Statistical Science
2011, Vol. 26, No. 1, 57–58
DOI: 10.1214/11-STS345D
Main article DOI: 10.1214/10-STS345

Discussion of "Feature Matching in Time Series Modeling" by Y. Xia and H. Tong

Qiwei Yao

Many congratulations to Professors Xia and Tong for another stimulating paper initiated from their own creative thinking. The base point of the proposed approach is the fact that most, if not all, statistical models are wrong. This not only applies to time series models, as a statistical model is, hopefully, a simplified representation of the truth. At the best it catches some features of the unknown underlying population. While the understanding of this nature is within the common wisdom, most statistical inference methods are confined to the framework which assumes that the true model is a member of the family of models concerned. The approach advocated in this paper acknowledges explicitly that the assumed model is not the truth, and indeed it is advantageous sometimes not to read too much into the assumed model. For example, the authors have articulated elegantly that if our interest lies in catching the linear dynamical structure, we should not use the (Gaussian) maximum likelihood estimation which effectively minimizes the one-step-ahead prediction errors only, and in fact a better fitted autocovariance is resulted from minimizing up to m-step-ahead predictions for $m > 1$.

Following the lead of the authors, it seems to make sense to take on board the concern for "wrong models" at the stage of the model selection, too, as hinted at the end of the paper. In the way, this has been actively researched in the context of model selection. However, a difference here is to use a different measure for "goodness of fit" instead of likelihood (or log-likelihood). Let us consider a simple case: fit a linear AR(p) model to observations $y_1, \ldots, y_n$ from a stationary time series with mean 0, where the order p is to be determined by the data, too. Let $\mathbf{y}_{t,p} = (y_t, y_{t-1}, \ldots, y_{t-p+1})'$. Based on an AR($p$) model (with independent innovations), the best predictor at the time t for a future value y_{t+m} should be a linear combination of the p components of $\mathbf{y}_{t,p}$. In fact the best linear predictor based on $\mathbf{y}_{t,p}$ is $\boldsymbol{\alpha}'_{m,p}\mathbf{y}_{t,p}$ with

$$\begin{aligned}\boldsymbol{\alpha}_{m,p} &= \boldsymbol{\Gamma}_p^{-1}\boldsymbol{\gamma}_{m,p}\\ &= \arg\min_{\boldsymbol{\gamma}} E\{(y_{t+m} - \boldsymbol{\alpha}'\mathbf{y}_{t,p})^2\},\end{aligned} \tag{1}$$

where $\boldsymbol{\Gamma}_p$ is a $p \times p$ matrix with $\gamma(j-i)$ as its (i,j)th element, $\boldsymbol{\gamma}_{m,p}$ is a $p \times 1$ vector with $\gamma(m+i-1)$ as its ith element, and $\gamma(\cdot)$ denotes the autocovariance function of y_t. In fact (1) holds for any stationary process. However, if we fit y_t with an AR(p), its autocovariance function $\gamma(\cdot)$ is then determined by $\boldsymbol{\theta}_p$—the parameters in an AR(p) model. Put $\boldsymbol{\alpha}_{m,p} = \boldsymbol{\alpha}_{m,p}(\boldsymbol{\theta}_p)$. Then $\mathbf{y}'_{t,p}\boldsymbol{\alpha}_{m,p}(\boldsymbol{\theta}_p)$ is the best predictor for y_{t+m} based on an AR(p) model. Using the "matching up-to-m-step-ahead point predictions" approach of Section 2.1, we estimate $\boldsymbol{\theta}_p$ (for p given) by

$$\widehat{\boldsymbol{\theta}}_p = \arg\min_{\boldsymbol{\theta}_p} Q_p(\boldsymbol{\theta}_p),$$

where

$$Q_p(\boldsymbol{\theta}_p) = \frac{1}{m}\sum_{k=1}^{m}\frac{1}{n-k-p+1} \cdot \sum_{t=p}^{n-k}\{y_{t+k} - \mathbf{y}'_{t,p}\boldsymbol{\alpha}_{k,p}(\boldsymbol{\theta}_p)\}^2.$$

However, we cannot choose p by minimizing $Q_p(\widehat{\boldsymbol{\theta}}_p)$, as $Q_p(\widehat{\boldsymbol{\theta}}_p)$ is likely to decrease as p increases.

To appreciate the difficulties involved, let us first consider the "ideal world" where the (true) distribution of $\{y_t\}$ is known. Then we should estimate $\boldsymbol{\theta}_p$ by

$$\widetilde{\boldsymbol{\theta}}_p = \arg\min_{\boldsymbol{\theta}_p} Q^*_p(\boldsymbol{\theta}_p),$$

where

$$Q^*_p(\boldsymbol{\theta}_p) = \frac{1}{m}\sum_{k=1}^{m} E[\{y_{t+k} - \mathbf{y}'_{t,p}\boldsymbol{\alpha}_{k,p}(\boldsymbol{\theta}_p)\}^2].$$

Unfortunately $Q^*_p(\widetilde{\boldsymbol{\theta}}_p)$ still decreases as p increases. The information (e.g., the variance) of the noise component of y_t is required in order to know when to stop. This is the standard problem in model selection even

Qiwei Yao is Professor, Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom (e-mail: q.yao@lse.ac.uk).