Expectation Maximization (EM) with Bridge Sampling

The governing equation for the problem in $\mathbb{R}^d$ is:

$$dX(t) = f(X(t))dt + \Gamma dW_t \tag{1}$$

with Γ equal to a constant diagonal matrix and W_t denoting Brownian motion in $\mathbb{R}^d$. Consider an additive model for $f(x)$:

$$f(x) = \sum_{k=1}^{M} \beta_k \phi_k(x) \tag{2}$$

Here $\{\phi_i\}$ is some family of functions we prescribe, e.g., tensor products of orthogonal polynomials. Each $\phi_i : \mathbb{R}^d \to \mathbb{R}$ should be fairly easy to compute.

We assume that $\Gamma = \operatorname{diag}\gamma$ and that there exists $\delta > 0$ such that $\gamma_i \geq \delta$ for all $i \in \{1, \ldots, d\}$. Under this condition, (1) should have a smooth density.

Suppose we have data in the form of a time series, $\mathbf{x}$, considered to be direct observations of $X(t)$ at discrete time points. For simplicity, let us assume the observations are collected at equispaced times, $j\Delta t$ for $0 \leq j \leq L$. Thus the observed data is $\mathbf{x} = x_0, x_1, \cdots, x_L$. Each $x_j \in \mathbb{R}^d$.

Our goal is to use the data to estimate the functional form of f and the constant vector γ.

To achieve this goal, we propose to use EM. Here we regard $\mathbf{x}$ as the incomplete data. The missing data $\mathbf{z}$ is thought of as data collected at a time scale $h \ll \Delta t$ that is fine enough such that the transition density of (1) is approximately Gaussian. That is, if we discretize (1) in time via Euler-Maruyama method, we obtain

$$\widetilde{X}_{n+1} = \widetilde{X}_n + f(\widetilde{X}_n; \beta)h + \gamma h^{1/2} Z_{n+1} \tag{3}$$

where Z_{n+1} is a standard normal, independent of X_n. Note that $\widetilde{X}_{n+1} | \widetilde{X}_n = v$ is multivariate Gaussian with mean vector $v + f(v)h$ and covariance matrix $h\Gamma^2$. Specifically, the density is

$$\left(\prod_{i=1}^{d} \frac{1}{\sqrt{2\pi h \gamma_i^2}} \right) \exp\left(-\frac{1}{2h} (x - v - h \sum_{k=1}^{M} \beta_k \phi_k(v))^T \Gamma^{-2} (x - v - h \sum_{\ell=1}^{M} \beta_\ell \phi_\ell(v)) \right).$$

As h decreases, this Gaussian will better approximate the transition density

$$X((n+1)h) | X(nh) = v,$$

where $X(t)$ refers to the solution of (1), not its time-discretization.

EM. The EM algorithm consists of two steps, computing the expectation of the log likelihood function (on the completed data) and then maximizing it with respect to the parameters $\boldsymbol{\theta} = (\beta, \gamma)$.

1. Start with an initial guess for the parameters, $\boldsymbol{\theta}^{(0)}$.

2. For the expectation (or E) step,
$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \mathbb{E}_{\mathbf{z}|\mathbf{x},\boldsymbol{\theta}^{(k)}}[\log p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta})] \tag{4}$$
Our plan is to evaluate this expectation via bridge sampling. That is, we will sample from diffusion bridges $\mathbf{z} \mid \mathbf{x}, \boldsymbol{\theta}^{(k)}$. Then $(\mathbf{x}, \mathbf{z})$ will be a combination of the original data together with sample paths.

3. For the maximization (or M) step, we start with the current iterate and a dummy variable $\boldsymbol{\theta}$ and define
$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) \tag{5}$$
It will turn out that we can maximize this quantity without numerical optimization. All we will need to do is solve a least-squares problem.

4. Iterate Step 2 and 3 until convergence.

Details. With a fixed parameter vector $\boldsymbol{\theta}^{(k)}$, the SDE (1) is specified completely, i.e., the drift and diffusion terms have no further unknowns. For this SDE, we assume a diffusion bridge sampler is available. We take F diffusion bridge steps to march from x_i to x_{i+1}; the time step will be $h = (\Delta t)/F$. We can think of this process as inserting $F-1$ *new* samples, $\{z_{i,j}\}_{j=1}^{F-1}$ between x_i and x_{i+1}.

Let $\mathbf{z}^{(r)}$ denote the r^{th} diffusion bridge sample path:
$$z^{(r)} \sim z \,|\, x, \beta^{(k)} \tag{6}$$
The observed and sampled data can be interleaved together to create a time series (completed data)
$$\mathbf{y}^{(r)} = \{y_j^{(r)}\}_{j=1}^{N}$$
of length $N = LF + 1$. Suppose we form R such time series. The expected log likelihood can then be approximated by

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= \mathbb{E}_{\mathbf{z}|\mathbf{x},\boldsymbol{\theta}^{(k)}}[\log p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta})] \\ &\approx \frac{1}{R} \sum_{r=1}^{R} \Bigg[\sum_{j=1}^{N} \left[\sum_{i=1}^{d} -\frac{1}{2} \log(2\pi h \gamma_i^2) \right] \\ &\qquad - \frac{1}{2h} (y_j^{(r)} - y_{j-1}^{(r)} - h \sum_{k=1}^{M} \beta_k \phi_k(y_{j-1}^{(r)}))^T \Gamma^{-2} (y_j^{(r)} - y_{j-1}^{(r)} - h \sum_{\ell=1}^{M} \beta_\ell \phi_\ell(y_{j-1}^{(r)})) \Bigg] \end{aligned}$$

To maximize Q over $\boldsymbol{\theta}$, we first assume $\Gamma = \operatorname{diag}\gamma$ is known and maximize over β. This is a least squares problem. The solution is given by forming the matrix

$$\mathcal{M}_{k,\ell} = \frac{1}{R}\sum_{r=1}^{R}\sum_{j=1}^{N} h\phi_k^T(y_{j-1}^{(r)})\Gamma^{-2}\phi_\ell^T(y_{j-1}^{(r)})$$

and the vector

$$\rho_k = \frac{1}{R}\sum_{r=1}^{R}\sum_{j=1}^{N} \phi_k^T(y_{j-1}^{(r)})\Gamma^{-2}(y_j^{(r)} - y_{j-1}^{(r)}).$$

We then solve the system

$$M\beta = \rho$$

for β. Now that we have β, we maximize Q over γ. The solution can be obtained in closed form:

$$\gamma_i^2 = \frac{1}{RNh}\sum_{r=1}^{R}\sum_{j=1}^{N}((y_j^{(r)} - y_{j-1}^{(r)} - h\sum_{\ell=1}^{M}\beta_\ell\phi_\ell(y_{j-1}^{(r)})) \cdot e_i)^2$$

where e_i is the i^{th} canonical basis vector in $\mathbb{R}^d$.

Remarks

1. In $d = 1$ dimension, when Γ is fixed and known, the procedure appears to work well even with $R = 10$ sample paths. This is for a problem where we generate data from a known model and then try to recover the β coefficients.

2. It is of interest to prove that the alternating β, Γ maximization approach increases the Q function. As long as the Q function increases from one iteration to the next, up to the sampling error (which should decay like $R^{-1/2}$), the EM algorithm should converge monotonically. That is, both the completed log likelihood and the log likelihood of the original data $\mathbf{x}$ should converge monotonically. The EM algorithm should yield a local maximizer.

3. If the alternating β, Γ maximization approach does not work, we could instead take a few gradient descent steps on the negative log likelihood. The gradients are simple to compute.