Data assimilation and inverse problems – Homework Set 9

One more particle filter, scaling of RWM with dimension and MCMC with "real" data

1. Consider the stochastic model

$$x_k = \frac{1}{2}x_{k-1} + 25\frac{x_{k-1}}{1+x_{k-1}^2} + 8\cos(1.2k) + v_k, \quad v_k \sim \mathcal{N}(0, 100), \text{ iid},$$

and the observations

$$y_k = \frac{1}{20}x_k^2 + \eta_k, \quad \eta_k \sim \mathcal{N}(0, 1), \text{ iid},$$

where $k = 1, 2, \ldots$. The initial state is $x_0 \sim \mathcal{N}(0, 100)$.

Build a standard particle filter with resampling for this model and use it to draw histograms of the posterior distribution $p(x_k|y_{1:k})$, $k = 1, \ldots, 100$ (compare to figure 1.2 of Chapter 1 of *Sequential Monte Carlo Methods in Practice*, Doucet, de Freitas, Gordon, (Editors), Springer, 2001).

2. Use a random walk Metropolis MCMC sampler to draw samples from an n-dimensional Gaussian $p(x) = \mathcal{N}(0, I_n)$ where I_n is the $n \times n$ identity matrix. Consider the cases $n = 10, 50, 100, 200, 500$. For fixed n, generate chains of length 10^5 using the optimal step-size ($\beta \propto n^{-1/2}$). Compute the average acceptance ratio (averaged over the chain) and the average integrated auto-correlation time (IACT) (averaged over the n-dimensions). Plot IACT as a function of n. Plot the acceptance ratio as a function of n. What do you conclude?

3. Consider the data in the file "data.txt". It contains the numbers of hare and lynx furs collected once per year, starting in 1927 and ending in 1937.

You can model hare and lynx populations using the Lotka-Volterra (LV) equations

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \alpha x - \beta xy, \quad \frac{\mathrm{d}y}{\mathrm{d}t} = -\gamma y + \delta xy,$$

where $\alpha, \beta, \gamma, \delta > 0$. You can identify x as a the number of hares and y as the number of lynx. Starting at $t = 0$ (equivalently in the year 1927), you can simulate the LV equations up to time $t = 10$ (equivalently 1937) and compare the hare and lynx populations of your model to the hare and lynx populations that somebody observed (the data). The hare and lynx populations of the model depend on the parameters $(\alpha, \beta, \gamma, \delta)$ and the initial conditions $(x(0), y(0))$. We want to find parameters that lead to lynx and hare populations similar to the observed ones and use MCMC and Bayesian modeling to do this.

To define the Bayesian inverse problem, we need a posterior distribution $p(\theta|d)$, where $\theta = (\alpha, \beta, \gamma, \delta, x(0), y(0))^T$ and d is a vector containing the data (lynx and hare furs at the given times). The posterior distribution is the product of a prior and a likelihood. We do not know much about the parameters, but we do know that they all must be positive. The prior is thus uniform over the cube $[0, 10]^6$. (The number 10 here is arbitrary, we hope that all parameters are in fact much smaller than 10). The likelihood is defined as follows. Let $d_i^{\mathcal{M}}$, $i = 1, 2, \ldots, 10$ be the solution of of the LV equation at time $t = i$, obtained by solving the LV equations numerically for $t = 10$ time units (you may need to think a little bit about how to solve the LV equations). Let d_i be a vector of size two whose elements are the number of

hare and lynx at time $t = i$. Assuming independent Gaussian errors with variance one, the likelihood is

$$p(d|\theta) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{10} ||d_i - d_i^{\mathcal{M}}||^2\right).$$

The posterior distribution is thus

$$p(\theta|d) \propto \begin{cases} \exp\left(-\frac{1}{2}\sum_{i=1}^{10} ||d_i - d_i^{\mathcal{M}}||^2\right) & \text{if } \theta \in [0,10]^6, \\ 0 \text{ otherwise.} \end{cases}$$

Build a random walk Metropolis MCMC sampler to sample this distribution. You can start the chain with $\theta = (0.5861, 0.2345, 0.7780, 0.1768, 2.5786, 3.8248)^T$. Compute integrated autocorrelation for all six variables and also the (chain averaged) acceptance ratio. You can tune the sampler via the covariance matrix of the proposal distribution.

Once you have determined that you have a good sampler, Draw a triangle plot of your chain and illustrate the posterior distribution as follows. Pick a parameter at random from your MCMC chain (e.g., by drawing a random integer and using the corresponding link in your chain). Use this parameter and simulate the LV equations for 10 years. Plot the result. Repeat this process (e.g., 100 times). The result is a plot that contains several trajectories of the LV equation. On top of these trajectories, plot the lynx and hare data.