Assignment 2, due September 28

Corrections (check the class message board): (none yet.)

1. Suppose that $Z_n \sim \mathcal{N}(0,1)$ and

$$X_{n+1} = \frac{1}{2} X_n - \frac{5}{16} X_{n-1} + Z_n \ .$$

Show that this Gaussian process is stable. Find the limiting joint distribution of X_n and X_{n-1} in the limit $n \to \infty$.

2. (*Covariance of the covariance*) Suppose $X \sim \mathcal{N}(0,C)$ is a d component normal with mean zero and covariance C. Suppose we have independent samples $X_k \in \mathbb{R}^d$ and wish to estimate C. We discuss estimating $C = E[XX^t]$ using

$$\widehat{C} = \frac{1}{N} \sum_{k=1}^{N} X_k X_k^t \ . \tag{1}$$

Of course, $\widehat{C}$ is a random variable with its own mean and covariance. The CLT implies that for large N, $\widehat{C}$ is approximately a multivariate normal.

(a) Show that the mean is $E[\widehat{C}] = C$.

(b) (*Wick's formula for degree 4*) Suppose $X \sim X \sim \mathcal{N}(0,C)$ with components $X_1, \ldots, X_d$ (note the change of notation). Show that

$$E[X_i X_j X_k X_l] = C_{ij} C_{kl} + C_{ik} C_{jl} + C_{il} C_{jk} \ . \tag{2}$$

We saw the one dimensional version of this in the previous assignment. The method there applies here too. Another approach is to write a formula

$$x_l e^{x^t H x/2} = - \sum_{m=1}^{d} A_{lm} \partial_{x_m} e^{-x^t H x/2} \ .$$

The matrix A is related to $H = C^{-1}$ in some way. Now integrate by parts in the integral

$$\begin{aligned} E[X_i X_j X_k X_l] &= \frac{1}{Z} \int_{\mathbb{R}^d} x_i x_j x_k x_l e^{-x^t H x/2} \, dx \\ &= \frac{1}{Z} \sum_{m=1}^{d} A_{lm} \int_{\mathbb{R}^d} \partial_{x_m} \left(x_i x_j x_k \right) e^{-x^t H x/2} \, dx \ . \end{aligned}$$

(c) Find a formula for

$$D_{ij,kl} = \text{cov}(\widehat{C}_{ij}, \widehat{C}_{kl}) \ .$$

3. (*Kalman filter*) In the terminology of linear Gaussian processes, suppose

$$\begin{aligned} X_{n+1} &= AX_n + U_n & &\text{(state dynamics)} \\ Y_n &= BX_n + V_n & &\text{(observation)} \\ R &= \text{cov}(U_n) & &\text{i.i.d., Gaussian} \\ S &= \text{cov}(V_n) & &\text{i.i.d., Gaussian} \end{aligned}$$

Suppose $X_0 = 0$. Let C_n be the conditional variance of X_n, conditioned on knowing $Y_{1:n}$. Let $\widehat{X}_n$ be the linear function of the observations $Y_{1:n}$ so that the prediction residual $X_n - \widehat{X}_n$ is independent of $Y_{1:n}$.

(a) Show that $X_n = \widehat{X}_n + W_n$, where W_n is multivariate normal with mean 0 and covariance C_n, and independent of $Y_{1:n}$.

(b) Find a matrix equation to solve for *gain* matrix K_{n+1} so that

$$\widehat{X}_{n+1} = A\widehat{X}_n + K_{n+1}\left(Y_{n+1} - \widehat{Y}_{n+1}\right) \ .$$

where $\widehat{Y}_{n+1}$ is the predicted observation based on $Y_{1:n}$, given by $\widehat{Y}_{n+1} = BA\widehat{X}_n$. The equation will involve the five relevant matrices: A, B, R, S, and C_n. Find a formula for C_{n+1}. Hint, once the prediction residual at time $n+1$ is uncorrelated to $Y_{n+1} - \widehat{Y}_{n+1}$, you have it, why?

(c) Assume $C_n \to C$ as $n \to \infty$ and $K_n \to K$ as $n \to \infty$. Find nonlinear equations for C and K. These are called *matrix Ricatti equations.* Control theory is full of them.

4. The US government commonly releases estimates of economic data on one month and the releases a revised estimate the following month. For example, in the government estimated that the "non-farm employment change" was 215,000. In August, they revised that estimate to 245,000[1]. Let us suppose the true number at period n is ξ_n, and that this true number is a Gaussian process of the form

$$\xi_{n+1} = \overline{\xi} + a(\xi_n - \overline{\xi}) + \epsilon_n^{(\xi)} \ ,$$

where $a \in (0,1)$ determines the rate at which ξ_n naturally returns to its long term mean $\overline{\xi}$, and the sequence $\epsilon_n^{(\xi)}$ consists of i.i.d. $\mathcal{N}(0, \sigma_\xi^2)$ random variables. Suppose the estimate at time n is a noisy observation of ξ_n taking the form $\eta_n = \xi_n + \epsilon_n^{(\eta)}$, where the $\epsilon_n^{(\eta)}$ are i.i.d. $\mathcal{N}(0, \sigma_\eta^2)$. Suppose at time $n+1$ we get a second observation of ξ_n that takes the form

[1]Source: http://www.bls.gov/news.release/empsit.nr0.htm .

$\zeta_n = \xi_n + \epsilon_n^{(\zeta)}$ where the sequence $\epsilon_n^{(\zeta)}$ consists of i.i.d. $\mathcal{N}(0, \sigma_\zeta^2)$ random variables. All observation errors ϵ are independent of all other observation errors. The data available at time n consists of the sequence $\eta_{[1:n]}$ and the sequence $\zeta_{[1:n-1]}$. The value of ζ_n becomes known at time $n+1$.

This exercise is to put this problem into the general framework of Kalman filtering to produce the best estimate of ξ_n available at time n, and the best estimate available at time $n+1$. For this, you should define a two component vector X_n whose components are ξ_n and ξ_{n-1}. You should define a two component vector Y_n that consists of the two new numbers that become available at time n. Then identify the matrices A, B, R, and S in terms of the numbers in this problem. Show that the new observation η_{n+1} would lead to a revision of ξ_n even if there were no new direct observations ζ_n.

The point of this exercise is that many problems may be formulated in general matrix terms even if they come formulated in a different way. There are many special instances of the general problem.

5. (nothing to hand in) Download the code `GaussianProcess.R` and see what it does.