Assignment #5
Due: 11:59pm April 19, 2019

Homework 5: Graphical Models, MDPs

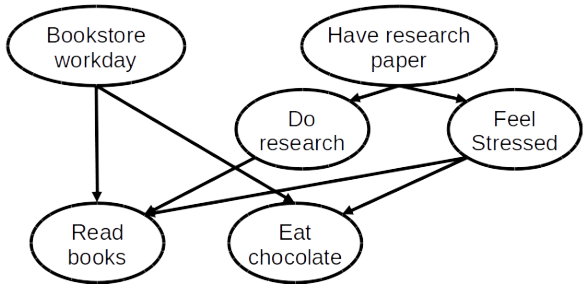
Introduction

There is a mathematical component and a programming component to this homework. Please submit your **tex, PDF, and Python files** to Canvas, and push all of your work to your GitHub repository. If a question requires you to make any plots, please include those in the writeup.

Bayesian Networks [7 pts]

Problem 1

In this problem we explore the conditional independence properties of a Bayesian Network. Consider the following Bayesian network representing a student's work. Each random variable is binary (true/false).



The random variables are:

- Have research paper: Does the student have a research paper?
- Do Research: Is the student doing research?
- Feel Stressed: Is the student feeling stressed?
- Bookstore Workday: Is the student working at the bookstore?
- Read Books: Is the student reading a book?
- Eat Chocolate: Is the student eating chocolate?

For the following questions, $A \perp B$ means that events A and B are independent and $A \perp B|C$ means that events A and B are independent conditioned on C. Use the concept of d-separation to answer the questions and show your work.

1. Is Bookstore Workday $\perp$ Have research paper? If NO, give intuition for why.
2. Is Bookstore Workday $\perp$ Have research paper | Read Books? If NO, give intuition for why.
3. Is Do Research $\perp$ Eat Chocolate? If NO, give intuition for why.
4. Is Do Research $\perp$ Eat Chocolate | Feel Stressed? If NO, give intuition for why.
5. Suppose the student has done some mindfulness exercises to avoid stress eating. Now, when they are stressed, they read (fun) books but don't eat chocolate. Draw the modified network.
6. For this modified network, is Do Research $\perp$ Eat Chocolate? If NO, give an intuition why. If YES, describe what observations (if any) would cause them to no longer be independent.

Solution

For the following, I used the picture from UC Berkeley's CS 188 posted on Piazza to classify paths as active or inactive.

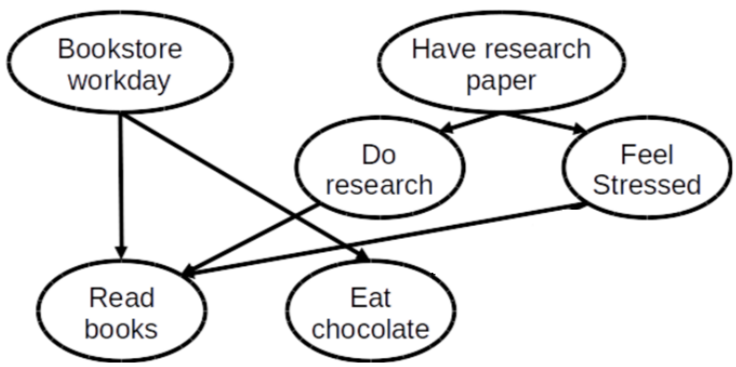
1. We have that $\boxed{\text{Bookstore Workday} \perp \text{Have research paper}}$ since there are no active paths between these nodes. Bookstore Workday points into Read books and Eat Chocolate, but both of these nodes are unobserved descendents of Have research paper. Hence, the path is inactive.
2. We have that $\boxed{\text{Bookstore Workday} \not\perp \text{Have research paper} \mid \text{Read Books}}$ since this creates an active path.

To gain intuition, consider that Read books is true and Bookstore Workday is false. Then, we know Read books must be true from Do research or Feel Stressed, which means Have research paper will be true. Hence, conditional on Read books, knowing something about Bookstore workday tells us something about Have research paper and vice versa.

3. We have that $\boxed{\text{Do Research} \not\perp \text{Eat Chocolate}}$ because there is an active path from Do Research to Have research paper to Feel stressed to Eat Chocolate. To gain intuition, consider that if Do research is true, then we know Have research paper is true, (or this becomes vary likely, depending on how we specify the network). Then, if Have research paper is true, we know that Feel stressed is also likely true. This in turn means that Eat chocolate is likely to be true. Hence, knowing information about Do Research gives us information about Eat chocolate, so these two are not independent.

4. We have that $\boxed{\text{Do Research} \perp \text{Eat Chocolate} \mid \text{Feel Stressed}}$ since conditioning on Feel Stressed kills the active path. To gain intuition[1], consider that the only inputs to Eat chocolate are Bookstore workday and Feel stressed. Do Research is independent of Bookstore workday, so it cannot influence Eat chocolate via this node. However, Do Research does have a common ancestor of Eat chocolate which is Have research paper. Hence, if Do Research is true, we expect Have research paper to be more likely to be true meaning Feel stressed is more likely to be true and similarly with Eat chocolate. Hence, Do Research tells us information about Eat chocolate. However, conditional on Feel stressed, learning about Do Research tells us nothing because we do not care if Feel stressed is more likely to be true or not: we know exactly what it is! Hence, conditional on Feel stressed, Do research and Eat chocolate are independent.

5. The modified network is depicted below. It is exactly the same, but now no edge exists between Feel stressed and Eat chocolate.



6. In this modified network, we have that $\boxed{\text{Do Research} \perp \text{Eat Chocolate}}$. However, if we observe $\boxed{\text{Read books}}$ these two nodes will become dependent since this would create an active path from Do Research to Read books to Bookstore workday to Eat chocolate. Specifically, if we observe that Read books is false and know that Do research is true, than it is likely that Bookstore workday is false meaning it is also likely that Eat chocolate is false. Hence, conditional on Read books, learning the value of Do research gives us information about Eat chocolate.

[1] oops, wasn't supposed to give intuition here...

Kalman Filters [7 pts]

Problem 2

In this problem, you will implement a one-dimensional Kalman filter. Assume the following dynamical system model:

$$\begin{aligned} z_{t+1} &= z_t + \epsilon_t \\ x_t &= z_t + \gamma_t \end{aligned}$$

where z are the hidden variables and x are the observed measurements. The random variables ϵ and γ are drawn from the following Normal distributions:

$$\begin{aligned} \epsilon_t &\sim N(\mu_\epsilon, \sigma_\epsilon) \\ \gamma_t &\sim N(\mu_\gamma, \sigma_\gamma) \end{aligned}$$

where $\mu_\epsilon = 0$, $\sigma_\epsilon = 0.05$, $\mu_\gamma = 0$ and $\sigma_\gamma = 1.0$

You are provided with the observed data x and the hidden data z in `kf-data.csv`, and the prior on the first hidden state is $p(z_0) = \mathrm{N}(\mu_p,\sigma_p)$ where $\mu_p = 5$ and $\sigma_p = 1$

1. The distribution $p(z_t|x_0...x_t)$ will be Gaussian $N(\mu_t, \sigma_t^2)$. Derive an iterative update for the mean μ_t and variance σ_t^2 given the mean and variance at the previous time step (μ_{t-1} and σ_{t-1}^2).
2. Implement this update and apply it to the observed data above (do not use the hidden data to find these updates). Provide a plot of μ_t over time as well as a $2\sigma_t$ interval around μ_t (that is $\mu_t \pm 2\sigma_t$). Does the Kalman filter "catch up" to the true hidden object?
3. Repeat the same process but change the observation at time $t = 10$ to $x_{t=10} = 10.2$ (an extreme outlier measurement). How does the Kalman filter respond to this outlier?
4. Comment on the misspecification of dynamical system model for these data. Based on the previous two parts, how does this misspecification affect the predictions?

Solution

1. We wish to find the mean and variance of $p(z_t|x_0 \dots x_t)$ given u_{t-1} and σ_{t-1}^2. By Bayes' rule, we have

$$p(z_t|x_0 \dots x_t) = \frac{p(x_t|x_0 \dots x_{t-1}, z_t)p(z_t|x_0 \dots x_{t-1})}{p(x_t|x_0 \dots x_{t-1})} \propto p(x_t|x_0 \dots x_{t-1}, z_t)p(z_t|x_0 \dots x_{t-1})$$

where I will ignore the denominator since it is a function of only the observed data. Now, using the fact that in our model x_t only depends on z_t, we can drop some of the above conditioning. We have that

$$p(z_t|x_0 \dots x_t) \propto \underbrace{p(x_t|z_t)}_{N(x_t; z_t+\mu_\gamma, \sigma_\gamma^2)} \; p(z_t|x_0 \dots x_{t-1})$$

Using LOTP and conditioning on all possible values of z_{t-1}, we can expand the second term as

$$p(z_t|x_0 \dots x_{t-1}) = \int_{-\infty}^{\infty} p(z_t|x_0 \dots x_{t-1}, z_{t-1})p(z_{t-1}|x_0 \dots x_{t-1})dz_{t-1}$$

Since z_t only depends on z_{t-1}, this becomes

$$= \int_{-\infty}^{\infty} \underbrace{p(z_t|z_{t-1})}_{N(z_t;z_{t-1}+\mu_\epsilon,\sigma_\epsilon^2)} \underbrace{p(z_{t-1}|x_0 \dots x_{t-1})}_{N(z_{t-1};\mu_{t-1},\sigma_{t-1}^2)} dz_{t-1}$$

Now, note that $N(z_t; z_{t-1} + \mu_\epsilon, \sigma_\epsilon^2) = N(-z_t; -z_{t-1} - \mu_\epsilon, \sigma_\epsilon^2) = N(z_{t-1}; z_t - \mu_\epsilon, \sigma_\epsilon^2)$. Hence, we have

$$p(z_t|x_0 \dots x_{t-1}) = \int_{-\infty}^{\infty} \underbrace{p(z_t|z_{t-1})}_{N(z_{t-1};z_t-\mu_\epsilon,\sigma_\epsilon^2)} \underbrace{p(z_{t-1}|x_0 \dots x_{t-1})}_{N(z_{t-1};\mu_{t-1},\sigma_{t-1}^2)} dz_{t-1}$$

At this point, it will help to evaluate the product of two general Gaussian PDF's evaluated at x. Specifically, for $N(x; a, \sigma_a^2)$ and $N(x; b, \sigma_b^2)$ Gaussians, the product of their PDFs is

$$\frac{1}{\sqrt{2\pi\sigma_a^2}} \exp\big(-\frac{(x-a)^2}{2\sigma_a^2}\big) \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\big(-\frac{(x-b)^2}{2\sigma_b^2}\big)$$

Now, focusing on the (negated) inside of the exponential terms and completing the square, we will get

$$\frac{(x-a)^2}{2\sigma_a^2} + \frac{(x-b)^2}{2\sigma_b^2} = \frac{\sigma_a^2+\sigma_b^2}{2\sigma_a^2\sigma_b^2}x^2 - \frac{a\sigma_b^2+b\sigma_a^2}{\sigma_a^2\sigma_b^2}x + \frac{a^2\sigma_b^2+b^2\sigma_a^2}{2\sigma_a^2\sigma_b^2}$$

$$= \underbrace{\frac{\sigma_a^2+\sigma_b^2}{2\sigma_a^2\sigma_b^2}\big(x - \frac{a\sigma_b^2+b\sigma_a^2}{\sigma_a^2+\sigma_b^2}\big)^2}_{\propto\text{Gaussian over x}} + \underbrace{\frac{(a-b)^2}{2(\sigma_a^2+\sigma_b^2)}}_{\propto\text{Gaussian over a}}$$

From the above, as indicated, we have two terms that are proportional to a Gaussian over x and a Gaussian over a, or equivalently b. In fact, we can actually conclude that these are both valid PDFs since they will need to normalize.[2] The Gaussian over x has mean and variance

$$\mu = \frac{a\sigma_b^2+b\sigma_a^2}{\sigma_a^2+\sigma_b^2}, \sigma^2 = \frac{\sigma_a^2\sigma_b^2}{\sigma_a^2+\sigma_a^2}$$

while the Gaussian over a has mean and variance

$$\mu = b, \sigma^2 = \sigma_a^2 + \sigma_b^2$$

Now, using this result to evaluate the integral in question above, we realize that the integrand can be rewritten as a Gaussian over z_{t-1} and a Gaussian over μ_{t-1}. Since we are integrating over all values of z_{t-1}, the first Gaussian will integrate to 1. The second Gaussian is a constant with respect to the integral, so it can be pulled out. Specifically, we will have

$$p(z_t|x_0 \dots x_{t-1}) = N(\mu_{t-1}; z_t-\mu_\epsilon, \sigma_\epsilon^2+\sigma_{t-1}^2) = N(-\mu_{t-1}; -z_t+\mu_\epsilon, \sigma_\epsilon^2+\sigma_{t-1}^2) = N(z_t; \mu_{t-1}+\mu_\epsilon, \sigma_\epsilon^2+\sigma_{t-1}^2)$$

[2] I'm not actually positive why this works out, but my results match those found in the Tina-Vision resource suggested by TF's and results in Bishop...so the normalization constants must work out!

Also, note that $p(x_t|z_t) = N(x_t; z_t + \mu_\gamma, \sigma_\gamma^2) = N(-x_t; -z_t - \mu_\gamma, \sigma_\gamma^2) = N(z_t; x_t - \mu_\gamma, \sigma_\gamma^2)$. Now, we realize that

$$p(z_t|x_0 \dots x_t) \propto \underbrace{p(x_t|z_t)}_{N(z_t;x_t-\mu_\gamma,\sigma_\gamma^2)} \underbrace{p(z_t|x_0 \dots x_{t-1})}_{N(z_t;\mu_{t-1}+\mu_\epsilon,\sigma_\epsilon^2+\sigma_{t-1}^2)}$$

Again, using the general formula for the product of two Gaussians evaluated at the same value, we realize that this is equivalent to a Gaussian over z_t multiplied by a Gaussian over $\mu_{t-1} + \mu_\epsilon$. Note that we are just working with proportionality here, so the second Gaussian can be ignored since it is constant wrt z_t and we are only concerned with the distribution of z_t. Hence, from my equations above, we conclude that $p(z_t|x_0 \dots x_t)$ is a Gaussian over z_t with the following mean and variance

$$\boxed{\mu_t = \frac{(x_t - \mu_\gamma)(\sigma_\epsilon^2 + \sigma_{t-1}^2) + (\mu_{t-1} + \mu_\epsilon)\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\epsilon^2 + \sigma_{t-1}^2}, \sigma_t^2 = \frac{\sigma_\gamma^2(\sigma_\epsilon^2 + \sigma_{t-1}^2)}{\sigma_\gamma^2 + \sigma_\epsilon^2 + \sigma_{t-1}^2}}$$

Plugging in the values we were given in this problem, these becomes

$$\boxed{\mu_t = \frac{x_t(0.0025 + \sigma_{t-1}^2) + \mu_{t-1}}{1.0025 + \sigma_{t-1}^2}, \sigma_t^2 = \frac{0.0025 + \sigma_{t-1}^2}{1.0025 + \sigma_{t-1}^2}}$$

So, we have found the desired iterative update. Now, we need to solve for our starting conditions. By Bayes' rule, we know

$$p(z_0|x_0) = \frac{p(x_0|z_0)p(z_0)}{p(x_0)} \propto p(x_0|z_0)p(z_0)$$

Again, we ignore the denominator since it is a function of the observed data. We are given that $p(z_0) = N(\mu_p, \sigma_p^2)$. Additionally, we know that $p(x_0|z_0) = N(x_0; z_0 + \mu_\gamma, \sigma_\gamma^2) = N(-x_0; -z_0 - \mu_\gamma, \sigma_\gamma^2) = N(z_0; x_0 - \mu_\gamma, \sigma_\gamma^2)$. Hence, we have that

$$p(z_0|x_0) \propto \underbrace{p(x_0|z_0)}_{N(z_0;x_0-\mu_\gamma,\sigma_\gamma^2)} \underbrace{p(z_0)}_{N(z_0;\mu_p,\sigma_p^2)}$$

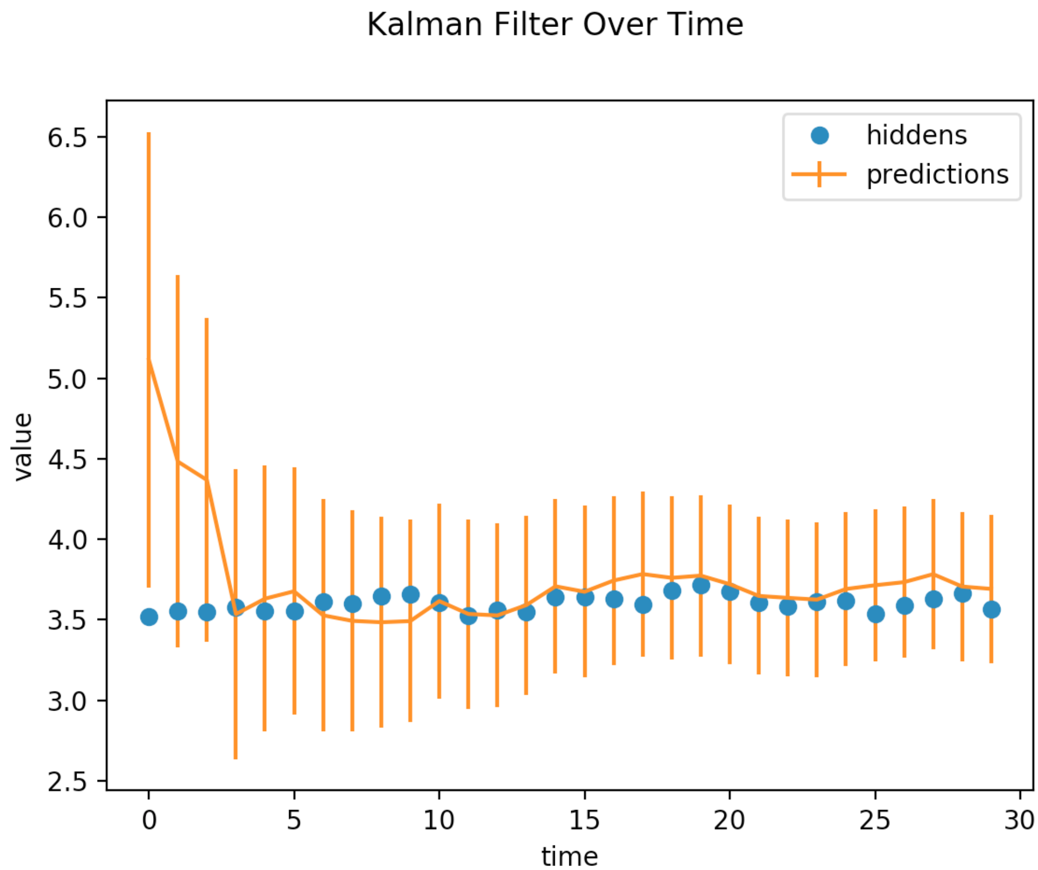
and from the formula I derived above, we conclude that this is equivalent to the product of a Gaussian over z_0 and a Gaussian over μ_p. We can ignore the second Gaussian since it is a constant wrt z_0 and we are working with a proportionality. Furthermore, since the r.h.s. is a valid PDF over z_0, we conclude the l.h.s. must be as well and the proportionality constant will work out (this is the same logic we used above). So, we have $p(z_0)$ is a Gaussian with mean and variance

$$\boxed{\mu_0 = \frac{(x_0 - \mu_\gamma)\sigma_p^2 + \mu_p\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_p^2}, \sigma_0^2 = \frac{\sigma_\gamma^2\sigma_p^2}{\sigma_\gamma^2 + \sigma_p^2}}$$

Plugging in the actual values for this problem, these become

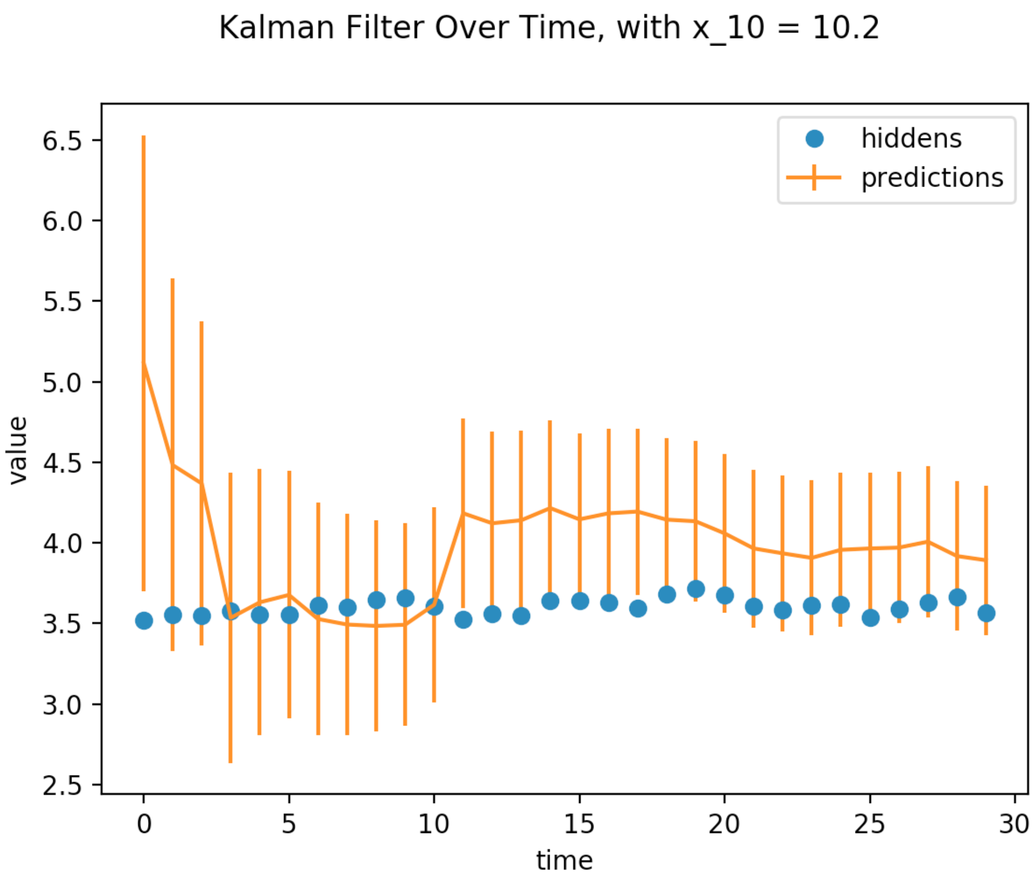
$$\boxed{\mu_0 = \frac{x_0 + 5}{2}, \sigma_0^2 = \frac{1}{2}}$$

2. Below is a graph of my implementation of the Kalman Filter on the observed data. The blue points are the hidden data and the orange line is the mean of $p(z_t|x_0 \dots x_t)$, μ_t, with $\pm 2\sigma_t$ errorbars.



As can be seen, the Kalman filter does indeed "catch up" to the true hiddens. By the fourth observation the Kalman Filter mean is practically equal to the hidden, and it stays very close to the hidden values as time progresses.

3. Below is the same graph but now $x_{t=10} = 10.2$, far outside of two standard deviations of the Kalman Filter mean.



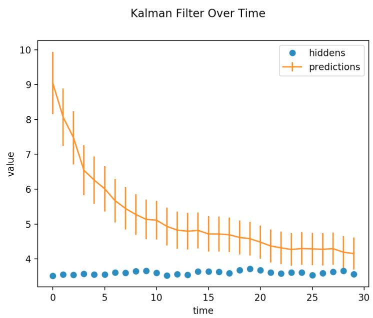
As can be seen, the Kalman filter responds poorly to this outlier and consistently overestimates the hiddens afterwards, even after nearly 20 additional observations. The Kalman filter mean does move closer to the hiddens by $t = 29$, but not by very much.

4. There are several ways this model can be misspecified. Specifically, our parameters are $\mu_p, \sigma_p, \mu_\epsilon, \sigma_\epsilon, \mu_\gamma$, and σ_γ. From looking at the above graphs, its clear that we have misspecified μ_p or σ_p since z_0 is well beyond two standard deviations below the mean, and we are using a Gaussian model. This means if our model is correct, we'd expect this to happen less than 5% of the time, roughly. So, it's likely we misspecified μ_p or σ_p or perhaps both. However, the mean of the Kalman Filter is able to catch up to the hiddens quickly. So with regard to the prior, it seems our model is not very sensitive to misspecification.

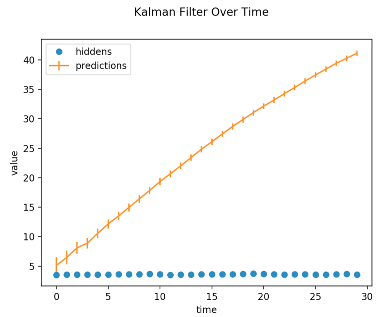
   In 2.3, we consider the scenario where we observe $x_{t=10} = 10.2$, which is far from what we would expect it to be since our model assumes $x_t \sim N(z_t + \mu_\gamma, \sigma_\gamma^2) = N(z_t, 1)$. In this scenario we have left z_t unchanged, and it is around 3.5. Hence, we realize that we have clearly misspecified μ_γ, σ_γ, or both of these parameters because the probability of $x_{t=10} = 10.2$ when $x_{t=10} \sim N(3.53, 1)$ is extremely small. Under this misspecification, our model does poorly to catch up to the true hiddens. Over the remaining 19 data point, the mean of $p(z_t|x_0 \ldots x_t)$ remains relatively far away from the hidden values.

   Note that this scenario of $x_{t=10} = 10.2$ could also have occurred if z_{10} had been a large value, which would have implied that μ_ϵ or σ_ϵ or both were misspecified since $z_{10} \sim N(z_9 + \mu_\epsilon, \sigma_\epsilon^2) = N(3.61, 1)$. Thus, we can also conclude that misspecification of μ_ϵ or σ_ϵ or both would lead to the mean of our Kalman Filter moving far away from the hidden values and failing to catch up for some time. All of this is to say that our model is sensitive to a misspecification of $\mu_\epsilon, \sigma_\epsilon, \mu_\gamma$, or σ_γ, or a combination of misspecifying these parameters. Our predictions become poor under such misspecification.

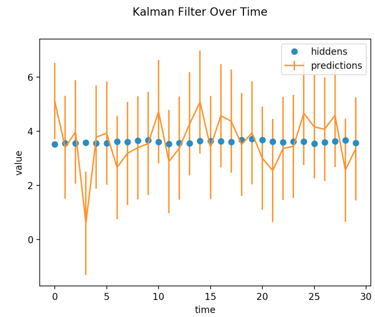
To explore this more, I considered a few extra scenarios. Note that these scenarios do not cover all possible misspecifications, but they do offer some insight.



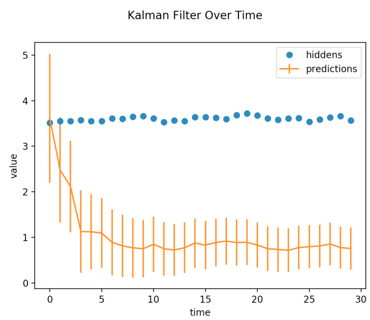
a) $\mu_p = 10, \sigma_p = 0.5$



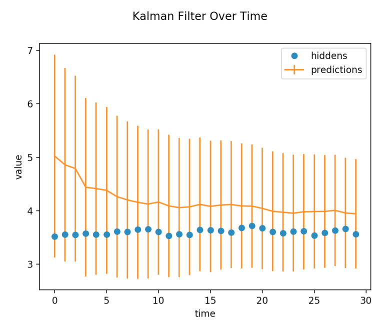
b) $\mu_\epsilon = 3$



c) $\sigma_\epsilon = 3$



d) $\mu_\gamma = 3$



e) $\sigma_\gamma = 3$

I listed the only altered parameters below each graph. All other parameters are the same as they are in the problem statement. As can be seen, the Kalman Filter mean does quite well to catch up to the data even under the extreme misspecification of the prior in (a). It is not fully able to catch up to the hiddens by $t = 29$, but it comes close. With all the other misspecifications, the Kalman Filter mean does a poor job getting close to the hiddens. It is especially bad under misspecifications of μ_ϵ and μ_γ. Misspecifications of σ_ϵ and σ_γ seem to affect the rate at which the Kalman Filter can catch up to the data, and it's ability to stay close to the hiddens once it has made a fairly accurate prediction.

Problem 3 (Explaining Away, 7 pts)

In this problem, you will carefully work out a basic example with the explaining away effect. There are many derivations of this problem available in textbooks. We emphasize that while you may refer to textbooks and other online resources for understanding how to do the computation, you should do the computation below from scratch, by hand.

We have three binary variables, rain r, grass-wet g, and sprinkler s. The conditional probability tables look like the following:

$$\begin{aligned} p(r=1) &= 0.25 \\ p(s=1) &= 0.5 \\ p(g=1|r=0,s=0) &= 0 \\ p(g=1|r=1,s=0) &= .75 \\ p(g=1|r=0,s=1) &= .75 \\ p(g=1|r=1,s=1) &= 1 \end{aligned}$$

1. You check on the sprinkler without checking on the rain or the grass. What is the probability that it is on?
2. You notice it is raining and check on the sprinkler without checking the grass. What is the probability that it is on?
3. You notice that the grass is wet and go to check on the sprinkler (without checking if it is raining). What is the probability that it is on?
4. You notice that it is raining and the grass is wet. You go check on the sprinkler. What is the probability that it is on?
5. What is the explaining away effect above?

Solution

1. The probability that the sprinkler is on in this scenario is simply its marginal probability. This is $\boxed{0.5}$.
2. We wish to find $P(s=1|r=1)$. Note that the rain and the sprinkler are independent. So, by the definition of conditional probability, we have

$$P(s=1|r=1) = \frac{P(s=1,r=1)}{P(r=1)} = \frac{P(s=1)P(r=1)}{P(r=1)} = P(s=1) = 0.5$$

   Thus, due to the independence between the sprinkler and the rain, this probability is also $\boxed{0.5}$.

3. We wish to find $P(s=1|g=1)$. Using Bayes' rule, this is

$$P(s=1|g=1) = \frac{P(g=1|s=1)P(s=1)}{p(g=1)}$$

   By LOTP, the numerator and denominator will be

$$\text{numerator: } \big(P(g=1|s=1,r=0)P(r=0|s=1) + P(g=1|s=1,r=1)P(r=1|s=1)\big)P(s=1)$$

$$\text{denominator: } \big(P(g=1|s=1,r=0)P(r=0|s=1) + P(g=1|s=1,r=1)P(r=1|s=1)\big)P(s=1)$$

$$+\big(P(g=1|s=0,r=0)P(r=0|s=0)+P(g=1|s=0,r=1)P(r=1|s=0)\big)P(s=0)$$

Then, since $P(s=1)=P(s=0)$, and since $s \perp r$, these become

$$\text{numerator: } P(g=1|s=1,r=0)P(r=0)+P(g=1|s=1,r=1)P(r=1)$$
$$\text{denominator: } P(g=1|s=1,r=0)P(r=0)+P(g=1|s=1,r=1)P(r=1)$$
$$+P(g=1|s=0,r=0)P(r=0)+P(g=1|s=0,r=1)P(r=1)$$

Plugging in the actual values for these, we have

$$P(s=1|g=1)=\frac{0.75^2+1\times 0.25}{0.75^2+1\times 0.25+0\times 0.75+0.75\times 0.25}$$
$$=\frac{13}{16}=0.8125$$

Hence, we see that the desired probability is $\boxed{0.8125}$.

4. We wish to find $P(s=1|g=1,r=1)$. Using conditional Bayes' rule, this is

$$P(s=1|g=1,r=1)=\frac{P(g=1|s=1,r=1)p(s=1|r=1)}{p(g=1|r=1)}$$
$$=\frac{1\times 0.5}{p(g=1|r=1,s=0)p(s=0|r=1)+p(g=1|r=1,s=1)p(s=1|r=1)}$$

Since $s \perp r$, this becomes

$$P(s=1|g=1,r=1)=\frac{1\times 0.5}{p(g=1|r=1,s=0)p(s=0)+p(g=1|r=1,s=1)p(s=1)}=\frac{1\times 0.5}{0.75\times 0.5+1\times 0.5}$$

$$=\frac{4}{7}\approx 0.571$$

Hence, desired probability is about $\boxed{0.571}$ or $\frac{4}{7}$.

5. The explaining away effect above is that conditioning on it raining in addition to the grass being wet in part 4 lowers the probability of the sprinkler being on (to 0.571) compared to when we found the probability of the sprinkler being on only conditioning on the grass being wet (which was 0.8125), found in part 3. This is because when we observe that the grass is wet, we know that this means that either the sprinkler is on, it is raining, or both. So, conditional on the grass being wet, the probability of the sprinkler being on and the probability of it raining are both high. However, when we observe that the grass is wet and that it is raining, the probability of the sprinkler being on goes down. This is because we have observed a reason for why the grass could be wet, and this observation “explains away” the probability of the sprinkler being on. It is still possible that the sprinkler is on in this scenario, but we expect it to be on with much less probability because we have already observed a reason for why the grass is wet.

- Name: Zachary Dietz
- Email: zdietz@college.harvard.edu
- Collaborators: Theo Walker
- Approximately how long did this homework take you to complete (in hours): 10