ESTIMASI PARAMETER MODEL *MIXTURE AUTOREGRESSIVE (MAR)* MENGGUNAKAN ALGORITMA EKSPEKTASI MAKSIMISASI (EM)

Mika Asrini[1] , Winita Sulandari[2], Santoso Budi Wiyono[3]
[1]Mahasiswa Jurusan Matematika FMIPA UNS
[2]Staf Pengajar Jurusan Matematika FMIPA UNS
[3]Staf Pengajar Jurusan Matematika FMIPA UNS

Abstract

Mixture autoregressive (MAR) Model is a mixture of K Gaussian autoregressive (AR) components. The mixture model is capable for modelling of nonlinear time series with multimodal conditional distributions. This paper discusses about the parameters estimation using EM algorithm. All possible models are then applied to national maize production data. In this case, the BIC is used for the MAR model selection.

Keywords : Mixture Autoregressive, EM Algorithm, BIC, Maize Production

1. Pendahuluan

Peramalan runtun waktu merupakan suatu peramalan yang didasarkan pada data masa lalu dalam variabel yang sama. Dalam hal ini, serangkaian data masa lalu dikumpulkan, dan dianalisis untuk membangun suatu model yang dapat mendeskripsikan hubungan antara data yang berurutan. Model yang terbentuk selanjutnya digunakan untuk ekstrapolasi data yang akan datang. Model peramalan yang paling populer dan relatif sering digunakan adalah model *ARIMA* (*autoregressive integrated moving average*). Model yang dipopulerkan oleh Box and Jenkins (1970) ini, mampu mewakili beberapa model runtun waktu yang lain, seperti *AR* (*autoregressive*), *MA* (*moving average*), *ARMA*, bahkan *exponential smoothing*[4]. Model *ARIMA* disajikan sebagai fungsi linear dari beberapa data masa lalu dan eror random. Oleh karena itu, model ini tidak mampu menangkap pola nonlinear dalam data. Sementara itu, tidak semua data runtun waktu bersifat linier terhadap nilai pengamatan yang lalu.

Wong and Li (2000) memperkenalkan model *mixture autoregressive* (*MAR*) yang merupakan model runtun waktu nonlinier. Model *MAR* merupakan gabungan dari beberapa komponen Gaussian *autoregressive* (*AR*). Kelebihan dari model ini yaitu pada kemampuannya dalam mengatasi sifat kemiringan data, leptokurtik, platikurtik, serta multimodal.

Menurut Wei (2006) metode yang dapat digunakan untuk estimasi parameter model *AR* adalah metode maksimum *likelihood*, karena praktis digunakan untuk mendapatkan nilai parameter yang tak bias dan bervariansi minimum. Meskipun model Model *MAR* merupakan model gabungan *AR*, namun parameternya tidak bisa diestimasi menggunakan maksimum *likelihood* secara langsung. Menurut Wong and Li (2000) parameter model *MAR* dapat diestimasi menggunakan algoritma ekspektasi maksimisasi (EM). Algoritma yang diperkenalkan oleh Dempster, Laird, and Rubin(1977) ini digunakan untuk menentukan nilai estimasi maksimum *likelihood* data gabungan. Ada dua tahap yang harus dilakukan dalam algoritma EM yaitu tahap ekspektasi dan tahap maksimisasi.

Artikel ini menjelaskan kembali mengenai estimasi parameter model *MAR* menggunakan algoritma EM dan selanjutnya diterapkan pada data produksi jagung nasional. Data ini dipilih karena memiliki pola nonlinier dan multimodal.

2. Model *Mixture Autoregressive (MAR)*

Model *mixture autoregressive* (*MAR*) dengan K komponen dinotasikan dengan $MAR(K;\ p_1,\ p_2,\ldots,p_K)$ dan didefinisikan sebagai

$$F(y_t|\mathcal{F}_{t-1}) = \sum_{k=1}^{K} \alpha_k \Phi\left(\frac{y_t - \phi_{k0} - \phi_{k1}y_{t-1} - \ \ldots\ - \phi_{kp_k}y_{t-p_k}}{\sigma_k}\right)$$

untuk $\sum_{k=1}^{K} \alpha_K = 1\,, \alpha_k > 0$ (1)

dengan:

$F(y_t|\mathcal{F}_{t-1})$: fungsi distribusi kumulatif y_t jika $y_{t-1}, \ldots, y_{t-p_k}$ diketahui
$\mathcal{F}_{t-1}$: informasi pada waktu $t-1$
$\phi(.)$: fungsi densitas probabilitas distribusi normal standar
$\Phi(.)$: distribusi kumulatif normal standar
σ_k : deviasi standar masing-masing komponen ke$-k$
α_k : proporsi masing-masing komponen gabungan ke$-k$
p_k : orde *AR* komponen ke – *k*.

3. Estimasi Parameter Model *MAR*

Langkah pertama yang harus dilakukan untuk mengestimasi parameter dalam model *MAR* adalah menentukan fungsi kepadatan probabilistik model *MAR*. Misalkan $Y = (y_1, y_2, \ldots, y_n)$ adalah data terobservasi yang dibangkitkan dari model MAR pada persamaan (2.1), sedangkan $Z = (z_1, z_2, \ldots, z_n)$ adalah variabel random tidak terobeservasi dengan Z_t adalah vektor berdimensi *K* dengan ketentuan sebagai berikut

$$Z_{t,k} = \begin{cases} 1, & \text{jika } y_t \text{ berasal dari komponen ke} - k \\ 0, & \text{jika } y_t \text{ bukan berasal dari komponen ke} - k. \end{cases}$$

Jika $X = (Y, Z)$ merupakan data lengkap, fungsi kepadatan probabilistik model *MAR* untuk data lengkap yaitu

$$f(X) = \sum_{k=1}^{K} \left[Z_{t,k} \frac{\alpha_k}{\sqrt{2\pi}\sigma_k} \mathrm{e}^{-\frac{1}{2}\left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}{\sigma_k}\right)^2} \right]. \quad (2)$$

Langkah selanjutnya adalah membentuk fungsi *likelihood* dari persamaan (2). Jika $X_1, X_2, \ldots, X_t$ adalah variabel random dengan fungsi kepadatan peluang $f(X_t; \boldsymbol{\theta})$ dengan $\boldsymbol{\theta} = (\alpha_k, \sigma_k, \phi_{k0}, \phi_{ki})$, maka fungsi *likelihood* data lengkap adalah

$$L(X_1, \ldots, X_t; \boldsymbol{\theta}) = \prod_{t=p+1}^{n} f(X; \alpha_k, \sigma_k, \phi_{k0}, \phi_{ki})$$

$$= \prod_{t=p+1}^{n} \sum_{k=1}^{K} \left[Z_{t,k} \frac{\alpha_k}{\sqrt{2\pi}\sigma_k} \mathrm{e}^{-\frac{1}{2}\left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}{\sigma_k}\right)^2} \right],$$

sehingga diperoleh fungsi log-*likelihood*

$$\log L(X, \boldsymbol{\theta}) = \sum_{t=p+1}^{n} \left[\sum_{k=1}^{K} Z_{t,k} \log(\alpha_k) - \sum_{k=1}^{K} Z_{t,k} \log(\sigma_k) - \sum_{k=1}^{K} \frac{Z_{t,k}}{2} \left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}{\sigma_k} \right)^2 \right] \quad (3)$$

Turunan pertama (3) terhadap $\boldsymbol{\theta}$ dijelaskan sebagai berikut.

a. Penurunan (3) terhadap parameter α_k

$$\frac{\partial L(X,\boldsymbol{\theta})}{\partial \alpha_k} = \frac{\partial \sum_{t=p+1}^{n}\left[\sum_{k=1}^{K} Z_{t,k}\log(\alpha_k)\right]}{\partial \alpha_k}$$

Oleh karena $\sum_{k=1}^{K}\alpha_k = 1$ maka $\alpha_K = 1 - \sum_{k=0}^{K-1}\alpha_k$, sehingga diperoleh

$$\frac{\partial L(X,\boldsymbol{\theta})}{\partial \alpha_k} = \frac{\partial \sum_{t=p+1}^{n}\left(\sum_{k=1}^{K-1} Z_{t,k}\log(\alpha_k) + Z_{t,K}\log(\alpha_K)\right)}{\partial \alpha_k}$$

$$= \frac{\sum_{t=p+1}^{n} Z_{t,k}}{\alpha_k} - \frac{\sum_{t=p+1}^{n} Z_{t,K}}{\alpha_K}\ ; k = 1,\dots,K-1 \tag{4}$$

b. Penurunan (3) terhadap ϕ_{k0}

$$\frac{\partial L(X,\boldsymbol{\theta})}{\partial \phi_{k0}} = -\frac{\partial \sum_{t=p+1}^{n}\left[\sum_{k=1}^{K}\frac{Z_{t,k}}{2}\left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k}\phi_{ki}y_{t-i}}{\sigma_k}\right)^2\right]}{\partial \phi_{k0}}$$

$$= -\frac{\sum_{t=p+1}^{n}\left[Z_{t,k}2(y_t - \phi_{k0} - \sum_{i=1}^{p_k}\phi_{ki}y_{t-i})(-1)\right]}{2\,(\sigma_k)^2}$$

$$= \sum_{t=p+1}^{n} Z_{t,k}\,y_t - \sum_{t=p+1}^{n} Z_{t,k}\phi_{k0} - \sum_{t=p+1}^{n}\sum_{i=1}^{p_k}\phi_{ki}y_{t-i}$$

dengan $k = 1,\dots,K$ (5)

c. Penurunan (3) terhadap ϕ_{ki}

$$\frac{\partial L(X,\boldsymbol{\theta})}{\partial \phi_{ki}} = -\frac{\partial \sum_{t=p+1}^{n}\left[\sum_{k=1}^{K}\frac{Z_{t,k}}{2}\left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k}\phi_{ki}y_{t-i}}{\sigma_k}\right)^2\right]}{\partial \phi_{ki}}$$

$$= \frac{-\sum_{t=p+1}^{n} Z_{t,k}\,2\left(y_t - \phi_{k0} - \sum_{i=1}^{p_k}\phi_{ki}y_{t-i}\right)(-y_{t-i})}{2\,(\sigma_k)^2}$$

$$= \sum_{t=p+1}^{n} Z_{t,k}\,y_t y_{t-i} - \sum_{t=p+1}^{n} Z_{t,k}\,\phi_{k0}y_{t-i} - \sum_{t=p+1}^{n} Z_{t,k}\,y_{t-i}\sum_{i=1}^{p_k}\phi_{ki}y_{t-i}$$

dengan $k = 1,\dots,K$ dan $i = 1,\dots,p_k$ (6)

d. Penurunan (3) terhadap σ_k

$$\frac{\partial L(X,\boldsymbol{\theta})}{\partial \sigma_k} = -\frac{\partial \sum_{t=p+1}^{n}\left[\sum_{k=1}^{K} Z_{t,k}\log(\sigma_k)\right]}{\partial \sigma_k} - \frac{\partial \sum_{t=p+1}^{n}\left[\sum_{k=1}^{K}\frac{Z_{t,k}}{2}\left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k}\phi_{ki}y_{t-i}}{\sigma_k}\right)^2\right]}{\partial \sigma_k}$$

$$= -\frac{\sum_{t=p+1}^{n} Z_{t,k}}{\sigma_k} + \frac{\sum_{t=p+1}^{n} Z_{t,k}\left(y_t - \phi_{k0} - \sum_{i=1}^{p_k}\phi_{ki}y_{t-i}\right)^2}{\sigma_k^3}$$

dengan $k = 1,\dots,K$. (7)

Algoritma EM digunakan untuk mengestimasi parameter $\boldsymbol{\theta}$ dengan cara memaksimalkan fungsi log-*likelihood* (3) melalui tahap ekspektasi dan tahap maksimisasi. Berikut ini adalah prosedur dalam EM.

a. Tahap ekspektasi. Misal $\boldsymbol{\theta}$ diketahui. Ekspektasi bersyarat komponen ke–k dari Z_t merupakan probabilitas bersyarat bahwa y_t berasal dari observasi berasal dari komponen ke-k dari distribusi gabungan, bersyarat pada $\boldsymbol{\theta}$ dan Y. Misalkan $\tau_{t,k}$ menyatakan ekspektasi bersyarat komponen ke-k dari Z_t, persamaan $\tau_{t,k}$ adalah

$$\tau_{t,k} = \frac{f(y_t,\theta)}{\sum_{k=1}^{K} f(y_t,\theta)} = \frac{\frac{\alpha_k}{\sigma_k} \mathrm{e}^{-\frac{1}{2}\left(\frac{y_t-\phi_{k0}-\sum_{i=1}^{p_k}\phi_{ki}y_{t-i}}{\sigma_k}\right)^2}}{\sum_{k=1}^{K}\frac{\alpha_k}{\sigma_k}\mathrm{e}^{-\frac{1}{2}\left(\frac{y_t-\phi_{k0}-\sum_{i=1}^{p_k}\phi_{ki}y_{t-i}}{\sigma_k}\right)^2}}, \qquad k = 1,\dots,K.$$

b. Tahap maksimisasi. Nilai estimasi parameter $\hat{\theta} = \left(\hat{\alpha}_k, \hat{\sigma}_k, \hat{\phi}_{k0}, \hat{\phi}_{ki}\right)$ dapat ditentukan dengan cara memaksimalkan fungsi log-*likelihood* (3), dengan cara menyamadengankan persamaan (4) – (7) dengan nilai 0. Berdasarkan persamaan (4), diperoleh

$$\frac{\sum_{t=p+1}^{n} Z_{t,k}}{\alpha_k} - \frac{\sum_{t=p+1}^{n} Z_{t,K}}{\alpha_K} = 0$$

$$\hat{\alpha}_k = \frac{\sum_{t=p+1}^{n} Z_{t,k}}{\frac{\sum_{t=p+1}^{n} Z_{t,K}}{\alpha_K}} = \frac{\sum_{t=p+1}^{n} Z_{t,k}}{(n-p)\,\alpha_K/\alpha_K} = \frac{\sum_{t=p+1}^{n} Z_{t,k}}{(n-p)}, \qquad k = 1,\dots,K$$

dan dari persamaan (5) dan (6) diperoleh

$$\sum_{t=p+1}^{n} Z_{t,k}\, y_t - \sum_{t=p+1}^{n} Z_{t,k}\phi_{k0} - \sum_{t=p+1}^{n}\sum_{i=1}^{p_k}\phi_{ki}y_{t-i} = 0$$

$$\hat{\phi}_{k0}\sum_{t=p+1}^{n}\tau_{t,k} = \sum_{t=p+1}^{n}\tau_{t,k}\, y_t - \sum_{t=p+1}^{n}\sum_{i=1}^{p_k}\hat{\phi}_{ki}y_{t-i} \tag{8}$$

dan

$$\sum_{t=p+1}^{n} Z_{t,k}\, y_t y_{t-i} - \sum_{t=p+1}^{n} Z_{t,k}\,\phi_{k0}y_{t-i} - \sum_{t=p+1}^{n} Z_{t,k}\, y_{t-i}\sum_{i=1}^{p_k}\phi_{ki}y_{t-i} = 0$$

$$\sum_{t=p+1}^{n}\tau_{t,k}y_{t-i}y_{t-i}\sum_{i=1}^{p_k}\hat{\phi}_{ki} = \sum_{t=p+1}^{n}\tau_{t,k}\, y_t y_{t-i} - \sum_{t=p+1}^{n}\tau_{t,k}\,\hat{\phi}_{k0}y_{t-i}. \tag{9}$$

Berdasarkan persamaan (8) dan (9) diperoleh

$$\mathrm{A}_k\hat{\phi}_{ki} = b_k$$

dengan

$$\mathrm{A}_k = \begin{bmatrix} \sum_{t=p+1}^{n}\tau_{t,k} & \cdots & \sum_{t=p+1}^{n}\tau_{t,k}y_{t-p_k} \\ \vdots & \ddots & \vdots \\ \sum_{t=p+1}^{n}\tau_{t,k}\, y_{t-p_k} & \cdots & \sum_{t=p+1}^{n}\tau_{t,k}y_{t-p_k}y_{t-p_k} \end{bmatrix},\ \hat{\phi}_{ki}^{T} = \left[\hat{\phi}_{k0},\dots,\hat{\phi}_{kp_k}\right]$$

dan

$$b_k^{T} = \left[\sum_{t=p+1}^{n}\tau_{t,k}\, y_t,\dots,\sum_{t=p+1}^{n}\tau_{t,k}y_t y_{t-p_k}\right].$$

Selanjutnya $\hat{\sigma}_k$ dapat ditentukan berdasarkan (7), yaitu

$$-\frac{\sum_{t=p+1}^{n} Z_{t,k}}{\sigma_k}+\frac{\sum_{t=p+1}^{n} Z_{t,k}\left(y_t-\phi_{k0}-\sum_{i=1}^{p_k}\phi_{ki}y_{t-i}\right)^2}{\sigma_k^3}=0$$

sehingga diperoleh

$$\hat{\sigma}_k=\left(\frac{\sum_{t=p+1}^{n}\tau_{t,k}\left(y_t-\hat{\phi}_{k0}-\sum_{i=1}^{p_k}\hat{\phi}_{ki}y_{t-i}\right)^2}{\sum_{t=p+1}^{n}\tau_{t,k}}\right)^{1/2}.$$

Parameter $\boldsymbol{\theta}$ dapat diestimasi dengan cara mengulangi kedua tahap di atas hingga diperoleh nilai yang konvergen.

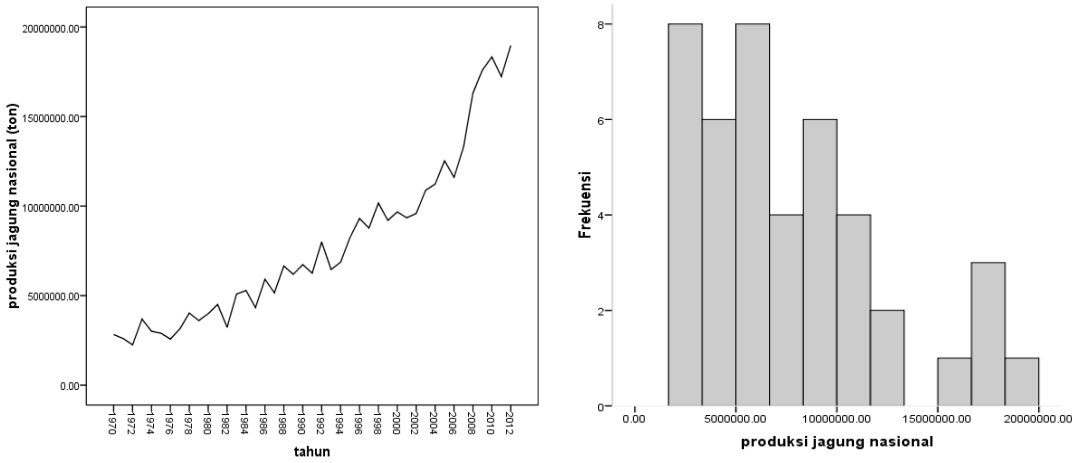
Dalam artikel ini, kriteria informasi yang digunakan untuk menentukan model *MAR* yang paling sesuai adalah *Bayes Information Criterion* (*BIC*). Penghitungan nilai *BIC* mengacu pada Schwarz (1978), yaitu

$$BIC=-2\log L(X;\theta)+\log(n-p_{max})\left(3K-1+\sum_{k=1}^{K}p_k\right)$$

dengan p_{max} adalah orde *AR* maksimal dari keseluruhan K komponen.

4. Contoh Kasus

Model *MAR* diterapkan pada produksi jagung nasional tahun 1970-2012 (dalam ton). Data diambil dari Basis Data Statistik Pertanian, Kementerian Pertanian Republik Indonesia untuk tahun 1970 – 2009 dan Badan Pusat Statistik untuk tahun 2010 - 2012.



Gambar 1. Grafik (kiri) dan histogram (kanan) data produksi jagung nasional 1970-2000

Berdasarkan grafik dan histogram pada Gambar 1 dan diperkuat dengan uji linearitas Harvey-Collier (Kraemer and Sonnberger, 1986) menggunakan *software* R dapat diambil kesimpulan bahwa data produksi jagung memiliki pola nonlinear, nonstasioner, dan multimodal. Dengan demikian, model *MAR* yang mungkin adalah *MAR* (2;1,1) dan *MAR* (3;1,1,1) tanpa konstanta untuk data terdiferensiasi 1.

Model *MAR* (2;1,1) dengan $\phi_{k0}=0\,(k=1,2)$ memberikan nilai *BIC* 591,9099 dan model *MAR* (3;1,1,1) dengan $\phi_{k0=}0(k=1,2,3)$ memberikan nilai *BIC* 594,5290, sehingga model *MAR* (2;1,1) dianggap sebagai model yang lebih sesuai. Berdasarkan hasil perhitungan menggunakan *software Ms Excel* diperoleh nilai estimasi parameter $(\alpha_1;\alpha_2;\phi_{11};\phi_{21};\sigma_1;\sigma_2)$ adalah

$$(0{,}5094;0{,}4906;-0{,}1041;-0{,}6784;1232160{,}2876;762096{,}0832).$$

Nilai ini merupakan estimasi parameter untuk data terdiferensiasi 1, untuk itu perlu dilakukan transformasi kembali ke dalam data awal (semula) yaitu

$$F(y_t|\mathcal{F}_{t-1}) = 0{,}5094\Phi\left(\frac{(y_t - y_{t-1}) - (-0{,}1041)(y_{t-1} - y_{t-2})}{1232160{,}2876}\right) + 0{,}4906\Phi\left(\frac{(y_t - y_{t-1}) - (-0{,}6784)(y_{t-1} - y_{t-2})}{762096{,}0832}\right)$$

sehingga diperoleh model *MAR* untuk dua komponen adalah

$$F(y_t|\mathcal{F}_{t-1}) = 0{,}5094\Phi\left(\frac{y_t - 0{,}8959y_{t-1} - 0{,}1041y_{t-2})}{1232160{,}2876}\right) + 0{,}4906\Phi\left(\frac{y_t - 0{,}3216y_{t-1} - 0{,}6784y_{t-2})}{762096{,}0832}\right)$$

Dengan demikian nilai peramalan produksi jagung nasional satu periode ke depan, yaitu untuk tahun 2013 pada interval kepercayaan 95% adalah antara 17994193,88 hingga 18592912,94 ton.

5. Kesimpulan

Model *MAR* dapat digunakan untuk memodelkan data runtun waktu yang nonlinier dan multimodal. Oleh karena model *MAR* adalah model gabungan, maka estimasi parameter dilakukan dengan menggunakan algoritma EM. Proses estimasi parameter model *MAR* diawali dengan identifikasi model dan inisiasi parameter. Identifikasi model merupakan tahap penentuan komponen dan orde pada setiap komponennya. Sedangkan inisiasi parameter adalah penentuan nilai awal parameter secara sembarang. Selanjutnya ditentukan fungsi log-*likelihood* dari model gabungan yang terdiri dari data terobservasi dan data hilang. Misalkan $Y = (y_1, y_2, \dots, y_n)$ adalah data terobservasi, $Z = (z_1, z_2, \dots, z_n)$ merupakan data hilang, dan $X = (Y, Z)$ merupakan data lengkap. Nilai parameter dapat dilakukan dengan maksimum log-*likelihood,* untuk mendapatkan nilai yang maksimum digunakan algoritma EM yang memiliki dua tahap inti yaitu tahap ekspektasi dan maksimisasi

DAFTAR PUSTAKA

1. Box, G., and Jenkins, G., *Time Series Analysis: Forecasting and Control*, Holden Day, San Francisco, 1994.
2. Dempster, A. P., Laird, N. M. and Rubin, D. B., Maximum Likelihood from Incomplete Data via the EM Algorithm, *J. Royal Statistical Society, Series B*, 1977, Vol. 39, No. 1: 1-38.
3. Kraemer, W. and Sonnberger, H. S., *The Linear Model Regression Under Test*, Physica-Verlag Heideberg Wien, 1986.
4. McKenzie, E. D., General Exponential Smoothing and The Equivalent ARMA process, *Journal of Forecasting*, 1984, Vol. 3: 333–344.
5. Schwarz, G., Estimating The Dimension of a Model, *Ann. Statist*., 1978, Vol. 6, No. 2: 461-464.
6. Wei, W. W. S., *Time Series Analysis: Univariate and Multivariate Methods*, Pearson Addison Wesley, 2006.
7. Wong, C. S and Li, W. K., On Mixture Autoregressive Model, *J. Royal Statistical Society, Series B*, 2000, Vol. 62, No. 1, 95-115.