Dynamic models

Dependent data

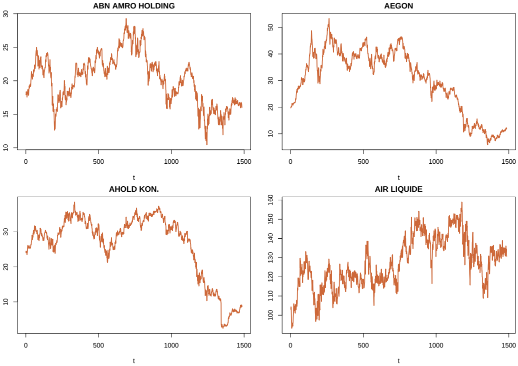
Huge portion of real-life data involving dependent datapoints

Example (Capture-recapture)

- capture histories
- capture sizes

Eurostoxx 50

First four stock indices of of the financial index Eurostoxx 50



Markov chain

Stochastic process $(x_t)_{t\in\mathcal{T}}$ where distribution of x_t given the past values $\mathbf{x}_{0:(t-1)}$ only depends on x_{t-1}.
Homogeneity: distribution of x_t given the past constant in $t \in \mathcal{T}$.

Corresponding likelihood

$$\ell(\theta|\mathbf{x}_{0:T}) = f_0(x_0|\theta) \prod_{t=1}^{T} f(x_t|x_{t-1},\theta)$$

[Homogeneity means f independent of t]

Stationarity constraints

Difference with the independent case: *stationarity* and *causality* constraints put restrictions on the parameter space

Stationarity processes

Definition (Stationary stochastic process)

$(x_t)_{t\in\mathcal{T}}$ is stationary if the joint distributions of $(x_1,\ldots,x_k)$ and $(x_{1+h},\ldots,x_{k+h})$ are the same for all h,k's.
It is *second-order stationary* if, given the autocovariance function

$$\gamma_x(r,s)=\mathbb{E}[\{x_r-\mathbb{E}(x_r)\}\{x_s-\mathbb{E}(x_s)\}],\quad r,s\in\mathcal{T}\,,$$

then

$$\mathbb{E}(x_t)=\mu \quad \text{and} \quad \gamma_x(r,s)=\gamma_x(r+t,s+t)\equiv\gamma_x(r-s)$$

for all $r,s,t\in\mathcal{T}$.

Imposing or not imposing stationarity

Bayesian inference on a non-stationary process can be [formaly] conducted

Debate

From a Bayesian point of view, to impose the *stationarity* condition is objectionable:stationarity requirement on finite datasets artificial *and/or* datasets themselves should indicate whether the model is stationary

Reasons for imposing stationarity:asymptotics (Bayes estimators are not necessarily convergent in non-stationary settings) causality, identifiability and ... common practice.

Unknown stationarity constraints

Practical difficulty: for complex models, stationarity constraints get quite involved to the point of being unknown in some cases

The AR(1) model

Case of linear Markovian dependence on the last value

$$x_t = \mu + \varrho(x_{t-1} - \mu) + \epsilon_t \,, \epsilon_t \overset{\text{i.i.d.}}{\sim} \mathscr{N}(0, \sigma^2)$$

If $|\varrho| < 1$, $(x_t)_{t\in\mathbb{Z}}$ can be written as

$$x_t = \mu + \sum_{j=0}^{\infty} \varrho^j \epsilon_{t-j}$$

and this is a stationary representation.

Stationary but...

If $|\varrho| > 1$, alternative stationary representation

$$x_t = \mu - \sum_{j=1}^{\infty} \varrho^{-j} \epsilon_{t+j} \,.$$

This stationary solution is criticized as artificial because x_t is correlated with *future* white noises $(\epsilon_t)_{s>t}$, unlike the case when $|\varrho| < 1$.
Non-causal representation...

Standard constraint

© Customary to restrict AR(1) processes to the case $|\varrho| < 1$

Thus use of a uniform prior on $[-1, 1]$ for ϱ

Exclusion of the case $|\varrho| = 1$ that leads to a random walk because the process is then a random walk *[no stationary solution]*

The AR(p) model

Conditional model

$$x_t | x_{t-1}, \ldots \sim \mathcal{N}\left(\mu + \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu), \sigma^2\right)$$

- Generalisation of AR(1)
- Among the most commonly used models in dynamic settings
- More challenging than the static models (stationarity constraints)
- Different models depending on the processing of the starting value x_0

Stationarity+causality

Stationarity constraints in the prior as a restriction on the values of θ.

Theorem

AR(p) model second-order stationary and causal iff *the roots of the polynomial*

$$\mathcal{P}(x) = 1 - \sum_{i=1}^{p} \varrho_i x^i$$

are all outside the unit circle

Initial conditions

Unobserved initial values can be processed in various ways

1. All $\mathbf{x}_{-i}$'s $(i > 0)$ set equal to μ, for computational convenience
2. Under stationarity and causality constraints, $(x_t)_{t\in\mathbb{Z}}$ has a stationary distribution: Assume $\mathbf{x}_{-p:-1}$ distributed from stationary $\mathscr{N}_p(\mu\mathbf{1}_p, \mathbf{A})$ distribution
   Corresponding marginal likelihood

$$\int \sigma^{-T} \prod_{t=0}^{T} \exp\left\{ \frac{-1}{2\sigma^2} \left(x_t - \mu - \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu) \right)^2 \right\}$$
$$f(\mathbf{x}_{-p:-1}|\mu, \mathbf{A})\,\mathrm{d}\mathbf{x}_{-p:-1}\,,$$

Initial conditions (cont'd)

3. Condition instead on the initial *observed* values $\mathbf{x}_{0:(p-1)}$

$$\ell^c(\mu, \varrho_1, \ldots, \varrho_p, \sigma | \mathbf{x}_{p:T}, \mathbf{x}_{0:(p-1)}) \propto \sigma^{-T} \prod_{t=p}^{T} \exp\left\{ -\left(x_t - \mu - \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu) \right)^2 \Big/ 2\sigma^2 \right\} .$$

Prior selection

For AR(1) model, Jeffreys' prior associated with the stationary representation is

$$\pi_1^J(\mu, \sigma^2, \varrho) \propto \frac{1}{\sigma^2} \frac{1}{\sqrt{1-\varrho^2}} .$$

Extension to higher orders quite complicated (ϱ part)!

Natural conjugate prior for $\theta = (\mu, \varrho_1, \ldots, \varrho_p, \sigma^2)$:
normal distribution on $(\mu, \varrho_1, \ldots, \varrho_p)$ and inverse gamma distribution on σ^2
... and for constrained ϱ's?

Stationarity constraints

Under stationarity constraints, complex parameter space: each value of ϱ needs to be checked for roots of corresponding polynomial with modulus less than 1

E.g., for an AR(2) process with autoregressive polynomial $\mathcal{P}(u) = 1 - \varrho_1 u - \varrho_2 u^2$, constraint is

$$\varrho_1 + \varrho_2 < 1, \quad \varrho_1 - \varrho_2 < 1 \quad \text{and} \quad |\varrho_2| < 1 \,.$$

Skip Durbin forward

A first useful reparameterisation

Durbin–Levinson recursion proposes a *reparametrisation* from the parameters ϱ_i to the *partial autocorrelations*

$$\psi_i \in [-1, 1]$$

which allow for a uniform prior on the hypercube.
Partial autocorrelation defined as

$$\begin{aligned}\psi_i = \mathsf{corr}\,(x_t - \mathbb{E}[x_t|x_{t+1}, \ldots, x_{t+i-1}],\\ x_{t+i} - \mathbb{E}[x_{t+1}|x_{t+1}, \ldots, x_{t+i-1}])\end{aligned}$$

[see also Yule-Walker equations]

Durbin–Levinson recursion

Transform

1. Define $\varphi^{ii} = \psi_i$ and $\varphi^{ij} = \varphi^{(i-1)j} - \psi_i \varphi^{(i-1)(i-j)}$, for $i > 1$ and $j = 1, \cdots, i-1$.
2. Take $\varrho_i = \varphi^{pi}$ for $i = 1, \cdots, p$.

Stationarity & priors

For AR(1) model, Jeffreys' prior associated with the stationary representation is

$$\pi_1^J(\mu, \sigma^2, \varrho) \propto \frac{1}{\sigma^2} \frac{1}{\sqrt{1-\varrho^2}} \,.$$

Within the non-stationary region $|\varrho| > 1$, Jeffreys' prior is

$$\pi_2^J(\mu, \sigma^2, \varrho) \propto \frac{1}{\sigma^2} \frac{1}{\sqrt{|1-\varrho^2|}} \sqrt{\left|1 - \frac{1-\varrho^{2T}}{T(1-\varrho^2)}\right|} \,.$$

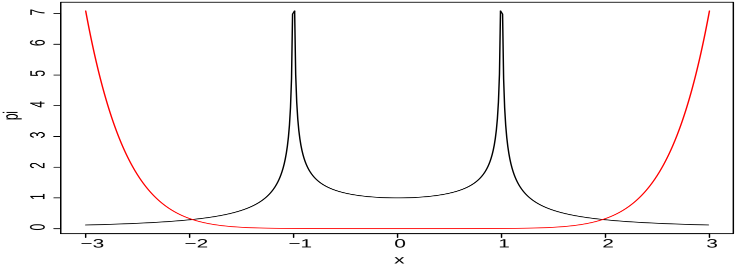
The dominant part of the prior is the non-stationary region!

Alternative prior

The reference prior π_1^J is only defined when the stationary constraint holds.

Idea Symmetrise to the region $|\varrho| > 1$

$$\pi^B(\mu, \sigma^2, \varrho) \propto \frac{1}{\sigma^2} \begin{cases} 1/\sqrt{1-\varrho^2} & \text{if } |\varrho| < 1, \\ 1/|\varrho|\sqrt{\varrho^2 - 1} & \text{if } |\varrho| > 1, \end{cases},$$



MCMC consequences

When devising an MCMC algorithm, use the Durbin-Levinson recursion to end up with single normal simulations of the ψ_i's since the ϱ_j's are linear functions of the ψ_i's

Root parameterisation

Skip Durbin back Lag polynomial representation

$$\left(\mathrm{Id} - \sum_{i=1}^{p} \varrho_i B^i\right) x_t = \epsilon_t$$

with (inverse) roots

$$\prod_{i=1}^{p} \left(\mathrm{Id} - \lambda_i B\right) x_t = \epsilon_t$$

Closed form expression of the likelihood as a function of the (inverse) roots

Uniform prior under stationarity

Stationarity The λ_i's are within the unit circle if in $\mathbb{C}$ [complex numbers] and within $[-1, 1]$ if in $\mathbb{R}$ [real numbers]

Naturally associated with a flat prior on either the unit circle or $[-1, 1]$

$$\frac{1}{\lfloor k/2 \rfloor + 1} \prod_{\lambda_i \in \mathbb{R}} \frac{1}{2} \mathbb{I}_{|\lambda_i| < 1} \prod_{\lambda_i \notin \mathbb{R}} \frac{1}{\pi} \mathbb{I}_{|\lambda_i| < 1}$$

where $\lfloor k/2 \rfloor + 1$ number of possible cases

- Term $\lfloor k/2 \rfloor + 1$ is important for reversible jump applications

MCMC consequences

In a Gibbs sampler, each λ_{i^*} can be simulated conditionaly on the others since

$$\prod_{i=1}^{p} (\mathsf{Id} - \lambda_i B)\, x_t = y_t - \lambda_{i^*} y_{t-1} = \epsilon_t$$

where

$$Y_t = \prod_{i \neq i^*} (\mathsf{Id} - \lambda_i B)\, x_t$$

Metropolis-Hastings implementation

1. use the prior π itself as a proposal on the (inverse) roots of $\mathcal{P}$, selecting one or several roots of $\mathcal{P}$ to be simulated from π;
2. acceptance ratio is likelihood ratio
3. need to watch out for real/complex dichotomy

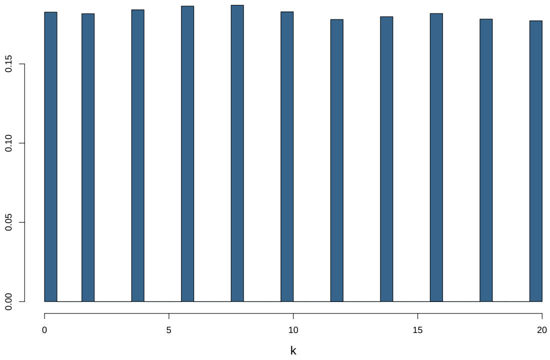
A [paradoxical] reversible jump implementation

- Define "model" $\mathfrak{M}_{2k}$ $(0 \le k \le \lfloor p/2 \rfloor)$ as corresponding to a number $2k$ of complex roots $o \le k \le \lfloor p/2 \rfloor)$
- Moving from model $\mathfrak{M}_{2k}$ to model $\mathfrak{M}_{2k+2}$ means that two real roots have been replaced by two conjugate complex roots.
- Propose jump from $\mathfrak{M}_{2k}$ to $\mathfrak{M}_{2k+2}$ with probability $1/2$ and from $\mathfrak{M}_{2k}$ to $\mathfrak{M}_{2k-2}$ with probability $1/2$ *[boundary exceptions]*
- accept move from $\mathfrak{M}_{2k}$ to $\mathfrak{M}_{2k+\text{ or }-2}$ with probability

$$\frac{\ell^c(\mu, \varrho_1^\star, \ldots, \varrho_p^\star, \sigma | \mathbf{x}_{p:T}, \mathbf{x}_{0:(p-1)})}{\ell^c(\mu, \varrho_1, \ldots, \varrho_p, \sigma | \mathbf{x}_{p:T}, \mathbf{x}_{0:(p-1)})} \wedge 1\,,$$

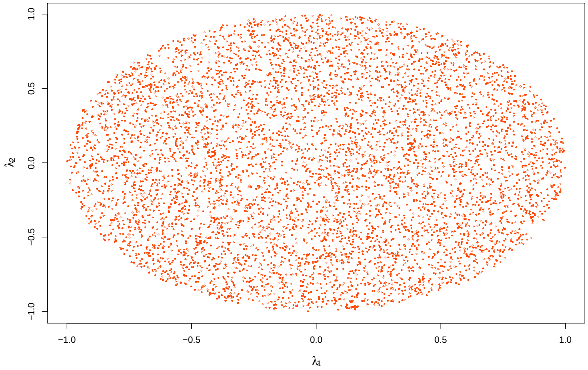
Checking your code

Try with no data and recover the prior



Checking your code

Try with no data and recover the prior



Order estimation

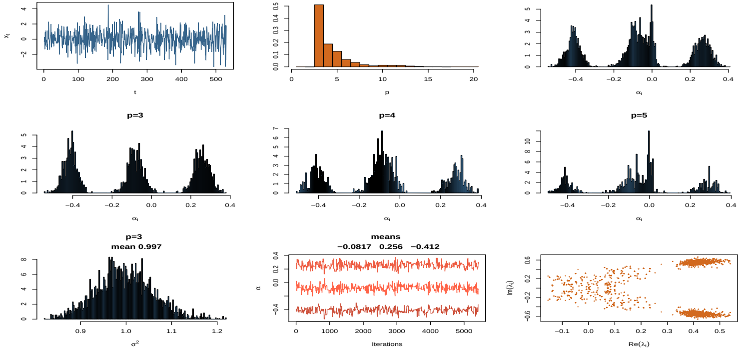
Typical setting for model choice: determine order p of $AR(p)$ model
Roots [may] change drastically from one p to the other.
No difficulty from the previous perspective: recycle above reversible jump algorithm

AR(?) reversible jump algorithm

Use (purely birth-and-death) proposals based on the uniform prior

- k $\longrightarrow$ k+1 [Creation of real root]
- k $\longrightarrow$ k+2 [Creation of complex root]
- k $\longrightarrow$ k-1 [Deletion of real root]
- k $\longrightarrow$ k-2 [Deletion of complex root]

Reversible jump output



$AR(3)$ simulated dataset of 530 points *(upper left)* with true parameters α_i $(-0.1, 0.3, -0.4)$ and $\sigma = 1$. *First histogram* associated with p, following histograms with the α_i's, for different values of p, and of σ^2. *Final graph:* scatterplot of the complex roots. *One before last:* evolution of $\alpha_1, \alpha_2, \alpha_3$.

The MA(q) model

Alternative type of time series

$$x_t = \mu + \epsilon_t - \sum_{j=1}^{q} \vartheta_j \epsilon_{t-j}\,, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

Stationary but, for identifiability considerations, the polynomial

$$\mathcal{Q}(x) = 1 - \sum_{j=1}^{q} \vartheta_j x^j$$

must have all its roots outside the unit circle.

Identifiability

Example

For the MA(1) model, $x_t = \mu + \epsilon_t - \vartheta_1 \epsilon_{t-1}$,

$$\mathrm{var}(x_t) = (1 + \vartheta_1^2)\sigma^2$$

can also be written

$$x_t = \mu + \tilde{\epsilon}_{t-1} - \frac{1}{\vartheta_1}\tilde{\epsilon}_t, \quad \tilde{\epsilon} \sim \mathcal{N}(0, \vartheta_1^2 \sigma^2)\,,$$

Both pairs (ϑ_1, σ) & $(1/\vartheta_1, \vartheta_1 \sigma)$ lead to alternative representations of the *same* model.

Properties of MA models

- Non-Markovian model (but special case of hidden Markov)
- Autocovariance $\gamma_x(s)$ is null for $|s| > q$

Representations

$\mathbf{x}_{1:T}$ is a normal random variable with constant mean μ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 & \gamma_1 & \gamma_2 & \dots & \gamma_q & 0 & \dots & 0 & 0 \\ \gamma_1 & \sigma^2 & \gamma_1 & \dots & \gamma_{q-1} & \gamma_q & \dots & 0 & 0 \\ & & & \ddots & & & & & \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & \gamma_1 & \sigma^2 \end{pmatrix},$$

with $(|s| \leq q)$

$$\gamma_s = \sigma^2 \sum_{i=0}^{q-|s|} \vartheta_i \vartheta_{i+|s|}$$

Not manageable in practice *[large T's]*

Representations (contd.)

Conditional on past $(\epsilon_0, \ldots, \epsilon_{-q+1})$,

$$L(\mu, \vartheta_1, \ldots, \vartheta_q, \sigma | x_{1:T}, \epsilon_0, \ldots, \epsilon_{-q+1}) \propto$$
$$\sigma^{-T} \prod_{t=1}^{T} \exp\left\{ - \left(x_t - \mu + \sum_{j=1}^{q} \vartheta_j \hat{\epsilon}_{t-j} \right)^2 \Big/ 2\sigma^2 \right\},$$

where $(t > 0)$

$$\hat{\epsilon}_t = x_t - \mu + \sum_{j=1}^{q} \vartheta_j \hat{\epsilon}_{t-j},\ \hat{\epsilon}_0 = \epsilon_0,\ \ldots,\ \hat{\epsilon}_{1-q} = \epsilon_{1-q}$$

Recursive definition of the likelihood, still costly $\mathrm{O}(T \times q)$

Recycling the AR algorithm

Same algorithm as in the AR(p) case when modifying the likelihood

Simulation of the past noises $\boldsymbol{\epsilon}_{-i}$ $(i = 1, \ldots, q)$ done via a Metropolis-Hastings step with target

$$f(\epsilon_0, \ldots, \epsilon_{-q+1} | \mathbf{x}_{1:T}, \mu, \sigma, \boldsymbol{\vartheta}) \propto \prod_{i=-q+1}^{0} e^{-\epsilon_i^2/2\sigma^2} \prod_{t=1}^{T} e^{-\widehat{\epsilon}_t^2/2\sigma^2} \,,$$

Representations (contd.)

Encompassing approach for general time series models

State-space representation

$$\begin{aligned} \mathbf{x}_t &= G\mathbf{y}_t + \boldsymbol{\varepsilon}_t\,, && (1)\\ \mathbf{y}_{t+1} &= F\mathbf{y}_t + \xi_t\,, && (2) \end{aligned}$$

(1) is the *observation equation* and (2) is the *state equation*

Note

As seen below, this is a special case of hidden Markov model

MA(q) state-space representation

For the MA(q) model, take

$$\mathbf{y}_t = (\epsilon_{t-q}, \ldots, \epsilon_{t-1}, \epsilon_t)'$$

and then

$$\begin{aligned}
\mathbf{y}_{t+1} &= \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ & & & \ldots & \\ 0 & 0 & 0 & \ldots & 1 \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix} \mathbf{y}_t + \epsilon_{t+1} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \\
x_t &= \mu - \begin{pmatrix} \vartheta_q & \vartheta_{q-1} & \ldots & \vartheta_1 & -1 \end{pmatrix} \mathbf{y}_t \,.
\end{aligned}$$

MA(q) state-space representation (cont'd)

Example

For the MA(1) model, observation equation

$$x_t = (1 \quad 0)\mathbf{y}_t$$

with

$$\mathbf{y}_t = (y_{1t} \quad y_{2t})'$$

directed by the state equation

$$\mathbf{y}_{t+1} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \mathbf{y}_t + \epsilon_{t+1} \begin{pmatrix} 1 \\ \vartheta_1 \end{pmatrix} .$$

ARMA extension

ARMA(p, q) model

$$x_t - \sum_{i=1}^{p} \varrho_i x_{t-1} = \mu + \epsilon_t - \sum_{j=1}^{q} \vartheta_j \epsilon_{t-j}\,, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

Identical stationarity and identifiability conditions for both groups $(\varrho_1, \ldots, \varrho_p)$ and $(\vartheta_1, \ldots, \vartheta_q)$

Reparameterisation

Identical root representations

$$\prod_{i=1}^{p}(\mathsf{Id} - \lambda_i B)x_t = \prod_{i=1}^{q}(\mathsf{Id} - \eta_i B)\epsilon_t$$

State-space representation

$$\mathbf{x}_t = x_t = \mu - \begin{pmatrix}\vartheta_{r-1} & \vartheta_{r-2} & \dots & \vartheta_1 & -1\end{pmatrix} \mathbf{y}_t$$

and

$$\mathbf{y}_{t+1} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ & & & \dots & \\ 0 & 0 & 0 & \dots & 1 \\ \varrho_r & \varrho_{r-1} & \varrho_{r-2} & \dots & \varrho_1 \end{pmatrix} \mathbf{y}_t + \epsilon_{t+1} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix},$$

under the convention that $\varrho_m = 0$ if $m > p$ and $\vartheta_m = 0$ if $m > q$.

Bayesian approximation

Quasi-identical MCMC implementation:

1. Simulate $(\varrho_1, \ldots, \varrho_p)$ conditional on $(\vartheta_1, \ldots, \vartheta_q)$ and μ
2. Simulate $(\vartheta_1, \ldots, \vartheta_q)$ conditional on $(\varrho_1, \ldots, \varrho_p)$ and μ
3. Simulate (μ, σ) conditional on $(\varrho_1, \ldots, \varrho_p)$ and $(\vartheta_1, \ldots, \vartheta_q)$

© Code can be recycled almost as is!

Hidden Markov models

Generalisation both of a mixture and of a state space model.

Example

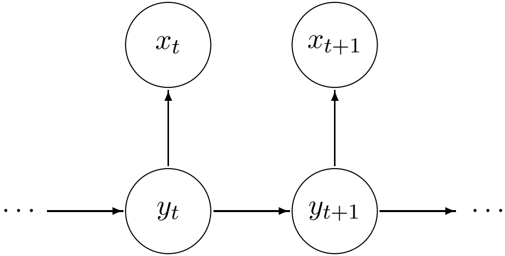
Extension of a *mixture* model with Markov dependence

$$x_t|z, x_j \; j \neq t \; \sim \mathcal{N}(\mu_{z_t}, \sigma^2_{z_t}), \qquad P(z_t = u|z_j \; , j < t) = p_{z_{t-1}u},$$

$(u = 1, \ldots, k)$

- ↯ Label switching also strikes in this model!

Generic dependence graph



$$(x_t, y_t)|\mathbf{x}_{0:(t-1)}, \mathbf{y}_{0:(t-1)} \sim f(y_t|y_{t-1})\, f(x_t|y_t)$$

Definition

Observable series $\{\mathbf{x}_t\}_{t\geq 1}$ associated with a second process $\{y_t\}_{t\geq 1}$, with a finite set of N possible values such that

1. indicators Y_t have an homogeneous **Markov dynamic**

$$p(y_t|\mathbf{y}_{1:t-1}) = p(y_t|y_{t-1}) = \mathbb{P}_{y_{t-1}y_t}$$

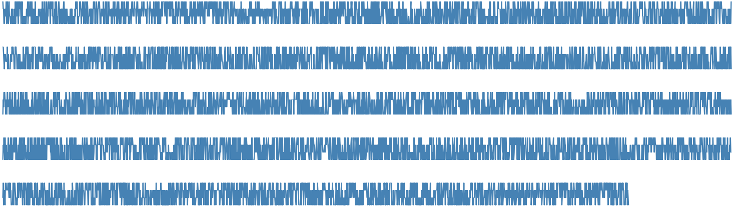
   where $\mathbf{y}_{1:t-1}$ denotes the sequence $\{y_1, y_2, \ldots, y_{t-1}\}$.

2. Observables x_t are independent conditionally on the indicators y_t

$$p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}) = \prod_{t=1}^{T} p(x_t|y_t)$$

Dnadataset

DNA sequence *[made of A, C, G, and T's]* corresponding to a complete HIV genome where A, C, G, and T have been recoded as $1, ..., 4$.



Possible modeling by a two-state hidden Markov model with

$$\mathscr{Y} = \{1, 2\} \quad \text{and} \quad \mathscr{X} = \{1, 2, 3, 4\}$$

Parameterization

- For the Markov bit, *transition matrix*

$$\mathbb{P} = [p_{ij}] \quad \text{where} \quad \sum_{j=1}^{N} p_{ij} = 1$$

and *initial distribution*

$$\varrho = \varrho \mathbb{P}$$

- for the observables,

$$f_i(x_t) = p(x_t|y_t = i) = f(x_t|\theta_i)$$

usually within the same parametrized class of distributions.

Finite case

When both hidden and observed chains are finite, with $\mathscr{Y} = \{1, \ldots, \kappa\}$ and $\mathscr{X} = \{1, \ldots, k\}$, parameter θ made up of p probability vectors $\mathbf{q}^1 = (q_1^1, \ldots, q_k^1), \ldots, \mathbf{q}^\kappa = (q_1^\kappa, \ldots, q_k^\kappa)$
Joint distribution of $(x_t, y_t)_{0 \le t \le T}$

$$\varrho_{y_0}\, q_{x_0}^{y_0} \prod_{t=1}^{T} p_{y_{t-1} y_t}\, q_{x_t}^{y_t}\,,$$

Bayesian inference in the finite case

Posterior of $(\theta, \mathbb{P})$ given $(x_t, y_t)_t$ factorizes as

$$\pi(\theta, \mathbb{P})\, \varrho_{y_0} \prod_{i=1}^{\kappa} \prod_{j=1}^{k} (q_j^i)^{n_{ij}} \times \prod_{i=1}^{\kappa} \prod_{j=1}^{p} p_{ij}^{m_{ij}} ,$$

where n_{ij} # of visits to state j by the x_t's when the corresponding y_t's are equal to i and m_{ij} # of transitions from state i to state j on the hidden chain $(y_t)_{t\in\mathbb{N}}$

Under a flat prior on p_{ij}'s and q_j^i's, posterior distributions are [almost] Dirichlet *[initial distribution side effect]*

MCMC implementation

Finite State HMM Gibbs Sampler

Initialization:

1. Generate random values of the p_{ij}'s and of the q_j^i's
2. Generate the hidden Markov chain $(y_t)_{0\le t\le T}$ by $(i=1,2)$

$$\mathbb{P}(y_t=i) \propto \begin{cases} p_{ii}\, q^i_{x_0} & \text{if } t=0\,, \\ p_{y_{t-1}i}\, q^i_{x_t} & \text{if } t>0\,, \end{cases}$$

and compute the corresponding sufficient statistics

MCMC implementation (cont'd)

Finite State HMM Gibbs Sampler

Iteration m $(m \geq 1)$:

1. Generate

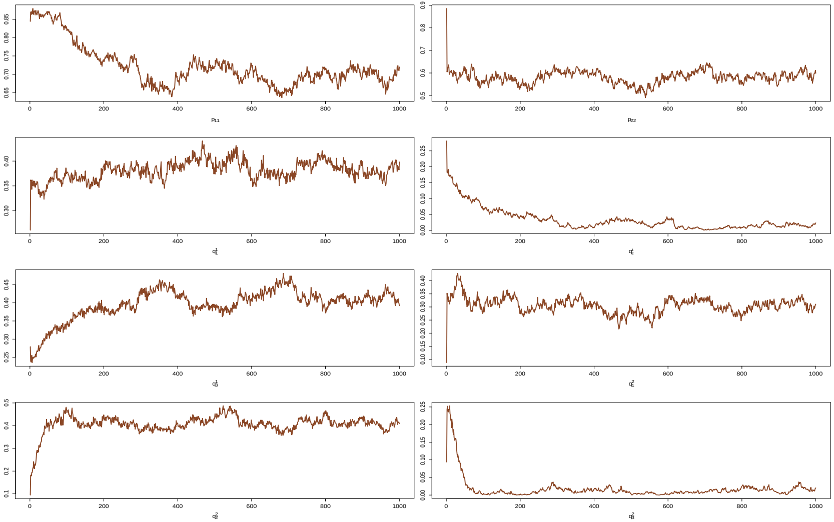
$$(p_{i1}, \ldots, p_{i\kappa}) \sim \mathscr{D}(1 + n_{i1}, \ldots, 1 + n_{i\kappa})$$
$$(q_1^i, \ldots, q_k^i) \sim \mathscr{D}(1 + m_{i1}, \ldots, 1 + m_{ik})$$

   and correct for missing initial probability by a MH step with acceptance probability $\varrho'_{y_0}/\varrho_{y_0}$
2. Generate successively each y_t $(0 \leq t \leq T)$ by

$$\mathbb{P}(y_t = i | x_t, y_{t-1}, y_{t+1}) \propto \begin{cases} p_{ii}\, q^i_{x_1}\, p_{iy_1} & \text{if } t = 0\,, \\ p_{y_{t-1}i}\, q^i_{x_t}\, p_{iy_{t+1}} & \text{if } t > 0\,, \end{cases}$$

   and compute corresponding sufficient statistics

Dnadataset



Forward-Backward formulae

Existence of a (magical) recurrence relation that provides the observed likelihood function in manageable computing time
Called *forward-backward* or *Baum–Welch* formulas

Observed likelihood computation

Likelihood of the *complete model* simple:

$$\ell^c(\boldsymbol{\theta}|\mathbf{x},\mathbf{y}) = \prod_{t=2}^{T} p_{y_{t-1}y_t}\, f(x_t|\theta_{y_t})$$

but likelihood of the *observed model* is not:

$$\ell(\boldsymbol{\theta}|\mathbf{x}) = \sum_{\mathbf{y}\in\{1,\ldots,\kappa\}^T} \ell^c(\boldsymbol{\theta}|\mathbf{x},\mathbf{y})$$

© $\mathrm{O}(\kappa^T)$ complexity

Forward-Backward paradox

It is possible to express the (observed) likelihood $L^O(\boldsymbol{\theta}|\mathbf{x})$ in

$$\mathrm{O}(T^2 \times \kappa)$$

computations, based on the Markov property of the pair (x_t, y_t).

Direct to backward smoothing

Conditional distributions

We have

$$p(\mathbf{y}_{1:t}|\mathbf{x}_{1:t}) = \frac{f(x_t|y_t)\, p(\mathbf{y}_{1:t}|\mathbf{x}_{1:(t-1)})}{p(x_t|\mathbf{x}_{1:(t-1)})}$$

[Smoothing/Bayes]

and

$$p(\mathbf{y}_{1:t}|\mathbf{x}_{1:(t-1)}) = k(y_t|\mathbf{y}_{t-1})p(\mathbf{y}_{1:(t-1)}|\mathbf{x}_{1:(t-1)})$$

[Prediction]

where $k(y_t|y_{t-1}) = p_{y_{t-1}y_t}$ associated with the matrix $\mathbb{P}$ and

$$f(x_t|y_t) = f(x_t|\theta_{y_t})$$

Update of predictive

Therefore

$$
\begin{aligned}
p(\mathbf{y}_{1:t}|\mathbf{x}_{1:t}) &= \frac{p(y_t|\mathbf{x}_{1:(t-1)})\, f(x_t|y_t)}{p(x_t|\mathbf{x}_{1:(t-1)})} \\
&= \frac{f(x_t|y_t)\, k(y_t|\mathbf{y}_{t-1})}{p(x_t|\mathbf{x}_{1:(t-1)})} p(\mathbf{y}_{1:(t-1)}|\mathbf{x}_{1:(t-1)})
\end{aligned}
$$

with the same order of complexity for $p(\mathbf{y}_{1:t}|\mathbf{x}_{1:t})$ as for $p(x_t|\mathbf{x}_{1:(t-1)})$

Propagation and actualization equations

$$p(y_t|\mathbf{x}_{1:(t-1)}) = \sum_{\mathbf{y}_{1:(t-1)}} p(\mathbf{y}_{1:(t-1)}|\mathbf{x}_{1:(t-1)})\, k(y_t|y_{t-1})$$

[Propagation]

and

$$p(y_t|x_{1:t}) = \frac{p(y_t|\mathbf{x}_{1:(t-1)})\, f(x_t|y_t)}{p(x_t|\mathbf{x}_{1:(t-1)})}\,.$$

[Actualization]

Forward–backward equations (1)

Evaluation of

$$p(y_t|\mathbf{x}_{1:T}) \quad t \le T$$

by *forward-backward algorithm*
Denote $t \le T$

$$\begin{aligned}
\gamma_t(i) &= P(y_t = i|x_{1:T}) \\
\alpha_t(i) &= p(\mathbf{x}_{1:t}, y_t = i) \\
\beta_t(i) &= p(\mathbf{x}_{t+1:T}|y_t = i)
\end{aligned}$$

Recurrence relations

Then

$$\begin{cases} \alpha_1(i) & = & f(x_1|y_t = i)\varrho_i \\ \alpha_{t+1}(j) & = & f(x_{t+1}|y_{t+1} = j)\sum_{i=1}^{\kappa} \alpha_t(i)p_{ij} \end{cases}$$

[Forward]

$$\begin{cases} \beta_T(i) & = & 1 \\ \beta_t(i) & = & \sum_{j=1}^{\kappa} p_{ij} f(x_{t+1}|y_{t+1} = j)\beta_{t+1}(j) \end{cases}$$

[Backward]

and

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{\kappa} \alpha_t(j)\beta_t(j)}$$

Extension of the recurrence relations

For

$$\xi_t(i,j) = P(y_t = i, y_{t+1} = j | \mathbf{x}_{1:T}) \qquad i,j = 1,\ldots,\kappa,$$

we also have

$$\xi_t(i,j) = \frac{\alpha_t(i)\mathbb{P}_{ij} f(x_{t+1}|y_t = j)\beta_{t+1}(j)}{\sum_{i=1}^{\kappa}\sum_{j=1}^{\kappa} \alpha_t(i)\mathbb{P}_{ij} f(x_{t+1}|y_{t+1} = j)\beta_{t+1}(j)}$$

Overflows and underflows

- On-line scalings of the $\alpha_t(i)$'s and $\beta_T(i)$'s for each t by

$$c_t = 1\Big/\sum_{i=1}^{\kappa} \alpha_t(i) \quad \text{and} \quad d_t = 1\Big/\sum_{i=1}^{\kappa} \beta_t(i)$$

avoid overflows or/and underflows for large datasets

Backward smoothing

Recursive derivation of conditionals

We have

$$p(y_s|y_{s-1},\mathbf{x}_{1:t}) = p(y_s|y_{s-1},\mathbf{x}_{s:t})$$

[Markov property!]

Therefore $(s = T, T-1, \ldots, 1)$

$$p(y_s|y_{s-1},\mathbf{x}_{1:T}) \propto k(y_s|y_{s-1})\, f(x_s|y_s) \sum_{y_{s+1}} p(y_{s+1}|y_s,\mathbf{x}_{1:T})$$

[Backward equation]

with

$$p(y_T|y_{T-1},\mathbf{x}_{1:T}) \propto k(y_T|y_{T-1}) f(x_T|y_t)\,.$$

End of the backward smoothing

The first term is

$$p(y_1|\mathbf{x}_{1:t}) \propto \pi(y_1)\, f(x_1|y_1) \sum_{y_2} p(y_2|y_1, \mathbf{x}_{1:t})\,,$$

with π stationary distribution of $\mathbb{P}$

The conditional for y_s needs to be defined for each of the κ values of y_{s-1}

$\boxed{\text{ⓒ } \mathrm{O}(t \times \kappa^2) \textbf{ operations}}$

Details

Need to introduce unnormalized version of the conditionals $p(y_t|y_{t-1},\mathbf{x}_{0:T})$ such that

$$
\begin{aligned}
p^\star_T(y_T|y_{T-1},\mathbf{x}_{0:T}) &= p_{y_{T-1}y_T} f(x_T|y_T) \\
p^\star_t(y_t|y_{t-1},\mathbf{x}_{1:T}) &= p_{y_{t-1}y_t} f(x_t|y_t) \sum_{i=1}^{\kappa} p^\star_{t+1}(i|y_t,\mathbf{x}_{1:T}) \\
p^\star_0(y_0|\mathbf{x}_{0:T}) &= \varrho_{y_0}\, f(x_0|y_0) \sum_{i=1}^{\kappa} p^\star_1(i|y_0,\mathbf{x}_{0:t})
\end{aligned}
$$

Likelihood computation

Bayes formula

$$p(\mathbf{x}_{1:T}) = \frac{p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T})p(\mathbf{y}_{1:T})}{p(\mathbf{y}_{1:T}|\mathbf{x}_{1:T})}$$

gives a representation of the likelihood based on the forward–backward formulae and an arbitrary sequence $\mathbf{x}^o_{1:T}$ (since the l.h.s. does *not* depend on $\mathbf{x}_{1:T}$).

Obtained through the $p^\star_t$'s as

$$p(\mathbf{x}_{0:T}) = \sum_{i=1}^{\kappa} p^\star_1(i|\mathbf{x}_{0:T})$$

Prediction filter

If

$$\varphi_t(i) = p(y_t = i|\mathbf{x}_{1:t-1})$$

Forward equations

$$\begin{aligned}\varphi_1(j) &= p(y_1 = j) \\ \varphi_{t+1}(j) &= \frac{1}{c_t}\sum_{i=1}^{\kappa} f(x_t|y_t = i)\varphi_t(i)p_{ij} \quad (t \geq 1)\end{aligned}$$

where

$$c_t = \sum_{k=1}^{\kappa} f(x_t|y_t = k)\varphi_t(k)\,,$$

Likelihood computation (2)

Follows the same principle as the backward equations
The (log-)likelihood is thus

$$
\begin{aligned}
\log p(\mathbf{x}_{1:t}) &= \sum_{r=1}^{t} \log \left[\sum_{i=1}^{\kappa} p(x_t, y_t = i | \mathbf{x}_{1:(r-1)}) \right] \\
&= \sum_{r=1}^{t} \log \left[\sum_{i=1}^{\kappa} f(x_t | y_t = i) \varphi_t(i) \right]
\end{aligned}
$$