CS 287, Fall 2012 Problem Set #3 Multivariate Gaussians, Kalman Filtering, Maximum Likelihood, EM

Deliverable: Reasonable number of pages write-up in pdf format. Due date/time: Monday November 12th, 23:59pm. Email to pabbeel@cs.berkeley.edu.

Please refer to the class webpage for the homework policy.

Various starter files are provided on the course website.

When making your write-up, make sure to answer all questions, and to include and discuss plots to demonstrate that you solved the problem.

---

1. Maximum Likelihood

The Poisson distribution is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event. Assume we obtain m i.i.d. samples $x^{(1)}, \ldots, x^{(m)}$ distributed according to the Poisson distribution $P(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ for $k = 0, 1, 2, 3, \ldots$. What is the maximum likelihood estimate of λ as a function of $x^{(1)}, \ldots, x^{(m)}$?

2. Linearity of Expectation, Positive Semi-definiteness

A matrix $A \in \mathbb{R}^{nxn}$ is positive semi-definite (often denoted by $A \succeq 0$) if and only if:

$$A_{ij} = A_{ji}$$
$$\forall z \in \mathbb{R}^n : z^\top A z \geq 0$$

Prove that covariance matrices, i.e., matrices of the form $\Sigma = \mathrm{E}[(X - EX)(X - EX)^\top]$ are guaranteed to be positive semi-definite.

3. Kalman Filtering, Smoothing, EM

(a) **Implementation of KF, Smoothing, EM.** In this question you will implement a Kalman Filter, a Kalman Smoother, and the EM algorithm to estimate the covariance matrices. Look at p3_a_starter.m for more detailed instructions.

(b) **Application to Species Population Size Estimation from Observations of Total Population Size.** Consider three species U, V, W that grow independently of each other, exponentially with growth rates: U grows 2% per hour, V grows 6% per hour, and C grows 11% per hour. The goal is to estimate the initial size of each population based on the measurements of total population.

Let $x_U(t)$ denote the population size of species U after t hours, for $t = 0, 1, \ldots$, and similarly for $x_V(t)$ and $x_W(t)$, so that

$$x_U(t+1) = 1.02x_U(t), \quad x_V(t+1) = 1.06x_V(t), \quad x_W(t+1) = 1.11x_W(t).$$

The total population measurements are $y(t) = x_U(t) + x_V(t) + x_W(t) + v(t)$, where $v(t)$ are IID, $\mathcal{N}(0, 0.36)$. (Thus the total population is measured with a standard deviation of 0.6).

The prior information is that $x_U(0), x_V(0), x_W(0)$ (which are what we want to estimate) are IID $\mathcal{N}(6, 2)$. (Obviously the Gaussian model is not completely accurate since it allows the initial populations to be negative with some small probability, but we'll ignore that.)

How long will it be (in hours) before we can estimate $x_U(0)$ with a variance less than 0.01? How long for $x_V(0)$? How long for $x_W(0)$?

(c) **Correlated Noise.** In many practical situations the noise is not independent. Consider the following stochastic system, for which the noise is not independent:

$$\begin{aligned}
&x_0 \sim N(\mu_0, \Sigma_0) \\
&x_{t+1} = Ax_t + w_t \\
&w_t = 0.3w_{t-1} + 0.2w_{t-2} + p_{t-1} \\
&p_t \sim \mathcal{N}(0, \Sigma_{pp}) \\
&y_t = Cx_t + v_t \\
&v_t = 0.8v_{t-1} + q_{t-1} \\
&q_t \sim \mathcal{N}(0, \Sigma_{qq}) \\
&p_{-1} = q_{-1} = v_{-1} = w_{-1} = w_{-2} = 0
\end{aligned}$$

Describe how, by choosing the appropriate state representation, the above setup can be molded into a standard Kalman filtering setup. In particular, describe the state, the dynamics model, and the measurement model such that the problem is transformed into the standard Kalman filtering setup with uncorrelated noise.

(d) **(Optional / Extra Credit) EM Equations for A, B, C, d.** Derive the EM update equations for A, B, D, d for the usual linear Gaussian system, which is of the form:

$$\begin{aligned}
x_{t+1} &= Ax_t + Bu_t + w_t \quad w_t \sim \mathcal{N}(0, \Sigma_w) \\
y_t &= Cx_t + d + v_t \quad v_t \sim \mathcal{N}(0, \Sigma_v)
\end{aligned}$$

where all w_t and v_t are independent. Show your work. Generate some data from a linear Gaussian system and report on the ability to learn A, B, C, d using EM.

4. Sensor Selection

We consider the following linear system:

$$\begin{aligned}
x_{t+1} &= Ax_t + w_t \\
z_t &= C_t x_t + v_t
\end{aligned}$$

where $A \in \mathbb{R}^{n \times n}$ is constant, but C_t can vary with time. The noise contributions are independent, and

$$x_0 \sim \mathcal{N}(0, \Sigma_0), \quad w_t \sim \mathcal{N}(0, \Sigma_w) \quad v_t \sim \mathcal{N}(0, \Sigma_v).$$

Here is the twist: the measurement matrix C_t at each time comes from the set $\mathcal{S} = \{S_1, \ldots, S_K\}$. In other words, at each time t, we have $C_t = S_{i_t}$. The sequence $i_0, i_1, i_2, \ldots$ specifies which of the K possible measurements is taken at time t. For example the sequence $2, 2, 2, \ldots$ means that $C_t = S_2$ for all t. The sequence $1, 2, \ldots, K, 1, 2, \ldots, K, \ldots$ is called round-robin: we cycle through the possible measurements, in order, over and over again.

Here is the interesting part: *you* get to choose the measurement sequence $i_0, i_1, i_2, \ldots$.

You will work with the following specific system:

$$A = \begin{bmatrix} 1.0007 & -0.0010 & 0.0160 \\ 0.0112 & 0.9944 & 0.0077 \\ -0.0003 & -0.0062 & 1.0009 \end{bmatrix}, \quad \Sigma_w = \mathrm{diag}(\begin{bmatrix} 0.1 & 0.1 & 1.0 \end{bmatrix}), \quad \Sigma_v = 0.1^2, \quad \Sigma_0 = I$$

and $K = 3$ with

$$S_1 = \begin{bmatrix} 1.0 & 1.0 & 0 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 0.2 & 0.2 & 0 \end{bmatrix}, \quad S_3 = \begin{bmatrix} 0 & 0 & 0.1 \end{bmatrix}.$$

(a) **Using One Sensor**. Plot $trace(\Sigma_{t|0:t})$ versus t for the three special cases when $C_t = S_1$ for all t, $C_t = S_2$ for all t, and $C_t = S_3$ for all t.

(b) **Round-robin.** Plot $trace(\Sigma_{t|0:t})$ versus t using the round-robin sensor sequence $1, 2, 3, 1, 2, 3, \ldots$.

(c) **Greedy Sensor Selection.** Plot $trace(\Sigma_{t|0:t})$ versus t using greedy sensor selection. In greedy sensor selection at time t the choice of $i_0, i_1, \ldots, i_{t-1}$ has already been made and it has determined $\Sigma_{t|0:t-1}$. Then $\Sigma_{t|0:t}$ depends on i_t only, i.e., which of $S_1, \ldots, S_K$ is chosen as C_t. Among these K choices you pick the one that minimizes $trace(\Sigma_{t|0:t})$.

(d) **(Optional / Extra Credit) h-Lookahead Sensor Selection.** In h-lookahead sensor selection, at time t the chioce of $i_0, i_1, \ldots, i_{t-1}$ has already been made and it has determined $\Sigma_{t|0:t-1}$. Then $\Sigma_{t-1+h|0:t-1+h}$ depends on $i_t, i_{t+1}, \ldots, i_{t-1+h}$ only. A search over all possible sensor selection sequences for $t, t+1, \ldots, t-1+h$ is run, and let's say $i_t^*, i_{t+1}^*, \ldots, i_{t-1+h}^*$ is the one that minimizes $trace(\Sigma_{t-1+h|0:t-1+h})$, then we choose $i_t = i_t^*$, and then we increment t by one and repeat the process. Note that greedy sensor selection corresponds to h=1. Plot $trace(\Sigma_{t|0:t})$ for $h = 1, 2, \ldots$ (however far your algorithm is able to look ahead in a reasonable amount of time).

In all parts show the plots over the interval $t = 0, \ldots, 50$ and report the steady-state ($t \to \infty$) values (if such a limit exists).

Note none of these require knowledge of the measurements $z_0, z_1, \ldots$.