INFORMATICA, 1992, Vol 3, No 1, 64–71

HYPOTHESES TESTING IN MIXTURES OF TIME SERIES II

Jiří MICHÁLEK

Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
182 08 Prague 8, Pod vodárenskou veží 4, Czechoslovakia

Abstract. The paper defines the decomposition problem of a mixture of time series into homogeneous components. First part deals with a solution based on Bayesian approach in the case of independent observations, the other part is devoted to a solution of on-line decomposition for a time series consisting of weakly stationary components.

Key words: Bayesian test, Yule–Walker estimates, I-divergence rate, autoregressive model.

We so far have been interested in Part I in the question about the distribution of an observed sequence in two components only. But, in practice one can expect larger number of possible orbits, usually unknown for the observer in advance. Even if we knew the number of orbits before our experiment the off-line distribution problem into individual orbits would be more complicated than the two components problem. The situation is much more getting worse if we admit a stochastic dependence among observations. Under this situation the on-line distribution problem seems to be better solvable. In practice one can easier assume that all the components are mutually stochastically independent. The on-line problem can be described as follows: we have obtained the last observation and we have learnt, let us say, M possible orbits constructed already from the previous observations. Now, we have to decide to which of these M orbits the last observation belongs with high probability or to establish a quite new orbit. Let us imagine we have learnt M components $\{x_1^j, x_2^j, \ldots, x_{j_k}^j\}$, $j = 1, 2, \ldots, M$ on the basis of previous observations and we have the last observation x and the statistical problem is to decide to which component x belongs. For simplicity we shall assume all the observations are mutually independent and every j-th component is determined by a density function

$$f^j(\mathbf{x}) = \prod_{j=1}^{j_k} f_j(x_i), \quad j = 1, 2, \ldots, M.$$

Under this situation there exists an optimal Bayesian procedure already mentioned advising to accept the decision x belongs to the j_0-th orbit if

$$p_{j_0} f_{j_0}(x) > \max_{\substack{1 \leqslant i \leqslant M \\ i \neq j_0}} \{p_j f_j(x)\}.$$

As long as the same maximum value is achieved by two or more orbits one could choose among them randomly. This approach can be used by time series where observations can be mutually dependent if we, of course knew the evolution of every orbit described by conditional distributions functions. Then, by use of the previous Bayesian procedure we can choose that orbit that possesses the maximal conditional density function of x under the history of the whole orbit. Namely, if $f_j(x|x_1^j, \ldots, x_{k_j}^j)$ will be the conditional density function describing the evolution of the j-th orbit, which can be briefly expressed as $f(x|j)$ then under prior distribution $\{p_j\}_{j=1}^M$ the aposterior density function $f(j|x)$ is given by Bayesian formula

$$f(j|x) = \frac{p_j f(x|j)}{\sum_{i=1}^M p_i f(x|i)}.$$

Choosing the orbit that maximazes the aposterior density function we do precisely the same as choosing that orbit j_0 satisfying

$$p_{j_0} f(x|j_0) > \max_{\substack{1 \leqslant i \leqslant M \\ i \neq j_0}} \{p_i f(x|i)\}.$$

This algorithm can be used, e.g., in the case of several different autoregressive models with Gaussian variables, where the last observation x belongs to that orbit or autoregressive model M_j, $j = 1, 2, \dots, M$ that minimizes the distance

$$\frac{x - \hat{x}_j}{\sigma_j^2}.$$

Here $\hat{x}_j$ is the best prediction of x one step ahead in the model M_j, σ_j^2 is the error of this optimal prediction.

In practice our situation is not so clear because usually we don't possess any models M_j, $j = 1, 2, \dots, M$ and as the first step one must identify them, i.e., to estimate unknown parameters using those observations that form on the basis of experience one orbit. One cannot expect in practice as well the knowledge of conditional density functions describing the evolution of orbits. From these practical reason it is reasonable to look for other types of distances, which could measure similarity between orbits. One such a class of distances is derived from entropy and usually known as the class of convex statistical distances, for detail see Liese, Vajda (1987), e.g. We shall consider the case where orbits can be approximated by weakly stationary sequences with different mean values and spectral density functions. Let us know from the previous decisions that $(x_1^j, x_2^j, \dots, x_{k_j}^j)$ form the j-th component or orbit and let x be again the last observation. The basic idea is to approximate the j-th orbit by a suitable autoregressive model, where the order of an autoregressive series is chosen by means of the Akaike criterion, e.g., and coefficient estimates are obtained by the least squares method or by means of Yule-Walker estimates. The mean value is estimated in usual by the arithmetic mean. The task is then understood as follows: to find in a certain sense the most similar model from the models given before to the model derived from every orbit completed by x. As one of possible distances can serve the AIR-distance (asymptotic I-divergence rate) defined by

$$AIR(\varphi_2, m_2 | \varphi_1, m_1) = \frac{(m_1 - m_2)^2}{\sigma_1^2} + \frac{1}{2} \int_{-\pi}^{\pi} \left(\frac{\varphi_2(\lambda)}{\varphi_1(\lambda)} - \ln \frac{\varphi_2(\lambda)}{\varphi_1(\lambda)} - 1 \right) d\lambda$$

where $\sigma_1^2 = 2\pi \exp\left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \varphi_1(\lambda)\, d\lambda \right\}$. This distance is derived from I-divergence between two Gaussian stationary measures given by mean values m_1, m_2 and spectral density function $\varphi_1(\cdot)$, $\varphi_2(\cdot)$; for details see Vajda (1989). It is necessary notice the AIR-distance is not symmetric. If we wished a symmetric distance we could consider, e.g.,

$$\frac{1}{2}[AIR(\varphi_2, m_2|\varphi_1, m_1) + AIR(\varphi_1, m_1|\varphi_2, m_2)]$$
$$= \frac{1}{2}\frac{(m_1 - m_2)^2}{\sigma_1^2} + \frac{1}{2}\frac{(m_1 - m_2)^2}{\sigma_2^2} + \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{(\varphi_1(\lambda) - \varphi_2(\varphi))^2}{\varphi_1(\lambda)\varphi_2(\lambda)}\,d\lambda.$$

Although this "similarity measure" is derived of Gaussian random variables one can recommend it in a case of non-Gaussian variables, too. We can namely in such a case understand the approximation of an observed orbit given by an autoregressive model based on Yule-Walker estimators as the closest projection into the class of Gaussian autoregressive models; for deeper information see, e.g., Vajda (1989), Michálek (1990). Now, let $\{1, a_1^{(i)}, \ldots, a_{p_i}^{(i)}, \sigma_i^2, m_i\}$ be the coefficients of the approximating autoregressive model of the order p_i for the i-th orbit. These coefficients are obtained by minimaxing $AIR(\hat{\varphi}_i, m_i|\varphi_{a\sigma}, m)$ over all the autoregressive models of the order p_i determined by mean value m and the spectral density function

$$\varphi_{a\sigma}(\lambda) = \frac{1}{2\pi}\frac{\sigma^2}{\left|\sum_{j=0}^{p_i} a_j e^{ij\lambda}\right|^2}, \quad a_0 = 1.$$

The spectral density function $\hat{\varphi}_i(\cdot)$ is given by the covariance coefficients estimates

$$\hat{R}_k^{(i)} = \frac{1}{k_i}\sum_{j=1}^{k_i - k}(x_{j+k}^{(i)} - m_i)(x_j^{(i)} - m_i), \quad k = 0, 1, \ldots, p_i$$
$$m_i = \frac{1}{k_i}\sum_{j=1}^{k_i} x_j^{(i)},$$

i.e.,

$$\hat{\varphi}_i(\lambda) = \hat{R}_0^{(i)} + 2\sum_{j=1}^{p_i}\hat{R}_j^{(i)}\cos(j\lambda).$$

As proved in Michálek (1990) the solution of this minimization problem is given by solving the Yule-Walker equations. One must demand automatically $p_i \ll k_i$ for every $i = 1, 2, \ldots, M$.

Let us consider for every orbit $\{x_1^{(i)}, x_2^{(i)}, \ldots, x_{k_i}^{(i)}\}$ a sliding window of a fixed length ℓ, where $\ell < k_i$, $i = 1, 2, \ldots, M$ and simultaneously $\ell \gg p_i$, $i = 1, 2, \ldots, M$. Calculate for every orbit the estimates of the covariance coefficients up to the order p_i derived from the observations forming that sliding window including the last observation x, i.e.,

$$\hat{R}_k^{(i)}(x) = \frac{1}{\ell+1} \sum_{j=k_i+1-\ell}^{k_i+1-k} \left(x_j^{(i)} - \overline{x}^{(i)}(x)\right)\left(x_{j+k}^{(i)} - \overline{x}^{(i)}(x)\right)$$

$$k = 0, 1, \ldots, p_i, \quad i = 1, 2, \ldots, M,$$

$$x_{k_i+1}^{(i)} = x \quad \overline{x}^{(i)}(x) = \frac{1}{\ell+1} \sum_{j=k_i+1-\ell}^{k_i+1} x_j^{(i)}.$$

At this moment we shall evaluate the AIR-distance between the autoregressive approximation of the i-th orbit and the corresponding sliding window. It is possible to prove that this AIR-distance, e.g., see Michálek (1990) can be written in the form

$$\frac{1}{2}\frac{\left(\overline{x}^{(i)}(x) - \overline{x}^{(i)}\right)^2}{\sigma_i^2} + \frac{1}{4\pi}\frac{A_0^{(i)} R_0^{(i)}(x) + 2\sum_{i=1}^{p_i} A_j^{(i)} R_j^{(i)}(x)}{\sigma_i^2}$$
$$+ \frac{1}{2}\ln\sigma_i^2 - \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln\varphi_x^{(i)}(\lambda)\, d\lambda, \tag{1}$$

where $A_k^{(i)} = \sum_{j=0}^{p_i-k} a_j^{(i)} a_{j+k}^{(i)}$, $k = 0, 1, \ldots, p_i$,

$$\varphi_x^{(i)}(\lambda) = \sum_{j=-p_i}^{p_i} e^{-i\lambda j} \hat{R}_j^{(i)}(x).$$

The only problem is to calculate the integral standing in (1).

Since the covariance coefficients $\hat{R}_j^i(x)$, $j = 0, 1, \ldots, p_i$ are chosen so that the Toeplitz matrix $\{\hat{R}_{j-k}^{(i)}(x)\}_{j,k=0}^{p_i}$ is positive definite with probability one then $\varphi_x^{(i)}(\cdot)$ as a Fourier image must be a non-negative trigonometric polynomial, and hence

$$\hat{\varphi}_x^i(\lambda) = R_0^{(i)}(x) + 2\sum_{j=1}^{p_i} \hat{R}_j^{(i)}(x)\cos\lambda_j = \Big|\sum_{j=0}^{p_i} r_j(x)e^{ij\lambda}\Big|^2,$$

where we want $r_0^{(i)}(x) > 0$. The relation between $r_k^{(i)}(x)$ and $\hat{R}_k^{(i)}(x)$ is given by the formula

$$\hat{R}_k^{(i)}(x) = \sum_{k=0}^{p_i-k} r_{j+k}^{(i)}(x) r_j^{(i)}(x), \quad k = 0, 1, \ldots, p_i.$$

In order to solve this system of non-linear equations we may use the Cholesky algorithm for the decomposition of the Toeplitz matrix into the product of two triangular matrices. One of these matrices will be $\{r_{k-\ell}^{(i)}(x)\}_{k,\ell=0}^{p_i}$, $k \leqslant \ell$. Then

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln \hat{\varphi}_x^{(i)}(\lambda)\, d\lambda = \ln\left[r_0^{(i)}(x)\right]^2.$$

Thus, the decision problem assigning the last observation x to one of the orbits is solved by minimizing AIR-distances between the autoregressive approximation and the corresponding sliding window.

We can proceed solving this decision problem in the following manner, too. We have approximated every orbit by a suitable autoregressive model

$$x_{n+1} + \sum_{j=1}^{p_i} a_j^{(i)} x_{n-1-j} = \sigma_i e_{n+1}.$$

Using this model we can construct a best predictor one step ahead, i.e.,

$$\hat{x}_{n+1} = -\sum_{j=1}^{p_i} a_j^{(i)} x_{n+1-j} + \left(\sum_{j=1}^{p_i} a_j^{(i)} + 1\right)\overline{x},$$

where

$$\overline{x} = \frac{1}{k_i}\sum_{j=1}^{k_i} x_{n+1-j},$$

and then we can compare this prediction derived from the observations forming the i-th orbit with the last observation x, i.e.,

$$\left|x - \hat{x}^{(i)}_{k_i+1}\right|.$$

As long as we can avoid some troubles with calculating the coefficients $r_k^{(i)}$, $k = 0, 1, \ldots, p_i$ we use the symmetric version of AIR-distance which in this case has a form

$$\frac{\left|\overline{x}^{(i)} - x^{(i)}(x)\right|^2}{\sigma_i^2} + \frac{\left|\overline{x}^{(i)} - x^{(i)}(x)\right|^2}{\sigma_i^2(x)} + \frac{A_0^{(i)}\hat{R}_0^{(i)}(x) + 2\sum_{j=1}^{p_i} A_j^{(i)}\hat{R}_j^{(i)}(x)}{\sigma_i^2}$$
$$+ \frac{A_0^{(i)}\hat{R}_0^{(i)}(x) + 2\sum_{j=1}^{p_i} A_j^{(i)}\hat{R}_j^{(i)}(x)}{\sigma_i^2(x)},$$

where $\sigma_i^2(x) = \hat{R}_0^{(i)}(x) + \sum_{j=1}^{p_i} \hat{R}_j^{(i)}(x) a_j^{(i)}(x)$ and

$$A_k^{(i)}(x) = \sum_{j=0}^{p_i-k} a_{j+k}^{(i)}(x) a_j^{(i)}(x), \quad k = 0, 1, \ldots, p_i.$$

The coefficients $a_j^{(i)}(x)$ with $a_0^{(i)}(x) = 1$ are determined on the basis of $\hat{R}_j^{(i)}(x)$, $j = 0, 1, \ldots, p_i$ by solving the Yule-Walker equations.

We explained in detail a method assigning the last observation x into the most similar subseries. This procedure can be carried out not only with the last observation x but with a group of observations bearing the latest informatio of evolution of an observed time series.

REFERENCES

Basseville, M., and A.Benveniste (1986). *Detection of Abrupt Changes in Signals and Dynamical Systems. Lectures Notes in Control and Information Sciences* 77. Springer Verlag, Berlin, Heidelberg.

Basseville, M. (1988). Detecting changes in signals and systems. *Automatica*, **24**(3), 309–326.

Hoel, P., and J.Peterson (1949). A solution to the problem of optimal classification. *Ann. Math. Stat.*, **20**, 433–438.

Kligienė ,N, and L.Telksnys (1983). Methods of detecting instants of change of random process properties. *Autom. Remote Control,* **44**, 1241–1283.

Liese, F., and I.Vajda (1987). *Convex Statistical Distances.* Teubner – Texte zur Mathematik, Band 95, Leipzig.

Michálek, J. (1990). Yule–Walker estimates and esymptotic *I*-divergence rate. *Problems of Control and Information Theory,* **19**(5–6), 387–398.

Page, E. (1954). Continuous inspection schemes. *Biometrika,* **41**, 100–115.

Vajda, I. (1989). *Theory of Statistical Inference and I-Information.* Kluwer Academic Publishers, Dordrecht, Boston, London.

Vajda, I. (1990). Generalization of discrimination-rate theorems of Chernoff and Stein. *Kybernetika,* **26**(4), 273–288.

Willsky, A.S. (1976). A survey of design methods for failure detection in dynamic systems. *Automatica,* **12**, 601–611.

Received December 1991

J. Michálek graduated from the Faculty of Mathematics and Physics of the Charles University of Prague in 1966. In 1968 he joined the Department of Information Theory in the Institute of Information Theory and Automation of Czechoslovak Academy of Sciences where he has been working up to nowadays. He received the Degree of Candidate of Mathematical Sciences in 1975. His professional interests are concentrated on statistical problems in random processes.