsenstrasse 10
D(W)-3400 Goettingen, Germany

H. Dette

Institute of Mathematical Stochastics
University of Goettingen
Lotzestrasse 13
D(W)-3400 Goettingen, Germany