

Lipari 2009 Biomathematics summer school.
Parameter estimation in physiological models.

Parameter estimation for stochastic differential equations
from noisy observations. Maximum likelihood and
filtering techniques.

V. Genon-Catalot
Laboratoire MAP5, CNRS UMR 8145
Université Paris Descartes
e-mail: Valentine.Genon-Catalot@parisdescartes.fr

Contents

1	I. Hidden Markov models. Exact likelihood.	3
1.1	Introduction	3
1.2	Hidden Markov models. The framework.	4
1.2.1	The hidden diffusion.	4
1.2.2	Observations.	4
1.2.3	Examples.	5
1.3	Filtering. Prediction. Marginals.	6
1.4	Likelihood.	9
1.4.1	Exact likelihood.	9
1.4.2	The Leroux method.	9
1.5	Asymptotic properties of the exact maximum likelihood estimator and related estimators.	10
2	Gaussian diffusions and additive noise.	15
2.1	Introduction	15
2.2	Model and observations.	15
2.3	Conditional distributions.	16
2.4	Exact likelihood and related contrasts.	18
2.5	ARMA property and consequences.	20
2.6	Non Gaussian noises.	21

2.7	Multidimensional extension.	21
2.8	Appendix	24
2.8.1	Iterations of the operators for predictive distribtuions.	25
3	Diffusions with multiplicative noise.	26
3.1	Introduction	26
3.2	Computable filters.	26
3.3	Hidden diffusion model. Prediction operator.	28
3.4	Noise distribution. Up-dating operator.	31
3.5	Algorithm for predictive distributions.	33
3.6	Exact likelihood.	34
3.7	Related models.	35
4	Other kernels.	37
4.1	Introduction	37
4.2	Conditional Poisson observations.	37
4.2.1	Algorithm for predictive distributions.	39
4.2.2	Exact likelihood.	40
4.2.3	Extension.	40
4.3	Wright-Fisher diffusion and conditional Bernoulli observations.	41

Chapter 1

I. Hidden Markov models. Exact likelihood.

1.1 Introduction

Consider a diffusion process $(x_t, t \geq 0)$ given as the solution of a stochastic differential equation with unknown parameters in the drift and diffusion coefficients to be estimated. For simplicity, we consider that (x_t) is one-dimensional but multidimensional processes may be considered too.

At times $0 \leq t_1 < \dots < t_n < \dots$, noisy observations $y_1, y_2, \dots, y_n, \dots$ of $x_{t_1}, \dots, x_{t_n}, \dots$ are taken: for instance, one observes $y_i = x_{t_i} + \varepsilon_i$ (additive noise) or $y_i = x_{t_i} \varepsilon_i$ (multiplicative noise). More generally, the distribution of y_i given x_{t_i} may be specified by a kernel $F(x_{t_i}, dy)$. We intend to describe the exact likelihood of such observations and focus on models where explicit computations are possible.

In our first lecture, we give the assumptions implying that $(y_i, i \geq 1)$ is a hidden Markov model. Under these assumptions, we explain how to compute the exact likelihood of (y_1, \dots, y_n) using the filtering-prediction algorithm. Some theoretical properties of the observed process $(y_i, i \geq 1)$ and of the exact maximum likelihood estimator are given.

In the second lecture, we study discrete observations of diffusions observed with additive noise. Explicit formulae for the likelihood can be obtained for Gaussian diffusions with Gaussian noise using the Kalman filter approach. Even when the noise is non Gaussian, the previous likelihood can be used as a contrast (quasi-likelihood) to obtain consistent and asymptotically Gaussian estimators.

In the third lecture, we consider discrete observations of diffusions observed with multiplicative noise. The square-root diffusion model (Cox-Ingersoll-Ross diffusion) with a specific multiplicative noise allows to obtain an explicit expression of the likelihood.

In the fourth lecture, we look at other kernels such as Poisson observations or binomial observations with stochastic parameters described by a diffusion.

1.2 Hidden Markov models. The framework.

1.2.1 The hidden diffusion.

Consider a one-dimensional diffusion process $(x_t, t \geq 0)$ described by:

$$dx_t = b(\theta, x_t)dt + \sigma(\theta, x_t)dW_t, \quad x_0 = \eta \quad (1.1)$$

where $(W_t)_{t \geq 0}$ is a Wiener process defined on a probability space (Ω, \mathcal{F}, P) , η is a random variable on Ω independent of (W_t) , $b(\theta, \cdot), \sigma(\theta, \cdot)$ are continuous functions on \mathbb{R} . The parameter θ is unknown and belongs to a parameter set $\Theta \subset \mathbb{R}^p$. Although d -dimensional diffusion processes could be considered too, we focus on the one-dimensional case for the sake of simplicity.

Classical assumptions on $b(\theta, \cdot), \sigma(\theta, \cdot)$ ensure that the stochastic differential equation (1.1) admits a unique strong solution and a unique stationary distribution π_θ (see *e.g.* Rogers and Williams (1990)). We assume that these assumptions hold and that (x_t) is in stationary regime, *i.e.* that the initial variable η has distribution π_θ . In this case, $(x_t, t \geq 0)$ is strictly stationary, ergodic, β -mixing and therefore α -mixing. For details, see *e.g.* Genon-Catalot *et al.* (2000). We denote by \mathcal{X} the state space of (x_t) which is an interval of the real line.

Let $P_{\theta,t}(x, dx')$ denote the transition semigroup of (x_t) . Under standard regularity assumptions on the functions $b(\theta, \cdot), \sigma(\theta, \cdot)$, we know that :

$$P_{\theta,t}(x, dx') = p_{\theta,t}(x, x')dx', \quad \pi_\theta(dx) = g_{\theta,t}(x)dx, \quad (1.2)$$

where dx denotes the Lebesgue measure on the state space \mathcal{X} and $p_{\theta,t}(x, x')$ is the transition density of $P_{\theta,t}(x, dx')$ (see *e.g.* Rogers and Williams (1990)). The transition semigroups of diffusion processes have very interesting properties (reversibility, spectral decomposition, eigenvalues, eigenfunctions, ergodicity ...).

1.2.2 Observations.

At equispaced instants $t_i = i\Delta$ with $\Delta > 0$, one takes measurements (x_{t_i}) but these measurements are not direct. We assume that, at time t_i , a random variable y_i is observed and that the following holds:

- (H1) (Conditional independence) Given $(x_{t_i}, i \geq 0)$, the random variables y_i are independent and the conditional distribution of y_i given $(x_{t_j}, j \geq 0)$ only depends on x_{t_i} .

- (H2) (Stationarity) The conditional distribution of y_i given $x_{t_i} = x$ does not depend on i .

Under these assumptions, the process (y_i) is called a hidden Markov model with hidden chain (x_{t_i}) (see *e.g.* Leroux (1992), Bickel and Ritov (1996), Cappé *et al.* (2005)). Now, we set a very important assumption concerning the kernel $F(x, dy)$ specifying the conditional distribution of y_i given $x_{t_i} = x$. We denote by \mathcal{Y} the state space of y_i .

- (H3) $F(x, dy) = f(y|x)\mu(dy)$ for some dominating positive measure μ on \mathcal{Y} .

This means that the observation y_i is a random function of x_{t_i} . The existence of the density $f(y|x)$ is crucial for the computation of the likelihood of (y_1, \dots, y_n) and of all the conditional distributions involved in filtering and prediction. Note that the transition operator of the hidden chain (x_{t_i}) is $P_\theta := P_{\theta, \Delta}$ where $(P_{\theta, t}, t \geq 0)$ is the transition semi-group of (x_t) . We denote below $p_\theta(x, x') := p_{\theta, \Delta}(x, x')$ the transition density of the hidden chain.

Let us now give some examples that will be detailed further on.

1.2.3 Examples.

- **Example 1: Additive noise.** Let (ε_i) be a sequence of real-valued i.i.d. random variables, having density $f(\cdot)$, independent of the whole process (x_t) . Assume that $y_i = x_{t_i} + \varepsilon_i, i \geq 0$. Then, assumptions are fulfilled and $f(y|x) = f(y - x)$. When (x_t) is Gaussian and f is a Gaussian density, we are in the Kalman filter model.
- **Example 2: Multiplicative noise.** With (ε_i) as in Example 1, consider $y_i = x_{t_i}\varepsilon_i$. Then, $f(y|x) = \frac{1}{x}f(\frac{y}{x})$. When ε_i has law $\mathcal{N}(0, 1)$, (y_i) is often called a stochastic volatility models (in discrete time). This model is investigated in Ruiz (1994), with (x_t) equal to the exponential of an Ornstein-Uhlenbeck process. With another kind of noise and with $(x_t = \sqrt{r_t})$ and (r_t) a Cox-Ingersoll-Ross diffusion, the model is treated in Chaleyat-Maurel and Genon-Catalot (2006). With (x_t) equal to the absolute value of an Ornstein-Uhlenbeck diffusion, the model is studied in Comte *et al.* (2008).
- **Example 3: Other kernels.** Suppose that, given (x_t) , y_i has Poisson distribution with parameter $\lambda(x_{t_i})$ for some continuous function $\lambda : \mathcal{X} \rightarrow (0, +\infty)$. Then $\mathcal{Y} = \mathbb{N}$ and $F(x, dy) = f(y|x)\mu(dy)$ with

$$f(y|x) = \exp(-\lambda(x)) \frac{\lambda(x)^y}{y!}, \quad \mu(y) = 1, y \in \mathbb{N}.$$

When $\lambda(x) = \lambda x + \mu$ (with $\lambda > 0, \mu \geq 0$) and $(x_t = r_t)$ with (r_t) a Cox-Ingersoll-Ross diffusion or the square of an Ornstein-Uhlenbeck process, explicit computations are possible (see Chaleyat-Maurel and Genon-Catalot (2006)). The continuous time version of this model is studied in Boel and Benes (1980).

Or, suppose that $\mathcal{X} = [0, 1]$ and that, given (x_t) , y_i has binomial distribution with parameters N, x_{t_i} . In this case, $\mathcal{Y} = \{0, 1, \dots, N\}$, $\mu(y) = 1$ for $y = 0, 1, \dots, N$ and $f(y|x) = \binom{N}{y} x^y (1-x)^{N-y}$. With (x_t) a Wright-Fisher diffusion process, the model is studied in Chaleyat-Maurel and Genon-Catalot (2008). With the same unobserved diffusion, geometric observations or negative binomial observations with parameter x also lead to explicit computations.

1.3 Filtering. Prediction. Marginals.

Recall the notations and assumptions. Since Δ is fixed, we set $x_i = x_{i\Delta}$. The following holds.

- (A0) The chain (x_i) is strictly stationary, ergodic, with transition operator $P_\theta(x, dx') = p_\theta(x, x') dx'$ and stationary distribution $\pi_\theta(dx) = g_\theta(x) dx$ where dx denotes the Lebesgue measure on the interval \mathcal{X} (state space of the hidden diffusion process).
- (A1) Given (x_i) , the random variables (y_i) are independent and the conditional distribution of y_i given $(x_i, j \geq 0)$ when $x_i = x$ is given by $F(x, dy) = f(y|x) \mu(dy)$ where μ is a positive measure on the state space \mathcal{Y} of y_i .

Proposition 1.3.1. *The joint process (x_i, y_i) is Markov with transition kernel*

$$Q(x, y; dx', dy') = p_\theta(x, x') f(y'|x') dx' \mu(dy').$$

Moreover, the process is strictly stationary, ergodic with marginal distribution $g_\theta(x) f(y|x) dx \mu(dy)$.

Proof. For $\psi : \mathcal{Y} \rightarrow \mathbb{R}^+$, set

$$h_\psi(x) = \int_{\mathcal{Y}} \psi(y) f(y|x) \mu(dy).$$

Consider positive functions $\varphi_i, i = 1, \dots, n$ on \mathcal{X} and positive functions $\psi_i, i = 1, \dots, n$ on \mathcal{Y} . Using successively the conditional independence property and the Markov property of (x_i) , we get:

$$\mathbb{E} \left(\prod_{i=1}^n \varphi_i(x_i) \psi_i(y_i) \right) = \mathbb{E} \left(\prod_{i=1}^n \varphi_i(x_i) h_{\psi_i}(x_i) \right) = \mathbb{E} \left(\prod_{i=1}^{n-1} \varphi_i(x_i) P_\theta(\varphi_n h_{\psi_n})(x_{n-1}) \right),$$

where

$$P_\theta(\varphi_n h_{\psi_n})(x) = \int_{\mathcal{X} \times \mathcal{Y}} \varphi_n(x') \psi_n(y') p_\theta(x, x') f(y'|x') dx' \mu(dy').$$

So, there appears the transition kernel $Q(x, y; dx', dy')$.

The stationarity is immediate. For the ergodicity (which is not immediate), we refer to *e.g.* Leroux (1992) for a direct proof or Genon-Catalot *et al.* (2000) where ergodicity is proved using α -mixing. \square

Now, we are in position to compute the following distributions:

- (Predictive distributions): $\nu_{i|i-1:1}(dx) = \mathcal{L}(x_i|y_{i-1}, \dots, y_1), i \geq 1$, with, by convention, $\nu_{1|0:1}(dx) = \mathcal{L}(x_1)$.
- (Filtering distributions): $\nu_{i|i:1}(dx) = \mathcal{L}(x_i|y_i, \dots, y_1), i \geq 1$
- (Marginal distributions): $\mu_{i|i-1:1}(dy) = \mathcal{L}(y_i|y_{i-1}, \dots, y_1), i \geq 1$ with, by convention, $\mu_{1|0:1}(dy) = \mathcal{L}(y_1)$.

Theorem 1.3.1. *The distributions $\nu_{i|i-1:1}(dx), \nu_{i|i:1}(dx)\mu_{i|i:1}(dx)$ can be recursively computed using the three following operators:*

1. *Up-dating operator: For ν a probability measure on \mathcal{X} , and $y \in \mathcal{Y}$, $\varphi_y(\nu)$ is the probability measure on \mathcal{X} defined by:*

$$\varphi_y(\nu)(dx) = \frac{f(y|x)\nu(dx)}{p_\nu(y)}, \quad \text{with} \quad p_\nu(y) = \int_{\mathcal{X}} \nu(dx)f(y|x). \quad (1.3)$$

2. *Prediction operator: For ν a probability measure on \mathcal{X} , $\psi_\theta(\nu) = \nu P_\theta$ is the probability on \mathcal{X} defined by:*

$$\nu P_\theta(dx') = \left(\int_{\mathcal{X}} \nu(dx)p_\theta(x, x') \right) dx' \quad (1.4)$$

3. *Marginal operator: For ν a probability measure on \mathcal{X} , we define the marginal distribution on \mathcal{Y} , $p_\nu(y)\mu(dy)$ with $p_\nu(y)$ given above.*

Then, the algorithm is as follows. Set $\nu_{1|0:1}(dx) = \mathcal{L}(x_1)$. we have, for all $i \geq 1$,

$$\begin{aligned} \text{(up-dating)} \quad \nu_{i|i:1} &= \varphi_{y_i}(\nu_{i|i-1:1}), & \text{(prediction)} \quad \nu_{i+1|i:1} &= \nu_{i|i:1} P_\theta, \\ \text{(marginal)} \quad \mu_{i|i-1:1}(dy) &= p_{\nu_{i|i-1:1}}(y)\mu(dy). \end{aligned}$$

Proof. To simplify notations, we denote by $p(z_1, \dots, z_n)$ the density of any n -tuple (z_1, \dots, z_n) of random variables and by $p(z_i|z_1, \dots, z_n)$ the conditional density of a random variable z_i given (z_1, \dots, z_n) . We use the symbol \propto to ignore constants in densities. For the first up-dating, it is immediate that:

$$\nu_{1|1:1}(dx) \propto \nu_{1|0:1}(dx)f(y|x).$$

Hence, $\nu_{1|1:1} = \varphi_{y_1}(\nu_{1|0:1})$. Then, we have

$$p(x_i|y_i, \dots, y_1) \propto p(x_i, y_i, \dots, y_1),$$

where

$$\begin{aligned} p(x_i, y_i, \dots, y_1) &= \int_{\mathcal{X}^{i-1}} g_\theta(x_1) f(y_1|x_1) \left(\prod_{j=2}^i p_\theta(x_{j-1}, x_j) f(y_j|x_j) \right) dx_1 \dots dx_{i-1} \\ &= p(x_i, y_{i-1}, \dots, y_1) f(y_i|x_i) \propto \nu_{i|i-1:1}(dx_i) f(y_i|x_i) \propto \varphi_{y_i}(\nu_{i|i-1:1})(dx_i). \end{aligned}$$

For the prediction, we have

$$p(x_{i+1}|y_i, \dots, y_1) \propto p(x_{i+1}, y_i, \dots, y_1),$$

with

$$\begin{aligned} p(x_{i+1}, y_i, \dots, y_1) &= \int_{\mathcal{X}^i} p(x_1, \dots, x_i, x_{i+1}, y_1, \dots, y_i) dx_1 \dots dx_i \\ &= \int_{\mathcal{X}^i} g_\theta(x_1) f(y_1|x_1) \left(\prod_{j=2}^i p_\theta(x_{j-1}, x_j) f(y_j|x_j) \right) p_\theta(x_i, x_{i+1}) dx_1 \dots dx_i \\ &= \int dx_i p(x_i, y_i, \dots, y_1) p_\theta(x_i, x_{i+1}) \propto \nu_{i|i:1} P_\theta(dx_{i+1}). \end{aligned}$$

For marginals, we have

$$\begin{aligned} p(y_i|y_{i-1}, \dots, y_1) &\propto p(y_1, \dots, y_i) = \int_{\mathcal{X}^i} p(x_1, \dots, x_i, y_1, \dots, y_i) dx_1 \dots dx_i \\ &= \int_{\mathcal{X}^i} g_\theta(x_1) f(y_1|x_1) \left(\prod_{j=2}^{i-1} p_\theta(x_{j-1}, x_j) f(y_j|x_j) \right) p_\theta(x_{i-1}, x_i) f(y_i|x_i) dx_1 \dots dx_i \\ &= \int dx_i p(x_i, y_{i-1}, \dots, y_1) f(y_i|x_i) \propto \int \nu_{i|i-1:1}(dx_i) f(y_i|x_i). \end{aligned}$$

Hence, the results. \square

Let us stress that, although the steps of the prediction-filtering algorithm are simple, there are few models where computations can be done explicitly. Let us define the two compound operators:

$$\Phi_y = \psi_\theta \circ \varphi_y, \quad \text{and} \quad \Psi_y = \varphi_y \circ \psi_\theta. \quad (1.5)$$

The iterations of Φ_y define the algorithm for predictive distributions. We have

$$\nu_{i|i-1:1} = \Phi_{y_{i-1}} \circ \dots \circ \Phi_{y_1}(\nu_{1|0:1}).$$

The iterations of Ψ_y define the algorithm for filtering distributions:

$$\nu_{i|i:1} = \Psi_{y_i} \circ \dots \circ \Psi_{y_2}(\nu_{1|1:1}).$$

In filtering theory, authors generally concentrate on the filtering distributions $\nu_{i|i:1}$. For statistical inference, the predictive distributions are more important because they give the marginal distributions used to compute the likelihood.

1.4 Likelihood.

1.4.1 Exact likelihood.

Several formulae are available for the exact likelihood of (y_1, \dots, y_n) . First, we can integrate the joint density of $(x_1, \dots, x_n, y_1, \dots, y_n)$ with respect to x_1, \dots, x_n . This gives:

$$p_n(\theta, y_1, \dots, y_n) = \int_{\mathcal{X}^n} dx_1 \dots dx_n g_\theta(x_1) f(y_1|x_1) \prod_{i=2}^n p_\theta(x_{i-1}, x_i) f(y_i|x_i). \quad (1.6)$$

Setting

$$p_n(\theta, y_1, \dots, y_n|x_1) = f(y_1|x_1) \int_{\mathcal{X}^{n-1}} dx_2 \dots dx_n \prod_{i=2}^n p_\theta(x_{i-1}, x_i) f(y_i|x_i) \quad (1.7)$$

which is the conditional density of (y_1, \dots, y_n) given x_1 , we obtain:

$$p_n(\theta, y_1, \dots, y_n) = \int_{\mathcal{X}} dx_1 g_\theta(x_1) p_n(\theta, y_1, \dots, y_n|x_1). \quad (1.8)$$

Finally, we have

$$p_n(\theta, y_1, \dots, y_n) = p_1(\theta, y_1) \prod_{i=2}^n p_i(\theta, y_i|y_{i-1}, \dots, y_1), \quad (1.9)$$

where $p_i(\theta, y_i|y_{i-1}, \dots, y_1) = p_{\nu_{i|i-1:1}}(y_i)$ is the conditional density of y_i given (y_{i-1}, \dots, y_1) . The latter formula gives a recursive way of computing the exact likelihood provided that the successive marginal distributions are explicitly computable.

1.4.2 The Leroux method.

Formula (1.8) suggests, following Leroux (1992), to consider other functions of (y_1, \dots, y_n) and θ that can be used as contrast functions to build maximum contrast estimators. For g a probability density on \mathcal{X} , set

$$p_n^g(\theta, y_1, \dots, y_n) = \int_{\mathcal{X}} dx_1 g(x_1) p_n(\theta, y_1, \dots, y_n|x_1). \quad (1.10)$$

Analogously, we have:

$$p_n^g(\theta, y_1, \dots, y_n) = p_1^g(\theta, y_1) \prod_{i=2}^n p_i^g(\theta, y_i|y_{i-1}, \dots, y_1), \quad (1.11)$$

where the successive terms of the product are computed via the filtering-prediction-marginal algorithm starting with the initial density g . The interest of this approach is that, for a well-chosen g , p_n^g may be much simpler to compute than the exact likelihood. Moreover, Leroux's

paper, completed by Genon-Catalot and Larédo (2006), proves that the estimators computed as maximising the p_n^g have the same asymptotic properties as the exact maximum likelihood estimator.

1.5 Asymptotic properties of the exact maximum likelihood estimator and related estimators.

Denote by Θ the parameter set. Let $\hat{\theta}_n$ (resp. $\tilde{\theta}_n^g$) be any solution of

$$p_n(\hat{\theta}_n, y_1, \dots, y_n) = \sup_{\theta \in \Theta} p_n(\theta, y_1, \dots, y_n),$$

(resp.

$$p_n^g(\tilde{\theta}_n^g, y_1, \dots, y_n) = \sup_{\theta \in \Theta} p_n^g(\theta, y_1, \dots, y_n).)$$

The study of $\hat{\theta}_n$ (and $\tilde{\theta}_n^g$) is a difficult problem that has come recently to a final solution obtained by Chen Der Fuh (2006). Before, several papers have dealt with the subject and we may quote:

- Leroux (1992) proves the strong consistency of $\hat{\theta}_n$ and $\tilde{\theta}_n^g$, under very mild assumptions, when the state space of the hidden chain is finite.
- Bickel *et al.* (1998) prove the asymptotic normality of $\hat{\theta}_n$ also for \mathcal{X} finite under additional assumptions.
- Douc and Matias (2001) prove the strong consistency and the asymptotic normality of $\hat{\theta}_n$ when \mathcal{X} is compact.
- Cheng Der Fuh (2006) proves the strong consistency and the asymptotic normality of $\hat{\theta}_n$ for a general state space \mathcal{X} (including the non compact case), under reasonable assumptions on the model.
- In between, Genon-Catalot and Larédo (2006) study, in the spirit of Leroux's paper, the likelihood (1.8) and the contrast (1.10) for a non compact state space \mathcal{X} and some related properties. A special attention is given to the Kalman filter model. In particular, an adequate choice of g in (1.10) simplifies considerably the formula and the study of the associated estimator and gives a clear identifiability assumption. See also Genon-Catalot *et al.* (2003).

Since Cheng Der Fuh (2006) contains the most complete result, we give below the set of assumptions ensuring consistency and asymptotic normality of the exact maximum likelihood estimator in hidden Markov models in the simplified case where (H3) holds, *i.e.* the conditional density of y_n given $x_n = x$ does not depend on unknown parameters.

The Markov chain (x_n) has state space \mathcal{X} , its transition kernel $P_\theta(x, dx')$ depends on an unknown parameter $\theta \in \Theta \subset \mathbb{R}^q$ and $P_\theta(x, dx') = p_\theta(x, x')m(dx')$ for some dominating measure m on \mathcal{X} . It admits a stationary distribution such that $\pi_\theta(dx) = g_\theta(x)m(dx)$. The observed process (y_n) takes values in \mathbb{R}^d and the conditional distribution of y_n given $x_n = x$ is given by $F(x, dy) = f(y|x)\mu(dy)$ for some dominating measure μ . Denote by \mathbb{E}_x^θ the conditional expectation given that $x_0 = x$. Then, the assumptions are the following:

- (C1) (x_n) is an irreducible, aperiodic Markov chain and for some weight function $w : \mathcal{X} \rightarrow [1, +\infty)$, $\int w(x)\pi_\theta(dx) < \infty$ for all θ . Moreover, for all θ ,

$$\lim_{n \rightarrow +\infty} \sup_{x \in \mathcal{X}} \{ |\mathbb{E}_x^\theta(h(x_n)) - \int h(x)\pi_\theta(dx)| / w(x) ; |h| \leq w \} = 0.$$

For some $p \geq 1$,

$$\sup_{x \in \mathcal{X}} \{ \mathbb{E}_x^\theta(w(x_p)) / w(x) \} < \infty.$$

For all $x, x' \in \mathcal{X}$, $0 < p_\theta(x, x') < \infty$, for all $y \in \mathbb{R}^d$, $0 < \sup_{x \in \mathcal{X}} f(y|x) < \infty$. Setting $h_\theta(y_1) = \sup_{x \in \mathcal{X}} \int p_\theta(x, x') f(y_1|x') m(dx')$, we assume, that, for some $p \geq 1$,

$$\sup_{x \in \mathcal{X}} \mathbb{E}_x^{\theta_0} \left[\log \left(h_\theta^p(y_1) \frac{w(x_p)}{w(x)} \right) \right] < 0,$$

$$\sup_{x \in \mathcal{X}} \mathbb{E}_x^{\theta_0} \left(h_\theta(y_1) \frac{w(x_1)}{w(x)} \right) < +\infty.$$

- (C2) The true value θ_0 of the parameter belongs to the interior of Θ . For all x, x' , $\theta \rightarrow p_\theta(x, x')$, $\theta \rightarrow g_\theta(x)$ are continuous on Θ and admit twice continuous derivatives in a neighborhood $N_\delta(\theta_0) = \{\theta, |\theta - \theta_0| < \delta\}$.

- (C3) For

$$\varphi_\theta(x) = \partial g_\theta / \partial \theta_i(x) \quad \text{or} \quad \partial^2 g_\theta / \partial \theta_i \partial \theta_j(x)$$

and

$$\psi_\theta(x, x') = \partial p_\theta / \partial \theta_i(x, x') \quad \text{or} \quad \partial^2 p_\theta / \partial \theta_i \partial \theta_j(x, x'),$$

with $i, j = 1, \dots, q$,

$$\int_{\mathcal{X}} \sup_{\theta \in N_\delta(\theta_0)} |\varphi_\theta(x)| m(dx) < \infty, \quad \int_{\mathcal{X}} \sup_{\theta \in N_\delta(\theta_0)} |\psi_\theta(x, y)| m(dy) < \infty.$$

- (C4)

$$\sup_{x \in \mathcal{X}} \mathbb{E}_x^{\theta_0} \left(\sup_{x', x'' \in \mathcal{X}} \frac{f(y_0|x') f(y_1|x')}{f(y_0|x'') f(y_1|x'')} \right)^2 < \infty$$

- (C5) $p_n(\theta, y_1, \dots, y_n) = p_n(\theta', y_1, \dots, y_n)$ for all $n \geq 0$ P_{θ_0} a.s. implies $\theta = \theta'$.

- (C6) For all $x \in \mathcal{X}$,

$$\mathbb{E}_x^{\theta_0}(|\log f(y_1|x)|) < \infty.$$

Condition (C1) states that the chain (x_n) is w -uniformly ergodic and implies

$$\begin{aligned} \exists \gamma > 0, 0 < \rho < 1, \forall h \in \mathbf{B} := \{h : \mathcal{X} \rightarrow C, \|h\|_w = \sup_x |h(x)|/w(x) < \infty\}, \\ \sup_{x \in \mathcal{X}} \{|\mathbb{E}_\theta h(x_n)|_{x_0 = x} - \int h(x)\pi_\theta(dx)\}/w(x)\} \leq \gamma \rho^n \|h\|_w. \end{aligned}$$

Condition (C2)-(C3) are standard regularity assumptions. Condition (C5) is the identifiability assumption. Condition (C4) is a technical assumption required for the existence of the Fisher information.

Denoting by $l_n(\theta) = \log p_n(\theta, y_1, \dots, y_n)$ for the loglikelihood, the usual properties hold for the score function and the Hessian:

$$\left(\frac{1}{\sqrt{n}}l'_{n,i}(\theta_0), i = 1, \dots, q\right) \rightarrow_{\mathcal{D}} \mathcal{N}_q(0, I(\theta_0)), \quad \left(\frac{1}{n}l''_{n,i,j}(\theta_0), i, j = 1, \dots, q\right) \rightarrow_{P_{\theta_0}} -I(\theta_0),$$

where the limiting Fisher information matrix $I(\theta_0)$ is expressed in a theoretical way (Lemma 5 in Cheng Der Fuh, 2006, p.23). The following theorem holds:

- Theorem 1.5.1.** • Assume (C1)-(C2) and (C5)-(C6) and let $\hat{\theta}_n$ be the MLE based on (y_1, \dots, y_n) . Then, $\hat{\theta}_n \rightarrow \theta_0$ IP_{θ_0} -a.s. as n tends to infinity.
- Assume (C1)-(C6). If the Fisher information matrix $I(\theta)$ is positive definite for θ in a neighborhood $N_\delta(\theta_0)$ of θ_0 , then, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to $\mathcal{N}_q(0, I^{-1}(\theta_0))$.

Bibliography

- [1] Bickel P.J., Ritov Y., Ryden T. (1998). Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.* **26** 1614-1635.
- [2] Boel R.K., Benes V.E. (1980). Recursive nonlinear estimation of a diffusion acting as the rate of an observed Poisson process *IEEE Transactions on information theory* **26** (5), 561-575.
- [3] Cappé O., Moulines E. and Rydèn T. (2005). *Inference in hidden Markov models*, Springer.
- [4] Chaleyat-Maurel M. and Genon-Catalot V. (2006). Computable infinite-dimensional filters with applications to discretized diffusion processes. *Stoch. Proc. and Applic.* **116**, 1447-1467.
- [5] Chaleyat-Maurel M. and Genon-Catalot V. (2008). Filtering the Wright-Fisher diffusion, to appear in *ESAIM P & S*.
- [6] Cheng-Der Fuh (2006). Efficient likelihood estimation in state space models. *Ann. Statist.* **34** (4), 2026-2068.
- [7] Comte F., Genon-Catalot V. and Kessler M. (2007). Multiplicative Kalman filtering, Prépublication 2007-16, MAP5, *Laboratoire de mathématiques appliquées de Paris Descartes*, submitted.
- [8] Douc R., Matias C. (2001). Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli* **7** (3) 381-420.
- [9] Genon-Catalot V. (2003). A non-linear explicit filter. *Statist. Probab. Lett.*, **61**, 145-154.
- [10] Genon-Catalot V., Jeantheau T. and Larédo C. (2000). Stochastic volatility models as hidden Markov models and statistical applications. *Bernoulli* **6**, 1051-1079.
- [11] Genon-Catalot V., Jeantheau T. and Larédo C. (2003). Conditional likelihood estimators for hidden Markov models and stochastic volatility models. *Scand. J. Statist.* **30** (2), 297-316.
- [12] Genon-Catalot V. and Kessler M. (2004). Random scale perturbation of an AR(1) process and its properties as a nonlinear explicit filter. *Bernoulli* (**10**) (4), 701-720.

- [13] Genon-Catalot V. and Larédo C. (2006). Leroux's method for general hidden Markov models. *Stoch. Proc. and Appl.* **116**, 222-243.
- [14] Leroux B.G. (1992). Maximum likelihood estimation for hidden Markov models, *Stoch. Proc. appl.* **40** 127-143.
- [15] Rogers L.C.G., Williams D. (1990). Diffusions, Markov processes and martingales. Volume 2. Itô Calculus. Wiley series in probability and mathematical statistics.
- [16] Ruiz E.(1994). Quasi-maximum likelihood estimation of stochastic volatility models. *Journal of Econometrics* **63**, 289-306.

Chapter 2

Gaussian diffusions and additive noise.

2.1 Introduction

In this second lecture, we consider a Gaussian diffusion, *i.e.* an Ornstein-Uhlenbeck process, observed in discrete time with an additive perturbation. When the perturbation is Gaussian, this model is the well-known Kalman filter, which is of common use in the field of filtering (on-line estimation of the unobserved data). Likelihood inference in such models is also rather standard. Nevertheless, it is indeed difficult and we intend to focus on some maybe less known features of this model. In particular, we insist on the links with hidden Markov models and the classical Gaussian likelihood theory. We give the expression of the exact likelihood and some related contrasts which yield estimators asymptotically equivalent to the exact maximum likelihood estimator. For simplicity, we consider the one-dimensional Ornstein-Uhlenbeck process first and give indications for the multidimensional case later on. There is a huge number of references on the subject. Therefore, it is difficult to give an exhaustive list. Since we rely mostly on hidden Markov models, we refer to Brockwell and Davies (1991), Cappé *et al.* (2005) for general properties and to specific papers quoted in the text.

2.2 Model and observations.

Let $(x(t), t \geq 0)$ be given by:

$$dx(t) = \alpha x(t)dt + \sigma dW_t, \quad x(0) = \eta \tag{2.1}$$

with η a real random variable independent of the Brownian motion W . We assume that $\alpha < 0$ and set $\theta = (\alpha, \sigma^2)$ for the unknown parameter. This process admits a stationary distribution $\pi_\theta(dx) = \mathcal{N}(0, \sigma_s^2(\theta))$ with

$$\sigma_s^2(\theta) = \frac{\sigma^2}{2|\alpha|}, \tag{2.2}$$

and we assume that the distribution of η is the stationary distribution so that the process $(x(t))$ is strictly stationary, Gaussian and ergodic. Solving (2.1) yields, for all $t, h \geq 0$:

$$x(t+h) = e^{\alpha h} x(t) + Z_{t,h}, \quad (2.3)$$

where

$$Z_{t,h} = \sigma e^{\alpha(t+h)} \int_t^{t+h} e^{-\alpha s} dW_s$$

is independent of $\mathcal{F}_t = \sigma(\eta, W_s, s \leq t)$ and has distribution $\mathcal{N}(0, \beta^2(h))$ with

$$\beta^2(h) = \sigma^2 \frac{e^{2\alpha h} - 1}{2\alpha}. \quad (2.4)$$

Considering a sampling interval Δ , the observation times $t_i = i\Delta$ and setting $x_i = x(i\Delta)$ for the discretized process, we get:

$$x_i = ax_{i-1} + \beta\eta_i, \quad x_0 = \eta, \quad (2.5)$$

where $a = e^{\alpha\Delta}$, $\beta = \beta(\Delta)$ and $(\eta_i \geq 1)$ is a sequence of i.i.d. random variables independent of η with distribution $\mathcal{N}(0, 1)$. Note that obviously, for all Δ ,

$$\sigma_s^2(\theta) = \frac{\beta^2}{1 - a^2}.$$

At time t_i , the observation is

$$y_i = x_i + \varepsilon_i, \quad (2.6)$$

where (ε_i) is a sequence of i.i.d. random variables with law $\mathcal{N}(0, \gamma^2)$ independent of $(x(t))$. We assume that γ^2 is known. The joint process (x_i, y_i) is a hidden Markov model (H1-H2 are immediate). The hidden chain (x_i) has state space $\mathcal{X} = \mathbb{R}$, transition kernel $P_\theta(x, dx') = \mathcal{N}(ax, \beta^2)(dx')$ and transition density

$$p_\theta(x, x') = \frac{1}{\beta\sqrt{2\pi}} \exp\left(-\frac{(x' - ax)^2}{2\beta^2}\right).$$

The observation space is $\mathcal{Y} = \mathbb{R}$ and $F(x, dy) = f(y|x)dy$ with

$$f(y|x) = \frac{1}{\gamma\sqrt{2\pi}} \exp\left(-\frac{(y-x)^2}{2\gamma^2}\right).$$

2.3 Conditional distributions.

To compute the exact likelihood associated with the observation (y_1, \dots, y_n) , we need to compute the predictive distributions $\nu_{i|i-1:1}^\theta(dx) = \mathcal{L}(x_i|y_{i-1}, \dots, y_1)$, $i \geq 1$, from which we derive the conditional densities

$$p_i(\theta, y_i|y_{i-1}, \dots, y_1) = p_{\nu_{i|i-1:1}^\theta}(y_i) = \int f(y_i|x)\nu_{i|i-1:1}^\theta(dx).$$

Usually, in the case of the Kalman filter model, these distributions are directly computed using the fact that they are all Gaussian. Hence, one can compute directly the conditional means and variances. This approach has a drawback. It cannot be generalized to non Gaussian models. On the contrary, the hidden Markov model approach is general. Moreover, in the case of the Kalman model, it is especially simple. Indeed, we only need to compute the up-dating operator φ_y , the prediction operator ψ_θ . Then, the compound operator $\Phi_y^\theta = \psi_\theta \circ \varphi_y$ gives the algorithm for predictive distributions. We recover the special feature of this model as these operators evolve within the family of Gaussian distributions (with the convention that a Dirac mass δ_x is considered as a Gaussian distribution with mean x and nul variance).

Proposition 2.3.1. • (Up-dating operator) If $\nu = \mathcal{N}(m, \sigma^2)$ and $y \in \mathbb{R}$, then, $\varphi_y(\nu) = \mathcal{N}(\hat{m}(y), \hat{\sigma}^2)$ with

$$\hat{m}(y) = \hat{\sigma}^2 \left(\frac{y}{\gamma^2} + \frac{m}{\sigma^2} \right), \quad \hat{\sigma}^2 = \frac{\sigma^2 \gamma^2}{\sigma^2 + \gamma^2}.$$

• (Prediction operator) If $\nu = \mathcal{N}(m, \sigma^2)$, then $\psi_\theta(\nu) = \nu P_\theta = \mathcal{N}(\bar{m}(y), \bar{\sigma}^2)$ with

$$\bar{m}(y) = am, \quad \bar{\sigma}^2 = \beta^2 + a^2 \sigma^2.$$

• (Marginal operator) If $\nu = \mathcal{N}(m, \sigma^2)$, $p_\nu(y) = \mathcal{N}(m, \sigma^2 + \gamma^2)$.

Proof. These results are elementary. For the first and third points, consider (X, Y) such that $X \sim \nu = \mathcal{N}(m, \sigma^2)$ and given $X = x$, $Y \sim \mathcal{N}(x, \gamma^2)$, then, $\varphi_y(\nu)$ is exactly the conditional distribution of X given $Y = y$ and $p_\nu(y)dy$ is the marginal distribution of Y . For the second point, consider (X, X') with $X \sim \nu = \mathcal{N}(m, \sigma^2)$ and $X' = aX + \eta_1$ with $\eta_1 \sim \mathcal{N}(0, \beta^2)$ independent of X , then $\psi_\theta(\nu)$ is the distribution of X' . \square

Corollary 2.3.1. (Operator for predictive distributions) If $\nu = \mathcal{N}(m, \sigma^2)$, then, $\Phi_y^\theta(\nu) = \psi_\theta \circ \varphi_y(\nu) = \mathcal{N}(\tilde{m}, \tilde{\sigma}^2)$ with

$$\tilde{m} := \Phi_y^\theta(m, \sigma^2) = a \left(m\delta(\sigma^2) + y(1 - \delta(\sigma^2)) \right), \quad \text{with} \quad \delta(\sigma^2) = \frac{\gamma^2}{\gamma^2 + \sigma^2},$$

$$\tilde{\sigma}^2 := \Phi^\theta(\sigma^2) = \beta^2 + a^2 \sigma^2 \delta(\sigma^2)$$

The corollary is an obvious consequence of the previous proposition. Let us stress some specific features which appear now. First, since only Gaussian distributions are involved, they are completely specified by their mean and variances. Hence, the operators acting on measures are simplified into operators acting on $\mathbb{R} \times \mathbb{R}^+$. This corresponds to the definition of a finite-dimensional filter. Moreover, note that the variance of $\Phi_y(\nu)$ only depends on σ^2 and neither on m nor on y). Now, we state the following useful lemma.

Lemma 2.3.1. The function

$$v \rightarrow \Phi^\theta(v) = \beta^2 + a^2 v \delta(v) = \beta^2 + a^2 v \frac{\gamma^2}{\gamma^2 + v}$$

giving the variance of the prediction algorithm is increasing from $I = [\beta^2, \frac{\beta^2}{1-a^2}]$ onto I and Lipschitz with constant a^2 . Consequently, Φ^θ admits a unique fixed point $\sigma_\infty^2(\theta) \in I$ defined as the solution of

$$\Phi^\theta(\sigma_\infty^2) = \sigma_\infty^2(\theta).$$

Moreover, for all $v \in I$, the n -th iterate $\Phi^\theta \circ \Phi^\theta \circ \dots \circ \Phi^\theta(v)$ tends with exponential rate a^{2n} to $\sigma_\infty^2(\theta)$ as n tends to infinity. In particular, this holds for $v = \frac{\beta^2}{1-a^2}$.

The proof is elementary and omitted. The fixed point can be explicitly computed since it solves a simple second degree equation. We do not need its explicit expression.

Now, for further use, we introduce the dependence on the initial distribution (distribution of x_1) in the notations. We denote by $\nu_{i|i-1:1}^\theta(\nu)$ the conditional distribution of x_i given y_{i-1}, \dots, y_1 when $\mathcal{L}(x_1) = \nu = \nu_{1|0:1}^\theta(\nu)$. With $\nu = \mathcal{N}(m, \sigma^2)$, we have:

$$\nu_{i|i-1:1}^\theta(\nu) = \mathcal{N}(m_{i|i-1:1}(\theta, (m, \sigma^2)), \sigma_{i|i-1:1}^2(\theta, \sigma^2)) = \Phi_{y_{i-1}}^\theta \circ \dots \circ \Phi_{y_1}^\theta(\nu). \quad (2.7)$$

The predictive conditional mean

$$\mathbb{E}_\nu^\theta(x_{i+1}|y_i, \dots, y_1) = m_{i+1|i:1}(\theta, (m, \sigma^2)) = \Phi_{y_i}^\theta(m_{i|i-1:1}(\theta, (m, \sigma^2))), \quad m_{1|0:1}(\theta, (m, \sigma^2)) = m \quad (2.8)$$

depends on the initial distribution (through (m, σ^2)) and the previous observations (y_i, \dots, y_1) . And, the conditional variance of x_{i+1} given y_i, \dots, y_1

$$\sigma_{i+1|i:1}^2(\theta, \sigma^2) = \Phi^\theta(\sigma_{i|i-1:1}^2(\theta, \sigma^2)), \quad \sigma_{1|0:1}^2(\theta, \sigma^2) = \sigma^2 \quad (2.9)$$

is deterministic. Using Lemma 2.3.1, it converges as i tends to infinity to the fixed point $\sigma_\infty^2(\theta)$ of Φ^θ . Finally, the conditional distribution of y_i given y_{i-1}, \dots, y_1 when the initial distribution is $\nu = \mathcal{N}(m, \sigma^2)$ is equal to

$$p_i^{(m, \sigma^2)}(\theta, y_i|y_{i-1}, \dots, y_1) = \mathcal{N}(m_{i|i-1:1}(\theta, (m, \sigma^2)), \gamma^2 + \sigma_{i|i-1:1}^2(\theta, \sigma^2)) \quad (2.10)$$

The distribution of y_1 is equal to

$$p_1^{(m, \sigma^2)}(\theta, y_1) = \mathcal{N}(m_{1|0:1}(\theta, (m, \sigma^2)), \gamma^2 + \sigma_{1|0:1}^2(\theta, \sigma^2)) = \mathcal{N}(m, \gamma^2 + \sigma^2).$$

2.4 Exact likelihood and related contrasts.

If the initial distribution is $\nu = \mathcal{N}(m, \sigma^2)$, then the exact likelihood is

$$p_n^{(m, \sigma^2)}(\theta, y_1, \dots, y_n) = p_1^{(m, \sigma^2)}(\theta, y_1) \prod_{i=2}^n p_i^{(m, \sigma^2)}(\theta, y_i|y_{i-1}, \dots, y_1).$$

Therefore, under our assumptions, the exact likelihood corresponds to $\nu = \pi_\theta$, *i.e.* $(m, \sigma^2) = (0, \sigma_s^2(\theta))$. For the exact likelihood, we simply omit the superscript in the notation and write

$$p_n^{(0, \sigma_s^2(\theta))}(\theta, y_1, \dots, y_n) = p_n(\theta, y_1, \dots, y_n).$$

The exact maximum likelihood estimator $\hat{\theta}_n$ is computed as a maximiser of the above function. Nevertheless, the function $p_n^{(m, \sigma^2)}(\theta, y_1, \dots, y_n)$ can be considered as a contrast function and the associated maximum contrast estimator can be studied.

Let us stress the fact that the expression of the exact likelihood is far from simple. Indeed, we have:

$$p_n(\theta, y_1, \dots, y_n) \propto \prod_{i=1}^n (\gamma^2 + \sigma_{i|i-1:1}^2(\theta))^{-1/2} \exp \left[-\frac{(y_i - m_{i|i-1:1}(\theta))^2}{\gamma^2 + \sigma_{i|i-1:1}^2(\theta)} \right], \quad (2.11)$$

where we have set for simplicity:

$$\begin{aligned} m_{i|i-1:1}(\theta) &= m_{i|i-1:1}(\theta, (0, \sigma_s^2(\theta))) = \Phi_{y_{i-1}}^\theta \circ \dots \circ \Phi_{y_1}^\theta((0, \sigma_s^2(\theta))), \\ \sigma_{i|i-1:1}^2(\theta) &= \sigma_{i|i-1:1}^2(\theta, \sigma_s^2(\theta)) = \Phi^\theta \circ \dots \circ \Phi^\theta(\sigma_s^2(\theta)). \end{aligned}$$

The iterations can be solved explicitly but lead to rather complicated sums. Nevertheless, the computation of the exact maximum likelihood estimator is numerically feasible. The score function and the hessian can be recursively computed.

To study the asymptotic behaviour of the exact maximum likelihood estimator, a first approach is to use Cheng Der Fuh (2006) and check the assumptions recalled in the first lecture. This is done in the latter paper (the weight function to be chosen is $w(x) = |x| + 1$). Consistency and asymptotic normality with rate \sqrt{n} hold. This is evidently well known and can be obtained by a classical approach for Gaussian likelihoods (see below).

Now, following Genon-Catalot and Larédo (2006), instead of looking at the exact likelihood, we consider the contrasts $p_n^{(m, \sigma^2)}(\theta, y_1, \dots, y_n)$ for $(m, \sigma^2) \neq (0, \sigma_s^2(\theta))$. It can be proved that the associated maximum contrast estimators are asymptotically equivalent to the exact maximum likelihood estimator whatever (m, σ^2) provided that $\sigma^2 > 0$. Consequently, we may choose in a convenient way the values m, σ^2 and study the associated contrast instead of the exact likelihood. The choice

$$m = 0, \quad \sigma^2 = \sigma_\infty^2(\theta)$$

leads to an noteworthy simplification. We denote by $\tilde{p}_n(\theta, y_1, \dots, y_n) = p_n^{(0, \sigma_\infty^2(\theta))}(\theta, y_1, \dots, y_n)$ the contrast corresponding to this choice. Let us note that, since the initial variance is equal to the fixed point of Φ^θ , this value remains unchanged along the iterations. As a consequence, the algorithm for the mean is simplified into:

$$H_y^\theta(m) := \Phi_y^\theta(m, \sigma_\infty^2(\theta)) = a \left(\frac{y}{\gamma^2} + \frac{m}{\sigma_\infty^2(\theta)} \right) \frac{\sigma_\infty^2(\theta) \gamma^2}{\sigma_\infty^2(\theta) + \gamma^2}.$$

Let us set

$$\rho = \frac{\sigma_\infty^2(\theta)}{\sigma_\infty^2(\theta) + \gamma^2} \in (0, 1).$$

Then,

$$H_y^\theta(m) = a(\rho y + m(1 - \rho)).$$

Hence,

$$H_{y_{i-1}}^\theta \circ \dots \circ H_{y_1}^\theta(0) = \sum_{j=0}^{i-2} a\rho(a(1 - \rho))^j y_{i-1-j} = x_i(\theta, y_{i-1}, \dots, y_1).$$

Thus

$$\tilde{p}_n(\theta, y_1, \dots, y_n) \propto (\gamma^2 + \sigma_\infty^2(\theta))^{-n/2} \exp\left(-\sum_{i=1}^n \frac{(y_i - x_i(\theta, y_{i-1}, \dots, y_1))^2}{2(\gamma^2 + \sigma_\infty^2(\theta))}\right).$$

This expression is much easier to handle. In particular, it allows to obtain an explicit and simple representation of the limit of the normalized log-likelihood. For details and further results, we refer to Genon-Catalot *et al.* (2003) and Genon-Catalot and Larédo (2006).

2.5 ARMA property and consequences.

We establish the links with the standard likelihood theory for stationary Gaussian processes. For this, refer to *e.g.* Dzhaparidze and Yaglom (1983), Dacunha-Castelle and Duflo (1986), Brockwell and Davies (1991).

Proposition 2.5.1. *The process (y_i) is ARMA(1, 1). Its spectral density is equal to*

$$f_\theta(\lambda) = \frac{\beta^2 + \gamma^2(1 + a^2) - 2a\gamma^2 \cos \lambda}{1 + a^2 - 2a \cos \lambda}.$$

Proof. Setting $\xi_i = y_i - ay_{i-1} = \beta\eta_i + \varepsilon_i - a\varepsilon_{i-1}$, we see that $\text{Cov}(\xi_i, \xi_{i+k}) = 0$ for $k \geq 1$. Thus, (ξ_i) is MA(1). The ARMA property follows. Since $\text{Var}\xi_i = \beta^2 + \gamma^2(1 + a^2)$ and $\text{Cov}(\xi_i, \xi_{i+1}) = -a\gamma^2$, using the ARMA property, we deduce the spectral density. \square

This simple result has some important consequences. First, it provides another way to deduce the asymptotic behaviour of the exact maximum likelihood estimator since the asymptotic behaviour of exact m.l.e. in stationary ARMA Gaussian processes is well known.

First, after some computations, we have the following identifiability property:

$$\forall \lambda, \quad f_\theta(\lambda) = f_{\theta'}(\lambda) \Rightarrow \theta = \theta'.$$

Then, setting $\theta = (\theta_1, \theta_2)$ (with $\theta_1 = a, \theta_2 = \beta^2$), the matrix

$$I(\theta) = \left(\int_{-\pi}^{\pi} f_\theta^{-2}(\lambda) \frac{\partial f_\theta}{\partial \theta_i}(\lambda) \frac{\partial f_\theta}{\partial \theta_j}(\lambda) \frac{d\lambda}{2\pi} \right)_{1 \leq i, j \leq 2}$$

is the asymptotic Fisher information matrix. Under some additional standard assumptions, $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution to $\mathcal{N}(0, I^{-1}(\theta))$.

For stationary Gaussian processes, the Whittle approximation of the likelihood defines a contrast which yields minimum contrast estimators which are asymptotically equivalent to the exact maximum likelihood estimator. Let us define the periodogram:

$$I_n(\lambda) = \frac{1}{n} \left| \sum_{j=1}^n y_j \exp(-ij\lambda) \right|^2.$$

The Whittle contrast is given by:

$$U_n(\theta) = \int_{-\pi}^{\pi} \left(\log f_{\theta}(\lambda) + \frac{I_n(\lambda)}{f_{\theta}(\lambda)} \right) \frac{d\lambda}{2\pi}.$$

The associated minimum contrast estimators are defined as minimisers of $U_n(\theta)$.

2.6 Non Gaussian noises.

Suppose that the noise ε_i is non Gaussian. Then, it is possible to use the Gaussian likelihood as a contrast. This yields consistent and asymptotically Gaussian estimators which are much better than simple moment estimators. This is demonstrated *e.g.* in Ruiz (1994).

2.7 Multidimensional extension.

Consider a d -dimensional Ornstein-Uhlenbeck process $(X(t))$ satisfying:

$$dX(t) = AX(t)dt + \Sigma dW_t, \quad X(0) = \eta,$$

with η a \mathbb{R}^d -valued random variable, independent of the Brownian motion W of \mathbb{R}^d and A is a (d, d) matrix. Assume that the observations are:

$$Y(t_i) = HX(t_i) + \varepsilon_i$$

where H is a known (k, d) -matrix ε_i is a sequence of i.i.d. variables with law $\mathcal{N}_k(0, Q)$, independent of $(X(t))$.

We make some simplifying assumptions. Assume that the matrix A is diagonalisable with negative eigenvalues $(\lambda_i, i = 1, \dots, d)$. Denote by P a matrix of eigenvectors such that $P^{-1}AP = D := \text{diag}((\lambda_i, i = 1, \dots, d))$. Then, the process $Z(t) = P^{-1}X(t)$ satisfies:

$$dZ(t) = DZ(t)dt + \Gamma dW_t, \quad \Gamma = P^{-1}\Sigma.$$

Since $(X(t))$ is not observed, we can change the model into:

$$Y_i = Y(t_i) = JZ(t_i) + \varepsilon_i, \quad J = HP.$$

It is worth noting that each column of P (eigenvector of P) is defined up to a multiplicative constant. In some cases, P can be chosen such that $J = HP$ does not depend on unknown parameters. Let us assume that this holds. Therefore, the unknown parameters are $\theta = (\lambda_i, i = 1, \dots, d, \gamma_{i,j}, i, j = 1, \dots, d)$ where the $\gamma_{i,j}$'s are the elements of Γ .

The discretization of $(Z(t))$ standardly yields a AR(1) process $Z_i = Z(i\Delta)$ such that

$$Z_{i+1} = \exp(D\Delta)Z_i + \exp(D(i+1)\Delta) \int_{i\Delta}^{(i+1)\Delta} \exp(-Ds) \Gamma dW_s.$$

Here, the matrix $\exp(D\Delta) = \text{diag}(\exp(\lambda_i\Delta), i = 1, \dots, d)$ is diagonal. The kernel of the hidden chain (Z_i) is

$$P_\theta(z, dz') = \mathcal{N}(\exp(D\Delta)z, R)$$

where

$$R = \int_0^\Delta \exp(Du) \Gamma \Gamma' \exp(Du) du = \left(\alpha_{i,j} \frac{e^{(\lambda_i + \lambda_j)\Delta} - 1}{\lambda_i + \lambda_j} \right),$$

where the $\alpha_{i,j}$'s are the elements of $\Gamma \Gamma'$. The process $(Z(t))$ admits a unique stationary distribution equal to the law $\mathcal{N}(0, V)$ with

$$V = \int_0^{+\infty} \exp(Du) \Gamma \Gamma' \exp(Du) du = \left(\frac{-\alpha_{i,j}}{\lambda_i + \lambda_j} \right).$$

The conditional distribution of Y_i given $Z_i = z$ is the distribution

$$F(z, dy) = \mathcal{N}(Jz, Q)(dy).$$

Hence, the up-dating and prediction operators can be easily computed. This allows to obtain the exact likelihood. We state the proposition analogous to Proposition 2.3.1:

Proposition 2.7.1. • (*Up-dating operator*) If $\nu = \mathcal{N}_d(m, K)$ and $y \in \mathbb{R}^k$, then $\varphi_y(\nu) = \mathcal{N}_d(\hat{m}(y), \hat{K})$ with

$$\hat{m}(y) = m + K J' (JK J' + Q)^{-1} (y - Jm), \quad \hat{K} = K - K J' (JK J' + Q)^{-1} JK.$$

• (*Prediction*) If $\nu = \mathcal{N}_d(m, K)$, then $\psi_\theta(\nu) = \nu P_\theta = \mathcal{N}_d(\bar{m}, \bar{K})$ with

$$\bar{m} = \exp(D\Delta)m, \quad \bar{K} = \exp(D\Delta)K \exp(D\Delta) + R.$$

• (*marginal operator*) If $\nu = \mathcal{N}_d(m, K)$, then $p_\nu(y) = \mathcal{N}_k(Jm, JK J' + Q)$.

The algorithm of predictive distributions is ruled by the compound operator $\Phi_y = \psi_\theta \circ \varphi_y$. The exact likelihood is obtained recursively. The score function and the Hessian are also obtained recursively.

Note that the following matrix relation holds:

$$(I + K J' Q^{-1} J)^{-1} = I - K J' (JK J' + Q)^{-1} J.$$

(Compare with the one-dimensional relation $\frac{1}{1+\frac{\alpha}{x}} = 1 - \frac{\alpha}{\alpha+x}$).

In Favetto and Samson (2008), a complete study of a partially model with biological motivation is treated. The unobserved process is a two-dimensional Ornstein-Uhlenbeck process and the observation is the sum of the components with additive noise. A special attention is given to the recursive computation of the score function and the Hessian and to the identifiability of parameters.

Note that Pedersen (1994) gives a description of the computation of the likelihood in the general case.

Bibliography

- [1] Brockwell P.J. and Davis R.A. (1991). *Time series: Theory and Methods*. Springer-Verlag.
- [2] Cappé O., Moulines E. and Rydèn T. (2005). *Inference in hidden Markov models*, Springer.
- [3] Cheng-Der Fuh (2006). Efficient likelihood estimation in state space models. *Ann. Statist.* **34** (4), 2026-2068.
- [4] Dacunha-Castelle D. and Duflo M. (1986). *Probability and Statistics. Volume 2*. Springer-Verlag.
- [5] Dzhaparidze K.O. and Yaglom A.M. (1983). Spectrum parameter estimation in time series analysis. In *Developments in statistics* (P.R. Krishnaiah, ed.) **4**, 1-181, Academic, New York.
- [6] Favetto B. and Samson A. (2008). Parameter estimation for a bidimensional partially observed Ornstein-Uhlenbeck process with biological application, Prepublication MAP5 2008-13 (Université Paris Descartes).
- [7] Genon-Catalot V., Jeantheau T. and Larédo C. (2003). Conditional likelihood estimators for hidden Markov models and stochastic volatility models. *Scand. J. Statist.* **30** (2), 297-316.
- [8] Genon-Catalot V. and Larédo C. (2006). Leroux's method for general hidden Markov models. *Stoch. Proc. and Appl.* **116**, 222-243.
- [9] Pedersen, A. (1994). Statistical analysis of Gaussian diffusion processes based on incomplete discrete observations. *Research report, Department of theoretical statistics, University of Aarhus* **297**.
- [10] Ruiz E.(1994). Quasi-maximum likelihood estimation of stochastic volatility models. *Journal of Econometrics* **63**, 289-306.

2.8 Appendix

In this section, we give some complementary results.

2.8.1 Iterations of the operators for predictive distributions.

We detail the iterations of the operator Φ_y^θ for predictive distributions given in Corollary 2.3.1. Starting with $\nu = \mathcal{N}(m, \sigma^2)$, let us set:

$$m_{1|0:1} = m, \quad \sigma_{1|0:1}^2 = \sigma^2,$$

$$m_{i|i-1:1}(m, \sigma^2) = \Phi_{y_{i-1}}^\theta \circ \dots \circ \Phi_{y_1}^\theta(m, \sigma^2), \quad \sigma_{i|i-1:1}^2(\sigma^2) = \Phi^\theta \circ \dots \circ \Phi^\theta(\sigma^2)$$

$$\delta_i = \frac{\gamma^2}{\gamma^2 + \sigma_{i|i-1:1}^2(\sigma^2)}.$$

Then,

$$m_{i|i-1:1}(m, \sigma^2) = a^{i-1} \delta_i \delta_{i-1} \dots \delta_1 m + a \sum_{l=1}^{i-1} y_{i-l} (1 - \delta_{i-l}) a^{l-1} \delta_{i-1} \delta_{i-2} \dots \delta_{i-l+1}.$$

The conditional distributions of y_i given y_{i-1}, \dots, y_1 is the Gaussian distribution with mean $m_{i|i-1:1}(m, \sigma^2)$ and variance $\sigma_{i|i-1:1}^2(\sigma^2) + \gamma^2$.

As seen above, if the iterations start with $m = 0$ and $\sigma^2 = \sigma_\infty^2(\theta)$, the expressions are considerably simpler as $\delta_i = \delta = 1 - \rho$ remains constant.

Chapter 3

Diffusions with multiplicative noise.

3.1 Introduction

Consider a one-dimensional diffusion process $(x(t), t \geq 0)$ and assume that the observations taken at times $0 \leq t_1 < \dots < t_n < \dots$ are of the form: $y_i = x(t_i)\varepsilon_i$ where (ε_i) is a sequence of i.i.d. random variables independent of the process $(x(t))$. The noise (ε_i) is multiplicative. As it is natural for scale perturbations, we assume that $(x(t))$ is non negative. As for the noise, we may assume that the ε_i 's are nonnegative or signed and symmetric. We present below models for which the up-dating, prediction and marginal operators can be computed explicitly. Consequently, the exact likelihood is also explicit. As for the Kalman model, explicit computations are obtained for a specific class of diffusion models and for a specific class of noises distributions. For this lecture, we refer to Genon-Catalot (2003), Genon-Catalot and Kessler (2004), Chaleyat-Maurel and Genon-Catalot (2006), Comte *et al.* (2007).

3.2 Computable filters.

Computations of the conditional distributions of $x(t_i)$, $x(t_{i+1})$ or y_{i+1} given y_1, \dots, y_i rely on iterations of the up-dating, prediction and marginal operators. These iterations are rapidly intractable unless both the up-dating and the prediction operators evolve within a relatively simple class of distributions on the state space of the hidden process. The ideal situation is when this class is a parametric family, *i.e.* a family of distributions specified by a fixed number of parameters. This is the case of the Kalman filter model: the up-dating and prediction operators both evolve within the family of Gaussian distributions. Hence, it is enough to specify recursively the means and variances. Such an ideal situation is not often encountered. Hence, the idea is to find a larger class built using mixtures of parametric distributions.

Recall the general notations. We have an unobserved Markov chain (x_i) (which we have

supposed to be a regular discrete sampling of a diffusion process) with state space \mathcal{X} and transition kernel $P_\theta(x, dx')$. Observations y_i are such that $\mathcal{L}(y_i|x_i = x) = f(y|x)\mu(dy)$. The up-dating operator acting on distributions on \mathcal{X} is the mapping:

$$\nu \rightarrow \varphi_y(\nu) \propto f(y|x)\nu(dx).$$

The proportionality coefficient is the marginal distribution $p_\nu(y)$. The prediction operator ψ_θ is the mapping:

$$\nu \rightarrow \psi_\theta(\nu) = \nu P_\theta.$$

In Chaleyat-Maurel and Genon-Catalot (2006), sufficient conditions are exhibited in order to obtain a ‘‘computable filter’’ model. We detail these conditions now.

Proposition 3.2.1. *Let $\mathcal{F} = \{\nu_c, c \in C\}$ be a parametric family of distributions on \mathcal{X} where $C \subset \mathbb{R}^p$ is a set of parameters (and $c \neq c'$ implies $\nu_c \neq \nu_{c'}$). Consider the enlarged class*

$$\bar{\mathcal{F}}_f = \left\{ \nu = \sum_{i=0}^L \alpha_i \nu_{c_i}, L \in \mathbb{N}, c_i \in C, \alpha_i \geq 0, i = 0, 1, \dots, L, \sum_{i=0}^L \alpha_i = 1 \right\}$$

composed of finite mixtures of distributions of \mathcal{F} and the following conditions:

- (C1) \mathcal{F} is a conjugate class for the family $y \rightarrow f(y|x), x \in \mathcal{X}$.
- (C2) If $\nu \in \mathcal{F}$, then $\psi_\theta(\nu) = \nu P_\theta \in \bar{\mathcal{F}}_f$.

Then, if $\nu \in \bar{\mathcal{F}}_f$, then, for all y , $\varphi_y(\nu)$ and $\psi_\theta(\nu)$ both belong to $\bar{\mathcal{F}}_f$.

Condition (C1) means that, if $c \in C$, then $\varphi_y(\nu_c)$ belongs to \mathcal{F} , i.e.:

$$\varphi_y(\nu_c) = \nu_{\varphi_y(c)} \tag{3.1}$$

where $\varphi_y(c) \in C$. Since we can identify a distribution ν_c to its parameter c , the operator φ_y is identified to the mapping:

$$c \in C \rightarrow \varphi_y(c) \in C. \tag{3.2}$$

Condition (C1) is classical in Bayesian statistics for obtaining explicit Bayes estimators. Indeed, interpreting ν as a prior on x for the parametric family of densities ($y \rightarrow f(y|x), x \in \mathcal{X}$), $\varphi_y(\nu)$ is the corresponding posterior distribution.

Condition (C2) means that, if $c \in C$, then $\psi_\theta(\nu_c)$ is a finite mixture of the form

$$\psi_\theta(\nu_c) = \sum_{i=0}^{L_c} \alpha_i(c) \nu_{\tau_i(c)}, \tag{3.3}$$

where $L_c, \alpha_i(c), \tau_i(c)$ depend on θ . Hence, the image of ν_c by the prediction operator ψ_θ is not an element of \mathcal{F} , but an element of $\bar{\mathcal{F}}_f$. This is why we are not in the ideal situation of a finite-dimensional filter and we need an extension of this notion, namely, the notion of computable filter.

The proof of Proposition 3.2.1 is simple algebra (using (3.1)-(3.3)) and omitted (see Chaleyat-Maurel and Genon-Catalot (2006)). Nevertheless, the result is important since it induces a recursive algorithm to compute the filtering or predictive distributions. The difficulty lies in finding the adequate class \mathcal{F} . This is what give in our examples.

Let us stress some facts. First, a distribution $\nu = \sum_{i=0}^L \alpha_i \nu_{c_i}$ in the class $\bar{\mathcal{F}}_f$ is specified by the finite sequence of parameters $(L, \alpha_0, \dots, \alpha_L, c_0, c_1, \dots, c_L)$. Hence, it is “computable”. Along the iterations of the prediction and up-dating operators, all these parameters will be up-dated. This includes the length L of the mixture. Hence, the situation is not the same as in the Kalman filter. The number of parameters to be specified may change along iterations. In this sense, a “computable filter” is not a finite-dimensional filter.

It is also important to note that (C1) only concerns the kernel $f(y|x)\mu(dy)$ whereas (C2) only concerns the kernel $P_\theta(x, dx')$. Therefore, they may be checked separately.

3.3 Hidden diffusion model. Prediction operator.

Now, we give a diffusion process $(x(t))$ and a class of distributions \mathcal{F} such that condition (C2) is fulfilled.

Consider the one-dimensional diffusion process $(x_t, t \geq 0)$ described by:

$$dx(t) = (2\theta x(t) + \delta\sigma^2)dt + 2\sigma\sqrt{x(t)}dW_t, \quad x(0) = \eta, \quad (3.4)$$

with η a random variable independent of the Brownian motion (W_t) . When η is nonnegative, $\theta \in \mathbb{R}$, $\sigma > 0$, $\delta \geq 0$, $x(t)$ is uniquely defined and $x(t) \geq 0$ for all $t \geq 0$. First, we give a representation of $(x(t))$ when δ is a positive integer.

Proposition 3.3.1. *Let $\delta \geq 1$ an integer, and consider $(\xi_t^i, i = 1, \dots, \delta)$ δ i.i.d. Ornstein-Uhlenbeck processes satisfying*

$$d\xi_t^i = \theta\xi_t^i dt + \sigma dW_t^i, \quad \xi_0^i = x^i,$$

where $(W^i, i = 1, \dots, \delta)$ are independent Wiener processes, $x^i, i = 1, \dots, \delta$ are real values. Then, setting $x(t) = \sum_{i=1}^{\delta} (\xi_t^i)^2$,

$$dx(t) = (2\theta x(t) + \delta\sigma^2)dt + 2\sigma\sqrt{x(t)}dB_t, \quad x(0) = \sum_{i=1}^{\delta} (x^i)^2,$$

where (B_t) is a Wiener process.

Proof. Applying the Ito formula, we get:

$$dx(t) = (2\theta x(t) + \delta\sigma^2)dt + 2\sigma \sum_{i=1}^{\delta} \xi_t^i dW_t^i.$$

Consider \tilde{B} a Wiener process independent of $(W^i, i = 1, \dots, \delta)$ and set

$$B_t = \int_0^t 1_{x(s)>0} \frac{\sum_{i=1}^{\delta} \xi_s^i dW_s^i}{x^{1/2}(t)} + \int_0^t 1_{x(s)=0} d\tilde{B}_s.$$

Since $\langle B \rangle_t = t$, B is a Wiener process and

$$\sqrt{x(t)} dB_t = \sum_{i=1}^{\delta} \xi_t^i dW_t^i,$$

which gives the result. \square

For general positive δ , the process $(x(t))$ has been largely popularized by its use for modelling interest rate data (see Cox *et al.* (1985)). Moreover, the CIR process is used to model the volatility in the stochastic volatility model proposed by Heston (1993). In the latter paper, a financial asset price $S(t)$ is described by:

$$d \log S(t) = \sqrt{x(t)} dB_t$$

with (B_t) a Wiener process independent of $(x(t))$. Therefore,

$$\log^2 \frac{S((i+1)\Delta)}{S(i\Delta)} \simeq x(i\Delta) \varepsilon_i,$$

where $\varepsilon_i = (B_{(i+1)\Delta} - B_{i\Delta})^2$. However, with such noises distribution, it is not possible to get explicit computations of the filters. This is why we introduce below other distributions for the noises.

Let us introduce

$$a(t) = e^{\theta t}, \quad \beta(t) = \sigma \left(\frac{e^{\theta t} - 1}{2\theta} \right)^{1/2}. \quad (3.5)$$

The transition density of $(x(t))$ can be written as a mixture of distributions: for $x \geq 0$,

$$p_t(x, x') = \sum_{i \geq 0} w_i \left(\frac{a^2(t)x}{\beta^2(t)} \right) g_{i, \beta(t)}(x'), \quad (3.6)$$

where, for $i \geq 0$, the weights w_i are given by

$$u \in \mathbb{R}, \quad w_i(u) = \exp(-u/2) (u/2)^i / i! \quad (3.7)$$

and the function

$$g_{i, \sigma}(x) = 1_{(x>0)} \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{x}{2\sigma^2}\right) \frac{x^{i-1+\frac{\delta}{2}}}{C_{2i+\delta-1} \sigma^{2i+\delta-1}}, \quad (3.8)$$

is the density of a Gamma distribution $G(i + \delta/2, 1/2\sigma^2)$ with scale parameter $1/2\sigma^2$ and location parameter $(i + \delta/2)$. The normalising constant C_a , is simply the absolute moment of order a of a Gaussian standard variable, i.e. , for $a \geq 0$ or $b \geq 1/2$,

$$C_a = \mathbb{E}(|X|^a) = \frac{2^{a/2}}{\sqrt{\pi}} \Gamma\left(\frac{a+1}{2}\right), \quad \Gamma(b) = \frac{\sqrt{2\pi}}{2^b} C_{2b-1}$$

for X a standard Gaussian variable (Γ is the usual Gamma function). The transition density $p_t(x, x')$ is exactly a $\beta^2(t)\chi'^2(\frac{\alpha^2(t)x}{\beta^2(t)}, \delta)$.

For $\delta = 1$, $x(t) = \xi_t^2$ is the square of an Ornstein-Uhlenbeck process. When $\theta < 0$, this process admits a unique stationary distribution π_ϑ equal to the distribution of ξ^2 with $\xi \sim \mathcal{N}(0, \sigma_s^2(\vartheta))$ and

$$\sigma_s^2(\vartheta) = \frac{\sigma^2}{2|\theta|}. \quad (3.9)$$

The invariant distribution is therefore

$$\pi_\vartheta(dx) = 1_{(x>0)} \frac{1}{\sigma_s(\theta)\sqrt{2\pi}} \exp\left(-\frac{x}{2\sigma_s^2(\theta)}\right) x^{-\frac{1}{2}} dx = G(1/2, \frac{1}{2\sigma_s^2(\theta)}).$$

When $\delta \geq 2$ and $\theta < 0$, $x(t)$ is positive recurrent on $(0, +\infty)$. In this case, its unique stationary distribution obtained by normalizing its speed measure is $\pi_\vartheta(dx) = G(\delta/2, \frac{1}{2\sigma_s^2(\theta)})$.

Let us note that formula (3.6) still holds in the case $x = 0$ and we have:

$$p_t(0, x') = g_{0,\beta(t)}(x').$$

The special structure of the transition density of $(x(t))$ suggests to introduce the class

$$\mathcal{F}^\delta = \{\nu_{i,\sigma} = g_{i,\sigma}(x)dx, i \in \mathbb{N}, \sigma \geq 0\}, \quad (3.10)$$

where, by convention $\nu_{i,0} = \delta_0$ for all i (note that, as $\sigma \rightarrow 0$, $\nu_{i,0} \Rightarrow \delta_0$). The stationary distribution belongs to \mathcal{F}^δ and corresponds to $i = 0, \sigma = \sigma_s(\theta)$. We define the extended class $\bar{\mathcal{F}}^\delta$ as the class of mixtures of distributions $\nu_{i,\sigma}$ having the same scale parameter:

$$\bar{\mathcal{F}}^\delta = \{\nu = \nu_{\alpha,\sigma} = \sum_{i \geq 0} \alpha_i \nu_{i,\sigma}, \alpha = (\alpha_i), \alpha_i \geq 0, \sum_{i \geq 0} \alpha_i = 1, \sigma \geq 0\}. \quad (3.11)$$

The sub-class $\bar{\mathcal{F}}_f^\delta$ is composed of all distributions $\nu_{\alpha,\sigma}$ having finite-length mixture parameter, *i.e.* such that

$$L(\alpha) = \sup\{i; \alpha_i > 0\} < \infty.$$

Note that the transition density belongs to the class $\bar{\mathcal{F}}^\delta$ of infinite mixtures.

Now, we state the result concerning the prediction operator.

Proposition 3.3.2. *(Prediction operator) Let $\vartheta = (\theta, \sigma^2)$ denote the unknown parameters. For $\Delta > 0$, let P_Δ^ϑ denote the transition operator of $(x(t))$ with step Δ . The prediction operator ψ_ϑ satisfies the following property.*

- If $\nu_{i,\sigma}$ belongs to \mathcal{F}^δ with $\sigma > 0$, then

$$\psi_\vartheta(\nu_{i,\sigma}) = \nu_{i,\sigma} P_\Delta^\vartheta = \sum_{l=0}^i \alpha_l^{i,\sigma} \nu_{l,\tau(\sigma)},$$

with $\tau^2(\sigma) = \beta^2(\Delta) + a^2(\Delta)\sigma^2$ and for $l = 0, \dots, i$,

$$\alpha_i^{i,\sigma} = \binom{i}{l} \left(1 - \frac{\beta^2(\Delta)}{\tau^2(\sigma)}\right)^l \left(\frac{\beta^2(\Delta)}{\tau^2(\sigma)}\right)^{i-l}.$$

If $\sigma = 0$, then, $\psi_\vartheta(\delta_0) = \nu_{0,\beta(\Delta)}$.

- If $\nu = \nu_{\alpha,\sigma}$ belongs to $\bar{\mathcal{F}}_f^\delta$ with $\sigma > 0$,

$$\psi_\vartheta(\nu_{\alpha,\sigma}) = \nu_{\alpha,\sigma} P_\Delta^\vartheta = \sum_{l \geq 0} \bar{\alpha}_l \nu_{l,\tau(\sigma)}$$

with, for $l \geq 0$,

$$\bar{\alpha}_l = \bar{\alpha}_l(\alpha, \sigma) = \sum_{i \geq l} \alpha_i \alpha_i^{i,\sigma}.$$

Moreover, $L(\bar{\alpha}) = L(\alpha)$. If $\sigma = 0$, then, $\psi_\vartheta(\delta_0) = \nu_{0,\beta(\Delta)}$ and $L(\bar{\alpha}) = 0$.

Hence, condition (C2) holds for the class \mathcal{F}^δ and the extended class $\bar{\mathcal{F}}_f^\delta$.

Note that the moments of a distribution $\nu_{\alpha,\sigma}$ are obtained easily: for $r \geq 0$,

$$\int x^r \nu_{\alpha,\sigma}(dx) = \sigma^{2r} \sum_{i \geq 0} \alpha_i \frac{C_{2(i+r)+\delta-1}}{C_{2i+\delta-1}}. \quad (3.12)$$

It can be simplified using the classical relation $C_{a+1} = aC_{a-1}$. In particular, for $r = 1$,

$$\int x \nu_{\alpha,\sigma}(dx) = \sigma^2 \sum_{i \geq 0} \alpha_i (2i + \delta) = \sigma^2 (\delta + 2 \sum_{i \geq 0} i \alpha_i).$$

For all $r \geq 1$,

$$\int x^r \nu_{\alpha,\sigma}(dx) = \sigma^{2r} \sum_{i \geq 0} \alpha_i ([2i + 2(r-1) + \delta][2i + 2(r-2) + \delta] \dots [2i + \delta]). \quad (3.13)$$

The variance is given by:

$$\text{Var} \nu_{\alpha,\sigma} = \sigma^4 [2\delta + 4 \sum_{i \geq 0} i(i+1) \alpha_i - 4(\sum_{i \geq 0} i \alpha_i)^2].$$

3.4 Noise distribution. Up-dating operator.

Recall that we observe $y_i = x_i \varepsilon_i$ with $x_i = x(i\Delta)$. In order to fulfill condition (C1), we choose the noise distribution as follows. The random variables ε_i are chosen to have Inverse Gamma

distribution $\text{Inv}G(k, \lambda)$ with scale parameter λ and location parameter k which must be a positive integer, *i.e.* the distribution of y_i given $x_i = x$ has density (w.r.t. dy)

$$f(y|x) = 1_{y>0} e^{-\frac{\lambda x}{y}} \frac{x^k \lambda^k}{y^{k+1} \Gamma(k)}. \quad (3.14)$$

This is the density of $\lambda x/G(k, 1)$.

It is worth noting that $\mathbb{E}(\varepsilon^r) < \infty$ holds if and only if $r < k$. In this case,

$$\mathbb{E}(\varepsilon^r) = \frac{\lambda^r \Gamma(k-r)}{\Gamma(k)}.$$

Proposition 3.4.1. (*up-dating and marginal operators*)

- If $\nu_{i,\sigma}$ belongs to \mathcal{F}^δ and $\sigma > 0$, then, for all positive y ,

$$\varphi_y(\nu_{i,\sigma}) = \nu_{i+k, \varphi_y(\sigma)}$$

with

$$\frac{1}{\varphi_y^2(\sigma)} = \frac{1}{\sigma^2} + \frac{2\lambda}{y}.$$

If $y = 0$ or $\sigma = 0$, $\varphi_y(\nu) = \delta_0$. Hence, the following holds, for all non negative y :

$$\varphi_y^2(\sigma) = \frac{\sigma^2 y}{y + 2\lambda\sigma^2}.$$

The corresponding marginal distribution has density (when $\sigma > 0$):

$$p_{\nu_{i,\sigma}}(y) = \frac{\lambda^k}{\Gamma(k)} \frac{C_{2(k+i)+\delta-1}}{C_{2i+\delta-1}} \frac{\sigma^{2k}}{y} \frac{y^{i+\frac{\delta}{2}}}{(y + 2\lambda\sigma^2)^{i+k+\frac{\delta}{2}}}$$

If y or σ equal 0, then, the marginal distribution is δ_0 .

- If $\nu = \nu_{\alpha,\sigma}$ belongs to $\bar{\mathcal{F}}_f^\delta$,

$$\varphi_y(\nu) = \sum_{i \geq 0} \hat{\alpha}_i \nu_{i, \varphi_y(\sigma)}$$

where

$$\hat{\alpha}_i = \hat{\alpha}_i(\alpha, \sigma) \propto 1_{i \geq k} \alpha_{i-k} \frac{C_{2i+\delta-1}}{C_{2(i-k)+\delta-1}} \left(\frac{\varphi_y^2(\sigma)}{\sigma^2} \right)^{i-k}$$

Moreover, $L(\hat{\alpha}) = L(\alpha) + k$ except if $y = 0$ in which case $\varphi_y(\nu) = \delta_0$ (with $\hat{\alpha}(k) = 1$).

For the marginal distribution, we have, when $\sigma > 0$,

$$p_\nu(y) = \sum_{i \geq 0} \alpha_i p_{\nu_{i,\sigma}}(y).$$

Condition (C1) holds for the class \mathcal{F}^δ and the extended class $\bar{\mathcal{F}}_f^\delta$.

3.5 Algorithm for predictive distributions.

Now, we have to compute the iterations of $\Phi_y^\vartheta = \psi_\vartheta \circ \varphi_y$. By the previous paragraph, when the distribution of x_1 belongs to the class $\bar{\mathcal{F}}_f^\delta$, for all i , the distributions $\nu_{i|i-1:1}(dx), \nu_{i|i:1}(dx)$ also belong to the class $\bar{\mathcal{F}}_f^\delta$. It is therefore enough to specify these distributions by their parameters $\alpha_{i|i-1:1}, \alpha_{i|i:1}$ (mixture parameters), $L_{i|i-1:1} = L(\alpha_{i|i-1:1}), L_{i|i:1} = L(\alpha_{i|i:1})$ (lengths of the mixture parameters), $\sigma_{i|i-1:1}, \sigma_{i|i:1}$ (scale parameters)

Proposition 3.5.1. *Assume that $\nu_{1|0:1}$ belongs to the class $\bar{\mathcal{F}}_f^\delta$. The scale and mixture and lengths parameters of $\nu_{i|i-1:1}(dx), \nu_{i|i:1}(dx)$ can be recursively computed as follows.*

- (up-dating) For $i \geq 1$,

$$\sigma_{i|i:1}^2 = \frac{\sigma_{i|i-1:1}^2 y_i}{y_i + 2\lambda \sigma_{i|i-1:1}^2}.$$

For $j \geq 0$,

$$\alpha_{i|i:1}(j) \propto 1_{j \geq k} \alpha_{i|i-1:1}(j-k) \frac{C_{2j+\delta-1}}{C_{2(j-k)+\delta-1}} \left(\frac{\sigma_{i|i:1}^2}{\sigma_{i|i-1:1}^2} \right)^{j-k}.$$

If $y_i \neq 0$,

$$L_{i|i:1} = L_{i|i-1:1} + k.$$

If $y_i = 0$, $\alpha_{i|i:1}(k) = 1$ and $L_{i|i:1} = k$.

- (prediction)

$$\sigma_{i+1|i:1}^2 = \beta^2(\Delta) + a^2(\Delta) \sigma_{i|i:1}^2.$$

For $j \geq 0$,

$$\alpha_{i+1|i:1}(j) = \sum_{l \geq j} \alpha_{i|i:1}(l) \kappa_j^{(l)}(\Delta)$$

with

$$\kappa_j^{(l)}(\Delta) = \binom{l}{j} \left(1 - \frac{\beta^2(\Delta)}{\sigma_{i+1|i:1}^2} \right)^j \left(\frac{\beta^2(\Delta)}{\sigma_{i+1|i:1}^2} \right)^{l-j}.$$

If $\sigma_{i|i:1}^2 > 0$, $L_{i+1|i:1} = L_{i|i:1}$. If $\sigma_{i|i:1}^2 = 0$, $L_{i+1|i:1} = 0$.

- (Marginal distributions):

$$p_{\nu_{i|i-1:1}}(y_i) = p_i(y_i | y_{i-1}, \dots, y_1) = \sum_{j=0}^{L_{i|i-1:1}} \alpha_{i|i-1:1}(j) p_{\nu_{j, \sigma_{i|i-1:1}}}(y_i).$$

Since it always hold that $\sigma_{i|i-1:1}^2 \geq \beta^2(\Delta) > 0$, this distribution always has density.

A special attention must be given to the length of the mixture parameters. Starting with a length $L_{1|0:1}$, if $y_1 \neq 0$, $L_{1|1:1} = L_{1|0:1} + k$ and $L_{2|1:1} = L_{1|1:1} = L_{1|0:1} + k$. If $y_1 = 0$, $\alpha_{1|1:1}(k) = 1$ and $\alpha_{2|1:1}(0) = 1$ which implies $L_{2|1:1} = 0$. More generally,

$$L_{i|i-1:1} \leq L_{1|0:1} + (i-1)k.$$

Each time a new observation is equal to 0, the filtering distribution is δ_0 . Then, the length of the mixture parameter in the predictive distribution is reset to 0 and the scale parameter is reset to $\beta^2(\Delta)$ (this means that the predictive distribution is not a mixture and is simply a Gamma $G(\delta/2, 1/2\beta^2(\Delta))$). Thus, the length of the mixture parameter may be much smaller than the above upper bound. This can be seen on simulated data where only two or three mixture parameters are significantly non nul.

The scale parameter of the predictive distributions is bounded from below and positive. The scale parameter of the filtering distributions is not bounded from below and may be nul.

It must be stressed that the scale parameter of predictive distributions evolve in a relatively simple way. Indeed, we have

$$\sigma_{i+1|i:1}^2 = F_{y_i}(\sigma_{i|i-1:1}^2) \quad \text{with} \quad F_y(v) = \beta^2(\Delta) + a^2(\Delta) \frac{vy}{y + 2\lambda v}.$$

When $\theta < 0$ ($a^2(\Delta) < 1$), the function F_y is increasing from $I = [\beta^2(\Delta), \frac{\beta^2(\Delta)}{1-a^2(\Delta)}]$ onto I and Lipschitz with constant $a^2(\Delta)$. This allows to obtain stability properties for the scale parameter $\sigma_{i|i-1:1}^2$ (see Genon-Catalot and Kessler (2004)).

Using the formula for moments (3.12), we obtain:

$$\begin{aligned} \mathbb{E}(x_i | y_{i-1}, \dots, y_1) &= \sigma_{i|i-1:1}^2 \left(\delta + 2 \sum_{j=0}^{L_{i|i-1:1}} j \alpha_{i|i-1:1}(j) \right), \\ \mathbb{E}(x_i | y_i, \dots, y_1) &= \sigma_{i|i:1}^2 \left(\delta + 2 \sum_{j=0}^{L_{i|i:1}} j \alpha_{i|i:1}(j) \right). \end{aligned}$$

The other conditional moments are obtained analogously.

3.6 Exact likelihood.

Let us now assume that $\theta < 0$ and either $\delta = 1$ ($x(t)$ is the square of an Ornstein-Uhlenbeck process) or $\delta \geq 2$ ($x(t)$ is positive recurrent on $(0, +\infty)$). Moreover, let us assume that the initial variable η has the stationary distribution $\pi_\vartheta = G(\delta/2, \frac{1}{2\sigma_s^2})$. Hence, the computation of the predictive distributions starts with

$$\nu_{1|0:1} = \pi_\vartheta = \nu_{0, \sigma_s(\vartheta)}$$

which belongs to $\bar{\mathcal{F}}_f^\delta$ and has parameters $L_{1|0:1} = 0$ and $\sigma = \sigma_s(\vartheta)$. The process (x_i, y_i) is strictly stationary. The exact likelihood is computed by:

$$p_n(\vartheta, y_1, \dots, y_n) = p_1(\vartheta, y_1) \prod_{i=2}^n p_i(\vartheta, y_i | y_{i-1}, \dots, y_1)$$

where $p_1(\vartheta, y_1)$ is the density of y_1 (and of all y_i 's) and is equal to

$$p_1(\vartheta, y_1) = p_{\nu_{1|0:1}}(y_1) = p_{\nu_{0, \sigma_s(\vartheta)}}(y_1) = \frac{\lambda^k}{\Gamma(k)} \frac{C_{2k+\delta-1}}{C_{\delta-1}} \frac{\sigma_s^{2k}(\vartheta) y^{\frac{\delta}{2}-1}}{(y + 2\lambda\sigma_s^2(\vartheta))^{k+\frac{\delta}{2}}}$$

and

$$p_i(\vartheta, y_i | y_{i-1}, \dots, y_1) = p_{\nu_{i|i-1:1}}(y_i) = \sum_{0 \leq j \leq L_{i|i-1:1}} \alpha_{i|i-1:1}(j) p_{\nu_{j, \sigma_{i|i-1:1}}} (y_i)$$

where $L_{i|i-1:1} \leq (i-1)k$ and

$$p_{\nu_{j, \sigma_{i|i-1:1}}}(y_i) = \frac{\lambda^k}{\Gamma(k)} \frac{C_{2(k+j)+\delta-1}}{C_{2j+\delta-1}} \frac{\sigma_{i|i-1:1}^{2k}}{y_i} \frac{y_i^{j+\frac{\delta}{2}}}{(y_i + 2\lambda\sigma_{i|i-1:1}^2)^{j+k+\frac{\delta}{2}}}$$

Of course, one can include δ among the unknown parameters and compute the maximum likelihood estimators of (θ, σ, δ) . The parameters k, λ come from the noise distribution. They are supposed to be known. The other parameters and the observations y_{i-1}, \dots, y_1 are included in the formulae of $\sigma_{i|i-1:1}^2, \alpha_{i|i-1:1}$.

Note that the density of y_i is equal to the density of G/G' with G, G' independent, $G \sim G(\delta/2, 1/2\sigma_s^2(\vartheta))$ and $G' \sim G(k, \lambda)$. This is the distribution of $2\lambda\sigma_s^2(\vartheta)F$ where F has Fisher distribution $F(\delta, 2k)$.

3.7 Related models.

Analogous results can be obtained for the following models:

$$y_i = \sqrt{x_i} \varepsilon_i$$

where ε_i has distribution $1/\sqrt{G(k, \lambda)}$. Or,

$$y_i = \sqrt{x_i} \varepsilon_i$$

where ε_i has the symmetric distribution $\varepsilon/\sqrt{G(k, \lambda)}$ with ε independent of $G(k, \lambda)$ and with symmetric Bernoulli distribution ($P(\varepsilon = 1) = P(\varepsilon = -1) = 1/2$).

Bibliography

- [1] Chaleyat-Maurel M. and Genon-Catalot V. (2006). Computable infinite-dimensional filters with applications to discretized diffusion processes. *Stoch. Proc. and Applic.* **116**, 1447-1467.
- [2] Comte F., Genon-Catalot V. and Kessler M. (2007). Multiplicative Kalman filtering, Prépublication 2007-16, MAP5, *Laboratoire de mathématiques appliquées de Paris Descartes*, submitted.
- [3] Cox J.C., Ingersoll J.E., Ross S.A., (1985). A theory of term structure of interest rates, *Econometrica* **53**, 385-407.
- [4] Genon-Catalot V. (2003). A non-linear explicit filter. *Statist. Probab. Lett.*, **61**, 145-154.
- [5] Genon-Catalot V. and Kessler M. (2004). Random scale perturbation of an AR(1) process and its properties as a nonlinear explicit filter. *Bernoulli* (**10**) (4), 701-720.
- [6] Heston S.L. (1993). A closed-form solution for options with stochastic volatility with applications to bonds and currency options, *Rev. Financial Studies* **6**, 327-343.

Chapter 4

Other kernels.

4.1 Introduction

In this lecture, we present two families of models:

- The hidden process is the same as in the previous lecture: $x(t)$ is either the square of an Ornstein-Uhlenbeck process or the CIR diffusion process. But, given the state $x(t_i)$, the observation y_i has Poisson distribution with parameter $\lambda(x(t_i)) = \lambda x(t_i)$.
- The hidden process ($x(t)$) is a Wright-Fisher diffusion process. Given the state $x(t_i)$, the observation y_i has Bernoulli distribution with parameter $x(t_i)$.

We refer to Chaleyat-Maurel and Genon-Catalot (2006, 2009) for more details.

4.2 Conditional Poisson observations.

Let us assume that observations y_i are such that the conditional distribution of y_i given $x(t_i) = x$ is $F(x, dy) = f(y|x)\mu(dy)$ where $\mu(dy)$ is the counting measure on \mathbb{N} simply given by $\mu(y) = 1$ for all $y \in \mathbb{N}$ and

$$f(y|x) = \exp(-\lambda x) \frac{(\lambda x)^y}{y!} \quad y \in \mathbb{N}.$$

The hidden diffusion process ($x(t)$) is given by

$$dx(t) = (2\theta x(t) + \delta\sigma^2)dt + 2\sigma\sqrt{x(t)}dW_t, \quad x(0) = \eta, \quad (4.1)$$

with η a random variable independent of the Brownian motion (W_t). We assume that $\delta = 1$ or $\delta \geq 2$.

We need to check the conditions (C1)-(C2) for computable filters presented in the previous lecture. Since the hidden process is the same, condition (C2) holds for the class

$$\mathcal{F}^\delta = \{\nu_{i,\sigma} = g_{i,\sigma}(x)dx, i \in \mathbb{N}, \sigma \geq 0\}, \quad (4.2)$$

where the density $g_{i,\sigma}(x)$ is given by

$$g_{i,\sigma}(x) = 1_{(x>0)} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x}{2\sigma^2}\right) \frac{x^{i-1+\frac{\delta}{2}}}{C_{2i+\delta-1}\sigma^{2i+\delta-1}}. \quad (4.3)$$

It is the density of a Gamma distribution $G(i + \delta/2, 1/2\sigma^2)$ with scale parameter $1/2\sigma^2$ and location parameter $(i + \delta/2)$. The normalising constant C_a , is simply the absolute moment of order a of a Gaussian standard variable.

The extended class $\bar{\mathcal{F}}^\delta$ is defined as the class of mixtures of distributions $\nu_{i,\sigma}$ having the same scale parameter:

$$\bar{\mathcal{F}}^\delta = \{\nu = \nu_{\alpha,\sigma} = \sum_{i \geq 0} \alpha_i \nu_{i,\sigma}, \alpha = (\alpha_i), \alpha_i \geq 0, \sum_{i \geq 0} \alpha_i = 1, \sigma \geq 0\}. \quad (4.4)$$

The sub-class $\bar{\mathcal{F}}_f^\delta$ is composed of all distributions $\nu_{\alpha,\sigma}$ having finite-length mixture parameter, *i.e.* such that

$$L(\alpha) = \sup\{i; \alpha_i > 0\} < \infty.$$

If $\sigma = 0$, $\nu\{\alpha, 0\} = \delta_0$ whatever α .

We only need to check condition (C1) for the same class of distributions.

Proposition 4.2.1. (*up-dating and marginal operators*)

- If $\nu_{i,\sigma}$ belongs to \mathcal{F}^δ , then, for all integer y

$$\varphi_y(\nu_{i,\sigma}) = \nu_{t_y(i), T_y(\sigma)}$$

with

$$t_y(i) = y + i \quad \text{and} \quad \frac{1}{T_y^2(\sigma)} = \frac{1}{T^2(\sigma)} = \frac{1}{\sigma^2} + 2\lambda$$

($T^2(\sigma) = \frac{\sigma^2}{1+2\lambda\sigma^2}$). The up-dating operator is therefore:

$$(i, \sigma) \rightarrow (t_y(i) = y + i, T(\sigma) = \frac{\sigma}{(1 + 2\lambda\sigma^2)^{1/2}}).$$

The corresponding marginal distribution has density, w.r.t. the counting measure on \mathbb{N} ,

$$p_{\nu_{i,\sigma}}(y) = \frac{(\lambda T^2(\sigma))^y}{y!} \frac{C_{2(y+i)+\delta-1}}{C_{2i+\delta-1}} \left(\frac{T(\sigma)}{\sigma}\right)^{2i+\delta}$$

- If $\nu = \nu_{\alpha, \sigma}$ belongs to $\bar{\mathcal{F}}_f^\delta$,

$$\varphi_y(\nu) = \sum_{i \geq 0} \hat{\alpha}_i \nu_{i, T(\sigma)}$$

where

$$\hat{\alpha}_i = \hat{\alpha}_i(\alpha, \sigma) \propto 1_{i \geq y} \alpha_{i-y} p_{\nu_{i-y, \sigma}}(y).$$

We have $L(\hat{\alpha}) = L(\alpha) + y$. Condition (C1) holds for the class \mathcal{F}^δ and the extended class $\bar{\mathcal{F}}_f^\delta$.

Proof. Noting that:

$$f(y|x)g_{i, \sigma}(x) \propto x^{y+i-1+\delta/2} \exp\left(-\left(\lambda + \frac{1}{2\sigma^2}\right)x\right),$$

we get the results. \square

Note that, here, the scale parameter does not depend on the new observation y .

4.2.1 Algorithm for predictive distributions.

We denote the parameters of a distribution $\nu_{i|i-1:1}(dx)$ by $\alpha_{i|i-1:1}$, $\sigma_{i|i-1:1}$ and $L_{i|i-1:1}$ for the length of the mixture parameter. We use analogous notations for $\nu_{i|i:1}(dx)$. Let us stress here the interest of conditions (C1)-(C2). Since (C2) only concerns the hidden Markov process, there are no changes at all for the prediction step.

Proposition 4.2.2. *Assume that $\nu_{1|0:1}$ belongs to the class $\bar{\mathcal{F}}_f^\delta$. The scale, mixture and length parameters of $\nu_{i|i-1:1}(dx)$, $\nu_{i|i:1}(dx)$ can be recursively computed as follows.*

- (up-dating) For $i \geq 1$,

$$\sigma_{i|i:1}^2 = \frac{\sigma_{i|i-1:1}^2}{1 + 2\lambda\sigma_{i|i-1:1}^2}.$$

For $j \geq 0$,

$$\alpha_{i|i:1}(j) \propto 1_{j \geq y_i} \alpha_{i|i-1:1}(j - y_i) \frac{C_{2j+\delta-1}}{C_{2(j-y_i)+\delta-1}} \left(\frac{\sigma_{i|i:1}^2}{\sigma_{i|i-1:1}^2} \right)^{j-y_i}.$$

The lengths of the mixture parameters satisfy $L_{i|i:1} = L_{i|i-1:1} + y_i$.

- (prediction)

$$\sigma_{i+1|i:1}^2 = \beta^2(\Delta) + a^2(\Delta)\sigma_{i|i:1}^2.$$

For $j \geq 0$,

$$\alpha_{i+1|i:1}(j) = \sum_{k \geq j} \alpha_{i|i:1}(k) \kappa_j^{(k)}(\Delta)$$

with

$$\kappa_j^{(k)}(\Delta) = \binom{k}{j} \left(1 - \frac{\beta^2(\Delta)}{\sigma_{i+1|i:1}^2}\right)^j \left(\frac{\beta^2(\Delta)}{\sigma_{i+1|i:1}^2}\right)^{k-j}.$$

The lengths of the mixture parameters satisfy $L_{i+1|i:1} = L_{i|i:1} = L_{1|0:1} + y_1 + \dots + y_i$.

- (Marginal distributions):

$$p_{\nu_{i|i-1:1}}(y_i) = p_i(y_i|y_{i-1}, \dots, y_1) = \sum_{j=0}^{L_{i|i-1:1}} \alpha_{i|i-1:1}(j) p_{\nu_{j, \sigma_{i|i-1:1}}}(y_i).$$

4.2.2 Exact likelihood.

We assume now that $\theta < 0$ and that the process $(x(t))$ is in stationary regime with marginal distribution

$$\nu_{1|0:1} = \pi_\vartheta = G\left(\delta/2, \frac{1}{2\sigma_s^2}\right) = \nu_{0, \sigma_s(\vartheta)}.$$

The joint process (x_i, y_i) is therefore strictly stationary and the distribution of y_i is given by

$$p_1(\vartheta, y) = p_{\nu_{1|0:1}}(y) = \frac{(\lambda T^2(\sigma_s(\vartheta)))^y C_{2y+\delta-1}}{y! C_{\delta-1}} \left(\frac{T(\sigma_s(\vartheta))}{\sigma_s(\vartheta)}\right)^\delta$$

The exact likelihood is computed by:

$$p_n(\vartheta, y_1, \dots, y_n) = p_1(\vartheta, y_1) \prod_{i=2}^n p_i(\vartheta, y_i|y_{i-1}, \dots, y_1)$$

where

$$p_i(\vartheta, y_i|y_{i-1}, \dots, y_1) = p_{\nu_{i|i-1:1}}(y_i) = \sum_{0 \leq j \leq L_{i|i-1:1}} \alpha_{i|i-1:1}(j) p_{\nu_{j, \sigma_{i|i-1:1}}}(y_i)$$

where $L_{i|i-1:1} = y_1 + \dots + y_{i-1}$ and

$$p_{\nu_{j, \sigma_{i|i-1:1}}}(y_i) = \frac{\lambda^{y_i} \sigma_{i|i-1:1}^{2y_i} C_{2(y_i+j)+\delta-1}}{y_i! C_{2j+\delta-1}} \left(\frac{\sigma_{i|i-1:1}}{\sigma_{i|i-1:1}}\right)^{2j+\delta}$$

4.2.3 Extension.

It is possible to consider a more general observation kernel. We may assume that the observations y_i are such that the conditional distribution of y_i given $x_i = x$ is Poisson with parameter $\lambda x_i + \lambda_0$ for $\lambda_0 > 0$ a constant. Then, $F(x, dy) = f(y|x)\mu(dy)$ with

$$\begin{aligned} f(y|x) &= \exp(-\lambda x + \lambda_0) \frac{(\lambda x + \lambda_0)^y}{y!} \\ &= \sum_{i=0}^y \exp(-\lambda_0) \frac{\lambda_0^{y-i} \lambda^i}{(y-i)! i!} x^i \exp(-\lambda x). \end{aligned}$$

The previous study can be generalized.

Note that this is a discrete time version of the continuous time model proposed in Boel and Benes (1980).

4.3 Wright-Fisher diffusion and conditional Bernoulli observations.

Bibliography

- [1] Boel R.K., Benes V.E. (1980). Recursive nonlinear estimation of a diffusion acting as the rate of an observed Poisson process *IEEE Transactions on information theory* **26** (5), 561-575.
- [2] Chaleyat-Maurel M. and Genon-Catalot V. (2006). Computable infinite-dimensional filters with applications to discretized diffusion processes. *Stoch. Proc. and Applic.* **116**, 1447-1467.
- [3] Chaleyat-Maurel M. and Genon-Catalot V. (2009). Filtering the Wright-Fisher diffusion, to appear in *ESAIM P & S*.