# Dynamic Statistical Models with Hidden Variables

Benjamin Poignard

CREST & PSL (Paris Dauphine University, CEREMADE)

# Agenda

**Chapter 1: Definitions and examples**

- Stationary Processes, ARMA and VARMA models.
- Random variance models, Hidden-Markov models.
- State space models.

**Chapter 2: The Kalman filter**

- General form.
- Prediction and Smoothing.
- Statistical inference.

# Agenda

**Chapter 3: Markov swithing models**

- Finite-state Markov chains.
- Hidden Markov models.
- Markov-switching ARMA models.
- Estimation of the MS-AR(p) model.

# Dynamic models, hidden variables

- Dynamic : time series tutorial.
  **Reference:** Brockwell, P. J. and R. A. Davis (2002). *Introduction to Time Series and Forecasting*, Springer.
- Hidden (or latent): variables that are not statistically observable.

# Hidden variables

In the standard multivariate linear regression model, we have

$$Y = X\beta + U,$$

- $Y \in \mathbb{R}^N$: endogenous/dependent variable.
- $X$: $N \times K$ matrix, whose columns are the exogenous/explanatory variables.
- $\beta \in \mathbb{R}^K$: vector of parameters.
- $U \in \mathbb{R}^N$: error term, which is unobserable and interpreted as a measurement error.

# Time series models

Many models can be described as

$$Y_t = f(Y_{t-1}, Y_{t-2}, \cdots, ; U_t),$$

where $U_t$ is an unobservable error term.

Exogenous variables can be included such that

$$Y_t = f(Y_{t-1}, Y_{t-2}, \cdots, ; X_t; U_t).$$

We can model latent (or unobserved) variables models as

$$\begin{array}{rcl} Y_t & = & f(Y_{t-1}, Y_{t-2}, \cdots, ; \alpha_t; U_t), \\ \alpha_t & = & f^\star(\alpha_{t-1}, \alpha_{t-2}, \cdots, ; V_t), \end{array}$$

where $U_t$ and $V_t$ are error terms.

# Motivation

- Dynamic properties of real series may be difficult to capture with classical models.
- Latent variables may have an economic interpretation.
- Allows to reduce the dimension of a statistical problem.

Entails two types of difficulties :

- **Probabilistic properties** of such models can be difficult to derive (existence of solutions ?...).
- **Standard statistical** tools can be inadequate.

# Strict stationarity

$(Y_t)_{t\in\mathbb{Z}}$ a stochastic process (discrete time) valued in $\mathbb{R}^d$.

## Definition

The process $(Y_t)$ is **strictly stationary** if

$$(Y_{t_1}, Y_{t_2}, \cdots, Y_{t_k})' \overset{d}{=} (Y_{t_1+h}, Y_{t_2+h}, \cdots, Y_{t_k+h})',$$

for all $k \in \mathbb{N}$, and $h, t_1, \cdots, t_k \in \mathbb{Z}$.

# Second-order stationarity

## Definition

The process $(Y_t)$ is **second-order stationary** if

$$(i) \forall t \in \mathbb{Z}, Y_t \in L^2, \text{ ie } \mathbb{E}[\|Y_t\|^2] < \infty,$$

$$(ii) \forall t \in \mathbb{Z}, \mathbb{E}[Y_t] = \mu,$$

$$(iii) \forall t, h \in \mathbb{Z}, \operatorname{cov}(Y_t, Y_{t+h}) = \Gamma(h).$$

$\Gamma(.)$ is the autocovariance function of $(Y_t)$.

# Concepts of noises

A (weak) white noise process $(\epsilon_t)$ is a sequence of centered and uncorrelated variables :

$$\epsilon_t \in L^2, \mathbb{E}[\epsilon_t] = 0, \operatorname{cov}(\epsilon_t, \epsilon_{t+h}) = 0, \forall h \neq 0.$$

A strong white noise process $(\epsilon_t)$ is a sequence of centered, independent variables belonging to $L^2$.

An i.i.d. (independent and identically distributed) noise is a strong white noise where the $(\epsilon_t)$ have the same distribution.

White noise are useful to construct more complex stationary processes.

# Two concepts of prediction

- Conditional expectation: If $(Y_t)$ is second order stationary, then
$$\mathbb{E}[Y_t | Y_{t-1}, Y_{t-2}, \cdots]$$
is the best approximation of $Y_t$ (in the $L^2$ sense) as a function of its past.
- Linear conditional expectation:
$$\mathbb{E}L[Y_t | Y_{t-1}, Y_{t-2}, \cdots]$$
is the best approximation of $Y_t$ as a linear function of its past. Notation: $Y_{\underline{t-1}}$ the past of $Y_t$ and $\mathcal{H}_Y(t-1)$ the linear past of $Y_t$.

# Two concepts of innovation

- Strong innovation:

$$\epsilon_t = Y_t - \mathbb{E}[Y_t | Y_{t-1}, Y_{t-2}, \cdots]$$

is orthogonal to any function of the past of $Y_t$, idest

$$\mathbb{E}[\epsilon_t' Z_{t-1}] = 0, \forall Z_{t-1} \in Y_{\underline{t-1}}.$$

- Linear innovation:

$$\epsilon_t^\star = Y_t - \mathbb{E}L[Y_t | Y_{t-1}, Y_{t-2}, \cdots],$$

where $(\epsilon_t^\star)$ is a white noise and $\epsilon_t^\star$ is orthogonal to any linear function of the past of $Y_t$, idest

$$\mathbb{E}[\epsilon_t^{\star'} Z_{t-1}] = 0, \forall Z_{t-1} \in \mathcal{H}_Y(t-1).$$

# Wold representation theorem, 1938

## Theorem

*Any* **purely non deterministic** [a] *real centered second-order stationary process admits an infinite Moving Average (MA) representation*

$$X_t = \epsilon_t + \sum_{i=1}^{\infty} c_i \epsilon_{t-i}, \sum_{i=1} c_i^2 < \infty,$$

*where* $(\epsilon_t)$ *is the* **linear innovation process** *of* $(X_t)$, *that is*

$$\epsilon_t = X_t - \mathbb{E}[X_t | \mathcal{H}_X(t-1)],$$

*where* $\mathcal{H}_X(t-1)$ *is the Hilbert space generated by the random variables* $X_{t-1}, X_{t-2}, \cdots$

[a] $\bigcap_{n=-\infty}^{\infty} \mathcal{H}_X(n) = \{0\}$, where $\mathcal{H}_X(n)$ denotes, in the Hilbert space of centered and square integrable variables, the sub-space of the linear combinations of the variables $X_{n-i}, i \geq 0$. See Brockwell and Davis (1991)

# Moving averages

By truncating the infinite sum, we obtain the process

$$X_t(q) = \epsilon_t + \sum_{i=1}^{q} c_i \epsilon_{t-i},$$

which is called Moving Average of order $q$ (MA(q)).

We have

$$\|X_t(q) - X_t\|_2^2 = \mathbb{E}[\epsilon_t^2 \sum_{i>q} c_i^2] \xrightarrow[q\to\infty]{} 0.$$

For parsimony reasons, it is however preferable to work with the larger class of Autoregressive Moving Average (ARMA) processes.

# VARMA processes

A VARMA process is any stationary solution (when existing) $(Y_t)$ of the stochastic recurrence equation

$$Y_t - \sum_{i=1}^{p} A_i Y_{t-i} = c + \epsilon_t + \sum_{i=1}^{q} B_j \epsilon_{t-j},$$

where $(\epsilon_t)$ is a white noise, $A_i$ and $B_j$ are real $d \times d$ matrices, and $c$ is a $d$-vector.

The model can be written as

$$\Phi(B) Y_t = c + \Psi(B)\epsilon_t,$$

where $\Phi(B)$ and $\Psi(B)$ are lag polynomials.

# Existence of a nonanticipative solution

A solution $(Y_t)$ is called causal, or nonanticipative, if $Y_t$ can be written as a measurable function of the $\epsilon_s, s \leq t$.

## Proposition

*If*

$$det(\Phi(z)) = 0 \Rightarrow |z| > 1,$$

*then the VARMA model admits a* **unique nonanticipative stationary** *solution* $(Y_t)$ *of the form*

$$Y_t = d + \sum_{j=1}^{\infty} C_j \epsilon_{t-j}.$$

# Strengths and weaknesses of VARMA

VARMA models remain very much used (mainly VAR models in econometrics, because VARMA seem too complex).

The class is very flexible, especially when weak noise are used for the errors (cf the Wold rep.).

VARMA provide adequate fits of many vector series, and can be extended to incorporate seasonalities (SARMA) and even some types of non stationarity (ARIMA).

However, they are unable to capture some phenomena called nonlinear : volatility effects, change of regimes, bubbles...

Various alternative models have been proposed, some of which including hidden variables (apart from the noise).

# Random variance

The idea is to make the standard deviation random:

$$\epsilon_t = \sigma_t \eta_t,$$

where

- $(\eta_t)$ is an i.i.d. process, centered and with variance one.
- $(\sigma_t)$ is a process called volatility, $\sigma_t > 0$.
- the variables $\sigma_t$ and $\eta_t$ are independent.

# Two classes of models

- GARCH (Generalized AutoRegressive Conditional Heteroskedasticity): $\sigma_t \in \underline{\epsilon_{t-1}}$. For instance

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2.$$

- Stochastic volatility (SV) models: $\sigma_t \notin \underline{\epsilon_{t-1}}$. For instance, $\log \sigma_t^2 \sim AR(1)$:

$$\log \sigma_t^2 = \omega + \beta \log \sigma_{t-1}^2 + \tau\nu_t,$$

where $\nu_t$ is an i.i.d. process, with $\nu_t \perp \eta_t$ .

# Regime-switching modeling

If $d$ regimes have to be accounted, a latent variable $\Delta_t \in \{1, ..., d\}$ can be introduced as

$$Y_t = f(Y_{t-1}, \cdots, ; \epsilon_t; \Delta_t).$$

For instance, a switching regime-AR(1)

$$Y_t = f(\Delta_t) Y_{t-1} + \epsilon_t.$$

The model has to be completed by specifying the dynamics of the process $(\Delta_t)$:

- Could be i.i.d. (but this assumption is often too restrictive).
- A Markov chain.
- More complex processes ? not much used.

# State-space models

Many models can be written under the form

$$\begin{cases} y_t & = & M_t\alpha_t + d_t + u_t, \text{ Measurement equation} \\ \alpha_t & = & T_t\alpha_{t-1} + c_t + R_t v_t, \text{ Transition equation} \end{cases}$$

where $y_t \in \mathbb{R}^N$, $\alpha_t \in \mathbb{R}^m$ (state vector), $(u_t)$ and $(v_t)$ are two sequences of independent variables, valued in $\mathbb{R}^N$ and $\mathbb{R}^K$, such that

$$\mathbb{E}[u_t] = \mathbf{0}_N, \mathbb{E}[v_t] = \mathbf{0}_K, \mathrm{Var}(u_t) = H_t, \mathrm{Var}(v_t) = Q_t,$$

where $M_t$, $T_t$ and $R_t$ are non-random $N \times m$, $m \times m$ and $m \times K$ matrices, $d_t \in \mathbb{R}^N$ and $c_t \in \mathbb{R}^m$ non-random vectors.

# Example: MA model

$$y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}.$$

We then have the state-space representation

$$\begin{cases} y_t & = & (1, \theta_1, \cdots, \theta_q)\alpha_t, \\ \alpha_t & = & T\alpha_{t-1} + (\epsilon_t, 0, \cdots, 0)'. \end{cases}$$

$$\alpha_t = \begin{pmatrix} \epsilon_t \\ \epsilon_{t-1} \\ \vdots \\ \vdots \\ \epsilon_{t-q} \end{pmatrix}, \quad T = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

# VAR model

$$y_t - \mu = \phi_1(y_{t-1} - \mu) + \cdots + \phi_p(y_{t-p} - \mu) + \epsilon_t.$$

Vector representation:

$$\begin{pmatrix} y_t - \mu \\ y_{t-1} - \mu \\ \vdots \\ \vdots \\ y_{t-p+1} - \mu \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{pmatrix} \begin{pmatrix} y_{t-1} - \mu \\ y_{t-2} - \mu \\ \vdots \\ \vdots \\ y_{t-p} - \mu \end{pmatrix} + \begin{pmatrix} \epsilon_t \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}$$

$$\Leftrightarrow \alpha_t = \Phi\alpha_{t-1} + v_t.$$

The measurement equation is

$$y_t = (I, 0, \cdots, 0)\alpha_t + \mu.$$

# Stochastic-trend models

Structural model : involves unobservable variables which can be interpreted.

For instance, the series can be decomposed as the sum of a trend and a noise as

$$y_t = \mu_t + \epsilon_t, \text{ where } (\mu_t) \perp (\epsilon_t).$$

The trend can be modelled as

$$\left\{ \begin{array}{lcl} \mu_t & = & \mu_{t-1} + \beta_{t-1} + \eta_t, \\ \beta_t & = & \beta_{t-1} + \zeta_t, \end{array} \right.$$

where $(\eta_t) \perp (\zeta_t)$ are white noise with variances $\sigma_\eta^2$ and $\sigma_\zeta^2$.
The second equation introduces a stochastic slope in the random walk followed by $\mu_t$.

# Stochastic-trend models

State-space representation of the model:

$$\begin{cases} y_t & = & (1,0)\begin{pmatrix}\mu_t\\ \beta_t\end{pmatrix} + \epsilon_t, \\ \begin{pmatrix}\mu_t\\ \beta_t\end{pmatrix} & = & \begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}\begin{pmatrix}\mu_{t-1}\\ \beta_{t-1}\end{pmatrix} + \begin{pmatrix}\eta_t\\ \zeta_{t-1}\end{pmatrix}. \end{cases}$$

Let $\Delta = 1 - B$ the difference operator. The latent variables $\mu_t$ and $\beta_t$ can be eliminated from the representation:

$$\Delta^2 y_t = \zeta_t + \Delta\eta_t + \Delta^2\epsilon_t.$$

# Random coefficients models

With temporal data, the classical linear model writes

$$Y_t = X_t\beta + U_t,$$

where $X_t \in \mathbb{R}^K$ is a vector of exogenous variables.
Structural changes can motivate the introduction of random coefficients indexed by time :

$$Y_t = X_t\beta_t + U_t$$

For instance an AR(1) model can be set on the coefficients :

$$\beta_t = \Phi\beta_{t-1} + V_t.$$

The latter equation is the transition equation in a state-space model.

# Canonical stochastic volatility model

$$\begin{cases} \epsilon_t & = & \sqrt{h_t}\eta_t, \\ \log h_t & = & \omega + \beta \log h_{t-1} + \sigma v_t. \end{cases}$$

- Similar to diffusion models used in the financial literature.
- Positivity of $(h_t)$ does not entail constraints on the coefficients.
- Interpretation of the coefficients : $\omega$ is the level parameter; $\beta$ is the persistence parameter, in general $\beta > 0$; $\sigma > 0$ is the volatility of the volatility.

Assuming $\mathbb{P}(\eta_t = 0) = 0$, we obtain the state-space model

$$\begin{cases} \log \epsilon_t^2 & = & \log h_t + \log \eta_t^2, \\ \log h_t & = & \omega + \beta \log h_{t-1} + \sigma v_t. \end{cases}$$

# Some reminders regarding statistical inference

*Statistical model*: pair $(\mathcal{Z}, \mathcal{P})$, with
$\mathcal{Z}$ the observation space
$\mathcal{P}$ the family of probability distribution on $\mathcal{Z}$.

Basic assumption: the true distribution $\mathbb{P}_0$ belongs to $\mathcal{P}$.
*z* is the observation or result of the random experiment of the random function $Z$, whose distribution is $\mathbb{P}_0$.
The statistician is interested in a parameter $\theta$.

An identifiable parameter of interest is a function $\theta$ defined on $\mathcal{P}$ valued in $\Theta \subset \mathbb{R}^d$:

$$\mathcal{P} \to \Theta.$$

$\theta_0 = \theta(\mathbb{P}_0)$ is called the true value of the parameter.

One should always make sure that the parameter is identifiable (one-to-one mapping):

$$\mathbb{P}_\theta = \mathbb{P}_{\theta'} \Rightarrow \theta = \theta'.$$

## Examples of statistical parametric model

$$(\mathcal{Z}, \mathcal{P}) = (\{0,1\}^n, \text{Bernoulli}(\theta)^{\otimes n}, \theta \in [0,1]).$$

$$(\mathcal{Z}, \mathcal{P}) = (\mathbb{R}^n, \mathcal{N}(m, \sigma^2)^{\otimes n}, \theta = (m, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+).$$

From the parametric model, we have a likelihood function (observations, parameters).

# Information

Let $\mathbb{P}^*$ and $\mathbb{P}$ two probability distribution on $\mathcal{Z}$ with pdf $f^*(z)$ and $f(z)$.

## Definition

The Kullback information between $\mathbb{P}^*$ and $\mathbb{P}$ is

$$\mathcal{I}(\mathbb{P}^*|\mathbb{P}) = \mathbb{E}_{\mathbb{P}^*}[\log \frac{f^*(Z)}{f(Z)}].$$

It satisfies: $\mathcal{I}(\mathbb{P}|\mathbb{P}^*) \geq 0$ and $\mathcal{I}(\mathbb{P}|\mathbb{P}^*) = 0 \Leftrightarrow \mathbb{P} = \mathbb{P}^*$.

# Fisher information

We consider a parametric conditional model.
That is for $x$ fixed at the observed value of $X$, the family of possible conditional distributions of $Y$ given $X = x$ denoted $\mathcal{P}_x$, is defined by the pdf

$$\{f(y|x;\theta), \theta \in \Theta\}$$

such that

$$f(y|x;\theta) = \prod_{i=1}^{n} \tilde{f}(y_i|x_i;\theta).$$

The Fisher information matrix at $x$ is

$$I_F^x(\theta) = \mathrm{V}_\theta[\nabla_\theta \log f(Y|x;\theta)].$$

$V_\theta$ is the conditional variance covariance matrix of the distribution defined by $f(y|x;\theta)$.

# Kullback and Fisher

For any $x$, we can define the Kullback information between $f(y|x;\theta_0)$ and $f(y|x;\theta_1)$ for any pair $(\theta_0,\theta_1)$ as

$$\mathcal{I}_F^x(\theta_1|\theta_0) = \mathcal{I}(f(y|x;\theta_1)|f(y|x;\theta_0)),$$

with $x$ fixed. Then

$$[\nabla^2_{\theta\theta'}\mathcal{I}(\theta|\theta_0)]_{\theta=\theta_0} = I_F^x(\theta_0).$$

# Extremal estimator

## Definition

A statistic is a function of $Z \in \mathcal{Z}$ and an estimator of $\theta$ is a statistic into $\theta$.

## Definition

An extremal estimator of $\theta$ is a statistic $\hat{\theta}_n(Z)$ satisfying

$$\hat{\theta}_n(Z) = \arg\max_{\theta} \mathcal{L}_n(Z;\theta).$$

$\mathcal{L}_n(.;.)$ is a real function defined on $\mathcal{Z} \times \Theta$. In statistics, we look at the asymptotic properties of $\hat{\theta}_n$.

# Consistency

If

- $\Theta$ is compact,
- $\mathcal{L}_n(Z;\theta)$ is continuous,
- $\sup_{\theta} |\mathcal{L}_n(Z;\theta) - \mathcal{L}_\infty(Z;\theta)| \overset{\mathbb{P}}{\to} 0$,
- $\mathcal{L}_\infty(Z;\theta)$ has a unique maximum at $\theta_0$,

then $\hat{\theta} \overset{\mathbb{P}}{\to} \theta_0$.

# Asymptotic normality

If

- $\mathcal{L}_n(Z;.)$ is twice continuously differentiable,
- $\nabla^2_{\theta\theta'}\mathcal{L}_n(Z;\theta) \xrightarrow{\mathbb{P}} -J(\theta)$, $\|\theta-\theta_0\| \leq \|\hat{\theta}-\theta_0\|$,
- $J(\theta_0)$ is invertible,
- $\sqrt{n}\nabla_\theta\mathcal{L}_n(Z;\theta_0) \xrightarrow{d} \mathcal{N}_{\mathbb{R}^d}(0, I(\theta_0))$,

then $\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} \mathcal{N}_{\mathbb{R}^d}(0, J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0))$.

# M-estimators

General case:

$$Q_n(Y, X; \theta) = \frac{1}{n} \sum_{i=1}^{n} l(Y_i, X_i; \theta).$$

The model is consistent if there is a unique maximum at $\theta_0$.

The maximum likelihood method (MLE) is defined as

$$\hat{\theta} = \arg \min_{\theta} Q_n(Y, X; \theta).$$