CS 476/676, Spring 2016 Problem Set #1a (Tentative: Due by 11:59pm on Monday, February 15)

---

1 Instructions

This assignment includes programming problems, modeling problems, and analytical questions. Please read the document carefully.

1.1 What to Hand In

All of your submission files should be handed in as a single archive named `hw1a-username.zip`, where `username` is replaced with your JHED ID. See the final `Deliverables` section for instructions.

Hand in the zip archive by creating a private note to the instructors on Piazza with the title *Submission 1a from <list of names of team members>*, and attaching your zip file to that note. The note should be submitted to the `submission1a` folder.

1.2 Submission Policies

Please note the following:

- **Collaboration:** Please work in groups of size 2 people. The homeworks are a way for you to work through the material you're learning in this class on your own. But, by working in a group, and debugging each other's solutions, you'll have a chance to learn the material in more depth. The recommended format for tackling these problem sets is the following. Write a high level sketch of the solution for all of the problems on your own. Meet as a group to brainstorm your solutions and converge on a solution as a group. It is important that you have a good understanding of how you'd have approached the problem independently before discussing your solution with the other group members. Developing this intuition will serve you well in the final exam where you will be required to work on your own. Persuant to your group meeting, write up the solutions on your own. Thereafter, meet as a group to clean up and submit a final write up as a group. By now, each of you should have a solid understand of the concepts involved, and by meeting as a group, you've had a chance to see common ways in which one can make mistakes. Submit your final solution as a final writeup for the group. Your submission should include the names of every team member. Also, name your file as `hw1a-username1-username2.zip`.
- **Late Submissions:** We allow each student to use up to 3 late days over the semester. You have late days, not late hours. This means that if your submission is late by any amount of time past the deadline, then this will use up a late day. If it is late by any amount beyond 24 hours past the deadline, then this will use a second late. **If you jointly submit an assignment as a team, then every team member will lose late days if the assignment is submitted late.** If you collaborate with team members but independently submit your own version, then late hours will only apply to you.

2 Problem Set 1a (90 Points)

2.1 Bayesian Linear Regression

Notation: In what follows, you may assume that bold symbols (e.g. $\mathbf{w}$) denote column vectors, plain symbols (e.g. y_i) denote constants, and bold capital letters (e.g. $\mathbf{X}$) denote matrices.

Model Description: You work at a prestigious investment bank, and your job as an analyst is to build a model to predict the price of company Y's stock given the current prices for companies A, B, C, and D.

You will be receiving new price data every millisecond. You would like to build a model that continuously updates its beliefs about $\boldsymbol{w}$ as you receive new data. You think that a Gaussian distribution approximately captures your beliefs *a priori* about the parameter $\boldsymbol{w}$. Suppose the prior distribution is parameterized with the mean vector μ_0 and covariance matrix Σ_0. The likelihood model is the standard linear regression likelihood:

$$y_i \sim \mathbf{x_i}^T \boldsymbol{w} + \mathcal{N}(0, \sigma^2) \tag{1}$$

(Just to be totally clear, we have $\boldsymbol{w} \in \boldsymbol{R}^4$, $\mu_0 \in \boldsymbol{R}^4$, $\Sigma_0 \in \boldsymbol{R}^{4\times 4}$, $y_i \in \boldsymbol{R}$, $\boldsymbol{x}_i \in \boldsymbol{R}^4$, and $\sigma \in \boldsymbol{R}$.) Also,let $\boldsymbol{y}$ be a column vector containing the y_is and $\boldsymbol{X}$ be a matrix whose rows are the $\mathbf{x}_i$s:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} — & \boldsymbol{x}_1^T & — \\ — & \boldsymbol{x}_2^T & — \\ & \vdots & \\ — & \boldsymbol{x}_n^T & — \end{pmatrix}$$

(Again, just to be totally clear, we have $\boldsymbol{y} \in \boldsymbol{R}^n$ and $\boldsymbol{X} \in \boldsymbol{R}^{n\times d}$.)

a. **10 points:** After n observations, what will be our prior distribution over $\boldsymbol{w}$(a column vector) at time $n+1$ (what is the distribution and what are its parameters)? In stead of barely listing final results, you have to show how you derived them step by step. Same with Question `b`, `f` and `h`.

b. **15 points:** Given X_{n+1} and the prior density you computed above, write down the full posterior predictive distribution over Y_{n+1}. (Hint: 1. The prior you computed in (a) is the distribution of $\boldsymbol{w}$ estimated from previous experience. From this perspective, the posterior at time n is being used as the prior at time (n+1). 2. Search `Woodbury matrix identity` online. You may need it.)

c. **5 points:** Suppose we are unsure about our choice of hyperparameters μ_0 and Σ_0. To account for our uncertainty, we can use hierarchical Bayes and place some prior $\pi(\mu_0, \Sigma_0)$ over the hyperparameters. What estimation technique could we use to approximate the prior in the full hierarchical Bayes model?

d. **5 points:** You need to be able to predict Y_{n+1} quickly, why is the full posterior predictive distribution a bad choice in this scenario? How could you approximate it? (Hint: Instead of a specific model, you just need answer the posterior predictive distributions in general) Under what condition is this approximation reasonable?

e. **10 points:** After observing $10,000$ examples, your boss comes to let you know that the PhDs in the back have suggested that there may be a correlation between the influences that the stock prices for companies A and B or for companies C and D have on Y's price, but they're not sure which. In fact, they're not even entirely sure that the correlation exists.

   Let $\mathcal{M}_0$ denote the model in which there are no correlations between any of the weights in the model, $\mathcal{M}_{AB}$ denote the model in which the weights for A and B are correlated, and $\mathcal{M}_{CD}$ denote the model in which the weights for C and D are correlated. Assuming that all variances are σ_0^2 and all covariances are γ_0^2, how would you encode the beliefs of these three models in three different prior distributions?

f. **15 points:** You'd like to compare the models by computing the posterior distribution over $\mathcal{M}_0, \mathcal{M}_{AB}$, and $\mathcal{M}_{CD}$. Recall that the posterior distribution over model $\mathcal{M}_i$ is

$$P(\mathcal{M}_i|y_{1:n}, x_{1:n}) \propto P(y_{1:n}|x_{1:n}, \mathcal{M}_i)P(\mathcal{M}_i) \tag{2}$$

   Derive a closed form expression for $P(y_{1:n}|x_{1:n}, \mathcal{M}_i)$ where Σ_i is the covariance matrix associated with model $\mathcal{M}_i$. (Hint: How closed-form your final result should be? Well, if your final result was much more complicated than question(b), please double check your answer to avoid mathematical errors.)

g. **15 points:** Using the data distributed with homework 1 (`stocks.csv` available on Piazza), compute the posterior distribution over the 3 models. Assume a uniform prior over the models, and let $\mu_0 = \mathbf{0}, \sigma^2 = 4$, $\sigma_0^2 = 1$ and $\gamma_0^2 = 1/2$. Include a table showing the posterior distribution over the three possible models. Which one would you choose? (Hint: If you are not sure you can implement it correctly, please include a detailed analysis to avoid 0 point)

h. **15 points:** Recent issues with the company's communications infrastructure has caused some unusually noisy observations to be collected. Assuming that the error for observation i is now defined by

$$\epsilon_i \sim \theta\mathcal{N}(0, \sigma^2) + (1 - \theta)\mathcal{N}(0, 50) \tag{3}$$

   (Hint: ϵ_i can be seen as the weighted sum of two independent random variables, whose probability distributions are both Gaussian.) Where $0 < \theta < 1$ is a constant, what is the new posterior distribution over $\boldsymbol{w}$ under this noise model at time n?

3 Deliverables

- Turn in your written solutions as a `pdf` file. The name of the file should be `writeup-username1-username2.pdf`. Also, put your names on the writeup file. (Latex is preferred but not required. For those who are not proficient with Latex, a website called `codecogs` is recommended.)
- Turn in an executable `modelcheck` that will check the three models from part `g` above and print the posterior probability of each.
- Turn in any source code that `modelcheck` relies on.

- Any programming language would be acceptable for this homework. But since hw2 and hw3 will need at least Matlab or python, those who can only use R may want to warm up by starting using Matlab or python(preferred) here.