4F7 Spectrum Estimation
Maximum Likelihood for ARMA model estimation

Sumeetpal Singh
Email : sss40@eng.cam.ac.uk

Modification history:
$\leq$ 2008: S. Godsill, P. Rayner
Revised by S.S. Singh

1 Maximum likelihood

- First a simplified example: you are given n independent samples z_i, $1 \leq i \leq n$, from a Normal distribution with mean μ and variance σ^2

- The *likelihood* of (μ, σ) or probability density of the observed data given (μ, σ) is

$$p(z_1, ..., z_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z_i - \mu)^2}{2\sigma^2}\right)$$

- Estimate (μ, σ^2) by maximising $\log p(z_1, ..., z_n)$ w.r.t. (μ, σ^2)

- The ARMA(P,Q) model is

$$x_n = \sum_{p=1}^{P} a_p\, x_{n-p} + \sum_{q=0}^{Q} b_q\, w_{n-q}$$

  Assume random variables w_n are i.i.d. Gaussian with mean zero and variance σ^2

- Given data $x_0, ..., x_{N-1}$ the model parameter estimates $\widehat{a}_i$, $\widehat{b}_i$, and $\widehat{\sigma^2}$ are

$$\arg \max_{\substack{a_1,...,a_P \\ b_0,...,b_Q \\ \sigma^2}} p(x_0, ..., x_{N-1})$$

- As $N \to \infty$ the estimates converge to the true values

- The difficulty is searching for the global maximizer

- Also, for the ARMA model the data is statistically dependent and the likelihood is more difficult to calculate

- We will use the probability chain rule for a collection of dependent random variables $z_1,\ z_2,\ ...,\ z_n$:

$$p(z_1,\ ...,\ z_n) = p(z_1) \prod_{i=2}^{n} p(z_i | z_1, ..., z_{i-1})$$

2 Maximum likelihood for AR(P)

- The AR(P) model is

$$x_n = \sum_{p=1}^{P} a_p \, x_{n-p} + w_n$$

  where w_n are i.i.d. Gaussian with mean zero and variance σ^2

- The probability chain rule applied to $p(x_P, ..., x_{N-1} | x_0, \ldots, x_{P-1})$

$$\prod_{i=P}^{N-1} p(x_i | x_0, ..., x_{i-1}) = \prod_{i=P}^{N-1} p(x_i | x_{i-P}, ..., x_{i-1})$$

- and $p(x_i | x_{i-P}, ..., x_{i-1})$ is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left(x_i - a_1 \, x_{i-1} - \cdots - a_P x_{i-P}\right)^2 \right)$$

- Let $e_i = x_i - a_1 \, x_{i-1} - \cdots - a_P x_{i-P}$. Thus

$\log p(x_P, ..., x_{N-1}|x_0, \ldots, x_{P-1})$ is

$$-0.5(N-P)\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=P}^{N-1} e_i^2$$

- To avoid having to compute $p(x_0, \ldots, x_{P-1})$ maximise $p(x_P, ..., x_{N-1}|x_0, \ldots, x_{P-1})$ instead
- This instance of Maximum likelihood is equivalent to least squares for the AR model
- First minimize $\sum_{i=P}^{N-1} e_i^2$ w.r.t. $(a_1, \ldots, a_P)$ to get $(a_1^*, \ldots, a_P^*)$
- Let $\mathcal{E} = \sum_{i=P}^{N-1} e_i^2$ evaluated at $(a_1^*, \ldots, a_P^*)$
- Now maximise this log-likelihood with respect to σ^2 by differentiating:

$$\frac{d}{d\sigma^2}\log p(x_P, ..., x_{N-1}|x_0, \ldots, x_{P-1}) = \frac{-0.5}{\sigma^2}(N-P) + \frac{0.5}{(\sigma^2)^2}\mathcal{E}$$

and hence at the maximising σ is

$$\sigma^* = \sqrt{\frac{\mathcal{E}}{N-P}}$$

which is an intuitive result.

- AR models are by far the simpler to estimate
- ARMA process may be well approximated by an AR process with 'sufficiently' large P. Hence practitioners very often work with large AR models, even when an ARMA structure is suspected
- To compute $p(x_0, \ldots, x_{P-1})$ write the AR(P) model in state-space form (see Examples paper)

$$\begin{bmatrix} x_n \\ \vdots \\ x_{n-P+1} \end{bmatrix} = \mathbf{\Lambda} \begin{bmatrix} x_{n-1} \\ \vdots \\ x_{n-P} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ \vdots \end{bmatrix} w_n \quad (1)$$

- When the model is stationary, $p(x_{n-P+1}, \ldots, x_n)$ is a Gaussian density with zero mean and covariance matrix $\mathbf{R}$ for any n. Computing the variance of the left and right-hand-side of (1) we get

$$\mathbf{R} = \mathbf{\Lambda R \Lambda}^{\mathrm{T}} + \sigma^2 \mathbf{b}\mathbf{b}^{\mathrm{T}} \quad (2)$$

where $\mathbf{b} = [1, 0, \cdots, 0]^{\mathrm{T}}$

- Let $r_{i,j} = [\mathbf{R}]_{i,j}$ then

$$r_{i,j} = \sum_{k=1}^{P} \sum_{l=1}^{P} \lambda_{i,k} r_{k,l} \lambda_{j,l}$$

$$r_{1,1} = \sigma^2 + \sum_{k=1}^{P} \sum_{l=1}^{P} \lambda_{1,k} r_{k,l} \lambda_{1,l}$$

where $\lambda_{i,j} = [\mathbf{\Lambda}]_{i,j}$

- For example, for an AR(2) model

$$\mathbf{\Lambda} = \begin{bmatrix} a_1 & a_2 \\ 1 & 0 \end{bmatrix}$$

$$\begin{aligned} r_{1,2} &= a_1 r_{1,1} + a_2 r_{2,1} \\ r_{2,1} &= a_1 r_{1,1} + a_2 r_{1,2} \\ r_{2,2} &= r_{1,1} \\ r_{1,1} &= \sigma^2 + a_1^2 r_{1,1} + a_1 a_2 (r_{1,2} + r_{2,1}) + a_2^2 r_{2,2} \end{aligned}$$

which gives

$$r_{1,1} = (1 - a_1^2 - \frac{2a_1^2 a_2}{1 - a_2} - a_2^2)^{-1} \sigma^2$$

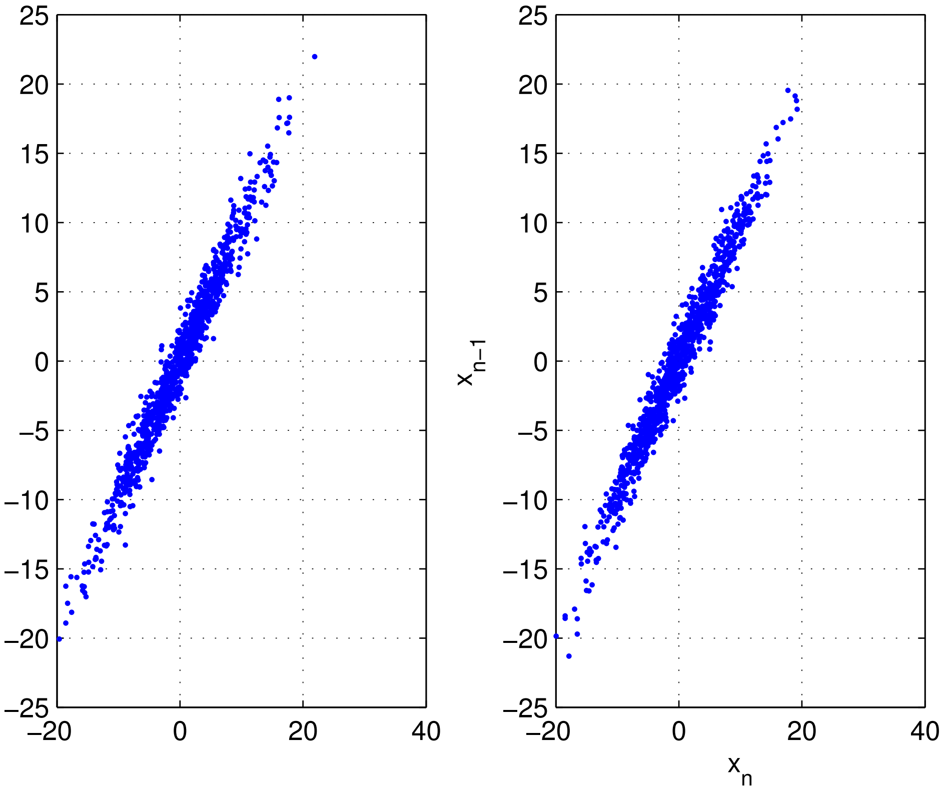
$$r_{1,2} = r_{2,1} = \frac{a_1}{1 - a_2} r_{1,1}$$

- Check: an AR(2) model with roots 0.9 and 0.7 will have transfer function

  $$1-a_1 z^{-1}-a_2 z^{-1} = \left(1-0.9z^{-1}\right)\left(1-0.7z^{-1}\right)$$

  which implies $a_1 = 1.6$, $a_2 = -0.63$. For $\sigma^2 = 1$, $r_{1,1} =$ 45.4634, $r_{1,2} = 44.6267$ and (2) will be satisfied

- To confirm the analysis, shown in the figure below are plots of samples from a Gaussian distribution with mean 0 and variance $[45.4634 \quad 44.6267; 44.6267 \quad 45.4634]$ (left-hand-side) and the plot of 1000 samples from the AR(2) model (1) for these same values of a_1, a_2 and σ^2 (each dot represents a value of (x_n, x_{n-1}))



3 Maximum likelihood for ARMA(P,Q)

- Special case: consider the ARMA(2,2) model

  $$x_n = a_1 x_{n-1} + a_2 x_{n-2} + b_0 w_n + b_1 w_{n-1}$$

  and lets first <u>assume</u> $x_i = 0$ and $w_i = 0$ for $i < 0$ for simplicity

- We can express variables x_n in terms of variables w_n explicitly

  $$\begin{aligned} x_0 &= b_0 w_0 \\ x_1 &= a_1 x_0 + b_0 w_1 + b_1 w_0 \\ &= (a_1 b_0 + b_1) w_0 + b_0 w_1 \\ x_2 &= (a_1^2 b_0 + a_1 b_1 + a_2) w_0 + (a_1 b_0 + b_1) w_1 + b_0 w_2 \end{aligned}$$

  and in general we will get

  $$[x_0, ..., x_n]^{\mathrm{T}} = \mathbf{L}[w_0, ..., w_n]^{\mathrm{T}}$$

  where $\mathbf{L}$ is a lower-triangular matrix with diagonal components all equal to b_0

- For any $n \geq 0$, given $x_0, ..., x_n$, then we also know $w_0, ..., w_n$

- Using $x_n = a_1 x_{n-1} + a_2 x_{n-2} + b_0 w_n + b_1 w_{n-1}$, $p(x_n | x_0, ..., x_{n-1})$ is

$$\frac{1}{\sqrt{2\pi\sigma^2 b_0^2}} \exp\left(-\frac{(x_n - a_1 x_{n-1} - a_2 x_{n-2} - b_1 w_{n-1})^2}{2\sigma^2 b_0^2}\right)$$

- The expression for $p(x_0, ..., x_{N-1})$ follows from the probability chain rule. There is a sequential way to evaluate $p(x_0, ..., x_{N-1})$ and its computational cost grows linearly with N

- We can evaluate the log of the likelihood for any value of parameter $(a_1, a_2, b_0, b_1, \sigma)$ and could use an optimization routine that only needs the function being optimized to be computable at any value of parameter

- The assumption $x_i = 0$ and $w_i = 0$ for $i < 0$ should have progressive less and less influence on the maximum likelihood parameter estimates as N grows and asymptotically have <u>no influence</u>

- We can express this ARMA(2,2) model in state-space form:

$$\mathbf{x}_n = \begin{bmatrix} x_n \\ z_n \end{bmatrix} = \begin{bmatrix} a_1 & 1 \\ a_2 & 0 \end{bmatrix} \begin{bmatrix} x_{n-1} \\ z_{n-1} \end{bmatrix} + \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} w_n$$
$$y_n = [1, 0]\mathbf{x}_n = \mathbf{c}^{\mathrm{T}}\mathbf{x}_n$$

  where $x_{-1} = z_{-1} = 0$. (Verify this)

- Let $\mathbf{A} = \begin{bmatrix} a_1 & 1 \\ a_2 & 0 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$

- Apply the Kalman filter to this state-space model to calculate

$$p(y_0, \ldots, y_{N-1}) = p(x_0, \ldots, x_{N-1})$$

  via the probability chain rule

Calculating $p(x_0, \ldots, x_{N-1})$ without assuming $x_i = 0$ for $i < 0$ is possible

- Initialization: $\hat{\mathbf{x}}_{-1} = [0, 0]^{\mathrm{T}}$ and $\mathbf{R}_{-1}$ is the solution to

$$\mathbf{R}_{-1} = \mathbf{A}\mathbf{R}_{-1}\mathbf{A}^{\mathrm{T}} + \mathbf{b}\mathbf{b}^{\mathrm{T}}\sigma^2$$

  Computation: for $n = 0, 1, \ldots$

- Prediction step

$$\begin{aligned} \hat{\mathbf{x}}_{n|n-1} &= \mathbf{A}\hat{\mathbf{x}}_{n-1} \\ \mathbf{R}_{n|n-1} &= \mathbf{A}\mathbf{R}_{n-1}\mathbf{A}^{\mathrm{T}} + \mathbf{b}\mathbf{b}^{\mathrm{T}}\sigma^2 \end{aligned}$$

- Gain calculation

$$\mathbf{K}_n = \mathbf{R}_{n|n-1}\mathbf{c} \times \left[\mathbf{c}^{\mathrm{T}}\mathbf{R}_{n|n-1}\mathbf{c}\right]^{-1}$$

- Update step

$$\begin{aligned} \hat{\mathbf{x}}_n &= \hat{\mathbf{x}}_{n|n-1} + \mathbf{K}_n\left[y_n - \mathbf{c}^{\mathrm{T}}\hat{\mathbf{x}}_{n|n-1}\right] \\ \mathbf{R}_n &= \left[\mathbf{I} - \mathbf{K}_n\mathbf{c}^{\mathrm{T}}\right]\mathbf{R}_{n|n-1} \end{aligned}$$

- Likelihood calculation

$$p(y_n|y_0,\ldots,y_{n-1})$$
$$= \left(2\pi\mathbf{c}^{\mathrm{T}}\mathbf{R}_{n|n-1}\mathbf{c}\right)^{-1/2}\exp\left(-\frac{(y_n-\mathbf{c}^{\mathrm{T}}\hat{\mathbf{x}}_{n|n-1})^2}{2\mathbf{c}^{\mathrm{T}}\mathbf{R}_{n|n-1}\mathbf{c}}\right)$$

- For a general ARMA(P,Q) model, let

$$r = \max(P, Q+1)$$

If $r > P$ set

$$a_{P+1} = \cdots = a_r = 0$$

If $r - 1 > Q$, set

$$b_{Q+1} = \cdots = b_{r-1} = 0$$

$\mathbf{x}_n$ is a $r \times 1$ vector,

$$\mathbf{A} = \begin{bmatrix} a_1 & 1 & 0 & \cdots & 0 \\ a_2 & 0 & 1 & 0 & \cdots \\ & & & & \\ a_{r-1} & 0 & \cdots & 0 & 1 \\ a_r & 0 & \cdots & & 0 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ \\ b_{r-1} \end{bmatrix}$$

(See Gardner *et. al.* (1980) An Algorithm for Exact Maximum Likelihood Estimation of Autoregressive-Moving Average Models by Means of Kalman Filtering, *Appl. Statist.*, **29**, 311-322.)