Modelling Non-linear and Non-stationary Time Series

Chapter 7: State Space Models
Maximum likelihood estimation

Henrik Madsen

Lecture Notes

September 2016

Contents

State Space Models

- Due to the (first order) *Markov Property* of the state vector the state space representation is very useful for describing *non-stationary* and *time varying* systems.
- It will be seen that even linear state space models can be used to describe some *non-linear* processes.
- State space models provides a flexible approach to *adaptive estimation*.
- A class of state space models (SDM) will be introduced to describe most of the non-linear time series models that has been introduced.

The non-linear, time-varying, discrete time state space model

Consider the *non-linear*, discrete time system

$$\boldsymbol{X}_{t+1} = \boldsymbol{f}(t, \boldsymbol{X}_t, \boldsymbol{U}_t) + \boldsymbol{v}_t \quad (1)$$

$$\boldsymbol{Y}_t = \boldsymbol{h}(t, \boldsymbol{X}_t, \boldsymbol{U}_t) + \boldsymbol{e}_t \quad (2)$$

where $\{\boldsymbol{e}\}$ and $\{\boldsymbol{v}\}$ are sequences of independent random variables with

$$\begin{aligned} \mathrm{E}[\boldsymbol{v}\boldsymbol{v}^T] &= \boldsymbol{R}_{1,t} \\ \mathrm{E}[\boldsymbol{e}\boldsymbol{e}^T] &= \boldsymbol{R}_{2,t} \\ \mathrm{E}[\boldsymbol{v}\boldsymbol{e}^T] &= \boldsymbol{0} \end{aligned}$$

The Extended Kalman Filter

Given the previously defined non-linear state space models the *Extended Kalman Filter (EKF)* estimate of $\boldsymbol{X}_{t+1}$ given $\boldsymbol{Y}_0, \boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_t$ is given by

$$\hat{\boldsymbol{X}}_{t+1|t} = \boldsymbol{f}(t, \hat{\boldsymbol{X}}_t) + \boldsymbol{K}_t[\boldsymbol{Y}_t - h(t, \hat{\boldsymbol{X}}_{t|t-1})] \tag{3}$$

where

$$\begin{aligned}
\boldsymbol{K}_t &= \boldsymbol{F}(t, \hat{\boldsymbol{X}}_{t|t-1})\boldsymbol{P}_t\boldsymbol{H}^T(t, \hat{\boldsymbol{X}}_{t|t-1})[\boldsymbol{H}(t, \hat{\boldsymbol{X}}_{t|t-1})\boldsymbol{P}_t\boldsymbol{H}^T(t, \hat{\boldsymbol{X}}_{t|t-1}) + \boldsymbol{R}_{2,t}]^{-1} \\
\boldsymbol{P}_{t+1} &= \boldsymbol{F}(t, \hat{\boldsymbol{X}}_{t|t-1})\boldsymbol{P}_t\boldsymbol{F}^T(t, \hat{\boldsymbol{X}}_{t|t-1}) + \boldsymbol{R}_{1,t} \\
&\quad -\boldsymbol{K}_t[\boldsymbol{R}_{2,t} + \boldsymbol{H}(t, \hat{\boldsymbol{X}}_{t|t-1})\boldsymbol{P}_t\boldsymbol{H}^T(t, \hat{\boldsymbol{X}}_{t|t-1})]\boldsymbol{K}_t^T
\end{aligned}$$

and where the Jacobians of $\boldsymbol{f}$ and $\boldsymbol{h}$ are given by

$$\begin{aligned}
\boldsymbol{F}(t, \hat{\boldsymbol{X}}) &= \frac{\partial}{\partial \boldsymbol{X}} \boldsymbol{f}(t, \boldsymbol{X})|_{\boldsymbol{X}=\hat{\boldsymbol{X}}} \\
\boldsymbol{H}(t, \hat{\boldsymbol{X}}) &= \frac{\partial}{\partial \boldsymbol{X}} \boldsymbol{h}(t, \boldsymbol{X})|_{\boldsymbol{X}=\hat{\boldsymbol{X}}}
\end{aligned}$$

This is often called the *discrete-discrete Extended Kalman Filter (EKF)*

Maximum likelihood estimation

- Consider the linear stochastic state space model:

$$\boldsymbol{X}_t = \boldsymbol{A}_t \boldsymbol{X}_{t-1} + \boldsymbol{B}_t \boldsymbol{u}_{t-1} + \boldsymbol{e}_{1,t}, \tag{4}$$

$$\boldsymbol{Y}_t = \boldsymbol{C}_t \boldsymbol{X}_t + \boldsymbol{e}_{2,t}, \tag{5}$$

  where $\{\boldsymbol{e}_{1,t}\}$ and $\{\boldsymbol{e}_{2,t}\}$ are mutually uncorrelated normal distributed white noise sequences with variance $\boldsymbol{\Sigma}_{1,t}$ and $\boldsymbol{\Sigma}_{2,t}$, respectively.

- Assume that $\dim \boldsymbol{Y} = m$ and that the model is **global identifiable**.
- The parameters are embedded in the matrices $\boldsymbol{A}_t$, $\boldsymbol{B}_t$, $\boldsymbol{C}_t$, $\boldsymbol{\Sigma}_{1,t}$ and $\boldsymbol{\Sigma}_{2,t}$, which might be *time varying*.
- As an example a missing observation at time t can be handled by putting

$$\boldsymbol{\Sigma}_{2,t} = \infty \tag{6}$$

ML estimates of state space models

Let $\boldsymbol{\theta}$ denote the unknown parameters, and $\mathcal{Y}_{\mathcal{N}} = (\boldsymbol{Y}_{\mathcal{N}}, \cdots, \boldsymbol{Y}_0)$ all $\mathcal{N}$ available observations in the time series of length N ($\mathcal{N} < N$).
The *conditional likelihood function* (conditioned on $\boldsymbol{Y}_0$) is

$$L(\boldsymbol{\theta}; \mathcal{Y}_{\mathcal{N}}) = f(\mathcal{Y}_N|\boldsymbol{\theta}) \tag{7}$$

$$= f(\boldsymbol{Y}_{\mathcal{N}}|\mathcal{Y}_{\mathcal{N}-1}, \boldsymbol{\theta}) f(\boldsymbol{Y}_{\mathcal{N}-1}|\mathcal{Y}_{\mathcal{N}-2}, \boldsymbol{\theta}) \cdots f(\boldsymbol{Y}_1|\boldsymbol{Y}_0, \boldsymbol{\theta}). \tag{8}$$

ML estimation for state space models

- Since the model is linear and all the random variables are Gaussian it follows that $\boldsymbol{Y}_{t+1}|\mathcal{Y}_t$ is Gaussian too.
- This conditional Gaussian density is characterized by the conditional mean and the conditional variance.
- Let us introduce the *one-step prediction error* (the *innovation*) and its covariance:

$$\tilde{\boldsymbol{Y}}_{t+1|t} = \boldsymbol{Y}_{t+1} - \hat{\boldsymbol{Y}}_{t+1|t}, \tag{9}$$

$$\begin{aligned} \boldsymbol{R}_{t+1} &= \mathrm{V}\left[\tilde{\boldsymbol{Y}}_{t+1|t}\right] = \mathrm{V}\left[\boldsymbol{Y}_{t+1}|\mathcal{Y}_t\right] = \boldsymbol{\Sigma}^{yy}_{t+1|t} \\ &= \boldsymbol{C}_t \boldsymbol{\Sigma}^{xx}_{t+1|t} \boldsymbol{C}_t^T + \boldsymbol{\Sigma}_{2,t}. \end{aligned} \tag{10}$$

- The conditional mean and covariance are simply calculated by an ordinary Kalman Filter.

ML estimates

- Given the mean and variance of $\boldsymbol{Y}_{t+1}|\mathcal{Y}_t$ we are able to state the conditional density function

$$f\left(\boldsymbol{Y}_{t+1}|\mathcal{Y}_t\right) = \left[(2\pi)^m \det \boldsymbol{R}_{t+1}\right]^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\tilde{\boldsymbol{Y}}_{t+1|t}^T \boldsymbol{R}_{t+1}^{-1} \tilde{\boldsymbol{Y}}_{t+1|t}\right] \tag{11}$$

- Using this conditional density function the conditional likelihood function writes

$$L\left(\boldsymbol{\theta}; \mathcal{Y}_{\mathcal{N}}\right) = \prod_{i=1}^{\mathcal{N}} \left[(2\pi)^m \det \boldsymbol{R}_i\right]^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\tilde{\boldsymbol{Y}}_{i|i-1}^T \boldsymbol{R}_i^{-1} \tilde{\boldsymbol{Y}}_{i|i-1}\right], \tag{12}$$

- As starting values we can take $\hat{\boldsymbol{X}}_{0|0} = 0$ and $\boldsymbol{\Sigma}_{0|0}^{xx} = \alpha \boldsymbol{I}$ where α is 'large'. Alternatively we can estimate the starting values using the ML approach.

ML estimates of state space models

- The maximization of the likelihood function is equivalent to maximization of

$$\log L\left(\boldsymbol{\theta}; \mathcal{Y}_{\mathcal{N}}\right) = -\frac{1}{2}\sum_{i=1}^{\mathcal{N}}[\log\det \boldsymbol{R}_i + \tilde{\boldsymbol{Y}}_{i|i-1}^{T}\boldsymbol{R}_i^{-1}\tilde{\boldsymbol{Y}}_{i|i-1}] + \text{const.} \quad (13)$$

- The ML-estimate of $\boldsymbol{\theta}$ is the argument which maximizes this, i.e.

$$\hat{\boldsymbol{\theta}} = \arg\left\{\max_{\boldsymbol{\theta}} \log L\left(\boldsymbol{\theta}; \mathcal{Y}_{\mathcal{N}}\right)\right\}. \quad (14)$$

- For the maximization a numerical method must be used.

Uncertainty of ML estimates

- As an approximation of the covariance of the parameter estimates we can use

$$\mathrm{V}\left[\hat{\boldsymbol{\theta}}\right] \simeq -\boldsymbol{H}^{-1}, \tag{15}$$

where

$$\{\boldsymbol{H}\}_{ij} = \left.\frac{\partial^2 \log L\left(\boldsymbol{\theta}; \mathcal{Y}_{\mathcal{N}}\right)}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{16}$$

- Note that $\boldsymbol{H}$ is used as an approximation of the Fisher information matrix.

ML estimates under stationary conditions

It is straight forward to seen that in a *stationary situation* the problem reduces to the following likelihood function

$$L(\boldsymbol{\theta}; \mathcal{Y}_N) = \prod_{t=1}^{N} [(2\pi)^m \det \boldsymbol{\Sigma}]^{-\frac{1}{2}} \exp\left[-\frac{1}{2} \tilde{\boldsymbol{Y}}_{t|t-1}^T \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{Y}}_{t|t-1}\right] \quad (17)$$

$$= [(2\pi)^m \det \boldsymbol{\Sigma}]^{N/2} \exp\left[-\frac{1}{2} \sum_{t=1}^{N} \tilde{\boldsymbol{Y}}_{t|t-1}^T \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{Y}}_{t|t-1}\right] \quad (18)$$

The assumption about stationarity implies that the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are constant. This also means that missing observations are not allowed.

Σ known

In this case maximization of the likelihood function is equivalent to minimization of $S_1(\boldsymbol{\theta})$ where

$$S_1(\boldsymbol{\theta}) = \frac{1}{N}\sum_{t=1}^{N} \tilde{\boldsymbol{Y}}_{t|t-1}^T \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{Y}}_{t|t-1} \tag{19}$$

where $\tilde{\boldsymbol{Y}}_{t+1|t} = \boldsymbol{Y}_{t+1} - \hat{\boldsymbol{Y}}_{t+1|t}$ is the one-step prediction error.
Let us introduce the *sample covariance of the predictions errors*, i.e.

$$\boldsymbol{D}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{t=1}^{N} \tilde{\boldsymbol{Y}}_{t|t-1} \tilde{\boldsymbol{Y}}_{t|t-1}^T \tag{20}$$

then the *cost function* S_1 can be written

$$S_1(\boldsymbol{\theta}) = \mathrm{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}(\boldsymbol{\theta}) \tag{21}$$

Σ unknown

- In this case the maximization of the likelihood function above is equivalent to minimization of

$$S(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \frac{1}{2} N \log \det \boldsymbol{\Sigma} + \frac{1}{2} \sum_{t=1}^{N} \tilde{\boldsymbol{Y}}_{t|t-1}^T \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{Y}}_{t|t-1} \tag{22}$$

- Differentiating with respect to $\boldsymbol{\Sigma}$ gives

$$\frac{\partial S}{\partial \boldsymbol{\Sigma}} = \frac{N}{2} \boldsymbol{\Sigma}^{-1} - \frac{1}{2} \boldsymbol{\Sigma}^{-1} \Big(\sum_{t=1}^{N} \tilde{\boldsymbol{Y}}_{t|t-1} \tilde{\boldsymbol{Y}}_{t|t-1}^T \Big) \boldsymbol{\Sigma}^{-1} \tag{23}$$

  which equals zero (ie. ML estimator) for

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{t=1}^{N} \tilde{\boldsymbol{Y}}_{t|t-1} \tilde{\boldsymbol{Y}}_{t|t-1}^T = \boldsymbol{D}(\boldsymbol{\theta}) \tag{24}$$

  i.e. for any given value of $\boldsymbol{\theta}$ then (24) minimizes (22). The problem can then be reduced by the constraint $\boldsymbol{\Sigma} = \boldsymbol{D}(\boldsymbol{\theta})$.

Σ unknown

By substituting this constraint into the above cost function gives

$$S_c(\boldsymbol{\theta}) = \frac{1}{2}\log\det \boldsymbol{D}(\boldsymbol{\theta}) + \frac{1}{2}\sum_{t=1}^{N} \tilde{\boldsymbol{Y}}_{t|t-1}^T \boldsymbol{D}^{-1}(\boldsymbol{\theta})\tilde{\boldsymbol{Y}}_{t|t-1} \tag{25}$$

or

$$\begin{aligned} S_c(\boldsymbol{\theta}) &= \frac{1}{2}\log\det \boldsymbol{D}(\boldsymbol{\theta}) + \frac{1}{2}\mathrm{tr}\left[\left(\sum_{t=1}^{N} \tilde{\boldsymbol{Y}}_{t|t-1}\tilde{\boldsymbol{Y}}_{t|t-1}^T\right)\boldsymbol{D}^{-1}(\boldsymbol{\theta})\right] && (26)\\ &= \frac{1}{2}\log\det \boldsymbol{D}(\boldsymbol{\theta}) + \frac{N}{2}\mathrm{tr} I_m && (27)\\ &= \frac{1}{2}\log\det \boldsymbol{D}(\boldsymbol{\theta}) + \frac{Nm}{2} && (28) \end{aligned}$$

Since the last term is constant the ML estimate is found by minimizing the *cost function*

$$S_2(\boldsymbol{\theta}) = \log\det \boldsymbol{D}(\boldsymbol{\theta}) \tag{29}$$

Estimation of some doubly stochastic models

- Consider the *non-linear* process:

$$Y_t = \Phi_t Y_{t-1} + \epsilon_t \quad (30)$$

$$\Phi_t - \mu = \phi(\Phi_{t-1} - \mu) + \zeta_t \quad (31)$$

  where $\{\epsilon_t\}$ and $\{\zeta_t\}$ are mutually uncorrelated Gaussian white noise processes with variances σ_ϵ^2 and σ_ζ^2, respectively.
- This model is some times referred to as an AR(1)-AR(1) model, since both the model for Y_t and the embedded model for the parameter variation is an AR(1) model.
- The parameters of the non-linear model are $(\phi, \mu, \sigma_\epsilon^2, \sigma_\zeta^2)$.

Estimation of some doubly stochastic models

By choosing a different parameterization of the model it can, however, be written as a linear state space model. The model for the parameter variation can be written

$$\Phi_t = \phi\Phi_{t-1} + \mu(1-\phi) + \zeta_t$$

Hence, by introducing $\delta = \mu(1-\phi)$ the variation of both $\{Y_t\}$ and $\{\Phi_t\}$ can be described by the linear state space model

$$\begin{pmatrix} \Phi_t \\ \delta_t \end{pmatrix} = \begin{pmatrix} \phi & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \Phi_{t-1} \\ \delta_{t-1} \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \zeta_t \tag{32}$$

$$Y_t = (Y_{t-1}, 0) \begin{pmatrix} \Phi_t \\ \delta_t \end{pmatrix} + \epsilon_t \tag{33}$$

Note that the new parameter is introduced as a state variable, and that there is a unique relation between the new parameters $(\phi, \delta, \sigma_\epsilon^2, \sigma_\zeta^2)$ and the parameters of the model (30).

A generalization

The principle outlined above is readily generalized. It is, for instance, rather easy to see that the following doubly stochastic model (an AR(1)-AR(2) model)

$$Y_t = \Phi_t Y_{t-1} + \epsilon_t \tag{34}$$

$$\Phi_t - \mu = \phi_1(\Phi_{t-1} - \mu) + \phi_2(\Phi_{t-1} - \mu) + \zeta_t \tag{35}$$

can be written as the following linear state space model

$$\begin{pmatrix} \Phi_t \\ \Phi_{t-1} \\ \delta_t \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \Phi_{t-1} \\ \Phi_{t-2} \\ \delta_{t-1} \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \zeta_t \tag{36}$$

$$Y_t = (Y_{t-1}, 0, 0) \begin{pmatrix} \Phi_t \\ \Phi_{t-1} \\ \delta_t \end{pmatrix} + \epsilon_t \tag{37}$$

where $\delta = \mu(1 - \phi_1 - \phi_2)$.

An example

The above outlined class of doubly stochastic processes provides a rather flexible description of non-linear phenomena.
Consider as an example a doubly stochastic process with two different parameter sets as described here:

Parameter	ϕ	μ	σ_ζ^2	σ_ϵ^2
True values No. 1	0.85	0.80	0.10^2	4.0^2
True values No. 2	0.95	0.80	0.10^2	4.0^2

An example (cont.)

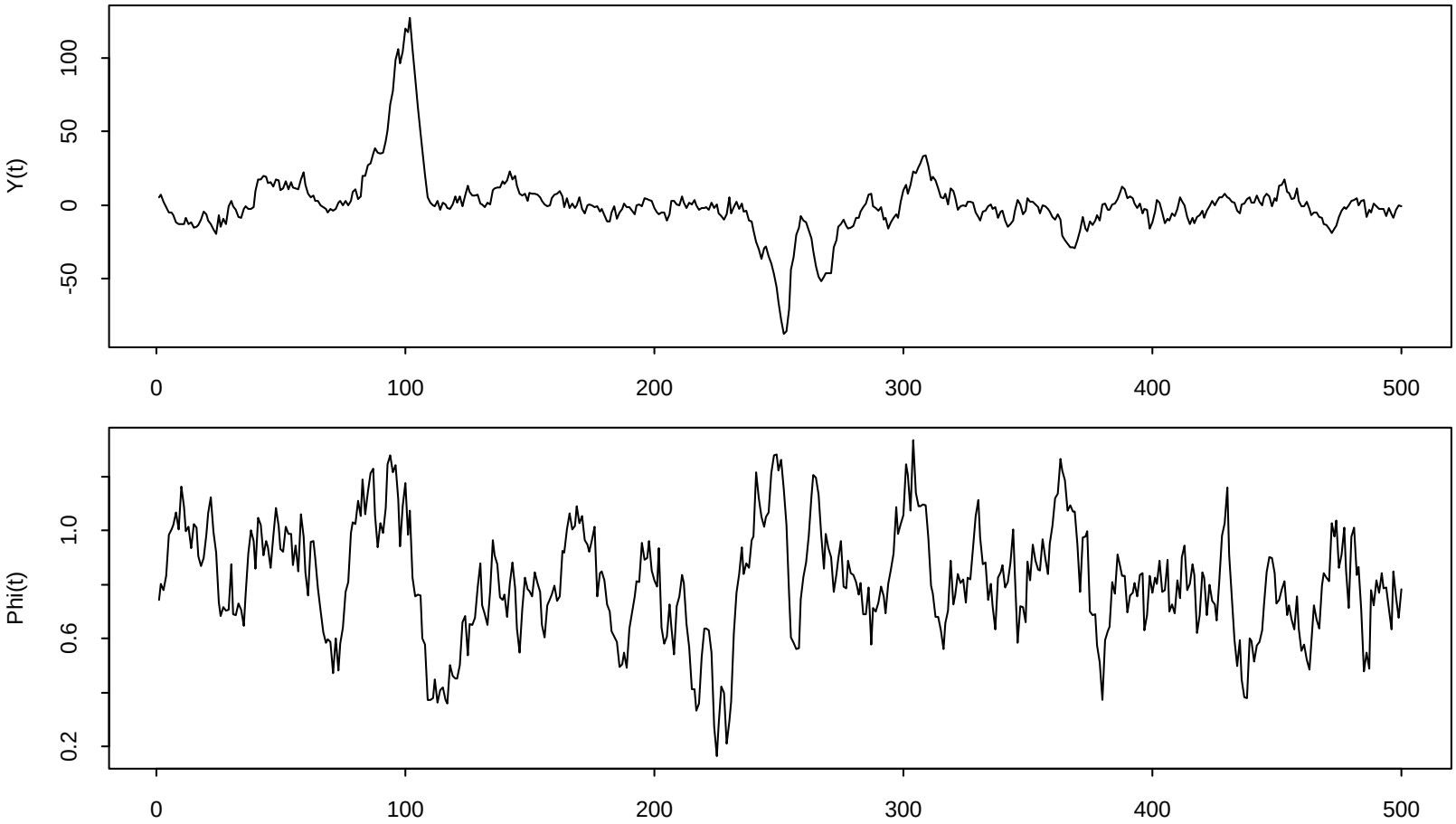

Figure : A simulation of a doubly stochastic process - $\phi = 0.85$

An example (cont.)

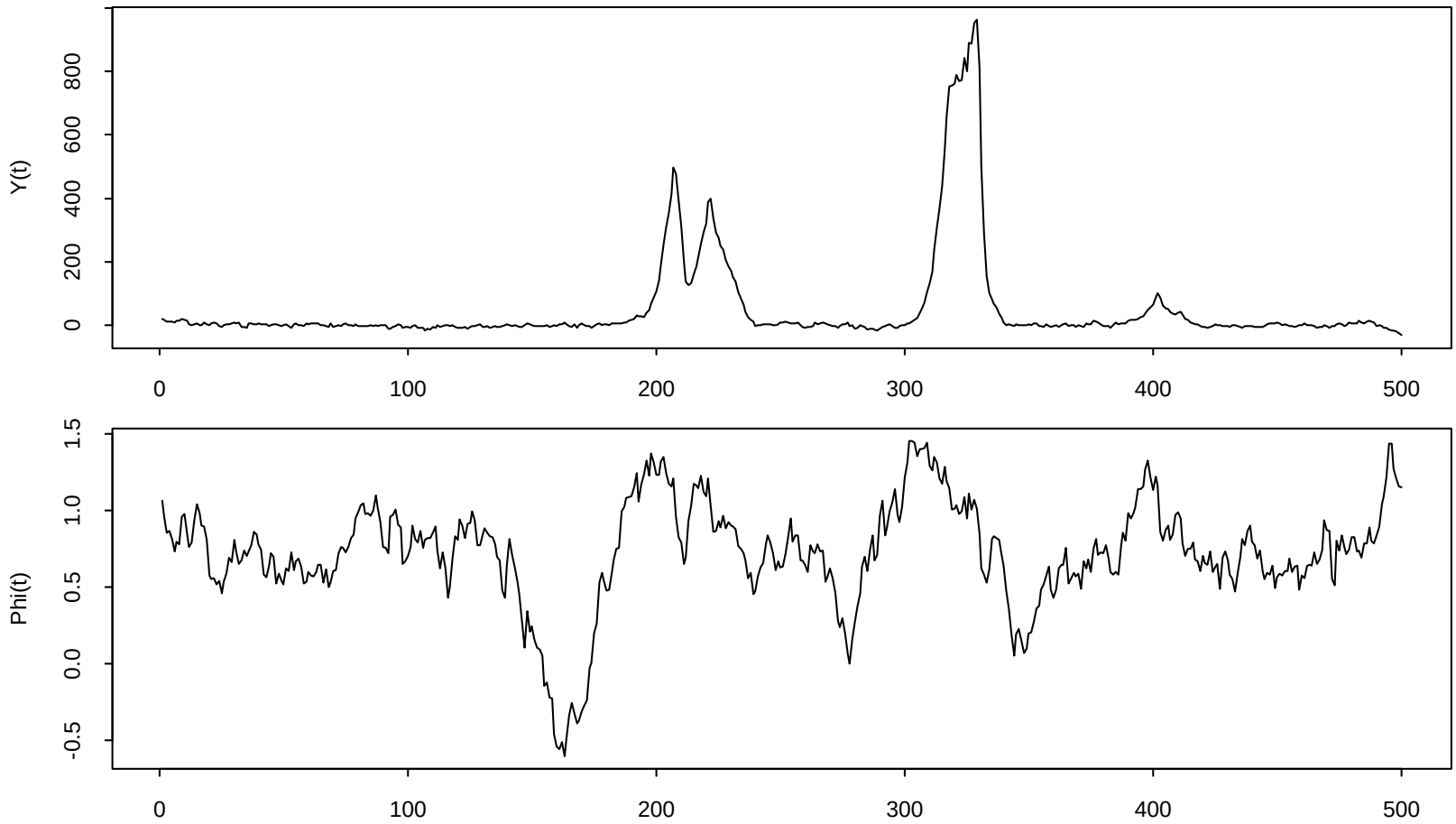

Figure : A simulation of a doubly stochastic process - $\phi = 0.95$.

Extended Kalman Filter for parameter estimation

Idea: *Include the unknown parameters in the state vector* and then use a method for state estimation. The resulting state space model is in general a non-linear state space model, and the *extended Kalman filter* is then most often used for the state estimation.
Consider the following state space model

$$\begin{aligned} \boldsymbol{X}_{t+1} &= \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{X}_t + \boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{u}_t + \boldsymbol{v}_t \\ \boldsymbol{Y}_t &= \boldsymbol{C}(\boldsymbol{\theta})\boldsymbol{X}_t + \boldsymbol{e}_t \end{aligned}$$

where the parameter vector $\boldsymbol{\theta}$ *is unknown*.

Extended Kalman Filter for parameter estimation

By including ($\boldsymbol{\theta} = \boldsymbol{\theta}_t$) in a new state vector

$$\boldsymbol{Z}_t = \begin{pmatrix} \boldsymbol{X}_t \\ \boldsymbol{\theta}_t \end{pmatrix}$$

this leads to the following non-linear state space model

$$\begin{aligned} \boldsymbol{Z}_{t+1} &= f(\boldsymbol{Z}_t, \boldsymbol{u}_t) + \begin{pmatrix} \boldsymbol{v}_t \\ 0 \end{pmatrix} \\ \boldsymbol{Y}_t &= h(\boldsymbol{Z}_t) + \boldsymbol{e}_t \end{aligned}$$

where

$$\begin{aligned} f(\boldsymbol{Z}_t, \boldsymbol{u}_t) &= \begin{pmatrix} \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{X}_t + \boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{u}_t \\ \boldsymbol{\theta}_t \end{pmatrix} \\ h(\boldsymbol{Z}_t) &= \boldsymbol{C}(\boldsymbol{\theta})\boldsymbol{X}_t \end{aligned}$$

Now an EKF state estimation will provide a simultaneous estimation of the original state vector and the parameter vector $\boldsymbol{\theta}$

The EM algorithm

For some problems the likelihood function for an extended (complete) data set is more easy to formulate that the likelihood function for the (incomplete) data at hand, which is assumed to be a subset of the complete data set.

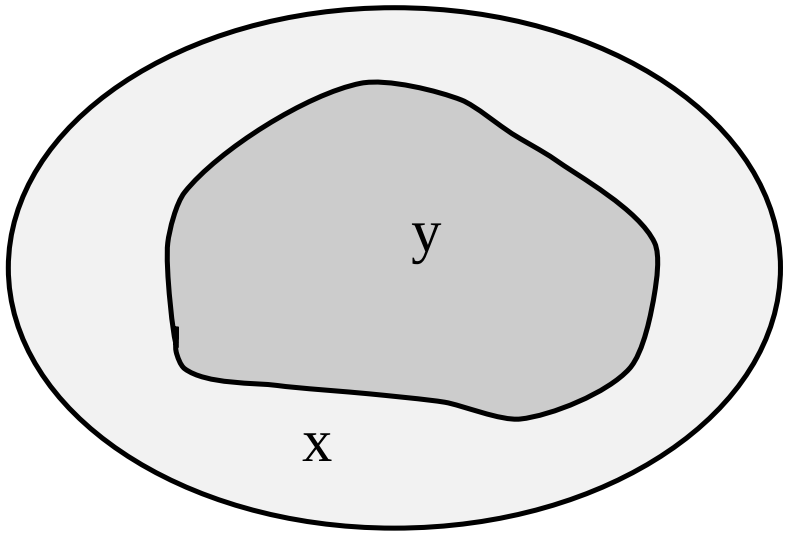

Figure : Complete (x) and incomplete data (y).

The incompleteness may be due to e.g. missing observations, or

The EM algorithm contains two steps:

1. The **E**xpectation step for estimating the complete likelihood function conditioning on the available data and the latest parameter estimates.
2. The **M**aximization step for finding the optimal parameter values in the current iteration.

Let $k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ be the conditional density of $\boldsymbol{x}$ given $\boldsymbol{y}$ and $\boldsymbol{\theta}$. Since $\boldsymbol{y}$ is a subset of $\boldsymbol{x}$ we have

$$k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) = \frac{g(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{y}|\boldsymbol{\theta})}.$$

or

$$\log f(\boldsymbol{y}|\boldsymbol{\theta}) = \log g(\boldsymbol{x}|\boldsymbol{\theta}) - \log k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$$

By multiplying both sides by the density $p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^{(p)})$ and integrate with respect to $\boldsymbol{x}$ we obtain the log likelihood

$$\begin{aligned} L(\boldsymbol{\theta}) &= \log f(\boldsymbol{y}|\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) \\ Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) &= \mathrm{E}\left[\log g(\boldsymbol{x}|\boldsymbol{\theta})|\boldsymbol{y},\boldsymbol{\theta}^{(p)}\right] \\ H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) &= \mathrm{E}\left[\log k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})|\boldsymbol{y},\boldsymbol{\theta}^{(p)}\right]. \end{aligned}$$

The EM-iteration

Now we can define the EM-iteration

- *E-step*: Calculate $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) = \mathrm{E}_p[\log L(\boldsymbol{x}, \boldsymbol{\theta}|y_1, \ldots, y_n, \boldsymbol{\theta}^{(p)})]$, as a function of $\boldsymbol{\theta}$.
- *M-step*: Determine $\boldsymbol{\theta}^{(p+1)}$ such that $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)})$ is maximized.

That is the EM algorithm. The EM algorithm thus finds the maximum of the likelihood function iteratively.

Sketch of the proof

Jensen inequality states that for a random variable X and a convex function $\phi(x)$

$$\mathrm{E}[\phi(X)] \geq \phi(\mathrm{E}[X])$$

That is (here: concave function)

$$\mathrm{E}\left[\log k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})|\boldsymbol{y},\boldsymbol{\theta}^{(p)}\right] \leq \log k(\boldsymbol{x}^{(p)}|\boldsymbol{y},\boldsymbol{\theta}^{(p)})$$

$$H(\boldsymbol{\theta},\boldsymbol{\theta}^{(p)}) \leq H(\boldsymbol{\theta}^{(p)},\boldsymbol{\theta}^{(p)})$$

And hence

$$L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^{(p)}) = Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(p)}) - Q(\boldsymbol{\theta}^{(p)},\boldsymbol{\theta}^{(p)}) + H(\boldsymbol{\theta}^{(p)},\boldsymbol{\theta}^{(p)}) - H(\boldsymbol{\theta},\boldsymbol{\theta}^{(p)}) \geq 0$$

if $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(p)})$ is maximized.
Hence, the EM algorithm iteration never decreases the log likelihood.
The key result is that in order to maximize L we only need to maximize Q.

EM-estimation of NLAR(1) model

$$Y_t = h(Y_{t-1}, \theta) + \epsilon_t \tag{38}$$

where $\{\epsilon_t\}$ is i.i.d. N(0,1).
Assume that the available observations are

$$\mathcal{Y}_N = (Y_1, Y_2, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_N) \tag{39}$$

i.e. Y_j is missing.

The i'th iteration of the EM-algorithm is

- *E-step* : Find $\mathrm{E}[Y_j|\mathcal{Y}_N,\theta^{(i)}] = Y_j^{(i)}, \mathrm{E}[Y_j^2|\mathcal{Y}_N,\theta^{(i)}] = (Y_j^2)^{(i)}, \mathrm{E}[h(Y_j)|\mathcal{Y}_N,\theta^{(i)}] = h(Y_j)^{(i)}, \mathrm{E}[h^2(Y_j)|\mathcal{Y}_N,\theta^{(i)}] = h^2(Y_j)^{(i)}$.
  In the evaluation it can be used that (for the NLAR(1)-model)

$$p(Y_j|\mathcal{Y}_N) = \frac{p(Y_j|Y_{j-1})p(Y_{j+1}|Y_j)}{p(Y_{j+1}|Y_{j-1})} \tag{40}$$

  Note also that $p(Y_{j+1}|Y_j) = \phi(Y_{j+1} - h(Y_j))$ where ϕ is the density function of N(0,1).

- *M-step* : Minimize with respect to θ the 'sum of squares'

$$\sum_{t=1}^{N} Y_t^2 - 2\sum_{t=1}^{N} Y_t h(Y_{t-1}) + \sum_{t=1}^{N} h^2(Y_{t-1}) \tag{41}$$

where we replace $Y_j, Y_j^2, h(Y_j), h^2(Y_j)$ by the values found under the E-step. This gives $\theta^{(i+1)}$.

State-dependent models (SDM)

- Previously we introduced the general non-linear model,

$$X_t = h'(\epsilon_t, \epsilon_{t-1}, \dots) \tag{42}$$

  which clearly is “infinite dimensional”, since infinitely many past variables are used.
- The idea behind the state-dependent models is to reduce this to a “finite dimensional” form.
- In a parsimonious finite description it might be a good idea to consider previous X and ϵ-values simultaneously.
  Hence consider:

$$X_t = h(X_{t-1}, \cdots, X_{t-p}, \epsilon_{t-1}, \cdots, \epsilon_{t-q}) + \epsilon_t \tag{43}$$

State-dependent models

- Define the "state vector"

$$\tilde{\boldsymbol{x}}_t = (\epsilon_{t-q+1}, \ldots, \epsilon_t, X_{t-p+1}, \ldots, X_t)^T$$

  Priestly (1980, 1988) uses the term "state vector" although the definition does not in general lead to a *minimal realization*[1]. It is for instance well-known that for an ARMA(p, q) model the minimal realization has a state vector of dimension $\max(p, q+1)$, whereas the state vector $\tilde{\boldsymbol{x}}_t$ has the dimension $(q+p)$.
- After a first-order Taylor expansion of h about some fixed time point we obtain the *State-Dependent Model (SDM) of order (p,q)*:

$$X_t + \sum_{i=1}^{p} \phi_i(\tilde{\boldsymbol{x}}_{t-1}) X_{t-i} = \mu(\tilde{\boldsymbol{x}}_{t-1}) + \epsilon_t + \sum_{i=1}^{q} \psi_i(\tilde{\boldsymbol{x}}_{t-1}) \epsilon_{t-i} \tag{44}$$

[1] Recall that a realization of a state space model (represented by the triplet of matrices ($\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$)) is minimal iff the corresponding system is completely controllable and completely observable

Some nonlinear models on SDM form

- *ARMA models*: μ, $\{\phi_i\}$ and $\{\psi_i\}$ constants.
- *Bilinear models*: μ, $\{\phi_i\}$ constants, and

$$\psi_i(\tilde{\mathbf{x}}_{t-1}) = c_i + \Sigma_{j=1}^{m} b_{ji} X_{t-j} \quad ; \quad i = 1, \cdots, q \tag{45}$$

  By selecting q properly and by putting some of c_i or b_{ji} equal to zero the general bilinear model is obtained.
- *Threshold models*:

$$\{\psi_i\} = 0 \tag{46}$$

$$\mu_t = a_0^{(J_t)} \tag{47}$$

$$\phi_i(\tilde{\mathbf{x}}_{t-1}) = -a_i^{(J_t)} \quad \text{if} \quad X_{t-d} \in R^{(J_t)} \tag{48}$$

- *Exponential AR models*: For $\mu = 0$, $\{\psi_i\} = 0$ and

$$\phi_i(\tilde{\mathbf{x}}_{t-1}) = -(\alpha_i + \beta_i e^{-\delta X_{t-1}^2}) \quad i = 1, \cdots, p \tag{49}$$

  then (44) reduces to the EXPAR model.

A state space from of SDM models

The "state vector"

$$\tilde{\boldsymbol{x}}_t = (\epsilon_{t-q+1}, \cdots, \epsilon_t, X_{t-p+1}, \cdots, X_t)^T$$

together with the innovation, ϵ_{t+1}, determine completely the evolution from time t to time $t+1$.
A formal state space representation is

$$\begin{aligned} \tilde{\boldsymbol{x}}_{t+1} &= \boldsymbol{\mu}(\tilde{\boldsymbol{x}}_t) + \boldsymbol{F}(\tilde{\boldsymbol{x}}_t)\tilde{\boldsymbol{x}}_t + \tilde{\boldsymbol{\epsilon}}_{t+1} && (50) \\ X_t &= \boldsymbol{C}\tilde{\boldsymbol{x}}_t && (51) \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\mu}(\tilde{\boldsymbol{x}}_{t-1}) &= (0, \cdots, 0 \vdots 0, \cdots, \mu(\tilde{\boldsymbol{x}}))^T && (52) \\ \tilde{\boldsymbol{\epsilon}}_t &= \epsilon_t(0, \cdots, 1 \vdots 0, \cdots, 1)^T && (53) \\ \boldsymbol{C} &= (0, \cdots, 0, 1) && (54) \\ & && (55) \end{aligned}$$

The F matrix is

$$
\boldsymbol{F}(\tilde{\boldsymbol{x}}_t) = \left[\begin{array}{ccccc:ccccc}
0 & 1 & 0 & \cdots & 0 & 0 & \cdots & \cdots & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 & 0 & \cdots & \cdots & \cdots & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & & & & \vdots \\
0 & 0 & 0 & \cdots & 1 & 0 & \cdots & \cdots & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & \cdots & \cdots & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 \\
0 & \cdots & \cdots & \cdots & 0 & 0 & 0 & 1 & \cdots & 0 \\
\vdots & & & & \vdots & \vdots & \vdots & \vdots & & \vdots \\
0 & \cdots & \cdots & \cdots & 0 & 0 & 0 & 0 & \cdots & 1 \\
\psi_q & \psi_{q-1} & & \cdots & \psi_1 & -\phi_p & -\phi_{p-1} & & \cdots & -\phi_1
\end{array}\right] \tag{56}
$$