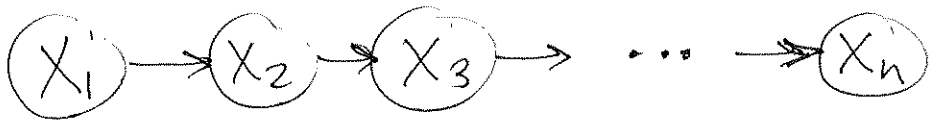
SOLUTIONS FOR STA360/601 FINAL EXAM.

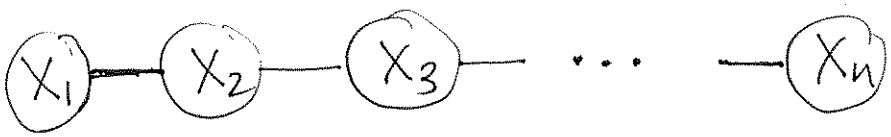
1. (14 points) Graphical models

   (a) (4 points) Draw the directed graphical model (DGM) for a Markov chain. Also draw the associated moral graph.

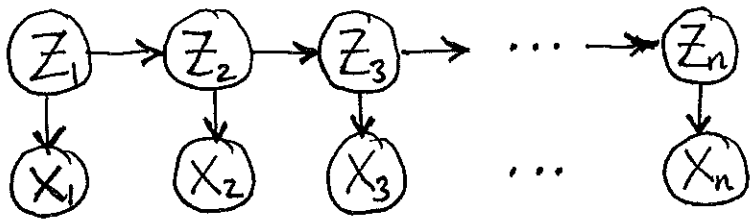
DGM:



Moral graph:



   (b) (4 points) A hidden Markov model is a distribution on $Z_1, \ldots, Z_n, X_1, \ldots, X_n$ that respects the DGM shown below. Write the factorization of $p(z_1, \ldots, z_n, x_1, \ldots, x_n)$ implied by this DGM.



$$p(z_1)p(x_1|z_1)\prod_{i=2}^{n} p(z_i|z_{i-1})p(x_i|z_i)$$

   (c) (6 points) Consider a distribution respecting the DGM in part 1b. For each question, circle either Yes, No, or Indeterminate. Indeterminate means that the answer cannot be determined from the information given. (You do not need to justify your answer.)

i. Is $X_1 \perp X_3 \mid Z_2$?	(Yes)	No	Indeterminate
ii. Is $X_1 \perp X_3$?	Yes	No	(Indeterminate)
iii. Is $X_1 \perp X_3 \mid X_2$?	Yes	No	(Indeterminate)

2. (10 points) (Semi-conjugate priors)
Given $S \in \mathbb{R}^{d\times d}$ symmetric positive definite, and $\nu > d-1$, the Wishart distribution with inverse scale S and ν degrees of freedom has density

$$W_d(X \mid S^{-1}, \nu) = \frac{|S|^{\nu/2}|X|^{(\nu-d-1)/2}\exp(-\frac{1}{2}\mathrm{tr}(SX))}{2^{\nu d/2}\Gamma_d(\nu/2)}$$

for $X \in \mathbb{R}^{d\times d}$ symmetric positive definite. Here, $\Gamma_d(\nu/2)$ is the multivariate gamma function (its definition is unimportant for this problem), and tr is the trace, i.e., $\mathrm{tr}(A) = \sum_{i=1}^{d} A_{ii}$.

Show that the Wishart distribution is a (semi-)conjugate prior for S. That is, if $X_1, \ldots, X_n | S \overset{\text{iid}}{\sim} W_d(S^{-1}, \nu)$ and $S \sim W_d(S_0^{-1}, \nu_0)$, then $p(S|X_{1:n})$ is a Wishart distribution. (Show your work.) (Hint: $\mathrm{tr}(SX) = \mathrm{tr}(XS)$.)

$$p(X_{1:n}|S) = \prod_{i=1}^{n} W_d(X_i|S^{-1},\nu) \qquad \mathrm{tr}(X_i S) = \mathrm{tr}(S X_i)$$

$$\propto_S \prod_{i=1}^{n} |S|^{\nu/2}\exp(-\tfrac{1}{2}\mathrm{tr}(SX_i)) \quad (\text{since } \mathrm{tr}(A+B) = \mathrm{tr}(A)+\mathrm{tr}(B))$$

$$= |S|^{n\nu/2}\exp\left(-\tfrac{1}{2}\mathrm{tr}\left(\sum_{i=1}^{n} X_i S\right)\right)$$

$$p(S) = W_d(S|S_0^{-1},\nu_0) \propto_S |S|^{\frac{\nu_0-d-1}{2}}\exp(-\tfrac{1}{2}\mathrm{tr}(S_0 S))$$

$$\Rightarrow p(S|X_{1:n}) \propto_S p(X_{1:n}|S)p(S)$$

$$\propto_S |S|^{\frac{\nu_0+n\nu-d-1}{2}}\exp\left(-\tfrac{1}{2}\mathrm{tr}\left((S_0+\sum_{i=1}^{n} X_i)S\right)\right)$$

$$\propto_S W_d\left(S \mid (S_0+\sum_{i=1}^{n} X_i)^{-1},\ \nu_0+n\nu\right).$$

3. (10 points) (Multivariate normal)

(a) (5 points) How can you transform a random vector $X \sim \mathcal{N}(\mu, C)$ into a $\mathcal{N}(0, I)$-distributed random vector? (Hint: Any symmetric positive definite matrix C can be factored as $C = U\Lambda U^T$ where U is orthogonal and Λ is diagonal with positive diagonal entries. Use this and the affine transformation property.)

(b) (5 points) Suppose $X \sim \mathcal{N}(a, C)$ and $Y \sim \mathcal{N}(b, D)$ independently, with means $a, b \in \mathbb{R}^d$ and covariance matrices $C, D \in \mathbb{R}^{d \times d}$ respectively. What is the distribution of $X + Y$? Justify your answer. (Hint: A fairly easy way to do this is to note that $\begin{pmatrix} X \\ Y \end{pmatrix}$ is multivariate normal, and use the affine transformation property.)

(a) Let $A = U\Lambda^{1/2}$ where $\begin{cases} C = U\Lambda U^T \\ \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix}, \quad \Lambda^{1/2} = \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_d} \end{bmatrix} \end{cases}$

Then $C = AA^T$.

If $X \sim N(\mu, C)$ then $X - \mu \sim N(0, C)$

and $A^{-1}(X - \mu) \sim N(0, I)$ since ~~$A^{-1}AA^T(A^{-1})^T$~~

$$A^{-1}C(A^{-1})^T = A^{-1}C(A^T)^{-1} = A^{-1}AA^T(A^T)^{-1} = I.$$

(b) $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} C & 0 \\ 0 & D \end{bmatrix}\right)$ since X and Y are independent.

(It is not required, but this can be shown by multiplying the densities.)

By the affine transformation property,

$$X + Y = \begin{bmatrix} I & I \end{bmatrix}\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} I & I \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} I & I \end{bmatrix}\begin{bmatrix} C & 0 \\ 0 & D \end{bmatrix}\begin{bmatrix} I \\ I \end{bmatrix}\right)$$

$$= N(a + b, C + D), \text{ (using block matrix notation).}$$

4. (14 points) (Linear regression, Variable selection)
Vibration analysis is often used to monitor the condition of large complicated factory equipment. To monitor a particular machine, you have a sensor that precisely measures the position of the machine over time. Suppose that at times $x_1, \ldots, x_n \in \mathbb{R}$, you obtain position measurements $y_1, \ldots, y_n \in \mathbb{R}$. Since the movement is periodic, you model the mean position at time x_i as

$$\sum_{k=0}^{m} c_k \cos(2\pi k x_i)$$

for some known m and some unknown $c_0, c_1, \ldots, c_m \in \mathbb{R}$.

(a) (5 points) Assuming Gaussian noise, write down the likelihood of a linear regression model for the position measurements. Assume a common variance σ^2 for all times.

(b) (4 points) Write down a (semi-)conjugate prior for the regression coefficients.

(c) (5 points) Each coefficient c_k represents the amount of vibration at a particular frequency. To analyze the machine, you want to know which frequencies it is vibrating at.

i. In words, describe how variable selection can be used to address this.

ii. You want to do Gibbs sampling for variable selection. Assume you have already derived an expression for the marginal likelihood $p(y|x, z)$, where $y = y_{1:n}$, $x = x_{1:n}$, and $z = z_{0:m}$ with $z_k \in \{0, 1\}$ being the indicator variable for whether c_k is included. Assuming a uniform prior on z, derive an expression for $\mathbb{P}(Z_k = 1 \mid z_{-k}, y, x)$, where z_{-k} denotes the variables in z other than z_k.

(a) $Y_i | x_i, c, \sigma^2 \sim N(c^T \varphi(x_i), \sigma^2)$ where $\varphi(x_i) = \begin{bmatrix} \cos(2\pi \cdot 0 \cdot x_i) \\ \cos(2\pi \cdot 1 \cdot x_i) \\ \vdots \\ \cos(2\pi m x_i) \end{bmatrix}$ and $c = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_m \end{bmatrix}$.

$\Rightarrow p(y | x, c, \sigma^2) = \prod_{i=1}^{n} N(y_i | c^T \varphi(x_i), \sigma^2)$.

(b) Any normal prior on c will be semi-conjugate.

e.g. $c_0, c_1, \ldots, c_m \overset{iid}{\sim} N(\mu_0, \sigma_0^2)$.

(Extra paper for question 4.)

(c)(i) By putting a prior on indicator variables $z_0, z_1, \ldots, z_m \in \{0,1\}$ indicating which coefficients $c_0, c_1, \ldots, c_m$ are to be included in the model, and then considering the posterior probability that $z_k = 1$ for a given k, we can assess the probability that the k^{th} frequency is occurring in the vibrations of the machine.

(ii) $p(z_k \mid z_{-k}, y, x) \underset{z_k}{\propto} p(y \mid z, x) p(z) \underset{z_k}{\propto} p(y \mid z, x)$ (since $p(z)$ is uniform)

$$\Rightarrow P(z_k = 1 \mid z_{-k}, y, x) = \frac{p(y \mid x, z_k = 1, z_{-k})}{\sum_{b \in \{0,1\}} p(y \mid x, z_k = b, z_{-k})}$$

5. (14 points) (Bayesian hypothesis testing)
Suppose you are conducting an experiment on the effects of exercise on cognitive performance. To assess performance, you measure the time required to solve a creative thinking task. You divide subjects into a control group of m subjects and a test group of n subjects. The test group exercises for 30 minutes before performing the task, while the control group does not exercise beforehand. You measure the time required by each individual to solve the task, yielding data $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$. You want to consider the evidence for the two hypotheses, H_0: no difference between groups, and H_1: the groups are different.

(a) (5 points) Modeling the data as Exponential, write down a model to address this using Bayesian hypothesis testing. Use conjugate priors, and assume the two hypotheses have equal prior probability.

(b) (5 points) Give a closed-form expression for the Bayes factor B_{01} in favor of H_0 over H_1.

(c) (4 points) Give an expression for the posterior probability of H_1 in terms of the Bayes factor B_{01}.

(a) $p(H_0) = p(H_1) = 1/2$

Given H_0: $X_1, \ldots, X_m, Y_1, \ldots, Y_n \mid \theta \overset{iid}{\sim} Exp(\theta)$

$\theta \sim Gamma(a, b)$.

Given H_1: $X_1, \ldots, X_m \mid \theta_1 \overset{iid}{\sim} Exp(\theta_1)$

$Y_1, \ldots, Y_n \mid \theta_2 \overset{iid}{\sim} Exp(\theta_2)$

$\theta_1, \theta_2 \overset{iid}{\sim} Gamma(a, b)$.

(b) If $z_1, \ldots, z_k \mid \theta \sim Exp(\theta)$ and $\theta \sim Gamma(a, b)$, then

$$p(z) = \int p(z \mid \theta) p(\theta) \, d\theta = \int_0^\infty \theta^k e^{-\theta \Sigma z_i} \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \, d\theta$$

$$= \frac{b^a}{\Gamma(a)} \int_0^\infty \theta^{a+k-1} e^{-(b+\Sigma z_i)\theta} \, d\theta = \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+k)}{(b+\Sigma z_i)^{a+k}}.$$

(Extra paper for question 5.)

$$\Rightarrow p(x,y|H_0) = \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+m+n)}{(b+\Sigma x_i+\Sigma y_i)^{a+m+n}}$$

(Taking $(z_1,\dots,z_k)=(x_1,\dots,x_m,y_1,\dots,y_n)$, $k=m+n$)

and $$p(x,y|H_1) = \left(\frac{b^a}{\Gamma(a)}\right)^2 \frac{\Gamma(a+m)\Gamma(a+n)}{(b+\Sigma x_i)^{a+m}(b+\Sigma y_i)^{a+n}}$$

$$\Rightarrow B_{01} = \frac{p(x,y|H_0)}{p(x,y|H_1)} = \frac{\Gamma(a)}{b^a} \frac{\Gamma(a+m+n)}{\Gamma(a+m)\Gamma(a+n)} \frac{(b+\Sigma x_i)^{a+m}(b+\Sigma y_i)^{a+n}}{(b+\Sigma x_i+\Sigma y_i)^{a+m+n}}.$$

(c) $$p(H_1|x,y) = \frac{p(x,y|H_1)p(H_1)}{\sum_{k\in\{0,1\}} p(x,y|H_k)p(H_k)} \quad \text{(by Bayes)}$$

$$= \frac{p(x,y|H_1)}{p(x,y|H_0)+p(x,y|H_1)}$$

$$= \frac{1}{\frac{p(x,y|H_0)}{p(x,y|H_1)}+1} = \frac{1}{B_{01}+1}.$$

6. (14 points) (Hierarchical models, Gibbs sampling)
In another experiment on cognitive performance, you have n subjects play the computer game Tetris, and you count how many "tetrises" they get, resulting in a nonnegative integer score. Each subject plays the game 5 times, yielding scores $x_{i1}, \ldots, x_{i5} \in \{0, 1, 2, \ldots\}$ for subject i.

(a) (5 points) Modeling the scores for each subject as Poisson, write down a hierarchical model for this data, using semi-conjugate priors. For the top level in your model, put a prior on only one of the hyperparameters (one for which there is a nice semi-conjugate prior).

(b) (4 points) Draw the directed graphical model (DGM) for your model.

(c) (5 points) Derive closed form expressions for all the full conditionals required to do Gibbs sampling. (Show the derivations.)

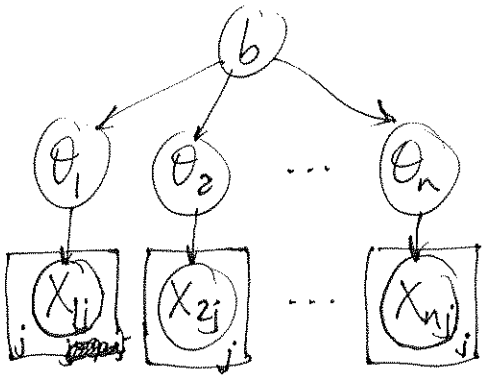
(a) $b \sim \text{Gamma}(r, s)$

$\theta_1, \ldots, \theta_n \mid b \overset{\text{iid}}{\sim} \text{Gamma}(a, b)$ with a fixed.

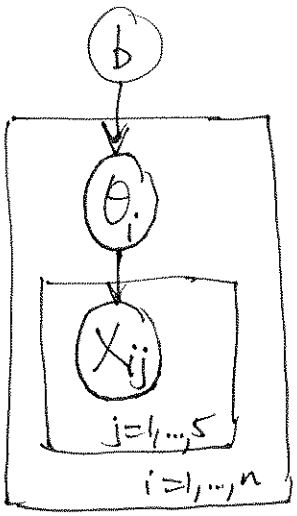
For $i = 1, \ldots, n$ independently, given $\theta_1, \ldots, \theta_n$:

$X_{i1}, \ldots, X_{i5} \mid \theta_i \overset{\text{iid}}{\sim} \text{Poisson}(\theta_i)$.

(b)



OR



(in plate notation)

(It is also ok to include a, r, and/or s, if they are enclosed in a square e.g. [a] or shaded, to indicate that they are conditioned on.)

(Extra paper for question 6.)

(c)

$$p(\theta_i \mid \cdots) \underset{\theta_i}{\propto} \left(\prod_{j=1}^{5} p(x_{ij} \mid \theta_i)\right) p(\theta_i \mid b)$$

$$= \prod_{j=1}^{5} \left(e^{-\theta_i} \frac{\theta_i^{x_{ij}}}{x_{ij}!}\right) \frac{b^a}{\Gamma(a)} \theta_i^{a-1} e^{-b\theta_i} \mathbb{1}(\theta_i > 0)$$

$$\underset{\theta_i}{\propto} \theta_i^{a + \sum_{j=1}^{5} x_{ij} - 1} e^{-(b+5)\theta_i} \mathbb{1}(\theta_i > 0)$$

$$\underset{\theta_i}{\propto} \text{Gamma}\left(\theta_i \mid a + \sum_{j=1}^{5} x_{ij},\ b + 5\right)$$

$$p(b \mid \cdots) \underset{b}{\propto} p(b) \prod_{i=1}^{n} p(\theta_i \mid b)$$

$$= \frac{s^r}{\Gamma(r)} b^{r-1} e^{-sb} \mathbb{1}(b > 0) \prod_{i=1}^{n} \frac{b^a}{\Gamma(a)} \theta_i^{a-1} e^{-b\theta_i} \mathbb{1}(\theta_i > 0)$$

$$\underset{b}{\propto} b^{r+na-1} e^{-\left(s + \sum_{i=1}^{n} \theta_i\right) b} \mathbb{1}(b > 0)$$

$$\underset{b}{\propto} \text{Gamma}\left(b \mid r + na,\ s + \sum_{i=1}^{n} \theta_i\right)$$

7. (10 points) (Markov chains)
Find a transition matrix T on three states (say, $\{1,2,3\}$) which gives rise to a Markov chain that is irreducible and has stationary distribution $\pi = (1/4, 1/2, 1/4)$, BUT is periodic (i.e., not aperiodic). (For full credit, you must explain why it has each of the properties stated above.)

We know that if it is irreducible and there is any x s.t. $T_{xx} > 0$, then it is aperiodic.

$\Rightarrow T_{xx} = 0$ for $x = 1, 2, 3$.

Write $T = \begin{bmatrix} 0 & a & b \\ c & 0 & d \\ e & f & 0 \end{bmatrix}$.

The rows must sum to 1. $\Rightarrow \begin{cases} a+b=1 \\ c+d=1 \\ e+f=1 \end{cases}$

$\pi T = \pi \Rightarrow 4\pi T = 4\pi \Rightarrow \begin{cases} 2c+e=1 \\ a+f=2 \\ b+2d=1 \end{cases}$

Since $0 \le a \le 1$ and $0 \le f \le 1$,

then $a+f=2$ implies $\underline{a=f=1} \Rightarrow \begin{cases} b=e=0 \\ c=d=1/2 \end{cases}$

$\Rightarrow T = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{bmatrix}$.

- It is irreducible since there is positive probability of moving between any two states, eventually.
- It has stationary distn π since $\pi T = \pi$ by construction.
- It is periodic since $\forall x$, $\{t : \mathbb{P}(X_t = x \mid X_0 = x) > 0\} = \{0, 2, 4, 6, \ldots\}$, so its GCD is 2.

8. (14 points) (Metropolis–Hastings MCMC)
Consider a target distribution $\pi(x)$ on a discrete space, and a transition matrix T.

(a) (4 points) What does it mean for π and T to have detailed balance? (Give the mathematical definition.)

(b) (5 points) Suppose T is the transition matrix corresponding to a Metropolis–Hastings move with proposal distribution $q_x(x^*)$. Write down an expression for T_{xy} when $x \neq y$. (Note: A Metropolis–Hastings move refers to a single propose–accept/reject iteration.)

(c) (5 points) Using your answers from the first two parts, show that Metropolis–Hastings moves always have detailed balance. Assume that $\pi(x) > 0$ and $q_x(x^*) > 0$ for all x, x^*. (Hint: You don't need an explicit expression for T_{xx} since the case of $x = y$ is trivial.)

(a) For all x and y, $\pi(x) T_{xy} = \pi(y) T_{yx}$.

(b) $T_{xy} = q_x(y) \min\left\{1, \frac{\pi(y) q_y(x)}{\pi(x) q_x(y)}\right\}$, when $x \neq y$,
(propose y) (accept y)
since y must be proposed from x, and it must be accepted, in order to go from x to y.

(c) $\pi(x) T_{xy} = \pi(x) q_x(y) \min\left\{1, \frac{\pi(y) q_y(x)}{\pi(x) q_x(y)}\right\}$ (when $x \neq y$).

$= \min\{\pi(x) q_x(y), \pi(y) q_y(x)\}$

$= \pi(y) q_y(x) \min\left\{\frac{\pi(x) q_x(y)}{\pi(y) q_y(x)}, 1\right\}$

$= \pi(y) T_{yx}$.

When $x = y$,
$\pi(x) T_{xy} = \pi(x) T_{xx} = \pi(y) T_{yx}$.