Bayesian Learning for Partially-Observed Dynamical Systems

Randal Douc and Sylvain Le Corff

Tutorial 2 : Maximum likelihood.

randal.douc@telecom-sudparis.eu   sylvain.le_corff@telecom-sudparis.eu

CHAPTER 2. MAXIMUM LIKELIHOOD ESTIMATION

EXERCICE 1 Let $p \in \mathbb{N}^*$ and consider the AR(p) model, $X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \sigma Z_t$, where $\{Z_t,\ t \in \mathbb{N}\}$ is a strong white Gaussian noise. The unknown parameter is $\theta = (\phi_1, \ldots, \phi_p, \sigma^2)$ and Θ is a compact subset of $\mathbb{R}^p \times \mathbb{R}_+$.

1. Write for all $n \geqslant p$ the conditional log-likelihood of the observations $\ln q^\theta(X_{p:n}|X_{0:p-1})$.
2. Prove that the maximum likelihood estimator of the regression coefficients explicitly as follows :

$$\begin{pmatrix} \hat\phi_{n,1} \\ \hat\phi_{n,2} \\ \vdots \\ \hat\phi_{n,p} \end{pmatrix} = \hat\Gamma_n^{-1} \begin{pmatrix} n^{-1}\sum_{t=p}^{n} X_t X_{t-1} \\ n^{-1}\sum_{t=p}^{n} X_t X_{t-2} \\ \vdots \\ n^{-1}\sum_{t=p}^{n} X_t X_{t-p} \end{pmatrix} \tag{1}$$

   where $\hat\Gamma_n$ is the $(p \times p)$ empirical covariance matrix for which the i,j-th element is defined by $\hat\Gamma_n(i,j) = n^{-1}\sum_{t=p}^{n} X_{t-i}X_{t-j}$.
3. Prove that the maximum likelihood estimator for the innovation variance is given by :

$$\hat\sigma_n^2 = \frac{1}{n-p+1}\sum_{t=p}^{n}\left(X_t - \sum_{j=1}^{p}\hat\phi_{n,j}X_{t-j}\right)^2 . \tag{2}$$

4. Assume that $(\phi_1, \phi_2, \ldots, \phi_p) \in \mathbb{R}^p$ is such that $\phi(z) = 1 - \sum_{j=1}^{p}\phi_j z^j \neq 0$ for $|z| \leqslant 1$. Set

$$\ln q^\theta(x_{0:p-1}, x_p) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(x_p - \sum_{j=1}^{p}\phi_j x_{p-j}\right)^2 .$$

   Compute the Fisher information matrix $\mathcal{J}(\theta) \stackrel{\text{def}}{=} -\mathbb{E}^\theta\left[\nabla^2 \ln q^\theta(X_{0:p-1}; X_p)\right]$.

EXERCICE 2 In the case where we are interested in estimating unknown parameters $\theta \in \mathbb{R}^m$ characterizing a model with missing data, the Expectation Maximization (EM) algorithm (Dempster et al. 1977) can be used when the joint distribution of the missing data X and the observed data Y is explicit. For all $\theta \in \mathbb{R}^m$, let p_θ be the probability density function of (X,Y) when the model is parameterized by θ with respect to a given reference measure μ. The EM algorithm aims at computing iteratively an approximation of the maximum likelihood estimator which maximizes the observed data loglikelihood :

$$\ell(\theta; Y) = \log p_\theta(Y) = \log\int p_\theta(x,Y)\mu(\mathrm{d}x) .$$

As this quantity cannot be computed explicitly in general cases, the EM algorithm finds the maximum likelihood estimator by iteratively maximizing the expected complete data loglikelihood. Start with an inital value $\theta^{(0)}$ and let $\theta^{(t)}$ be the estimate at the t-th iteration for $t \geqslant 0$, then the next iteration of EM is decomposed into two steps.

1. **E step**. Compute the expectation of the complete data loglikelihood, with respect to the conditional distribution of the missing data given the observed data parameterized by $\theta^{(t)}$:

$$Q(\theta, \theta^{(t)}) = \mathbb{E}_{\theta^{(t)}}\left[\log p_\theta(X,Y)|Y\right] .$$

2. **M step**. Determine $\theta^{(t+1)}$ by maximizing the function Q :

$$\theta^{(t+1)} \in \mathrm{argmax}_\theta Q(\theta, \theta^{(t)}) .$$

1. Prove the following crucial property motivates the EM algorithm. For all $\theta, \theta^{(t)}$,

$$\ell(Y;\theta) - \ell(Y;\theta^{(t)}) \geqslant Q(\theta, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)}) .$$

In the following, $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ where $(X_i)_{0 \leqslant i \leqslant n}$ is a Markov chain taking values in $\{1, \ldots, r\}$ with transition matrix $\mathrm{Q} = (q_{i,j})_{1 \leqslant i,j \leqslant r}$ and, for all $1 \leqslant k \leqslant n$, the conditional distribution of Y_k given the σ-field generated by $(X_{1:n}, Y_{1:k-1})$ is a Gaussian distribution with mean $\mu_{X_k} \in \mathbb{R}$ and variance $\vartheta_{X_k} \in \mathbb{R}_+^*$. In this case, the unknown parameter $\theta = (\mu_{1:k}, \vartheta_{1:k}, \mathrm{Q})$

1. Write the complete data loglikelihood $\theta \mapsto \log p_\theta(X_{1:n}, Y_{1:n}|X_0)$.
2. Let $\theta^{(t)}$ be the current parameter estimate. Compute $\theta \mapsto Q(\theta, \theta^{(t)})$ using $\mathbb{P}_{\theta^{(t)}}(X_k = i|Y_{1:n})$ and $\mathbb{P}_{\theta^{(t)}}(X_{k-1} = i, X_k = j|Y_{1:n})$ for all $1 \leqslant i, j \leqslant r$.
3. Compute $\theta^{(t+1)}$.

EXERCICE 3 Assume that the observations $\{Y_t,\ t \in \mathbb{Z}\}$ are a strict-sense stationary ergodic process associated to

$$\mathbb{P}\left[Y_t \in A \,|\, \mathcal{F}_{t-1}\right] = Q^\star(X_{t-1}, A) = \int_A q^\star(X_{t-1}, y)\mu(\mathrm{d}y) , \quad \text{for any } A \in \mathcal{B}(\mathsf{Y}) ,$$
$$X_t = f^{\theta^*}_{Y_t}(X_{t-1}) , \quad t \in \mathbb{Z} .$$

The observations are used to fit the following observation-driven model

$$\mathbb{P}\left[Y_t \in A \,|\, \mathcal{F}_{t-1}\right] = Q(X_{t-1}, A) , \quad \text{for any } A \in \mathcal{B}(\mathsf{Y}) ,$$
$$X_t = f^{\theta}_{Y_t}(X_{t-1}) , \quad (t, \theta) \in \mathbb{Z} \times \Theta .$$

where $Q(x, \cdot)$ is assumed to belong to the class of exponential family distributions. More precisely, we assume that for all $(x,y) \in \mathsf{X} \times \mathsf{Y}$, $q(x,y) = \exp(xy - A(x))h(y)$ for some twice differentiable function $A : \mathsf{X} \to \mathbb{R}$ and some measurable function $h : \mathsf{Y} \to \mathbb{R}^+$.

1. Prove that for all x, $\int Q(x^\star, \mathrm{d}y) \frac{\partial^2 \ln q(x,y)}{\partial x^2} \leqslant 0$, and show that A is convex.
2. Deduce the maximum of $x \mapsto \int Q^\star(x, \mathrm{d}y) \ln q(x,y)$.
3. Apply the consistency result for observation driven models in the case of a log-linear Poisson autoregression model where

$$q(x,y) = \exp(xy - \mathrm{e}^x)/y! ,$$

i.e. provide an assumption on $Q^\star$ to obtain consistency of the Quasi Maximum Likelihood Estimators.