# Dynamic Models with Latent Variables

Jean-Michel Zakoian

CREST-ENSAE

Markov-switching models

# Time series and breaks

- Many economic time series occasionally exhibit dramatic breaks in their behavior. Such breaks may concern the level, the variability, the serial correlations..
- They are often associated with events such as financial crises or abrupt changes in government policy. For instance, many economic variables tend to behave differently during economic downturns.
- When models are fitted over sub-periods of long times series, one often detects significant changes in the estimated parameter values.

## How can we model dramatic changes ?

Until the end of the 80's, the main approach consisted in fitting different models over different subperiods.

For instance, for AR(1) models :

$$\begin{array}{rcllll} y_t &=& \phi_1 y_{t-1} + \epsilon_t, & \text{if} & t \leq t_0, \\ y_t &=& \phi_2 y_{t-1} + \epsilon_t, & \text{if} & t > t_0. \end{array}$$

**Drawbacks :**

- Sub-models are not related
- Arbitrary choice of the break date $t_0$
- Assumption of non stationarity

# A simple way to model dramatic changes

AR(1) model with random AR coefficient :

$$y_t = \phi(\Delta_t)y_{t-1} + \epsilon_t,$$

where $\Delta_t$ is a discrete random variable.

A complete description of the probability law governing the data would then require a probabilistic model for $\Delta_t$.

The simplest such specification is that $\Delta_t$ is the realization of a **finite-state Markov chain**.

Such a model is called **Markov-Switching (MS)** AR(1) model.

# References

The introduction of **MS models** in the econometrics literature is due to Hamilton

- Hamilton, J. D. (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57, 357–384.

In the statistics literature, related models called Hidden Markov Models (HMM) had been studied much earlier :

- Baum, L. E., et T. Petrie (1966) Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics* 30, 1554-1563.
- Baum, L. E., Petrie, T., Soules, G. et Weiss, N. (1970) A maximization technique in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical statistics* 41, 164-171.

1. Hidden Markov Model
   - Finite-state Markov chains
   - Properties of Hidden Markov Models
2. MS-ARMA$(p,q)$ process
3. Estimation of MS-AR models

# Definition

Let $\Delta_0, \Delta_1, \ldots \in \mathscr{E} = \{1, ..., d\}$ a sequence of random variables. The sequence is a *Markov chain* if

$$\begin{aligned} & P(\Delta_t = j | \Delta_{t-1} = i) \\ = \quad & P(\Delta_t = j | \Delta_{t-1} = i, \Delta_{t-2} = e_{t-2}, \ldots, \Delta_0 = e_0) = p(i,j) \end{aligned}$$

for any $t$ and any $(i, j, e_{t-2}, \ldots, e_0) \in \mathscr{E}^{t+1}$ for which $P(\Delta_{t-1} = i, \Delta_{t-2} = e_{t-2}, \ldots, \Delta_0 = e_0) > 0$.

The set $\mathscr{E}$ is called the state space of the process and the $p(i,j)$ are the transition probabilities.

# Law

The law of the Markov chain is entirely characterized by

1. the *initial probabilities*

$$\pi_0(i) = P(\Delta_0 = i), \quad \pi_0(i) \geq 0, \qquad i = 1, \ldots, d, \quad \sum_{i=1}^{d} \pi_0(i) = 1$$

2. the *transition probability matrix* $P = \left(p(i,j)\right)_{1 \leq i,j \leq d}$.

# Higher-order transitions

The $k$th power of the transition matrix, $P^k = \left(p^{(k)}(i,j)\right)_{1\le i,j\le d}$, provides the $k$-step transition probabilities :

$$p^{(k)}(i,j) = P(\Delta_t = j|\Delta_{t-k} = i), \quad i,j \in \mathscr{E}, k \ge 0.$$

Let

$$\pi_0 = \begin{pmatrix} \pi_0(1) \\ \vdots \\ \pi_0(d) \end{pmatrix} \quad \text{and} \quad \pi_n = \begin{pmatrix} P(\Delta_n = 1) \\ \vdots \\ P(\Delta_n = d) \end{pmatrix}.$$

We have

$$\pi_n = P'\pi_{n-1}, \quad \pi_n = P'^{n}\pi_0, \quad n \ge 0.$$

# Invariant probability

A probability $\pi$ on $\mathscr{E}$ is called invariant probability if

$$\pi = P'\pi, \quad \pi'\mathbf{1} = 1 \quad (\text{with } \mathbf{1}' = (1,\ldots,1)).$$

- If the limit law $\pi_\infty := \lim_{n\to\infty} \pi_n$ exists then it is an invariant probability.
- An invariant probability always exists (for a finite state space).
- If $\pi_0 = \pi$ where $\pi$ is an invariant probability, then $\pi_n = \pi$ for all $n \geq 0$.

# Irreducibility, aperiodicity

It is possible for a chain starting in $i$ to reach $j$ if and only if

$$p^{(n)}(i,j) > 0, \quad \text{for some } n.$$

If it is true for all $i$ and $j$, the Markov chain is called **irreducible**.

A state $i$ is called **aperiodic** if

$$1 = \gcd\{n; \quad p^{(n)}(i,i) > 0\}.$$

If all states verify this condition, the chain is aperiodic.

# Exponential convergence to the stationary law and ergodicity

## Proposition

*If the chain is* **irreducible** *and* **aperiodic**, *there is a stationary distribution* $\pi$ *and there exists* $K \geq 0$ *and* $0 < \rho < 1$ *such that*

$$|p^{(n)}(i,j) - \pi(j)| \leq K\rho^n,$$

*for all states* $i$ *and* $j$*.*

Under these conditions, we have the ergodic property :

$$\frac{1}{n}\sum_{t=1}^{n} f(\Delta_t) \overset{n\to\infty}{\to} E_\pi\{f(\Delta_t)\} = \sum_{i=1}^{d} \pi_i f(i), \quad a.s.$$

for any function $f$.

An irreducible, aperiodic and stationary Markov chain is called **ergodic**.

## Definition

A process $(X_t)_{t\geq 0}$ follows a HMM if

1. conditionally to a (hidden) Markov chain $(\Delta_t)$, the variables $X_0, X_1, \ldots$ are independent ;
2. the conditional law of $X_s$ given $(\Delta_t)$ only depends on $\Delta_s$.

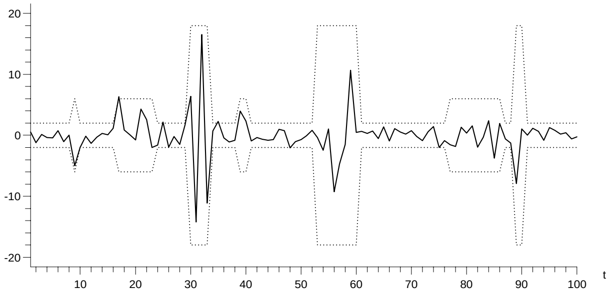
Canonical HMM :

$$\epsilon_t = \sigma(\Delta_t)\eta_t,$$

where

- $0 < \sigma(1) < \ldots < \sigma(d)$,
- $(\eta_t)$ is an **iid sequence** of variables, $E(\eta_t) = 0$, $\mathrm{var}(\eta_t) = 1$,
- $(\Delta_t)$ is an **ergodic Markov chain** on $\mathscr{E} = \{1, ..., d\}$.
- the sequences $(\eta_t)$ and $(\Delta_t)$ are independent.

# Simulations of length 100 of a HMM




3 regimes Model with $\eta_t \sim \mathcal{N}(0,1)$ and

$$\sigma(1) = 1, \quad \sigma(2) = 3, \quad \sigma(3) = 9, \qquad P = \begin{pmatrix} 0.85 & 0.1 & 0.05 \\ 0.3 & 0.7 & 0 \\ 0.3 & 0 & 0.7 \end{pmatrix}.$$

# Unconditional law

If $\eta_t \sim \mathcal{N}(0,1)$, the law of $\epsilon_t$ is a mixture of centered Gaussian distributions : its density is given by

$$f(x) = \sum_{i=1}^{d} \pi(i) \frac{1}{\sigma(i)} \phi\left(\frac{x}{\sigma(i)}\right).$$

The law is **not Gaussian** (except when $\sigma(i) = \sigma$ for all $i$).

For any law of $\eta_t$, the marginal moments of $\epsilon_t$ can be obtained : for any $r > 0$,

$$E(\epsilon_t^r) = E\sigma^r(\Delta_t) E(\eta_t^r) = \sum_{i=1}^{d} \sigma^r(i) \pi(i) E(\eta_t^r).$$

In particular, $E(\epsilon_t) = 0$.

# Correlations

We have

$$\text{Corr}(\epsilon_t, \epsilon_{t-k}) = 0, \quad \text{for all } k > 0.$$

Thus $(\epsilon_t)$ is a white noise with variance

$$E\epsilon_t^2 = \sum_{i=1}^{d} \sigma^2(i)\pi(i).$$

The dependence cannot be seen on the 2nd order structure.

## Correlations of squares : case $d=2$

- Eigenvalues of $P$ : 1 and $\lambda = p(1,1)+p(2,2)-1.$
- $-1<\lambda<1$ because the chain is irreducible and aperiodic.
- The entries of $P^k$ have the form

$$p^{(k)}(i,j) = a_1(i,j) + a_2(i,j)\lambda^k, \quad k \geq 0.$$

- We have $a_1(i,j) = \pi(j)$, and $a_1(i,j)+a_2(i,j) = \mathbf{1}_{\{i=j\}}$.
- Thus, for $j=1,2$ and $i \neq j$

$$p^{(k)}(i,j) = \pi(j)(1-\lambda^k), \quad p^{(k)}(j,j) = \pi(j) + \lambda^k(1-\pi(j)),$$

and, for $i,j=1,2,\ k\geq 0$

$$p^{(k)}(i,j) - \pi(j) = \lambda^k \left[\{1-\pi(j)\}\,\mathbf{1}_{\{i=j\}} - \pi(j)\,\mathbf{1}_{\{i\neq j\}}\right].$$

# Correlations of squares : case $d=2$

We have, for $k>0$,

$$
\begin{aligned}
\mathrm{Cov}(\epsilon_t^2,\epsilon_{t-k}^2) &= E\left\{\sigma^2(\Delta_t)\sigma^2(\Delta_{t-k})\right\}-\{E\sigma^2(\Delta_t)\}^2 \\
&= \sum_{i,j=1}^{2} p^{(k)}(i,j)\pi(i)\sigma^2(i)\sigma^2(j)-\left\{\sum_{i}^{2}\pi(i)\sigma^2(i)\right\}^2 \\
&= \sum_{i,j=1}^{2}\{p^{(k)}(i,j)-\pi(j)\}\pi(i)\sigma^2(i)\sigma^2(j) \\
&= \lambda^k\left\{\sum_{j=1}^{2}(1-\pi(j))\pi(j)\sigma^4(j)-\sum_{i\neq j}\pi(i)\pi(j)\sigma^2(i)\sigma^2(j)\right\} \\
&= \lambda^k\{\sigma^2(1)-\sigma^2(2)\}^2\pi(1)\pi(2).
\end{aligned}
$$

# Correlations of squares : case $d=2$

$$\mathrm{Cov}(\epsilon_t^2, \epsilon_{t-k}^2) \quad = \quad \lambda^k \{\sigma^2(1) - \sigma^2(2)\}^2 \pi(1)\pi(2), \qquad k > 0.$$

- Covariances have the sign of $\lambda^k$.
- The dependence increases with $|\lambda| = |p(1,1) + p(2,2) - 1|$ and with $|\sigma^2(1) - \sigma^2(2)|$.
- $(\epsilon_t^2)$ is an ARMA(1,1) process, since

$$\gamma_{\epsilon^2}(k) = \lambda \gamma_{\epsilon^2}(k-1), \quad k > 1.$$

A similar computation shows that

$$\mathrm{Var}(\epsilon_t^2) = \{\sigma^2(1) - \sigma^2(2)\}^2 \pi(1)\pi(2) + \{\sigma^4(1)\pi(1) + \sigma^4(2)\pi(2)\} \mathrm{Var}(\eta_t^2).$$

# Correlations of squares : general case

The transition matrix $P$ may not be diagonalizable but 1 remains an eigenvalue, with corresponding eigenspace of dimension 1 (otherwise there would exist several invariant distributions).
Let $\lambda_1,\ldots\lambda_m$ the eigenvalues different from 1 and $n_1,\ldots,n_m$ the dimensions of the corresponding eigenspace $(n_1+\cdots+n_m=d-1)$.
We use the Jordan representation

$$P = SJS^{-1}, \qquad \text{where } S \text{ is nonsingular,}$$

$$J = \begin{pmatrix} J_{n_1}(\lambda_1) & 0 & \ldots & & 0 \\ 0 & J_{n_2}(\lambda_2) & 0 & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & \ddots & J_{n_m}(\lambda_m) & 0 \\ 0 & \ldots & & 0 & 1 \end{pmatrix}, \qquad J_l(\lambda) = \lambda I_l + N_l(1)$$

where $N_l(i)$ the $l\times l$ matrix whose elements are null, except the $i$-th superdiagonal filled with 1.

# Correlations of squares : general case

We have $J_l^k(\lambda) = \lambda^k P^{(l)}(k)$, for some polynomial $P^{(l)}$ of degree $l-1$,

$$\Rightarrow \quad P^k = S \begin{pmatrix} \lambda_1^k P^{(n_1)}(k) & 0 & \dots & & 0 \\ 0 & \lambda_2^k P^{(n_2)}(k) & 0 & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \lambda_m^k P^{(n_m)}(k) & 0 \\ 0 & \dots & & 0 & 1 \end{pmatrix} S^{-1}.$$

It follows that

$$p^{(k)}(i,j) = \pi(j) + \sum_{l=1}^{m} \lambda_l^k p_{i,j}^{(n_l)}(k), \qquad d^o(p_{i,j}^{(n_l)}) = n_l - 1$$

and, for some polynomials $q^{(n_l)}(k)$ of degree $n_l - 1$,

$$\mathrm{Cov}(\epsilon_t^2, \epsilon_{t-k}^2) = \sum_{i,j=1}^{d} \sum_{l=1}^{m} \lambda_l^k p_{i,j}^{(n_l)}(k) := \sum_{l=1}^{m} \lambda_l^k q^{(n_l)}(k), \quad k > 0.$$

# Definition

The MS-ARMA$(p,q)$ model is defined by

$$\begin{cases} X_t &= c(\Delta_t) + \sum_{i=1}^{p} a_i(\Delta_t) X_{t-i} + \epsilon_t + \sum_{j=1}^{q} b_j(\Delta_t)\epsilon_{t-j}, \\ \\ \epsilon_t &= \epsilon_t(\Delta_t) = \sigma(\Delta_t)\eta_t, \qquad (\eta_t) \overset{iid}{\sim} (0,1), \end{cases}$$

where $(\Delta_t)$ is an ergodic (aperiodic, irreducible, stationary) Markov chain on $\mathscr{E} = \{1,2,...,d\}$, which is independent of $(\eta_t)$, $a_i(\cdot), b_j(\cdot), c(\cdot) \in \mathbb{R}$, $\sigma(\cdot) > 0$.

This model contains the HMM as a particular case.

Except in the case $p = 0$ (MS-MA$(q)$), the **existence of stationary solutions** require additional conditions.

# Notations

For any function $f : \mathscr{E} \to \mathscr{M}_{n\times n'}(\mathbb{R})$, where $\mathscr{M}_{n\times n'}(\mathbb{R})$ is the space of real $n \times n'$ matrices, and for all positive integers $i, n$ et $n'$, let

$$\mathbb{P}^{(i)}(f) = \begin{pmatrix} p^{(i)}(1,1)f(1) & \cdots & p^{(i)}(d,1)f(1) \\ \vdots & & \vdots \\ p^{(i)}(1,d)f(d) & \cdots & p^{(i)}(d,d)f(d) \end{pmatrix}, \qquad \Pi(f) = \begin{pmatrix} \pi(1)f(1) \\ \vdots \\ \pi(d)f(d) \end{pmatrix}$$

- When $i = 1$, write $\mathbb{P}(f) = \mathbb{P}^{(1)}(f)$,
- for $f \equiv 1$, let $\mathbb{P} = \mathbb{P}(1) = (p(j,i))$, the transpose of the transition matrix.

# A useful result for computing expectations

## Lemma

*Let $f_0,\dots,f_k$ functions $\mathscr{E} \mapsto \mathscr{M}_{n\times n}(\mathbb{R})$. For $k>0$,*

$$E\{f_0(\Delta_t)f_1(\Delta_{t-1})\dots f_k(\Delta_{t-k})\} = \boldsymbol{I}\mathbb{P}(f_0)\dots\mathbb{P}(f_{k-1})\Pi(f_k)$$

*where $\boldsymbol{I}=(I_n,\dots,I_n)$ is a $n\times nd$ matrix and $I_n$ is the identity matrix of size $n$.*

**Proof $(k=1)$.** We have

$$\begin{aligned} E\{f_0(\Delta_t)f_1(\Delta_{t-1})\} &= E[E\{f_0(\Delta_t)\mid \Delta_{t-1}\}f_1(\Delta_{t-1})] \\ &= \sum_{j=1}^{d} E\{f_0(\Delta_t)\mid \Delta_{t-1}=j\}f_1(j)\pi(j) \\ &= \sum_{j=1}^{d}\sum_{i=1}^{d} f_0(i)p(j,i)f_1(j)\pi(j) = \boldsymbol{I}\mathbb{P}(f_0)\Pi(f_1). \end{aligned}$$

# MS-AR(1) without intercept

$$X_t \quad = \quad a(\Delta_t)X_{t-i} + \sigma(\Delta_t)\eta_t$$

**Problems :**

- Existence of a strictly stationary solution.
- Existence of a 2nd-order stationary solution.

We are interested in **non-anticipative solutions**, i.e. solutions of the form

$$X_t = f(\eta_t, \Delta_t, \eta_{t-1}, \Delta_{t-1}, \ldots).$$

# MS-AR(1) without intercept

By successive replacements, for $k \geq 1$,

$$\begin{aligned} X_t &= a(\Delta_t)\dots a(\Delta_{t-k+1})X_{t-k} \\ &\quad + \sum_{i=0}^{k-1} a(\Delta_t)\dots a(\Delta_{t-i+1})\sigma(\Delta_{t-i})\eta_{t-i} \end{aligned}$$

(with by convention $a(\Delta_t)\dots a(\Delta_{t-i+1}) = 1$ if $i = 0$).

Solutions should be given by

$$\tilde{X}_t = \sum_{n=0}^{\infty} a(\Delta_t)\dots a(\Delta_{t-n+1})\sigma(\Delta_{t-n})\eta_{t-n},$$

**provided the series converges almost surely.**

# Cauchy root test

To derive an absolute convergence condition, we will use the $n$th root (Cauchy) test.*

Let $u_n = a(\Delta_t)\dots a(\Delta_{t-n+1})\sigma(\Delta_{t-n})\eta_{t-n}$. We have

$$|u_n|^{1/n} = \exp\left\{\frac{1}{n}\sum_{k=1}^{n}\log|a(\Delta_{t-k+1})| + \frac{1}{n}\log\{\sigma(\Delta_{t-n})|\eta_{t-n}|\}\right\}.$$

Because $\overline{\lim}\, n^{-1}\log\{\sigma(\Delta_{t-n})|\eta_{t-n}|\} = 0$ a.s., we have, by the ergodic theorem

$$\overline{\lim}\,|u_n|^{1/n} = \exp\{E\log|a(\Delta_t)|\}.$$

$\Rightarrow$ **Condition :** $\quad E\log|a(\Delta_t)| < 0.$

---

$*$. For a non-negative sequence $(u_n)$, the series $\sum u_n$ converges if $\overline{\lim}\sqrt[n]{u_n} < 1$.

# Uniqueness

Under the previous condition,

$$\tilde{X}_t = \sum_{n=0}^{\infty} a(\Delta_t)\dots a(\Delta_{t-n+1})\sigma(\Delta_{t-n})\eta_{t-n}, \quad a.s.$$

is well defined. Moreover it is a solution of the MS-AR(1) model. Now suppose there exists another strictly stationary solution $(X_t^*)$.

$$X_t^* \quad = \quad a(\Delta_t)\dots a(\Delta_{t-k+1})X_{t-k}^* + \sum_{i=0}^{k-1} a(\Delta_t)\dots a(\Delta_{t-i+1})\sigma(\Delta_{t-i})\eta_{t-i}.$$

Hence

$$|X_t - X_t^*| \quad \leq \quad |a(\Delta_t)\dots a(\Delta_{t-k+1})||X_{t-k}^*| + r_{t,k},$$

where $r_{t,k} \to 0$ and $|a(\Delta_t)\dots a(\Delta_{t-k+1})| \to 0$ a.s. as $k \to \infty$. Hence $X_t = X_t^*$ a.s.

# Strict stationarity condition

## Proposition

*There exists a unique strictly stationary solution if*

$$E \log |a(\Delta_t)| = \sum_{i=1}^{d} \log |a(i)| \pi(i) < 0.$$

*Moreover, this solution is non anticipative.*

# Strict stationarity condition

**Remarks :**

- $$\sum_{i=1}^{d} \log|a(i)|\pi(i) < 0 \quad \Leftrightarrow \quad \prod_{i=1}^{d} |a(i)|^{\pi(i)} < 1.$$
- The stationarity condition of an AR(1), $|a| < 1$, has to be satisfied "in average" over the different regimes. Regimes with $|a(i)| > 1$ are allowed (but they must not be visited too often).
- Only depends on the stationary probabilities of the Markov chain, not on the transition probabilities.
- If $a(i) = 0$ for at least one regime $i$, there is a stationary solution.

# Second-order stationarity of the MS-AR(1) model

**Problem :** existence in $L^2$ of

$$\tilde{X}_t = \sum_{n=0}^{\infty} r_{t,n}, \qquad \text{where} \quad r_{t,n} = a(\Delta_t)\dots a(\Delta_{t-n+1})\sigma(\Delta_{t-n})\eta_{t-n}.$$

Norm on $L^2$ : $\|X\| = \{E(X^2)\}^{1/2}$.

We have

$$\|\tilde{X}_t\|_{L^2} \leq \sum_{n=0}^{\infty} \|r_{t,n}\|_{L^2},$$

and

$$\begin{aligned} Er_{t,n}^2 &= Ea^2(\Delta_t)\dots a^2(\Delta_{t-n+1})\sigma^2(\Delta_{t-n}) \\ &= (1,\dots,1)\mathbb{P}^n(a^2)\Pi(\sigma^2). \end{aligned}$$

# Second-order stationarity of the MS-AR(1) model

**Proposition**

*If*

$$\rho\{\mathbb{P}(a^2)\} = \rho \begin{pmatrix} p(1,1)a^2(1) & \cdots & p(d,1)a^2(1) \\ \vdots & & \vdots \\ p(1,d)a^2(d) & \cdots & p(d,d)a^2(d) \end{pmatrix} < 1,$$

$(\tilde{X}_t)$ *is the unique 2nd-order stationary and non anticipative solution of the MS-AR(1) model.*

*Conversely, if* $\rho\{\mathbb{P}(a^2)\} \geq 1$ *there is no 2nd-order stationary and non anticipative solution.*

# Proof

- **Sufficient part :** We have, for some $K>0$,

$$(1,\dots,1)\mathbb{P}^n(a^2)\Pi(\sigma^2)\le K[\rho\{\mathbb{P}(a^2)\}]^i$$

- **Uniqueness :** same arguments as for the strict stationarity.
- **Necessary part :** Let $(X_t)$ a 2nd-order stationary, non anticipative solution. We have

$$\begin{aligned} EX_t^2 &= E\{a(\Delta_t)\dots a(\Delta_{t-k+1})X_{t-k}\}^2 + E\left\{\sum_{n=0}^{k-1} r_{t,n}\right\}^2 \\ &\ge \sum_{n=0}^{k-1} E\{a(\Delta_t)\dots a(\Delta_{t-n+1})\sigma(\Delta_{t-n})\}^2 E\eta_t^2, \qquad \forall k>0. \end{aligned}$$

Therefore, since $EX_t^2<\infty$,

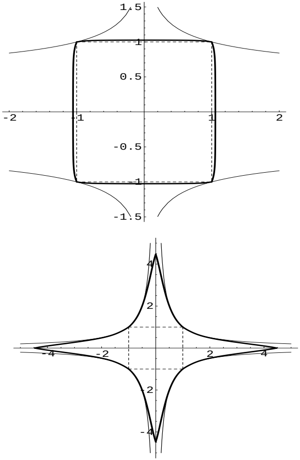
$$\sum_{n=0}^{\infty} E\{a(\Delta_t)\dots a(\Delta_{t-n+1})\sigma(\Delta_{t-n})\}^2 = \sum_{i=0}^{\infty}(1,\dots,1)\mathbb{P}^i(a^2)\Pi(\sigma^2)<\infty.$$

# Remarks

- The condition involves the transition probabilities $p(i,j)$ (not only the stationary probabilities $\pi(i)$).
- When the Markov chain is independent, we have $p(i,j) = p(k,j)$ for all $i,j,k$. Thus, the columns of $\mathbb{P}(a^2)$ are identical. The only non-zero eigenvalue is $\sum_{i=1}^{d} \pi(i)a^2(i)$ and the condition reduces to

$$\sum_{i=1}^{d} \pi(i)a^2(i) < 1.$$

# Stationarity regions when $d=2$ in the $(a(1),a(2))$ plane




$p(1,1)=0.8, p(2,2)=0.95$

$p(1,1)=p(2,2)=0.05$

# Autocovariance function

Under the 2nd-order stationarity condition, the autocovariance of $(X_t)$ can be explicitly computed.
We have

$$EX_t = \sum_{i=0}^{\infty} E\{a(\Delta_t)\dots a(\Delta_{t-i+1})\sigma(\Delta_{t-i})\}E\eta_{t-i} = 0,$$

and

$$\begin{aligned} EX_t^2 &= \sum_{i=0}^{\infty} E\{a(\Delta_t)\dots a(\Delta_{t-i+1})\sigma(\Delta_{t-i})\}^2 \\ &= \sum_{i=0}^{\infty} (1,\dots,1)\mathbb{P}^i(a^2)\Pi(\sigma^2) = (1,\dots,1)\{I_d - \mathbb{P}(a^2)\}^{-1}\Pi(\sigma^2). \end{aligned}$$

# Autocovariance function

For $h > 0$,

$$
\begin{aligned}
EX_tX_{t-h} &= \sum_{i=0}^{\infty} E\{a(\Delta_t)\dots a(\Delta_{t-h+1})a^2(\Delta_{t-h})\dots a^2(\Delta_{t-h-i+1})\sigma^2(\Delta_{t-h-i})\} \\
&= \sum_{i=0}^{\infty}(1,\dots,1)\mathbb{P}^h(a)\mathbb{P}^i(a^2)\Pi(\sigma^2) \\
&= (1,\dots,1)\mathbb{P}^h(a)\{I_d - \mathbb{P}(a^2)\}^{-1}\Pi(\sigma^2).
\end{aligned}
$$

# Example with $d = 2$

$$X_t = \begin{cases} \epsilon_t & \text{f} \quad \Delta_t = 1, \\ aX_{t-1} + \sigma\epsilon_t & \text{if} \quad \Delta_t = 2. \end{cases}$$

The model admits a strictly stationary solution whatever $a$.
We have

$$\mathbb{P}(a^2) = \begin{pmatrix} 0 & 0 \\ p(1,2)a^2 & p(2,2)a^2 \end{pmatrix},$$

thus $\rho\{\mathbb{P}(a^2)\} = p(2,2)a^2$ and the 2nd-order stationarity condition is

$$p(2,2)a^2 < 1.$$

Under this condition

$$EX_t^2 = \frac{\{1 - a^2 p(2,2) + a^2 p(1,2)\}\pi(1) + \sigma^2 \pi(2)}{1 - a^2 p(2,2)}.$$

## Example with $d=2$

We have $\mathbb{P}^h(a)=\{ap(2,2)\}^{h-1}\mathbb{P}(a)$ for all $h>0$. Therefore

$$\gamma(h):=EX_tX_{t-h}=\frac{p(1,2)\pi(1)+\sigma^2p(2,2)\pi(2)}{1-a^2p(2,2)}a^hp^{h-1}(2,2),\quad h>0.$$

Note that

$$\gamma(h)=ap(2,2)\gamma(h-1),\qquad h>1,$$

which entails that $X_t$ admits an ARMA representation, of the form

$$X_t-ap(2,2)X_{t-1}=u_t-\theta u_{t-1}$$

where $(u_t)$ is a white noise and $\theta$ is a coefficient (which can be obtained from $\gamma(0)$ and $\gamma(1)$).

This property, existence of an ARMA representation, is general to MS-ARMA processes.

# Markov representation : $\boldsymbol{Z}_t = \boldsymbol{A}_t \boldsymbol{Z}_{t-1} + \boldsymbol{B}_t$

$$\boldsymbol{B}_t = \boldsymbol{C}_t + \underline{\epsilon}_t = \boldsymbol{C}_t + \boldsymbol{\Sigma}_t \eta_t,$$

$$\boldsymbol{C}_t = \begin{pmatrix} c(\Delta_t) \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{p+q}, \quad \boldsymbol{Z}_t = \begin{pmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-p+1} \\ \epsilon_t \\ \epsilon_{t-1} \\ \vdots \\ \epsilon_{t-q+1} \end{pmatrix} \in \mathbb{R}^{p+q}, \quad \boldsymbol{\Sigma}_t = \begin{pmatrix} \sigma(\Delta_t) \\ 0 \\ \vdots \\ 0 \\ \sigma(\Delta_t) \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

$$\boldsymbol{A}_t = \begin{pmatrix} \boldsymbol{a}(\Delta_t) & \boldsymbol{b}(\Delta_t) \\ \boldsymbol{0} & \boldsymbol{J} \end{pmatrix} \in \mathscr{M}_{(p+q)\times(p+q)}(\mathbb{R})$$

# Markov representation : $\boldsymbol{Z}_t = \boldsymbol{A}_t \boldsymbol{Z}_{t-1} + \boldsymbol{B}_t$

$$\boldsymbol{a}(\Delta_t) = \begin{pmatrix} a_1(\Delta_t) & & \cdots & & a_p(\Delta_t) \\ 1 & 0 & \cdots & & 0 \\ 0 & 1 & \cdots & & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & & \ldots & 1 & 0 \end{pmatrix},$$

$$\boldsymbol{b}(\Delta_t) = \begin{pmatrix} b_1(\Delta_t) & & \cdots & & b_q(\Delta_t) \\ 0 & 0 & \cdots & & 0 \\ 0 & 0 & \cdots & & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & & \ldots & 1 & 0 \end{pmatrix}, \quad \boldsymbol{J} = \begin{pmatrix} 0 & 0 & \cdots & & 0 \\ 1 & 0 & \cdots & & 0 \\ 0 & 1 & \cdots & & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix}$$

# Top-Lyapunov exponent

The **top-Lyapunov exponent** of the sequence $\{\boldsymbol{a}(\Delta_t)\}$ is defined by

$$\begin{aligned}\gamma &= \inf_{t>0} E\left(\frac{1}{t}\log\|\boldsymbol{a}(\Delta_t)\boldsymbol{a}(\Delta_{t-1})\cdots\boldsymbol{a}(\Delta_1)\|\right)\\ &\overset{a.s.}{=} \lim_{t\to\infty}\frac{1}{t}\log\left\|\prod_{i=1}^{t}\boldsymbol{a}(\Delta_{t-i})\right\|\end{aligned}$$

for any norm on $\mathscr{M}_{p\times p}(\mathbb{R})$.

# Strict stationarity condition

## Proposition

*Suppose $\gamma_a < 0$. Then, for any $t \in \mathbb{Z}$, the series*

$$\boldsymbol{Z}_t = \boldsymbol{B}_t + \sum_{k=1}^{\infty} \boldsymbol{A}_t \cdots \boldsymbol{A}_{t-k+1} \boldsymbol{B}_{t-k}$$

*converges a.s. and the process $(\boldsymbol{X}_t)$, defined as the first component of $(\boldsymbol{Z}_t)$, is the **unique strictly stationary solution** of the MS-ARMA$(p,q)$ model.*

# Remarks

- The strict stationarity condition only depends on the coefficients $a_j(k)$ (as for standard ARMA).
- In particular for a MS-ARMA$(1,q)$, the condition reduces to

$$\sum_{i=1}^{d} \pi(i) \log|a_1(i)| < 0.$$

- The condition is only sufficient in general. The coefficients $b_j(k)$ and $c(k)$ may matter for the strict stationarity.

# Second-order stationarity

Let $\otimes$ the Kronecker product. For any matrix function $f$ defined on $\{1,\ldots,d\}$, let $f^{\otimes 2}(k) = f(k) \otimes f(k)$.

**Proposition**

*If*

$$\rho\{\mathbb{P}(a^{\otimes 2})\} < 1,$$

*the strict stationary solution is also second-order stationary.*

It can be shown that this condition is also necessary in the case $p = q = 1$. However this is not general.

# Example 1 : A two-regime MS-$AR$(2)

$$X_t = \begin{cases} \eta_t & \text{if} \quad \Delta_t = 1 \\ \eta_t + aX_{t-2} & \text{if} \quad \Delta_t = 2 \end{cases}$$

Then

$$\mathbb{P}(A^{\otimes 2}) = \begin{pmatrix} 0_2 & 0_2 & 0_2 & 0_2 \\ P_1 & 0_2 & P_2 & 0_2 \\ 0_2 & aP_3 & 0_2 & aP_4 \\ P_3 & 0_2 & P_4 & 0_2 \end{pmatrix}$$

where $0_2$ is the null $2 \times 2$ matrix and

$$P_1 = \begin{pmatrix} 0 & 0 \\ p(1,1) & 0 \end{pmatrix}, \quad P_2 = \begin{pmatrix} 0 & 0 \\ p(2,1) & 0 \end{pmatrix},$$

$$P_3 = \begin{pmatrix} 0 & ap(1,2) \\ p(1,2) & 0 \end{pmatrix}, \quad P_4 = \begin{pmatrix} 0 & ap(2,2) \\ p(2,2) & 0 \end{pmatrix}.$$

# Example 1 : A two-regime MS-$AR$(2)

We have

$$\rho(\{\mathbb{P}(a^{\otimes 2})\}) = \max\left\{|a|p(2,2)^{1/2}, |a|\left\{p(2,2)^2 + p(1,2)p(2,1)\right\}^{1/2}\right\}.$$

But the expansion

$$X_t = \eta_t + \sum_{k=1}^{\infty} a^k \eta_{t-2k} \mathbf{1}_{\Delta_t=2,\dots,\Delta_{t-2k+2}=2}$$

shows that the necessary and sufficient 2nd-order stationarity condition is simply

$$|a|\left\{p(2,2)^2 + p(1,2)p(2,1)\right\}^{1/2} < 1,$$

which shows that **the condition $\rho(\{\mathbb{P}(a^{\otimes 2})\}) < 1$ may be too strong**.

# Example 2 : local stationarity is neither necessary nor sufficient

It is clear (see for instance Example 1) that the **local stationarity** (i.e. within each regime) is not necessary to ensure the (global) 2nd-order stationarity.

The next example also shows that it is not sufficient.

$$X_t = \begin{cases} a_1(1)X_{t-1} + a_2(1)X_{t-2} + \eta_t & \text{if} \quad \Delta_t = 1 \\ a_1(2)X_{t-1} + \eta_t & \text{if} \quad \Delta_t = 2 \end{cases}$$

where $(\eta_t)$ is i.i.d. $\mathcal{N}(0,1)$.

If $(X_t)$ is 2nd-order stationary then $E(X_t^2|\Delta_t = 1, \Delta_{t-1} = 2)$ exists and is independent of $t$.

# Example 2 (continued)

Thus

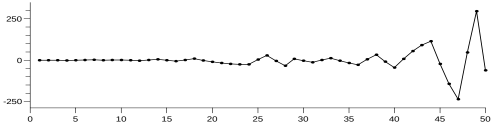
$$
\begin{aligned}
& E(X_t^2 | \Delta_t = 1, \Delta_{t-1} = 2) \\
= \quad & E\Big( \big[ \{a_1(1)a_1(2) + a_2(1)\} X_{t-2} + \eta_t + a_1(1)\eta_{t-1} \big]^2 | \Delta_t = 1, \Delta_{t-1} = 2 \Big) \\
\geq \quad & \{a_1(1)a_1(2) + a_2(1)\}^2 E\left( X_{t-2}^2 | \Delta_t = 1, \Delta_{t-1} = 2 \right) \\
\geq \quad & \{a_1(1)a_1(2) + a_2(1)\}^2 E\left( X_{t-2}^2 | \Delta_t = 1, \Delta_{t-1} = 2, \Delta_{t-2} = 1, \Delta_{t-3} = 2 \right) \\
& \times P(\Delta_{t-2} = 1, \Delta_{t-3} = 2 | \Delta_t = 1, \Delta_{t-1} = 2) \\
= \quad & \{a_1(1)a_1(2) + a_2(1)\}^2 p(2,1)p(1,2) E\left( X_{t-2}^2 | \Delta_{t-2} = 1, \Delta_{t-3} = 2 \right)
\end{aligned}
$$

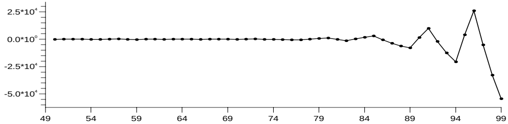
This is not possible when

$$\{a_1(1)a_1(2) + a_2(1)\}^2 p(2,1)p(1,2) > 1.$$

But this condition is compatible with the local stationarity.

# Example 2 (continued) : simulation




For instance if $a_1(1) = 1.8,\ a_2(1) = -0.9,\ a_1(2) = -0.2,\ p(1,1) = 0.2$ and $p(2,2) = 0.1$.

# MS-AR$(p)$ model

$$X_t = \sum_{i=1}^{p} a_i(\Delta_t) X_{t-i} + \sigma(\Delta_t)\eta_t, \qquad (\eta_t) \overset{iid}{\sim} \mathcal{N}(0,1),$$

**Parameter vector :**

$$\boldsymbol{\theta} = (p(1,1),\dots,p(1,d-1),p(2,1),\dots,p(d,d-1),\sigma(1),\dots,\sigma(d))'.$$

The likelihood can be written by conditioning with respect to all possible paths

$$(e_1,\dots,e_n)$$

of the Markov chain, where $e_i \in \mathcal{E} = \{1,\dots,d\}$. The probability of this path is

$$P(e_1,\dots,e_n) = P(\Delta_1 = e_1,\dots,\Delta_n = e_n) = \pi(e_1)p(e_1,e_2)\dots p(e_{n-1},e_n).$$

# Likelihood

For any path, we a have a likelihood given by

$$L^{(e_1,\ldots,e_n)}(X_1,\ldots,X_n) = \prod_{t=1}^{n} \phi_{e_t}\left(X_t - \sum_{i=1}^{p} a_i(e_t)X_{t-i}\right),$$

where $\phi_i(\cdot)$ is the density of $\mathcal{N}\left\{0,\sigma^2(i)\right\}$.

Finally the likelihood of the observations is

$$L_{\boldsymbol{\theta}}(X_1,\ldots,X_n) = \sum_{(e_1,\ldots,e_n)\in\mathcal{E}^n} L_{\sigma}^{(e_1,\ldots,e_n)}(X_1,\ldots,X_n)P(e_1,\ldots,e_n).$$

However, such a formula **cannot be used** in practice.

Several algorithms can be used to compute the likelihood.

# Algorithm based on a matrix product

**Idea : decompose the likelihood by conditioning on the last state**

$$L_{\boldsymbol{\theta}}(X_1,\ldots,X_n) = \sum_{i=1}^{d} L_{\boldsymbol{\theta}}(X_1,\ldots,X_n \mid \Delta_n = i)\pi(i).$$

Let

$$F_k(i) = g_k(X_1,\ldots,X_k|\Delta_k = i)\pi(i)$$

where $g_k(\cdot|\Delta_k = i)$ is the density of $(X_1,\ldots,X_k)$ given $\{\Delta_k = i\}$.

We have (with initial values $X_0 = X_{-1} = \ldots = X_{1-p} = 0$)

$$F_1(i) \quad = \quad g_1(X_1 \mid \Delta_1 = i)\pi(i) = \phi_i(X_1)\pi(i).$$

# Algorithm based on a matrix product

$$
\begin{aligned}
F_k(i) &= g(X_k|X_1,\ldots,X_{k-1},\Delta_k=i)\,g_{k-1}(X_1,\ldots,X_{k-1}|\Delta_k=i)\pi(i) \\
&= \phi_i\left(X_k-\sum_{\ell=1}^{p}a_\ell(i)X_{k-\ell}\right)\sum_{j=1}^{d}g_{k-1}(X_1,\ldots,X_{k-1}|\Delta_{k-1}=j,\Delta_k=i) \\
&\qquad\qquad\qquad\qquad \times P(\Delta_{k-1}=j\,|\,\Delta_k=i)\pi(i) \\
&= \frac{1}{\sigma(i)}\phi\{\eta_k(i)\}\sum_{j=1}^{d}F_{k-1}(j)p(j,i)
\end{aligned}
$$

where

$$
\eta_k(i)=\frac{1}{\sigma(i)}\left(X_k-\sum_{\ell=1}^{p}a_\ell(i)X_{k-\ell}\right).
$$

# Algorithm based on a matrix product

Finally, the algorithm is given by

$$
\begin{aligned}
F_1(i) &= \phi_i(X_1)\pi(i), \quad i=1,\dots,d, \\
F_k(i) &= \frac{1}{\sigma(i)}\phi\{\eta_k(i)\}\sum_{j=1}^{d} F_{k-1}(j)p(j,i), \qquad i=1,\dots,d, \quad k>1
\end{aligned}
$$

and the likelihood is obtained as

$$
L_\theta(X_1,\dots,X_n) = \sum_{i=1}^{d} F_n(i).
$$

## Algorithm based on a matrix product

In matrix form :

$$F_k := (F_k(1),\dots,F_k(d))' = M(X_k,\dots,X_{k-p})F_{k-1},$$

where

$$M(X_k,\dots,X_{k-p}) = \begin{pmatrix} p(1,1)\frac{\phi(\eta_k(1))}{\sigma(1)} & \cdots & p(d,1)\frac{\phi(\eta_k(1))}{\sigma(1)} \\ \vdots & & \vdots \\ p(1,d)\frac{\phi(\eta_k(d))}{\sigma(d)} & \cdots & p(d,d)\frac{\phi(\eta_k(d))}{\sigma(d)} \end{pmatrix}.$$

Hence

$$L_\theta(X_1,\dots,X_n) = \mathbf{1}'M(X_n,\dots,X_{n-p})M(X_{n-1},\dots,X_{n-1-p})\cdots M(X_2,\dots,X_{2-p})F_1.$$

which is numerically tractable ($O(d^2n)$ multiplications).

However, the algorithm can be unstable if $n$ is large (underflows due to a very small likelihood).

# Forward-Backward Algorithm [Baum (1972)]

Let

$$B_k(i) = g_{n-k}(X_{k+1},\ldots,X_n|\Delta_k = i, X_1,\ldots,X_k).$$

We have

$$\begin{aligned} L_{\boldsymbol{\theta}}(X_1,\ldots,X_n) &= \sum_{i=1}^{d} L_{\boldsymbol{\theta}}(X_1,\ldots,X_n \mid \Delta_k = i)\pi(i) \\ &= \sum_{i=1}^{d} B_k(i) F_k(i). \end{aligned}$$

Forward formulas allow to compute $F_k(i)$ for $k = 1,2,\ldots$.

**Backward formulas** allow to compute $B_k(i)$ for $k = n-1, n-2, \ldots$

# Forward-Backward Algorithm

$$
\begin{aligned}
B_k(i) &= \sum_{j=1}^{d} g_{n-k}(X_{k+1},\ldots,X_n|\Delta_{k+1}=j,\cancel{\Delta_k=i},X_1,\ldots,X_k)P(\Delta_{k+1}=j\mid\Delta_k=i)\\
&= \sum_{j=1}^{d} g_{n-k-1}(X_{k+2},\ldots,X_n|\Delta_{k+1}=j,X_1,\ldots,X_k,X_{k+1})\\
&\qquad\qquad \times g(X_{k+1}|\Delta_{k+1}=j,X_1,\ldots,X_k)P(\Delta_{k+1}=j\mid\Delta_k=i)\\
&= \sum_{j=1}^{d} B_{k+1}(j)\frac{\phi\{\eta_{k+1}(j)\}}{\sigma(j)}p(i,j)
\end{aligned}
$$

# Forward-Backward Algorithm

## Backward formulas :

$$
\begin{aligned}
B_n(i) &= 1, \\
B_k(i) &= \sum_{j=1}^{d} B_{k+1}(j) \frac{\phi\{\eta_{k+1}(j)\}}{\sigma(j)} p(i,j), \quad k < n.
\end{aligned}
$$

Compare with the Forward formulas :

$$
\begin{aligned}
F_1(i) &= \phi_i(X_1)\pi(i), \\
F_k(i) &= \frac{\phi\{\eta_k(i)\}}{\sigma(i)} \sum_{j=1}^{d} F_{k-1}(j) p(j,i), \quad k > 1.
\end{aligned}
$$

For any $k = 1, \ldots, n$, the likelihood is given by

$$
L_{\boldsymbol{\theta}}(X_1, \ldots, X_n) = \sum_{i=1}^{d} B_k(i) F_k(i).
$$

# Hamilton filter

Introduced by Hamilton (1989, Econometrica).

Main differences with respect to the previous algorithms :

- Based on the log-likelihood
- Provides filtered probabilities of the regimes

We have (neglecting the distribution of $X_1$) :

$$\log L_\theta(X_1,\ldots,X_n) = \sum_{t=1}^{n} \log f_t(X_t \mid X_{t-1},\ldots,X_1),$$

where

$$f_t(X_t \mid X_{t-1},\ldots,X_1) = \sum_{j=1}^{d} f_t(X_t \mid X_{t-1},\ldots,X_1,\Delta_t = j) P(\Delta_t = j \mid X_{t-1},\ldots,X_1).$$

# Hamilton filter

Let

$$\begin{aligned}
\pi_{t|t-1}(j) &= P(\Delta_t = j \,|\, X_{t-1}, \ldots, X_1), \\
\pi_{t|t}(j) &= P(\Delta_t = j \,|\, X_t, \ldots, X_1).
\end{aligned}$$

We have (here $g$ denotes a generic density)

$$\begin{aligned}
\pi_{t+1|t}(j) &= \sum_{j=1}^{d} P(\Delta_{t+1} = j \,|\, \Delta_t = i, X_t, \ldots, X_1)\pi_{t|t}(i) = \sum_{j=1}^{d} p(i,j)\pi_{t|t}(i), \\
\pi_{t|t}(j) &= \frac{g(X_t, \ldots, X_1 \,|\, \Delta_t = j)\pi(j)}{g(X_t, \ldots, X_1)},
\end{aligned}$$

using the formula $P(A|X=x) = \frac{f(x|A)}{f(x)} P(A)$.

# Hamilton filter

$$
\begin{aligned}
\pi_{t|t}(j) &= \frac{g(X_t,\ldots,X_1 \mid \Delta_t = j)\pi(j)}{g(X_t,\ldots,X_1)} \\
&= \frac{\frac{\phi\{\eta_t(j)\}}{\sigma(j)} g(X_{t-1},\ldots,X_1 \mid \Delta_t = j)\pi(j)}{g(X_t,\ldots,X_1)} \\
&= \frac{\frac{\phi\{\eta_t(j)\}}{\sigma(j)} P(\Delta_t = j \mid X_{t-1},\ldots,X_1) g(X_{t-1},\ldots,X_1)}{g(X_t \mid X_{t-1},\ldots,X_1) g(X_{t-1},\ldots,X_1)} \\
&= \frac{\frac{\phi\{\eta_t(j)\}}{\sigma(j)} \pi_{t|t-1}(j)}{g(X_t \mid X_{t-1},\ldots,X_1)} \\
&= \frac{\frac{\phi\{\eta_t(j)\}}{\sigma(j)} \pi_{t|t-1}(j)}{\sum_{i=1}^{d} \frac{\phi\{\eta_t(i)\}}{\sigma(i)} \pi_{t|t-1}(i)}
\end{aligned}
$$

# Hamilton filter

Finally, the sequences

$$\pi_{t|t-1}(j) = P(\Delta_t = j \mid X_{t-1}, \ldots, X_1), \quad \pi_{t|t}(j) = P(\Delta_t = j \mid X_t, \ldots, X_1)$$

are obtained recursively from

$$\begin{cases} \pi_{t|t}(j) = \dfrac{\frac{\phi\{\eta_t(j)\}}{\sigma(j)}\pi_{t|t-1}(j)}{\sum_{i=1}^d \frac{\phi\{\eta_t(i)\}}{\sigma(i)}\pi_{t|t-1}(i)}, \\ \pi_{t+1|t}(j) = \sum_{j=1}^d p(i,j)\pi_{t|t}(i), \end{cases}$$

with initial values $\pi_{1|0}(i) = \pi(i)$ (or any other distribution).

Numerically, the filter generates less underflows and tends to perform better than the (standard) forward-backward algorithm.

# Hamilton filter in matrix form

Let

$$\boldsymbol{\pi}_{t|t} = \begin{pmatrix} \pi_{t|t}(1) \\ \vdots \\ \pi_{t|t}(d) \end{pmatrix}, \quad \boldsymbol{\pi}_{t+1|t} = \begin{pmatrix} \pi_{t+1|t}(1) \\ \vdots \\ \pi_{t+1|t}(d)) \end{pmatrix}, \quad \Phi_t = \begin{pmatrix} \frac{\phi\{\eta_t(1)\}}{\sigma(1)} \\ \vdots \\ \frac{\phi\{\eta_t(d)\}}{\sigma(d)}) \end{pmatrix}.$$

$\odot$ : Hadamard product (componentwise).

We have

$$\pi_{t|t} = \frac{\pi_{t|t-1} \odot \Phi_t}{(1,\ldots,1)\left\{\pi_{t|t-1} \odot \Phi_t\right\}}, \qquad \pi_{t+1|t} = \mathbb{P}\pi_{t|t}$$

# EM algorithm

Maximisation of the (log-)likelihood can be achieved either using a classical optimization procedure, or using the **EM (Expectation–Maximization) algorithm.**

The EM algorithm, proposed by Dempster, Laird and Rubin (1977) † is used to find ML estimators of parameters in general statistical models involving latent variables.

†. Maximum Likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* 39, 1–38.

# EM algorithm for general models with incomplete data

Parameter of interest : $\boldsymbol{\theta}$

Likelihood inference based on the observed data, $Y$, is **intractable**.

Likelihood inference based on a completed data set, $(X,Y)$, becomes **tractable**.

The log-likelihood of the observation $Y$ can be decomposed as

$$\log \ell_{\boldsymbol{\theta}}(Y) = \log \ell_{\boldsymbol{\theta}}(X,Y) - \log \ell_{\boldsymbol{\theta}}(X|Y).$$

Multiplying both sides by $\ell_{\tilde{\boldsymbol{\theta}}}(X|Y)$ and integrating w.r.t. $X$ yields

$$\log \ell_{\boldsymbol{\theta}}(Y) = Q(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) - H(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}})$$

where

$$Q(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) = E_{\tilde{\boldsymbol{\theta}}}\left[\log \ell_{\boldsymbol{\theta}}(X,Y)|Y\right] \quad \text{and} \quad H(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) = E_{\tilde{\boldsymbol{\theta}}}\left[\log \ell_{\boldsymbol{\theta}}(X|Y)|Y\right].$$

# EM algorithm for general models with incomplete data

The difference

$$H(\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\theta}}) - H(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) = -E_{\tilde{\boldsymbol{\theta}}}\left[\log \frac{\ell_{\boldsymbol{\theta}}(X|Y)}{\ell_{\tilde{\boldsymbol{\theta}}}(X|Y)}|Y\right]$$

is the Kullback-Leibler divergence between the conditional distributions $\ell_{\boldsymbol{\theta}}(X|Y)$ and $\ell_{\tilde{\boldsymbol{\theta}}}(X|Y)$.

Thus,

$$\log \ell_{\boldsymbol{\theta}}(Y) - \log \ell_{\tilde{\boldsymbol{\theta}}}(Y) \geq Q(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\theta}}),$$

where the inequality is strict unless if $\ell_{\tilde{\boldsymbol{\theta}}}(\cdot|Y) = \ell_{\boldsymbol{\theta}}(\cdot|Y)$, a.e. Moreover, under regularity conditions,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log \ell_{\boldsymbol{\theta}}(Y) = \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) \mid_{\tilde{\boldsymbol{\theta}}=\boldsymbol{\theta}} .$$

# EM algorithm for general models with incomplete data

These results suggest that $\boldsymbol{\theta} \mapsto Q(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}})$ can be used as a surrogate for the log-likelihood function $\log \ell_{\boldsymbol{\theta}}(Y)$.

The EM algorithm uses this idea to maximize the (incomplete) likelihood $\log \ell_{\boldsymbol{\theta}}(Y)$, by iteratively maximizing the auxiliary function $\boldsymbol{\theta} \mapsto Q(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}})$.

Starting from an initial value $\boldsymbol{\theta}_0$, the procedure iterates between 2 steps for computing a new parameter value $\boldsymbol{\theta}^{(k)}$ from $\boldsymbol{\theta}^{(k-1)}$ :

- **(E)** Compute $\boldsymbol{\theta} \mapsto Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(k-1)})$ ;
- **(M)** Compute $\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(k-1)})$.

# EM algorithm for general models with incomplete data

In view of

$$\log \ell_{\boldsymbol{\theta}}(Y) - \log \ell_{\tilde{\boldsymbol{\theta}}}(Y) \geq Q(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}),$$

the sequence $\{\log \ell_{\boldsymbol{\theta}^{(k)}}(Y)\}$ of log-likelihoods is **non decreasing.**

Under appropriate regularity conditions, it can be shown that **the sequence $(\boldsymbol{\theta}^{(k)})$ converges** to some value $\boldsymbol{\theta}^*$ (though it may be a local maximum of the likelihood).

# EM algorithm for the HMM

We only consider the **HMM case** :

$$\epsilon_t = \sigma(\Delta_t)\eta_t.$$

Observations : $\epsilon_1, \ldots, \epsilon_n$

Parameters : $\boldsymbol{\theta}$ and the initial distribution $\boldsymbol{\pi}_0 = (\pi_0(1), \ldots, \pi_0(d))$.

Recall that likelihood of the observations is

$$L_{\boldsymbol{\theta},\boldsymbol{\pi}_0}(\epsilon_1, \ldots, \epsilon_n) = \sum_{(e_1,\ldots,e_n)\in\mathscr{E}^n} L_\sigma^{(e_1,\ldots,e_n)}(\epsilon_1, \ldots, \epsilon_n) P(e_1, \ldots, e_n).$$

**If**, in addition, **one could observe** $(\Delta_1, \ldots, \Delta_n)$, $\boldsymbol{\theta}$ and $\boldsymbol{\pi}_0$ could be easily estimated by ML.

# Maximizing the likelihood as if the chain was observed

Indeed

$$\begin{aligned}
& \log L_{\boldsymbol{\theta},\boldsymbol{\pi}_0}(\epsilon_1,\ldots,\epsilon_n,\Delta_1,\ldots,\Delta_n) \\
= \; & \sum_{t=1}^{n} \log \phi_{\Delta_t}(\epsilon_t) + \log \pi_0(\Delta_1) + \sum_{t=2}^{n} \log p(\Delta_{t-1},\Delta_t) = a_1 + a_2 + a_3
\end{aligned}$$

where

$$\begin{aligned}
a_1 &= a_1(\sigma^2) = \sum_{i=1}^{d}\sum_{t=1}^{n} \log \phi_i(\epsilon_t)\,\mathbf{1}_{\{\Delta_t = i\}}, \\
a_2 &= a_2(\boldsymbol{\pi}_0) = \log \pi_0(\Delta_1), \\
a_3 &= a_3(P) = \sum_{i=1}^{d}\sum_{j=1}^{d} \log p(i,j) \sum_{t=2}^{n} \mathbf{1}_{\{\Delta_{t-1}=i,\Delta_t=j\}}.
\end{aligned}$$

# Maximizing the likelihood as if the chain was observed

- Maximization of

$$a_1(\sigma^2) = \sum_{i=1}^{d} \sum_{t=1}^{n} \log\left\{ \frac{1}{\sigma(i)} \phi\left( \frac{\epsilon_t}{\sigma(i)} \right) \right\} \mathbf{1}_{\{\Delta_t = i\}}$$

with respect to the $\sigma^2(i)$, yields the « estimators »

$$\tilde{\sigma}^2(i) = \frac{1}{\sum_{t=1}^{n} \mathbf{1}_{\{\Delta_t=i\}}} \sum_{t=1}^{n} \epsilon_t^2 \mathbf{1}_{\{\Delta_t=i\}}.$$

- Maximization of $a_2$ in $\pi_0(1),\dots,\pi_0(d)$, under the constraint $\sum_{i=1}^{d} \pi_0(i) = 1$, yields

$$\tilde{\pi}_0(i) = \mathbf{1}_{\{\Delta_1=i\}}.$$

# Maximizing the likelihood as if the chain was observed

- For $i=1,\dots,d$ : maximization in $p(i,1),\dots,p(i,d)$, under the constraint $\sum_{j=1}^{d} p(i,j)=1$, of

$$\sum_{j=1}^{d} \log p(i,j) \frac{\sum_{t=2}^{n} \mathbf{1}_{\{\Delta_{t-1}=i,\Delta_t=j\}}}{\sum_{t=2}^{n} \mathbf{1}_{\{\Delta_{t-1}=i\}}}$$

yields

$$\tilde{p}(i,j) = \frac{1}{\sum_{t=2}^{n} \mathbf{1}_{\{\Delta_t=i\}}} \sum_{t=2}^{n} \mathbf{1}_{\{\Delta_{t-1}=i,\Delta_t=j\}}.$$

Indeed, if $p_1,\dots,p_n$ are positive numbers such that $\sum_i p_i = 1$, under the constraint $\sum_{i=1}^{d} \pi_i = 1$, the global maximum of the function $(\pi_1,\dots,\pi_d) \to \sum_i p_i \log \pi_i$ is reached at $(\pi_1,\dots,\pi_d) = (p_1,\dots,p_d)$.

# EM algorithm : E-step

In practice, $(\Delta_t)$ is not observed and the previous estimators cannot be used.

The idea is to **replace the quantities depending on the MC** by their **expectation**.

Suppose that at some step $k$, we have an estimator $(\theta^{(k)}, \pi_0^{(k)})$.

The unknown likelihood is approximated by its expectation given the observations $(\epsilon_1, \ldots, \epsilon_n)$, computed under the law of parameter $(\theta^{(k)}, \pi_0^{(k)})$.

# EM algorithm : E-step

We get the criterion

$$
\begin{aligned}
Q(\theta,\pi_0|\theta^{(k)},\pi_0^{(k)}) &= E_{\theta^{(k)},\pi_0^{(k)}}\left\{\log L_{\theta,\pi_0}(\epsilon_1,\dots,\epsilon_n,\Delta_1,\dots,\Delta_n)|\epsilon_1,\dots,\epsilon_n\right\} \\
&= A_1(\sigma)+A_2(\pi_0)+A_3(P),
\end{aligned}
$$

where

$$
\begin{aligned}
A_1(\sigma) &= \sum_{i=1}^{d}\sum_{t=1}^{n}\log\phi_i(\epsilon_t)P_{\theta^{(k)},\pi_0^{(k)}}\{\Delta_t=i|\epsilon_1,\dots,\epsilon_n\}, \\
A_2(\pi_0) &= \sum_{i=1}^{d}\log\pi_0(i)P_{\theta^{(k)},\pi_0^{(k)}}\{\Delta_1=i|\epsilon_1,\dots,\epsilon_n\}, \\
A_3(P) &= \sum_{i,j}\log p(i,j)\sum_{t=2}^{n}P_{\theta^{(k)},\pi_0^{(k)}}\{\Delta_{t-1}=i,\Delta_t=j|\epsilon_1,\dots,\epsilon_n\}.
\end{aligned}
$$

# M-step

In this step we solve

$$\max_{(\theta,\pi_0)} Q(\theta,\pi_0|\theta^{(k)},\pi_0^{(k)}).$$

- $$\hat{\sigma}^2(i) = \frac{\sum_{t=1}^n \epsilon_t^2 P_{\theta^{(k)},\pi_0^{(k)}}\{\Delta_t = i|\epsilon_1,\ldots,\epsilon_n\}}{\sum_{t=1}^n P_{\theta^{(k)},\pi_0^{(k)}}\{\Delta_t = i|\epsilon_1,\ldots,\epsilon_n\}}.$$

- $$\hat{\pi}_0(i) = P_{\theta^{(k)},\pi_0^{(k)}}\{\Delta_1 = i|\epsilon_1,\ldots,\epsilon_n\},$$

- $$\hat{p}(i,j) = \frac{\sum_{t=2}^n P_{\theta^{(k)},\pi_0^{(k)}}\left\{\Delta_{t-1} = i, \Delta_t = j|\epsilon_1,\ldots,\epsilon_n\right\}}{\sum_{t=2}^n P_{\theta^{(k)},\pi_0^{(k)}}\{\Delta_{t-1} = i|\epsilon_1,\ldots,\epsilon_n\}}.$$

# M-step

Starting from an initial value $(\theta^{(0)}, \pi_0^{(0)})$, these formulas allow to build a sequence $\left(\theta^{(k)}, \pi_0^{(k)}\right)_k$ which increases the likelihood.

Require to compute the smoothed probabilities

$$\pi_{t|n} = (P\{\Delta_t = i | \epsilon_1, \ldots, \epsilon_n\})'_{1 \le i \le d} \in \mathbb{R}^d$$

and

$$\pi_{t-1,t|n} = \left(P\left\{\Delta_{t-1} = i, \Delta_t = j | \epsilon_1, \ldots, \epsilon_n\right\}\right)'_{1 \le i,j \le d} \in \mathbb{R}^d \times \mathbb{R}^d,$$

(still evaluated with the parameters $\theta^{(k)}, \pi_0^{(k)}$).

Such probabilities are obtained from Hamilton's algorithm.

## Smoothed probabilities

The Markov property entails that given $\Delta_t$, the observations $\epsilon_t, \epsilon_{t+1}, \ldots$ do not convey information on $\Delta_{t-1}$.
Thus

$$P(\Delta_{t-1} = i | \Delta_t = j, \epsilon_1, \ldots, \epsilon_n) = P(\Delta_{t-1} = i | \Delta_t = j, \epsilon_1, \ldots, \epsilon_{t-1})$$

and

$$\begin{aligned}
\pi_{t-1,t|n}(i,j) &= P(\Delta_{t-1} = i | \Delta_t = j, \epsilon_1, \ldots, \epsilon_{t-1}) \pi_{t|n}(j) \\
&= \frac{p(i,j)\pi_{t-1|t-1}(i)\pi_{t|n}(j)}{\pi_{t|t-1}(j)}.
\end{aligned}$$

Moreover, for $t = n, n-1, \ldots, 2$,

$$\pi_{t-1|n}(i) = \sum_{j=1}^{d} \pi_{t-1,t|n}(i,j) = \sum_{j=1}^{d} \frac{p(i,j)\pi_{t-1|t-1}(i)\pi_{t|n}(j)}{\pi_{t|t-1}(j)}$$

# EM algorithm

Starting from initial values for the parameters $\boldsymbol{\pi}_0$, $p(i,j)$, $\sigma_i$, the EM algorithm consists in repeating until convergence the steps :

1. Set $\pi_{1|0} = \pi_0$ and

$$\pi_{t|t} = \frac{\pi_{t|t-1} \odot \Phi_t}{(1,\ldots,1)\left\{\pi_{t|t-1} \odot \Phi_t\right\}}, \quad \pi_{t+1|t} = \mathbb{P}\pi_{t|t}, \quad \text{for } t = 1,\ldots,n.$$

2. Compute the smoothed probabilities $\pi_{t|n}(i)$ and $\pi_{t-1,t|n}(i,j)$ :

$$\begin{aligned} \pi_{t-1|n}(i) &= \sum_{j=1}^{d} \frac{p(i,j)\pi_{t-1|t-1}(i)\pi_{t|n}(j)}{\pi_{t|t-1}(j)} \quad \text{for } t = n, n-1, \ldots, 2, \\ \pi_{t-1,t|n}(i,j) &= \frac{p(i,j)\pi_{t-1|t-1}(i)\pi_{t|n}(j)}{\pi_{t|t-1}(j)}. \end{aligned}$$

3. Replace the previous parameter values by $\pi_0 = \pi_{1|n}$,

$$p(i,j) = \frac{\sum_{t=2}^{n} \pi_{t-1,t|n}(i,j)}{\sum_{t=2}^{n} \pi_{t-1|n}(i)} \quad \text{and} \quad \sigma^2(i) = \frac{\sum_{t=1}^{n} \epsilon_t^2 \pi_{t|n}(i)}{\sum_{t=1}^{n} \pi_{t|n}(i)}.$$

# CAC40 and SP500 series

Daily series of the CAC40 and SP500 over the period : March 1st 1990 to December 29 2006.

HMM model with 4 regimes : 60 iterations of the EM algorithm yield

- SP500

$$\hat{\omega}_{SP} = \begin{pmatrix} 0.26 \\ 0.62 \\ 1.28 \\ 4.8 \end{pmatrix}, \quad \hat{P}_{SP} = \begin{pmatrix} 0.981 & 0.019 & 0.000 & 0.000 \\ 0.018 & 0.979 & 0.003 & 0.000 \\ 0.000 & 0.003 & 0.986 & 0.011 \\ 0.000 & 0.000 & 0.055 & 0.945 \end{pmatrix}$$

- CAC40

$$\hat{\omega}_{CAC} = \begin{pmatrix} 0.51 \\ 1.19 \\ 2.45 \\ 8.4 \end{pmatrix}, \quad \hat{P}_{CAC} = \begin{pmatrix} 0.993 & 0.003 & 0.002 & 0.002 \\ 0.003 & 0.991 & 0.003 & 0.003 \\ 0.000 & 0.020 & 0.977 & 0.003 \\ 0.004 & 0.000 & 0.032 & 0.963 \end{pmatrix}.$$

# CAC40 and SP500 series

Estimated probabilities of the 4 regimes :

$$\hat{\pi}_{SP} = (0.30, 0.32, 0.32, 0.06)', \qquad \hat{\pi}_{CAC} = (0.26, 0.49, 0.19, 0.06)',$$

Expected duration of the different regimes (equal to $1/\{1 - p(i,i)\}$) :

$$D_{SP} = (53, 48, 71, 18)', \qquad D_{CAC} = (140, 107, 43, 27)'.$$

# CAC40 and SP500 series