Expectation Maximization calculation for the DTQ method

We consider a parameteric SDE model:

$$dX_t = f(X_t; \boldsymbol{\theta})dt + g(X_t; \boldsymbol{\theta})dW_t \tag{1}$$

In this model $f(X_t; \boldsymbol{\theta})$ is the drift function and $g(X_t; \boldsymbol{\theta})$ is the diffusion function. A concrete example of such an SDE is the Ornstein-Uhlenback SDE (linear in nature),

$$dX_t = \theta_1(\theta_2 - X_t)dt + \theta_3 dW_t \tag{2}$$

We start with the parameter inference problem where we have data available as a time series, denoted by $\mathbf{x} = (x_0, x_1, \cdots, x_N)$. Since the observed data might be have large inter-observation times, we consider intermediate points which we consider as *missing data points*, denoted by $\mathbf{z}$. On the interval $[t_i, t_{i+1}]$, we have 2 observed data points, $X_{t_i} = x_i$ and $X_{t_{i+1}} = x_{i+1}$. We consider F missing data points on this interval, denoted by $z_{i,F}$, the first subscript corresponding to the interval and the second subscript for the missing data point on the interval. Thus the missing data on an interval $[t_i, t_{i+1}]$, can be represented as $\mathbf{z}_i = (z_{i1}, z_{i2}, \cdots, z_{iF})$. The complete data on this interval would thus become $(x_i, z_{i1}, z_{i2}, \cdots, z_{iF}, x_{i+1})$, comprising of the observed data and the unknown missing data that we introduced.

1 EM algorithm

The Expectation-Maximization algorithm consists of 2 steps, computing the expectation of the log likelihood function and maximizing this value with respect to the parameters.

1. Start with an initial guess for the parameter, $\boldsymbol{\theta}^{(0)}$
2. For the expectation step,

$$\begin{align} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= \mathbb{E}_{\mathbf{z}|\mathbf{x},\boldsymbol{\theta}^{(k)}}[\log p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta})] \tag{3} \\ &= \sum_{\mathbf{z}} \underbrace{\log p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta})}_{\text{Part I}} \cdot \underbrace{p(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)})}_{\text{Part II}} \tag{4} \end{align}$$

3. For the maximization step,

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) \tag{5}$$

   We can either use a numerical optimizer for the optimization step or differentiate the $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ function with respect to $\boldsymbol{\theta}$ vector and equate it to zero to get the maximal value.

4. Iterate Step 2 and 3 until convergence.

1.1 Computation of the complete log likelihood

The first part of the expectation is the complete likelihood, $\log p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta})$, which can be expanded as,

$$\log p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta}) = \log p(x_0 \mid \boldsymbol{\theta}) + \underbrace{\sum_{i=0}^{N-1} \log p(z_{i1} \mid x_i, \boldsymbol{\theta})}_{(1)} + \underbrace{\sum_{i=0}^{N-1} \sum_{j=1}^{F-1} \log p(z_{i,j+1} \mid z_{ij}, \boldsymbol{\theta})}_{(2)} \tag{6}$$

$$+ \underbrace{\sum_{i=0}^{N-1} \log p(x_{i+1} \mid z_{iF}, \boldsymbol{\theta})}_{(3)} \tag{7}$$

The expression can be simplified under the assumption that F is sufficiently large so that we can make an assumption that one-step transition densities in (1), (2) and (3) follow Gaussian distribution.

1.2 Computation of the density of the missing data points

Looking back at the expectation equation (3), the expected value if computed by summing over all $\mathbf{z}$ values which is a nested integral. Since the log likelihood can be expanded in 4 terms, so the density $p(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)})$ gets multiplied by each of these terms. Upon summing over all the values of $\mathbf{z}$, there will be 3 steps of terms remaining, corresponding to the respective terms in the equation (6), as described below,

1. Corresponding to term (1), we have the term $p(z_{i1} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)})$. Using Bayes theorem we get,

$$p(z_{i1}, \mathbf{x} \mid \boldsymbol{\theta}^{(k)}) = p(z_{i1} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)}) \cdot p(\mathbf{x} \mid \boldsymbol{\theta}^{(k)}) \tag{8}$$