A Note on Likelihood Estimation of Missing Values in Time Series
Author(s): Daniel Peña and George C. Tiao
Source: *The American Statistician*, Vol. 45, No. 3 (Aug., 1991), pp. 212-213
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: http://www.jstor.org/stable/2684292
Accessed: 24-11-2015 09:37 UTC

A Note on Likelihood Estimation of Missing Values in Time Series

DANIEL PEÑA and GEORGE C. TIAO*

Missing values in time series can be treated as unknown parameters and estimated by maximum likelihood or as random variables and predicted by the expectation of the unknown values given the data. The difference between these two procedures is illustrated by an example. It is argued that the second procedure is, in general, more relevant for estimating missing values in time series.

KEY WORDS: ARIMA models; Interpolation; Mean squared error.

The use of the maximum likelihood method to estimate missing observations or, in general, unobserved values of random variables is a controversial topic because different authors use different likelihoods to obtain the estimators [see Bayarri, DeGroot, and Kadane (1986) and discussion and Fuller (1987)].

To illustrate the problem in time series models, suppose that the time series $\{z_t\}$ follows the stationary first-order autoregressive process

$$z_t = \phi z_{t-1} + a_t, \qquad |\phi| < 1,$$

where the a_t's are iid $N(0, \sigma^2)$. For simplicity, let us assume that ϕ and σ^2 are known. Suppose that out of n observations z_t ($t = 1, \ldots, n$), the observation z_T is missing ($1 \le T \le n$). Then, denoting Z_n by the $n \times 1$ vector $Z_n = (z_1, \ldots, z_n)'$ and $Z_{(T)}$ by the $(n-1) \times 1$ vector obtained from Z_n by dropping z_T, the joint density function of the available data $Z_{(T)}$ for given z_T is

$$f(Z_{(T)} \mid z_T) = \frac{f(Z_n)}{f(z_T)}, \tag{1}$$

where

$$f(Z_n) = (2\pi\sigma^2)^{-(1/2)n}(1-\phi^2)^{1/2} \times \exp\left\{-\frac{1}{2\sigma^2}\left[(1-\phi^2)z_1^2 + \sum_{t=2}^{n}(z_t - \phi z_{t-1})^2\right]\right\} \tag{2}$$

and

$$f(z_T) = (2\pi\sigma^2)^{-1/2}(1-\phi^2)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}(1-\phi^2)z_T^2\right\}. \tag{3}$$

In (1), z_T is now an unknown parameter of the model for $Z_{(T)}$. The likelihood function of z_T can be written as

$$l(z_T \mid Z_{(T)}) = (2\pi\sigma^2)^{-1/2(n-1)}\exp\left\{-\frac{1}{2\sigma^2} \times \left[\sum_{t=1}^{T-1}(z_t - \phi z_{t+1})^2 + \sum_{t=T+1}^{n}(z_t - \phi z_{t-1})^2\right]\right\}, \tag{4}$$

where it is understood that in the exponent a term disappears if the range of summation is not positive.

Hence the maximum likelihood estimator of z_T takes the form

$$\begin{aligned}\hat{z}_T &= \phi^{-1}z_{T+\delta}, & T &= 1 \text{ or } n\\ &= (2\phi)^{-1}(z_{T+1} + z_{T-1}), & 1 &< T < n,\end{aligned} \tag{5}$$

with $\delta = 1$ if $T = 1$ and $\delta = -1$ if $T = n$. It is easy to verify from (4) that the mean squared error (MSE) of $\hat{z}_T$ is

$$\begin{aligned}\text{MSE}(\hat{z}_T) &= \phi^{-2}\sigma^2, & T &= 1 \text{ or } n\\ &= (2\phi^2)^{-1}\sigma^2, & 1 &< T < n.\end{aligned} \tag{6}$$

Several authors have computed least squares or "maximum likelihood" estimators by maximizing the joint density function $f(Z_n)$ in (2) with respect to missing observations for known data. [See, for instance, Brubacher and Tunnicliffe-Wilson (1976)]. Then, in the present case, it is easy to verify that the estimator is given by

$$\begin{aligned}\tilde{z}_T &= \phi z_{T+\delta}, & T &= 1 \text{ or } n\\ &= (1+\phi^2)^{-1}\phi(z_{T+1} + z_{T-1}), & 1 &< T < n.\end{aligned} \tag{7}$$

This estimator $\tilde{z}_t$ cannot be called a maximum likelihood estimator, however, because the function $f(Z_n)$ for z_T considered as an unknown parameter is not a joint density function, and, therefore, $f(Z_n)$ for unobserved z_T and known $Z_{(T)}$ is *not* a likelihood function as it is usually defined in standard texts.

To interpret the meaning of (7), let us consider z_T as a random variable following the probabilistic structure in (2). Then, the distribution of z_T given the data $Z_{(T)}$ is

$$f(z_T \mid Z_{(T)}) = \frac{f(Z_n)}{f(Z_{(T)})}, \tag{8}$$

where $f(Z_{(T)})$ can be obtained by integrating out z_T from $f(Z_n)$. As is well known [see, for example, Peña (1987)], the distribution in (8) is normal with

$$\begin{aligned}E(z_T \mid Z_{(T)}) &= \tilde{z}_T,\\ \text{var}(z_T \mid Z_{(T)}) &= \sigma^2, & T &= 1 \text{ or } n\\ &= (1+\phi^2)^{-1}\sigma^2, & 1 &< T < n.\end{aligned} \tag{9}$$

We see that the conditional expectation, $E(z_T \mid Z_{(T)})$, is equal to $\tilde{z}_T$ in (7); this is because $f(Z_n)$ is proportional

*Daniel Peña is Professor of Statistics, Laboratorio de Estadística, ETSII, Universidad Politécnica de Madrid, Madrid 28006, Spain. George C. Tiao is W. Allen Wallis Professor of Statistics, Graduate School of Business, University of Chicago, Chicago, IL 60637.

to $f(z_t \mid Z_{(T)})$ and, for the present example, the mode and the mean of the distribution $f(z_T \mid Z_{(T)})$ are identical. More important, we see that $E(z_T \mid Z_{(T)})$, which is the minimum MSE estimator of z_T, can be very different from the maximum likelihood estimator $\hat{z}_T$ in (5). Indeed, the MSE of $\hat{z}_T$ in (6) always exceeds, and can be very much larger than, $\operatorname{var}(z_T \mid Z_{(T)})$ in (9), which is, of course, also the MSE of the estimator $E(z_T \mid Z_{(T)})$.

The difference between the two estimators, $\tilde{z}_T$ and $\hat{z}_T$, is not surprising if we look at the problem from a Bayesian point of view. The estimator $\tilde{z}_T$ is the mean (or mode) of the posterior distribution $f(z_T \mid Z_{(T)})$ in (8), which is proportional to the product of the likelihood function $l(z_T \mid Z_{(T)})$ in (4) and the prior distribution $f(z_T)$ in (3). On the other hand, the estimator $\hat{z}_T$ can be regarded as the mean (or mode) of a posterior distribution of z_T proportional to the product $l(z_T \mid Z_{(T)})p_o(z_T)$, where $p_o(z_T)$ is a "locally uniform" or noninformative prior distribution (Box and Tiao 1973). Thus, in the stationary case, $|\phi| < 1$, the two means can be very different because very different prior distributions are employed. This also explains the fact that when ϕ goes to 1 (the model approaches a nonstationary one) the difference between these two estimators goes to 0 simply because in this case the prior distribution $f(z_T)$ also becomes nearly locally uniform.

It may be argued from a frequentist point of view that the optimal properties of $\hat{z}_T$ and $\tilde{z}_T$ in (6) and (9), respectively, are not really comparable, because they are obtained under very different assumptions. For the maximum likelihood estimator $\hat{z}_T$ the unknown observation z_T is regarded as a fixed parameter, and the $\operatorname{MSE}(\hat{z}_T)$ in (6) is obtained under (or at least motivated by) such an assumption; for the estimator $\tilde{z}_T$, however, the $\operatorname{MSE}(\tilde{z}_T) = \operatorname{var}(z_T \mid Z_{(T)})$ in (9) is obtained when z_T is regarded as random following the structure in (2). Indeed, it can be verified from (4) that, for fixed z_T, the MSE of $\tilde{z}_T$ is

$$\begin{aligned} \operatorname{MSE}^*(\tilde{z}_T) &= \phi^2\sigma^2 + (1-\phi^2)^2 z_T^2, && T = 1 \text{ or } n \\ &= (1+\phi^2)^{-2}\{2\phi^2\sigma^2 + (1-\phi^2)^2 z_T^2\}, && 1 < T < n, \end{aligned} \tag{10}$$

so for some values of z_T, $\operatorname{MSE}^*(\tilde{z}_T)$ can be larger than $\operatorname{MSE}(\hat{z}_T)$.

The point of the foregoing discussion is to show that, in estimating missing values in time series, the method of maximum likelihood can lead to results very different from those obtained by optimal prediction under stationary assumptions. Except for the initial value at $t = 0$, we do not think, however, that it is appropriate to treat missing *observations* as fixed parameters. This seems almost a contradiction in terms. In time series analysis we believe that it is natural in most applications to regard the missing observations as random variables following the same probabilistic structure as the remaining ones and hence adopt the conditional expectation or posterior mean as their optimal estimator.

[*Received July 1989. Revised May 1990.*]

REFERENCES

Bayarri, M. J., DeGroot, M. H., and Kadane, J. B. (1986), "What Is the Likelihood Function?" in *Proceedings of the Fourth Purdue Symposium on Statistical Decision Theory and Related Topics*, eds. S. S. Gupta and J. Berger, New York: Springer-Verlag, pp. 3–27.

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.

Brubacher, S. R., and Tunnicliffe-Wilson, G. (1976), "Interpolating Time Series With Application to the Estimation of Holiday Effects on Electricity Demand," *Applied Statistics*, 25, 107–116.

Fuller, W. A. (1987), *Measurement Errors Models*, New York: John Wiley.

Peña, D. (1987), "Measuring the Importance of Outliers in ARIMA Models," in *New Perspectives in Theoretical and Applied Statistics*, eds. M. L. Puri, J. P. Vilaplana, and W. Wertz, New York: John Wiley, pp. 109–118.