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

The correction of this exercise may be found in the lecture notes.

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 $\theta, \theta^{(t)}$,

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

This may be proved by noting that

$$\ell(Y;\theta) = \log\left(\frac{p_\theta(X,Y)}{p_\theta(X|Y)}\right) \; .$$

Considering the conditional expectation of both terms given Y when the parameter value is $\theta^{(t)}$ yields

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

Then,

$$\ell(Y;\theta) - \ell(Y;\theta^{(t)}) = Q(\theta,\theta^{(t)}) - Q(\theta^{(t)},\theta^{(t)}) + H(\theta,\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)}) \; ,$$

where

$$H(\theta,\theta^{(t)}) = -\mathbb{E}_{\theta^{(t)}}[\log p_\theta(X|Y)|Y] \; .$$

The proof is completed by noting that

$$H(\theta,\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)}) \geqslant 0 \; ,$$

as this difference if a Kullback-Leibler divergence.

Therefore, any value θ which improves $Q(\theta,\theta^{(t)})$ beyond reference value $Q(\theta^{(t)},\theta^{(t)})$, increases the observed-data likelihood.

In the following, $X = (X_1,\ldots,X_n)$ and $Y = (X_0,Y_1,\ldots,Y_n)$ where $(X_i)_{0\leqslant i\leqslant n}$ is a Markov chain taking values in $\{1,\ldots,r\}$ with transition matrix $\mathsf{Q} = (q_{i,j})_{1\leqslant i,j\leqslant r}$ and, for all $1 \leqslant k \leqslant n$, the conditional distribution of Y_k given the σ-field generated by $(X_{0:n},Y_{1:k-1})$ 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 $\theta = (\mu_{1:k},\vartheta_{1:k},\mathsf{Q})$

1. Write the complete data loglikelihood $\theta \mapsto \log p_\theta(X_{1:n}, Y_{1:n}|X_0)$.
   The complete data loglikelihood is given by

$$\begin{aligned}
\log p_\theta\left(X_{1:n}, Y_{1:n}|X_0\right) &= \log p_\theta\left(X_{1:n}|X_0\right) + \log p_\theta\left(Y_{1:n}|X_{0:n}\right) \,, \\
&= \sum_{k=1}^{n} \left(\log p_\theta\left(X_k|X_{k-1}\right) + \log p_\theta\left(Y_k|X_k\right)\right) \,, \\
&= \sum_{k=1}^{n} \left(\log p_\theta\left(X_k|X_{k-1}\right) - \frac{1}{2}\log(2\pi\vartheta_{X_k}) - \frac{1}{2\vartheta_{X_k}}\left(Y_k - \mu_{X_k}\right)^2\right) \,, \\
&= \sum_{1\leqslant i,j\leqslant r} \log q_{i,j} \left\{\sum_{k=1}^{n} \mathbb{1}_{X_{k-1}=i}\mathbb{1}_{X_k=j}\right\} - \sum_{1\leqslant i\leqslant r} \frac{1}{2}\log(2\pi\vartheta_i) \sum_{k=1}^{n} \mathbb{1}_{X_k=i} \\
&\qquad - \sum_{1\leqslant i\leqslant r}\sum_{k=1}^{n} \mathbb{1}_{X_k=i} \frac{1}{2\vartheta_i}\left(Y_k - \mu_i\right)^2 \,.
\end{aligned}$$

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 $1 \leqslant i, j \leqslant r$.
   The intermediate quantity of the EM algorithm is given, for all θ*, by*

$$\begin{aligned}
Q(\theta, \theta^{(t)}) = \sum_{1\leqslant i,j\leqslant r} \log q_{i,j} \sum_{k=1}^{n} \mathbb{P}^{\theta^{(t)}}\left(\mathbb{1}_{X_{k-1}=i}\mathbb{1}_{X_k=j}\middle|Y_{1:n}\right) - \sum_{1\leqslant i\leqslant r} \frac{1}{2}\log(2\pi\vartheta_i) \sum_{k=1}^{n} \mathbb{P}^{\theta^{(t)}}\left(X_k = i|Y_{1:n}\right) \\
- \sum_{1\leqslant i\leqslant r}\sum_{k=1}^{n} \mathbb{1}_{X_k=i} \frac{1}{2\vartheta_i}\left(Y_k - \mu_i\right)^2 \,,
\end{aligned}$$

   where $\mathbb{P}^{\theta^{(t)}}$ *denotes the distribution induced when of the model parameterized with* $\theta^{(t)}$*.*

3. Compute $\theta^{(t+1)}$.
   Maximizing $\theta \to Q(\theta, \theta^{(t)})$ *yields*

$$\begin{aligned}
\mu_i^{(t+1)} &= \frac{\sum_{k=1}^{n} \mathbb{P}^{\theta^{(t)}}\left(\mathbb{1}_{X_k=i}|Y_{1:n}\right) Y_k}{\sum_{k=1}^{n} \mathbb{P}^{\theta^{(t)}}\left(\mathbb{1}_{X_k=i}|Y_{1:n}\right)} \,, 1 \leqslant i \leqslant r \,, \\
\vartheta_i^{(t+1)} &= \frac{\sum_{k=1}^{n} \mathbb{P}^{\theta^{(t)}}\left(\mathbb{1}_{X_k=i}|Y_{1:n}\right)\left(Y_k - \mu_i\right)^2}{\sum_{k=1}^{n} \mathbb{P}^{\theta^{(t)}}\left(\mathbb{1}_{X_k=i}|Y_{1:n}\right)} \,, 1 \leqslant i \leqslant r \,, \\
q_{i,j}^{(t+1)} &= \frac{\sum_{k=1}^{n} \mathbb{P}^{\theta^{(t)}}\left(\mathbb{1}_{X_{k-1}=i}\mathbb{1}_{X_k=j}\middle|Y_{1:n}\right)}{\sum_{k=1}^{n} \mathbb{P}^{\theta^{(t)}}\left(\mathbb{1}_{X_{k-1}=i}\middle|Y_{1:n}\right)} \,, 1 \leqslant i, j \leqslant r \,.
\end{aligned}$$

   Note that updating all parameters require for instance to compute $\mathbb{P}^{\theta^{(t)}}\left(\mathbb{1}_{X_{k-1}=i}\mathbb{1}_{X_k=j}\middle|Y_{1:n}\right)$*. While the parameter* $\theta^{(t)}$ *is known (current parameter estimate) computing such probability under the distribution of the unobserved chain given the observations is not tractable. Such quantities may be estimated with Markov Chain Monte Carlo methods...*

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

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

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

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

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, $\int Q(x,\mathrm{d}y)\frac{\partial^2 \ln q(x,y)}{\partial x^2} \leqslant 0$, and show that A is convex.

   $$\frac{\partial^2 \ln q(x,y)}{\partial x^2} = \frac{\partial^2_{xx} q(x,y)}{q(x,y)} - \left(\frac{\partial \ln q(x,y)}{\partial x}\right)^2 .$$

   Then, note that for all x,

   $$\int Q(x,\mathrm{d}y)\frac{\partial^2_{xx} q(x,y)}{q(x,y)} = \int \partial^2_{xx} q(x,y)\mu(\mathrm{d}y) = \partial^2_{xx}\int q(x,y)\mu(\mathrm{d}y) = 0\ ,$$

   which concludes the proof (since it yields $A''(x) \geqslant 0$).

2. Deduce the maximum of $x \mapsto \int Q^\star(x,\mathrm{d}y)\ln q(x,y)$.

   Using the previous question, the function

   $$\begin{aligned}x \mapsto \int Q^\star(x^\star,\mathrm{d}y)\ln q(x,y) &= \int Q^\star(x^\star,\mathrm{d}y)\left(xy - A(x) + \ln h(y)\right)\\ &= x\int Q^\star(x^\star,\mathrm{d}y)y - A(x) + \int Q^\star(x^\star,\mathrm{d}y)\ln h(y)\ ,\end{aligned}$$

   is convex. The maximum of this function can thus be obtained by cancelling the derivatives with respect to x, which yields $\int Q^\star(x^\star,\mathrm{d}y)y - A'(x) = 0$.