Expectation Maximization 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-Uhlenbeck SDE,

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

We start with the parameter inference problem where we have data available as a time series, denoted by $\mathbf{x} = (x_0, x_1, \ldots, x_N)$. Suppose the observed data has large inter-observation times. Then we consider observations at intermediate times as *missing data points*, denoted by $\mathbf{z}$. On the interval $[t_i, t_{i+1}]$, we have two 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 $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iF})$. The complete data on this interval would thus become $(x_i, z_{i1}, z_{i2}, \ldots, z_{iF}, x_{i+1})$, comprising the observed data and the unknown missing data that we introduced.

1 EM algorithm

The Expectation-Maximization algorithm consists of two 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,

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \mathbb{E}_{\mathbf{z}|\mathbf{x},\boldsymbol{\theta}^{(k)}}[\log p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta})] \tag{3}$$

$$= \sum_{\mathbf{z}} \underbrace{\log p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta})}_{\text{Part I}} \cdot \underbrace{p(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)})}_{\text{Part II}} \tag{4}$$

   In the above expression and in what follows, we use $\sum_{\mathbf{z}}$ to stand for either summation or integration over the random vector $\mathbf{z}$, depending on whether $\mathbf{z}$ is discrete or continuous, respectively.

3. For the maximization step, we start with the current iterate and a dummy variable $\boldsymbol{\theta}$, so that the next iterate of the parameters would be the maximal value of $\boldsymbol{\theta}$

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

   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 Why does this work?

We begin with

$$p(\mathbf{x},\mathbf{z} \mid \boldsymbol{\theta}) = p(\mathbf{z} \mid \mathbf{x},\boldsymbol{\theta})p(\mathbf{x} \mid \boldsymbol{\theta}).$$

This is an identity, valid by the laws of probability. If we assume $p(\mathbf{z} \mid \mathbf{x},\boldsymbol{\theta}) \neq 0$, then we can divide both sides by the same and then take the log of both sides to arrive at

$$\log p(\mathbf{x} \mid \boldsymbol{\theta}) = \log p(\mathbf{x},\mathbf{z} \mid \boldsymbol{\theta}) - \log p(\mathbf{z} \mid \mathbf{x},\boldsymbol{\theta}).$$

We now multiple both sides by the density $p(\mathbf{z} \mid \mathbf{x},\boldsymbol{\theta}^{(k)})$ and sum over $\mathbf{z}$—this is the same as computing, on both sides, the conditional expectation of $\mathbf{z}$ given $\mathbf{x}$ and $\boldsymbol{\theta}^{(k)}$. The result is

$$\log p(\mathbf{x} \mid \boldsymbol{\theta}) = \sum_{\mathbf{z}} \log p(\mathbf{x},\mathbf{z} \mid \boldsymbol{\theta})p(\mathbf{z} \mid \mathbf{x},\boldsymbol{\theta}^{(k)}) - \sum_{\mathbf{z}} \log p(\mathbf{z} \mid \mathbf{x},\boldsymbol{\theta})p(\mathbf{z} \mid \mathbf{x},\boldsymbol{\theta}^{(k)}).$$

Let us define

$$H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = -\sum_{\mathbf{z}} \log p(\mathbf{z} \mid \mathbf{x},\boldsymbol{\theta})p(\mathbf{z} \mid \mathbf{x},\boldsymbol{\theta}^{(k)}).$$

Then we have

$$\log p(\mathbf{x} \mid \boldsymbol{\theta}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) + H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}).$$

If we insert $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$ into this last equation, we obtain

$$\log p(\mathbf{x} \mid \boldsymbol{\theta}^{(k)}) = Q(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(k)}) + H(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(k)}).$$

Now we subtract the last equation from the previous one to obtain

$$\log p(\mathbf{x} \mid \boldsymbol{\theta}) - \log p(\mathbf{x} \mid \boldsymbol{\theta}^{(k)}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) - Q(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(k)}) + H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) - H(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(k)}).$$

Let $a = p(\mathbf{z} \mid \mathbf{x},\boldsymbol{\theta}^{(k)})$ and $b = p(\mathbf{z} \mid \mathbf{x},\boldsymbol{\theta})$. Then we have

$$\begin{aligned}
H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) - H(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(k)}) &= \sum_{\mathbf{z}} a \log a - a \log b \\
&= -\sum_{\mathbf{z}} a \log \frac{b}{a} = \mathbb{E}_a[\phi(b/a)] \\
&\geq -\log \sum_{\mathbf{z}} a \cdot \frac{b}{a} = \phi(\mathbb{E}_a[b/a]) \\
&\geq -\log \sum_{\mathbf{z}} b = -\log 1 = 0.
\end{aligned}$$

The inequality above is Jensen's inequality applied to the convex function $\phi(x) = -\log x$. The overall inequality for generic densities a and b is often called Gibbs' inequality.

With this fact, we have

$$\log p(\mathbf{x} \mid \boldsymbol{\theta}) - \log p(\mathbf{x} \mid \boldsymbol{\theta}^{(k)}) \geq Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) - Q(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(k)})$$

Now as long as we choose $\boldsymbol{\theta}$ such that

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) \geq Q(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(k)}),$$

then we are guaranteed that

$$\log p(\mathbf{x} \mid \boldsymbol{\theta}) \geq \log p(\mathbf{x} \mid \boldsymbol{\theta}^{(k)}).$$

In words, what this says is that if $\boldsymbol{\theta}$ improves the value of the Q function, then the same $\boldsymbol{\theta}$ improves the value of the (incomplete) log likelihood function as well.

1.2 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,

$$\begin{aligned}\log p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta}) = \log p(x_0 \mid \boldsymbol{\theta}) &+ \underbrace{\sum_{i=0}^{N-1} \log p(z_{i1} \mid x_i, \boldsymbol{\theta})}_{(1)} + \underbrace{\sum_{i=0}^{N-1} \sum_{j=1}^{F-1} \log p(z_{i,j+1} \mid z_{ij}, \boldsymbol{\theta})}_{(2)} \\ &+ \underbrace{\sum_{i=0}^{N-1} \log p(x_{i+1} \mid z_{iF}, \boldsymbol{\theta})}_{(3)} \end{aligned} \qquad (6)$$

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. Thus all the terms, $p(z_{i1} \mid x_i, \boldsymbol{\theta}), p(z_{i,j+1} \mid z_{ij}, \boldsymbol{\theta})$ and $p(x_{i+1} \mid z_{iF}, \boldsymbol{\theta})$ can be expressed with a Gaussian function

$$G^h_{\boldsymbol{\theta}}(x, y) = p(x \mid y, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi g^2(y, \boldsymbol{\theta}) h}} \exp\left(-\frac{1}{2g^2(y, \boldsymbol{\theta}) h} (x - y - f(y, \boldsymbol{\theta}) h)^2 \right) \qquad (7)$$

In what follows, we will return to the question of how to evaluate this function on a spatially discrete grid.

1.3 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

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

   This implies

   $$p(z_{i1} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)}) = \frac{p(z_{i1}, \mathbf{x} \mid \boldsymbol{\theta}^{(k)})}{p(\mathbf{x} \mid \boldsymbol{\theta}^{(k)})} = \frac{p(z_{i1}, \mathbf{x} \mid \boldsymbol{\theta}^{(k)})}{p(x_0 \mid \boldsymbol{\theta}^{(k)}) \prod_{j=0}^{N-1} p(x_{j+1} \mid x_j, \boldsymbol{\theta}^{(k)})} \qquad (8)$$

The numerator can be expanded using the Markov property:

$$p(z_{i1}, \mathbf{x} \mid \boldsymbol{\theta}^{(k)}) = p(z_{i1}, x_0, x_1, \cdots, x_N \mid \boldsymbol{\theta}^{(k)})$$
$$= p(x_0 \mid \boldsymbol{\theta}^{(k)}) \prod_{j=i+1}^{N-1} p(x_{j+1} \mid x_j, \boldsymbol{\theta}^{(k)})\, p(x_{i+1} \mid z_{i1}, \boldsymbol{\theta}^{(k)})\, p(z_{i1} \mid x_i, \boldsymbol{\theta}^{(k)}) \prod_{j=0}^{i-1} p(x_{j+1} \mid x_j, \boldsymbol{\theta}^{(k)})$$

Substituting this expansion into (8), we obtain

$$p(z_{i1} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)}) = \frac{p(x_{i+1} \mid z_{i1}, \boldsymbol{\theta}^{(k)})\, p(z_{i1} \mid x_i, \boldsymbol{\theta}^{(k)})}{p(x_{i+1} \mid x_i, \boldsymbol{\theta}^{(k)})} \tag{9}$$

2. Corresponding to the F internal steps represented by term (2), we have the terms $p(z_{i,j+1}, z_{ij} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)})$. We again use Bayes' theorem to get

$$p(z_{i,j+1}, z_{ij} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)}) = \frac{p(z_{i,j+1}, z_{ij}, \mathbf{x} \mid \boldsymbol{\theta}^{(k)})}{p(x_0 \mid \boldsymbol{\theta}^{(k)}) \prod_{j=0}^{N-1} p(x_{j+1} \mid x_j, \boldsymbol{\theta}^{(k)})} \tag{10}$$

The numerator can be expanded using the Markov property as follows

$$p(z_{i,j+1}, z_{ij}, \mathbf{x} \mid \boldsymbol{\theta}^{(k)}) = p(x_0 \mid \boldsymbol{\theta}^{(k)}) \prod_{j=0}^{i-1} p(x_{j+1} \mid x_j, \boldsymbol{\theta}^{(k)}) \cdot p(z_{ij} \mid x_i, \boldsymbol{\theta}^{(k)}) \cdot p(z_{i,j+1} \mid z_{ij}, \boldsymbol{\theta}^{(k)})$$
$$p(x_{i+1} \mid z_{i,j+1}, \boldsymbol{\theta}^{(k)}) \prod_{j=1}^{N-1} p(x_{j+1} \mid x_j, \boldsymbol{\theta}^{(k)})$$

We insert this into the numerator of (10) to obtain

$$p(z_{i,j+1}, z_{ij} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)}) = \frac{p(x_{i+1} \mid z_{i,j+1}, \boldsymbol{\theta}^{(k)})\, p(z_{i,j+1} \mid z_{ij}, \boldsymbol{\theta}^{(k)})\, p(z_{ij} \mid x_i, \boldsymbol{\theta}^{(k)})}{p(x_{i+1} \mid x_i, \boldsymbol{\theta}^{(k)})} \tag{11}$$

3. The last term (3) corresponds to $p(z_{iF} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)})$. Using the same techniques described above, we can write this term as

$$p(z_{iF} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)}) = \frac{p(x_{i+1} \mid z_{iF}, \boldsymbol{\theta}^{(k)})\, p(z_{iF} \mid x_i, \boldsymbol{\theta}^{(k)})}{p(x_{i+1} \mid x_i, \boldsymbol{\theta}^{(k)})} \tag{12}$$

1.4 DTQ Computations

Now for the part where we explain how the DTQ method enables us to calculate the terms above. We focus on two terms in particular. In what follows, we will need several definitions. Let $\Delta x > 0$ denote the spatial grid spacing. Then the spatial grid is defined by $y_j = j\Delta x$ for $j = -M, \ldots, M$. In this scheme, there are $2M+1$ grid points. Recall the definition of $G^h_{\boldsymbol{\theta}}$ from (7). This gives rise to the $(2M+1) \times (2M+1)$ matrix K where

$$K_{ab} = (\Delta x) G^h_{\boldsymbol{\theta}}(a\Delta x, b\Delta x)$$

for $a, b \in \{-M, \ldots, M\}$. Note the prefactor of Δx. We also define the $(2M+1) \times 1$ vector $\mathbf{\Lambda}^i$ by

$$\mathbf{\Lambda}^i_a = G^h_{\boldsymbol{\theta}}(a\Delta x, x_i).$$

Finally, we define the $1 \times (2M+1)$ vector $\mathbf{\Gamma}^i$ by

$$\mathbf{\Gamma}^i_a = (\Delta x) G^h_{\boldsymbol{\theta}}(x_{i+1}, a\Delta x).$$

With these definitions, we have the DTQ approximation to the likelihood:

$$p(x_{i+1} \mid x_i, \boldsymbol{\theta}^{(k)}) \approx \mathbf{\Gamma}^i K^{F-1} \mathbf{\Lambda}^i.$$

Type I (forward). The term we have in mind here is

$$p(z_{ij} \mid x_i, \boldsymbol{\theta}^{(k)}).$$

To get from time t_i to the time associated with z_{ij} requires j steps of size h. The first such step is handled analytically and the result is $\mathbf{\Lambda}^i$. Hence only $j-1$ steps remain to reach the time associated with z_{ij}. Therefore,

$$p(z_{ij} \mid x_i, \boldsymbol{\theta}^{(k)}) \approx K^{j-1} \mathbf{\Lambda}^i.$$

To compute the right-hand side, we start with the column vector $\mathbf{\Lambda}^i$ and repeatedly multiply on the left by the matrix K.

Type II (backward). The term we have in mind here is

$$p(x_{i+1} \mid z_{i,j+1}, \boldsymbol{\theta}^{(k)}).$$

Here we think of initializing $(2M+1)$ parallel trajectories at the time associated with $z_{i,j+1}$. The a-th such trajectory is initialized with a Dirac delta centered at $a\Delta x$. Then, at the time associated with $z_{i,j+2}$, we have a matrix of Gaussians. Let us denote this matrix by

$$A = \left[G^h_{\boldsymbol{\theta}}(\mathbf{z}, -M\Delta x) \cdots G^h_{\boldsymbol{\theta}}(\mathbf{z}, M\Delta x) \right].$$

To be perfectly clear, the j-th column of A is the $(2M+1) \times 1$ column vector

$$G^h_{\boldsymbol{\theta}}(\mathbf{z}, j\Delta x)$$

obtained by evaluating $G^h_{\boldsymbol{\theta}}(a\Delta x, j\Delta x)$ for each value of $a \in \{-M, \ldots, M\}$. At this point, we recognize that

$$A_{ab} = G^h_{\boldsymbol{\theta}}(a\Delta x, b\Delta x),$$

implying that

$$A = (\Delta x)^{-1} K.$$

Now that we have the A matrix, we can write

$$p(x_{i+1} \mid z_{i,j+1}, \boldsymbol{\theta}^{(k)}) \approx \mathbf{\Gamma}^i K^{F-(j+2)} A = (\Delta x)^{-1} \mathbf{\Gamma}^i K^{F-(j+1)}.$$

To compute the quantity on the right-hand side, we start with the row vector $\mathbf{\Gamma}^i$. We repeatedly multiply on the right by the matrix K. Finally, we multiply on the right by the matrix A.

Note that we need 1 step to get from $z_{i,j+1}$ to $z_{i,j+2}$, $F-(j+2)$ steps to get from $z_{i,j+2}$ to z_{iF}, and then one last $\mathbf{\Gamma}^i$ step to get to time t_{i+1}.

Important Point. In the expressions we have given for $\mathbf{\Gamma}^i$, $\mathbf{\Lambda}^i$, K, and A, we have not specified the actual value to use for the parameter vector $\boldsymbol{\theta}$. This is to be determined by context. If we are conditioning on $\boldsymbol{\theta}^{(k)}$, then we should use $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$ consistently in $\mathbf{\Gamma}^i$, $\mathbf{\Lambda}^i$, K, and A.

Self-Consistency. Let us examine the self-consistency of our derivations. Using the laws of probability and the Markov property, we can write

$$p(x_{i+1} \mid x_i, \boldsymbol{\theta}^{(k)}) = \int_{z_{ij}} p(x_{i+1} \mid z_{ij}, \boldsymbol{\theta}^{(k)}) p(z_{ij} \mid x_i, \boldsymbol{\theta}^{(k)})\, dz_{ij}.$$

Suppose we now insert all of the approximations we have derived. The integral becomes a summation and we replace dz_{ij} with Δx. On the right-hand side, we obtain the approximation

$$(\Delta x)^{-1} \mathbf{\Gamma}^i K^{F-j} \cdot K^{j-1} \mathbf{\Lambda}^i \cdot \Delta x = \mathbf{\Gamma}^i K^{F-1} \mathbf{\Lambda}^i,$$

which is exactly identical to our approximation of the left-hand side. Hence our method is self-consistent. One consequence of this is that densities such as

$$p(z_{ij} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)})$$

computed using our method will be correctly normalized.

How to Actually Compute What We Want. The first term we actually want is (9). Given two vectors $\mathbf{a}$ and $\mathbf{b}$ of the same length—whether they are row or column vectors—we define their component-wise (or Hadamard) product to be the vector $\mathbf{a} \circ \mathbf{b}$ given by

$$(\mathbf{a} \circ \mathbf{b})_i = a_i b_i.$$

The vector $(\mathbf{a} \circ \mathbf{b})$ can be either a column or row vector, as appropriate. With this notation, we have

$$\begin{aligned} p(z_{i1} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)}) &= \frac{p(x_{i+1} \mid z_{i1}, \boldsymbol{\theta}^{(k)})\, p(z_{i1} \mid x_i, \boldsymbol{\theta}^{(k)})}{p(x_{i+1} \mid x_i, \boldsymbol{\theta}^{(k)})} \\ &= \frac{(\Delta x)^{-1} \mathbf{\Gamma}^i K^{F-1} \circ \mathbf{\Lambda}^i}{\mathbf{\Gamma}^i K^{F-1} \mathbf{\Lambda}^i}. \end{aligned}$$

This is term (9) above. We can handle (12) similarly:

$$\begin{aligned} p(z_{iF} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)}) &= \frac{p(x_{i+1} \mid z_{iF}, \boldsymbol{\theta}^{(k)})\, p(z_{iF} \mid x_i, \boldsymbol{\theta}^{(k)})}{p(x_{i+1} \mid x_i, \boldsymbol{\theta}^{(k)})} \\ &= \frac{(\Delta x)^{-1} \mathbf{\Gamma}^i \circ K^{F-1} \mathbf{\Lambda}^i}{\mathbf{\Gamma}^i K^{F-1} \mathbf{\Lambda}^i}. \end{aligned}$$

There is a slight increase in complexity when we seek to compute (11). Let us start by recognizing that the quantity

$$p(z_{i,j+1}, z_{ij} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)})$$

will be represented by a $(2M+1) \times (2M+1)$ matrix. Based on the right-hand side of (11), we start with

$$p(z_{i,j+1} \mid z_{ij}, \boldsymbol{\theta}^{(k)}) = G^h_{\boldsymbol{\theta}^{(k)}}(z_{i,j+1}, z_{ij}).$$

Discretizing the variables $z_{i,j+1}$ and z_{ij} using our equispaced grid results in the same matrix $A = (\Delta x)^{-1} K$ discussed above. Now let us add in the remaining pieces of (11). Consider

$$\begin{aligned} z_{i,j+1} &= a(\Delta x) \\ z_{ij} &= b(\Delta x) \end{aligned}$$

for $a, b \in \{-M, \dots, M\}$. Then our approximation of $p(z_{i,j+1}, z_{ij} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)})$ is the matrix Ψ given by

$$\Psi_{a,b} = (\Delta x)^{-1} \frac{\left[\boldsymbol{\Gamma}^i K^{F-(j+1)}\right]_a A_{ab} \left[K^{j-1} \boldsymbol{\Lambda}^i\right]_b}{\boldsymbol{\Gamma}^i K^{F-1} \boldsymbol{\Lambda}^i}. \tag{13}$$

For vectors $\mathbf{a}$ and $\mathbf{b}$, let $\mathbf{a} \otimes \mathbf{b}$ denote the Kronecker product. This product is defined by $[\mathbf{a} \otimes \mathbf{b}]_{ij} = a_i b_j$. Then we can write:

$$p(z_{i,j+1}, z_{ij} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)}) \approx \Psi = \frac{(\Delta x)^{-1}}{\boldsymbol{\Gamma}^i K^{F-1} \boldsymbol{\Lambda}^i} \left(\left[\boldsymbol{\Gamma}^i K^{F-(j+1)}\right] \otimes \left[K^{j-1} \boldsymbol{\Lambda}^i\right] \right) \circ A.$$

Here the Hadamard product $\circ$ is an element-wise product between two matrices of equal size.

1.5 Expectation Step

Combining the terms from Section 1.2 and Section 1.3, we can form a complete expression for the expectation. Going back to Section 1.2, we recall that the transition densities can be assumed to be Gaussian for sufficiently large enough F. Thus, the expectation expression can be rewritten as,

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \sum_{\mathbf{z}} \Big\{ \log p(x_0 \mid \boldsymbol{\theta}) &+ \sum_{i=0}^{N-1} \log G(z_{i1}, x_i, \boldsymbol{\theta}) \cdot p(z_{i1} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)}) \\ &+ \sum_{i=0}^{N-1} \sum_{j=1}^{F-1} \log G(z_{i,j+1}, z_{ij}, \boldsymbol{\theta}) \cdot p(z_{i,j+1}, z_{ij} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)}) \\ &+ \sum_{i=0}^{N-1} \log G(x_{i+1}, z_{iF}, \boldsymbol{\theta}) \cdot p(z_{iF} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)}) \Big\} \end{aligned} \tag{14}$$

1.6 Maximization Step

For the maximization step, there are 2 ways to maximize $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$, either through numerical optimizers or through equating the derivative with respect to the parameters to zero and using a root-finding solver if required.

Since both these methods require the gradient of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ so we specify the gradients below. The derivative of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ with respect to the $\boldsymbol{\theta}$ parameters would then be,

$$\begin{aligned} 0 = \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})}{\partial \theta_\ell} = \frac{p'(x_0 \mid \boldsymbol{\theta})}{p(x_0 \mid \boldsymbol{\theta})} &+ \sum_{i=0}^{N-1} \frac{H_\ell(z_{i1}, x_i, \boldsymbol{\theta})}{G(z_{i1}, x_i, \boldsymbol{\theta})} \cdot p(z_{i1} \mid x_i, \boldsymbol{\theta}^{(k)}) \\ &+ \sum_{i=0}^{N-1} \sum_{j=1}^{F-1} \frac{H_\ell(z_{i,j+1}, z_{ij}, \boldsymbol{\theta})}{G(z_{i,j+1}, z_{ij}, \boldsymbol{\theta})} \cdot p(z_{i,j+1}, z_{ij} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)}) \\ &+ \sum_{i=0}^{N-1} \frac{H_\ell(x_{i+1}, z_{iF}, \boldsymbol{\theta})}{G(x_{i+1}, z_{iF}, \boldsymbol{\theta})} \cdot p(z_{iF} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)}) \end{aligned}$$

where,

$$H_\ell(x, y, \boldsymbol{\theta}) = \frac{\partial G(x, y, \boldsymbol{\theta})}{\partial \theta_\ell} = \frac{\partial G}{\partial f} \cdot \frac{\partial f}{\partial \theta_\ell} + \frac{\partial G}{\partial g} \cdot \frac{\partial g}{\partial \theta_\ell}$$

1. With respect to θ_1, θ_2

$$\frac{H_1}{G}(x, y, \boldsymbol{\theta}) = \frac{\partial f}{\partial \theta_1} \left[\frac{(x - y - f(y)h)}{g^2(y)} \right], \frac{H_2}{G}(x, y, \boldsymbol{\theta}) = \frac{\partial f}{\partial \theta_2} \left[\frac{(x - y - f(y)h)}{g^2(y)} \right]$$

2. With respect to θ_3

$$\frac{H_3}{G}(x, y, \boldsymbol{\theta}) = \frac{\partial g}{\partial \theta_3} \left[\frac{(x - y - f(y)h)^2}{hg^3(y)} - \frac{1}{g(y)} \right]$$

1.7 Summary of terms that need to be computed

1. $p(x_{i+1} \mid z_{i1}, \boldsymbol{\theta}^{(k)}) \cdot p(z_{i1} \mid x_i, \boldsymbol{\theta}^{(k)})$
2. $p(z_{ij} \mid x_i, \boldsymbol{\theta}^{(k)}) \cdot p(z_{i,j+1} \mid z_{ij}, \boldsymbol{\theta}^{(k)}) \cdot p(x_{i+1} \mid z_{i,j+1}, \boldsymbol{\theta}^{(k)})$
3. $p(x_{i+1} \mid z_{iF}, \boldsymbol{\theta}^{(k)}) \cdot p(z_{iF} \mid x_i, \boldsymbol{\theta}^{(k)})$
4. $p(x_{i+1} \mid x_i, \boldsymbol{\theta}^{(k)})$ - the complete DTQ computation
5. $\partial G / \partial f, \partial G / \partial g$ - computed in Dtheta function
6. $\partial f / \partial \theta_1, \partial f / \partial \theta_2, \partial g / \partial \theta_3$ - computed in Dtheta function