DS-GA 3001.001 Special Topics in Data Science: Modeling Time Series
Homework 2

Due date: March 2, by midnight

Problem 1. LDS model, 10p
Consider a special case of LDS with $\mathbf{C} = \mathbf{I}$ and $\mathbf{R} = \sigma^2\mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix. Show that in the limit where there is no observation noise the best estimate for latent $\mathbf{z}_i$ is to simply use the observation $\mathbf{x}_i$: formally, in the limit when $\sigma^2 \to 0$ the posterior for $\mathbf{z}_i$ has mean $\mathbf{x}_i$ and vanishing variance.

Solution:
The prediction for the latent state $\mathbf{z}_i$ can be determined using the Kalman filter equations, with the particular setting of $\mathbf{C}$ considered here:

$$\begin{aligned}\mu_{i|i} &= \mu_{i|i-1} + \mathbf{K}_i\left(x_i - \mathbf{C}\mu_{i|i-1}\right) = \mu_{i|i-1} + \mathbf{K}_i\left(x_i - \mu_{i|i-1}\right)\\ \Sigma_{i|i} &= \Sigma_{i|i-1} - \mathbf{K}_i\mathbf{C}\Sigma_{i|i-1} = \Sigma_{i|i-1} - \mathbf{K}_i\Sigma_{i|i-1}\end{aligned}$$

where the Kalman gain is:

$$\begin{aligned}\mathbf{K}_i &= \Sigma_{i|i-1}\mathbf{C}^{\mathrm{t}}\left(\mathbf{C}\Sigma_{i|i-1}\mathbf{C}^{\mathrm{t}} + R\right)^{-1}\\ &= \Sigma_{i|i-1}\left(\Sigma_{i|i-1} + \sigma^2\mathbf{I}\right)^{-1}\end{aligned}$$

In the limit $\sigma^2 \to 0$, the gain $\mathbf{K}_i \to \mathbf{I}$, which translates into limits for the estimates

$$\lim_{\sigma^2\to 0}\mu_{i|i} = \mu_{i|i-1} + \mathbf{I}\left(\mathbf{x}_i - \mu_{i|i-1}\right) = \mathbf{x}_i \quad (1)$$
$$\lim_{\sigma^2\to 0}\Sigma_{i|i} = \Sigma_{i|i-1} - \mathbf{I}\Sigma_{i|i-1} = \mathbf{0} \quad (2)$$

Hence, in the limit of very low observation noise it is best to listen to the observations and completely ignore temporal dependencies.

Problem 2. LDS prediction, 20p
Given the standard parametrization of the LDS model, and the Kalman filtering estimates $\mu_{i|i}$ and $\Sigma_{i|i}$, obtained for a dataset $\mathbf{x}_{1:t}$ write down the expressions for predicting the following 2 observations in the sequence $\mathbf{x}_{t+1}$ and $\mathbf{x}_{t+2}$.

Solution:
Starting from the posterior marginal for $\mathbf{z}_t$, which is multivariate gaussian with mean $\mu_{t|t}$ and covariance $\Sigma_{t|t}$, the marginal predictive distribution for $\mathbf{z}_{t+1}$ and $\mathbf{z}_{t+2}$ is multivariate normal with parameters (using the properties of a linear gaussian model):

$$\mu_{t+1|t} = \mathbf{A}\mu_{t|t} \quad (3)$$
$$\Sigma_{t+1|t} = \mathbf{A}\Sigma_{t|t}\mathbf{A}^{\mathrm{t}} + \mathbf{Q} \quad (4)$$
$$\mu_{t+2|t} = \mathbf{A}\mu_{t+1|t} = \mathbf{A}^2\mu_{t|t} \quad (5)$$
$$\Sigma_{t+2|t} = \mathbf{A}\Sigma_{t+1|t}\mathbf{A}^{\mathrm{t}} + \mathbf{Q} = \mathbf{A}(\mathbf{A}\Sigma_{t|t}\mathbf{A}^{\mathrm{t}} + \mathbf{Q})\mathbf{A}^{\mathrm{t}} + \mathbf{Q} \quad (6)$$
$$= \mathbf{A}^2\Sigma_{t|t}\mathbf{A}^{2\mathrm{t}} + \mathbf{A}\mathbf{Q}\mathbf{A}^{\mathrm{t}} + \mathbf{Q} \quad (7)$$

Marginalizing out the unknown $\mathbf{z}_{t+1}$ yields the prediction for $\mathbf{x}_{t+1}$:

$$\mu_{x,t+1} = \mathbf{C}\mu_{t+1|t} = \mathbf{C}\mathbf{A}\mu_{t|t} \quad (8)$$
$$\Sigma_{x,t+1} = \mathbf{C}(\mathbf{A}\Sigma_{t|t}\mathbf{A}^{\mathrm{t}} + \mathbf{Q})\mathbf{C}^{\mathrm{t}} + \mathbf{R} = \mathbf{C}\mathbf{A}\Sigma_{t|t}(\mathbf{C}\mathbf{A})^{\mathrm{t}} + \mathbf{C}\mathbf{Q}\mathbf{C}^{\mathrm{t}} + \mathbf{R} \quad (9)$$

Similarly, marginalizing out $\mathbf{z}_{t+2}$ we get the prediction for $\mathbf{x}_{t+2}$:

$$\mu_{x,t+2} = \mathbf{C}\mu_{t+1|t} = \mathbf{C}\mathbf{A}^2\mu_{t|t} \quad (10)$$
$$\Sigma_{x,t+2} = \mathbf{C}\Sigma_{t+2|t}\mathbf{C}^{\mathrm{t}} + \mathbf{R} \quad (11)$$

Problem 3. LDS inference coding, 20p
Given the model parameters:
$A = \begin{bmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{bmatrix}$, $C = \begin{bmatrix} 1.1 & 0.2 \\ 0.1 & 0.95 \end{bmatrix}$, $Q = \sigma_Q^2 \mathbf{I}$, $R = \sigma_R^2 \mathbf{I}$ with default parameter values $\sigma_q^2 = 0.4$, $\sigma_r^2 = 0.001$ and initial condition parameters $\mu_0 = [00]^{\mathrm{t}}$, $\Sigma_0 = \mathbf{I}$. Use the lab4 code as a starting point for the following:

- draw a sequence of length $t = 10$ observations from this model.
- using this artificial data and the true model parameters, run kalman filtering and smoothing; display the data, observations, and corresponding marginal posteriors (filtering, and smoothing) using the conventions in Bishop figure 13.22: true latent trajectory in blue, observations in green, red cross and ellipse for mean and cov of Kalman filter, same in yellow for the kalman smoother.
- Manipulate parameters $\sigma_{q,r}^2$ such that a) the latent dynamics are close to deterministic and b) the observation noise is very small. Re-plot inference figure for each parameter configuration and comment on results.
- What happens to the forward vs. backward estimates when the latent noise term is large or very small?

Solution:
The general trends one should see when paying with the parameters: in the limit of very small noise the estimates should be precise, and determined by the current observation. With a lot of observation noise the the dependencies across time become important for prediction. This is especially true when the latent space noise is small and the trajectory is close to deterministic. When observation noise is large, and there is noise in the latent space such that it helps integrating inf over a broader window: the early kalman filtering estimates will be more uncertain, as we haven't seen much of the series, and this may result in a more pronounced reduction in uncertainty after the backward pass. This last effect will likely be small though.

Problem 4. LDS dimensionality reduction coding, 10p
Given the model parameters:
$A = 0.25 \begin{bmatrix} 3 & 0.5 \\ 0.8 & 1.7 \end{bmatrix}$, $C = \begin{bmatrix} 1.1 & 0.2 \\ 0.5 & 0.95 \\ 1 & 1 \end{bmatrix}$, $Q = \mathbf{I}_2$, $R = 0.001\mathbf{I}_3$ with initial condition parameters $\mu_0 = [00]^{\mathrm{t}}$, $\Sigma_0 = \mathbf{I}$. Visualize $t = 100$ data points sampled from the model then for the first 10 elements in the sequence show the predicted distributions for $\mathbf{x}_{i+1}|\mathbf{x}_{1:i}$.
The point here is just to think about high dimensional data with low dimensional latent structure might look like.

Problem 5. EM, 20p
For the same LDS model used in Problem 3, draw t samples from model build your own EM implementation (can rely on code from lab4) to estimate the model parameters. Do your parameter estimates depend strongly on the initial value of the parameters? Comment why (not). How do the parameter estimates improve with more data ($t = 100$ vs. $t = 500$)? For a fixed amount of data $t = 100$, how accurate are the parameter reconstructions as you vary the observation noise from $\sigma_r^2 = 0.001$ to $\sigma_r^2 = 0.1$?
Solution:
With more observation noise, it may be harder to pinpoint the parameters of the latent dynamics - requiring more data for a precise reconstruction. Depending on the exact EM implementation there may be additional caveats. Namely, due to the parameter degeneracy (as discussed during the lecture) it is possible to converge to parameter values that are different from those used for generating the data, but which are still completely equivalent. For the same reason, the solution might depend on initial conditions. How do we know we recovered the right thing: either rescale and rotate the axes back into the canonical form with $\mathbf{Q}$ identity and do the comparison there, or for the application minded folk: one could simply test if the recovered parameters are just as good as the ground truth ones at predicting future observations.

Problem 6. Particle filtering, 20p
For the same LDS model used in Problem 3, implement a full step of the particle filtering algorithm: use the closed form solution for $\mathrm{P}\left(z_i|\mathbf{x}_{1:i-1}\right)$ to draw an ensemble of N particles, combine with observation $\mathbf{x}_i$ and build a sampling-based approximation for the predicting distribution $\mathrm{P}\left(z_{i+1}|\mathbf{x}_{1:i}\right)$, compare the analytical posterior mean (Kalman filtering) to the sampling based estimates, for $N = 10, 100, 1000, 10000$.

Solution:
Start with a set of initial particles from the prior $\mathrm{P}\left(\mathbf{z}_1\right) = \mathcal{N}(\mathbf{A}\mu_0, \mathbf{A}\Sigma_0\mathbf{A}^{\mathrm{t}} + \mathbf{Q})$. The recursion then proceeds as follows: given N particles $\mathbf{z}_i^{(n)} \sim \mathrm{P}\left(\mathbf{z}_i|\mathbf{x}_{1:i-1}\right)$, we can incorporate the observation $\mathbf{x_i}$ by computing the importance weights:

$$w_n = \frac{\exp\left(-0.5\left(\mathbf{x}_i - \mathbf{C}\mathbf{z}_i^{(n)}\right)^{\mathrm{t}}\mathbf{R}^{-1}\left(\mathbf{x}_i - \mathbf{C}\mathbf{z}_i^{(n)}\right)\right)}{\sum_m \exp\left(-0.5\left(\mathbf{x}_i - \mathbf{C}\mathbf{z}_i^{(m)}\right)^{\mathrm{t}}\mathbf{R}^{-1}\left(\mathbf{x}_i - \mathbf{C}\mathbf{z}_i^{(m)}\right)\right)} \tag{12}$$

These weights combined with the transition probabilities $\mathrm{P}\left(\mathbf{z}_{i+1}|\mathbf{z}_i\right) = \mathcal{N}(\mathbf{z}_{i+1}; A\mathbf{z}_i, \mathbf{Q})$ define a mixture model that we need to sample from for building a new generation of particles. This is done in 2 steps: first we draw the class labels $c = 1, 2, \ldots, N$ according to probability $\mathbf{w}$ then for each of the samples $c = n$ we draw a possible transition into $\mathbf{z}_{i+1}$ from the corresponding multivariate gaussian $\mathcal{N}(\mathbf{z}_{i+1}; A\mathbf{z}_i^{(n)}, \mathbf{Q})$.

In general, the number of samples needed for a good approximation will depend on the dimensionality of the latent space and the sharpness of the posterior (which in turn depends on the magnitude of the noise sources, as specified by $\mathbf{R}$ and $\mathbf{Q}$).