STT 872, 867-868 Spring Preliminary Examination
Monday, August 15, 2022
12:30 - 5:30 pm

INSTRUCTIONS:

1. This examination is closed book. Every statement you make must be substantiated. You may do this either by quoting a theorem/result and verifying its applicability or by proving things directly. You may use one part of a problem to solve the other part, even if you are unable to solve the part being used. A complete and clearly written solution of a problem will get a more favorable review than a partial solution.

2. You must start solution of each problem on a separate page. Be sure to put the number assigned to you on the top left corner of every page of your solution. Also please number the pages with "n/m" (top right corner), where n is the current page number and m is the total number of pages, to keep the ordering and to avoid missing any pages during scanning.

3. In ZOOM, the video must be turned on for the whole duration of the exam, while the microphone must be muted for the whole duration of the exam. There should be no other people present in the room during the exam. DO NOT use virtual background. The camera should show a wide angle with you and the desk where your work is visible.

4. If you have questions during the exam (e.g. bathroom break requests) you can send a chat message in ZOOM to the host. Email/cell phone communication with Tami would be a back-up method to ZOOM/ D2L if they fail.

5. The exam will last 5 hours. Additional 30 minutes will be allowed to organize the paper solution (write your assigned number and the page number (n/m) on each page), scan it and upload to D2L. Submit your solution as a PDF file. Before the submission, make sure the PDF is clearly readable and it contains all your answers (check on your laptop). Failing to do so may result in substantial loss of points. Keep your paper solution until the examination result is out. If you run into any upload issues, email your solutions to Tami directly.

6. Please refrain from discussing the exam in any way before the results are made available.

1. Let X and Y be independent Poisson variables, X with mean θ, and Y with mean θ^2, respectively, where $\theta \in (0, \infty)$.

(a) (2 pts) Find a minimal sufficient statistic for the family of joint distributions.

(b) (2 pts) Is the above minimal sufficient statistic complete? Explain your reasoning clearly.

2. (3 pts) Let $X_1, \cdots, X_n$ be i.i.d. absolutely continuous random variables with common density $f_\theta(x) = \theta e^{-\theta x}, x > 0$. Let $X_{(1)} \leq \cdots \leq X_{(n)}$ be the order statistics and $X = (X_1 + \cdots + X_n)/n$ be the sample average. Show that $\bar{X}$ and $X_{(1)}/X_{(n)}$ are independent.

3. (4 pts) Let X_1, X_2, X_3 be i.i.d. geometric variables with common mass function $f_\theta(x) = P_\theta(X_i = x) = \theta(1-\theta)^{x-1}$, $x = 1, \cdots$. Find the UMVU estimator of θ^2.

Hint: Let $X_i, i = 1, 2, \cdots, n$, be independent random variables of geometric distribution, that is, let $P(X_i = m) = \theta(1-\theta)^{m-1}$, then $P(\sum_{i=1}^n X_i = m) = \binom{m-1}{n-1}\theta^n(1-\theta)^{m-n}$.

4. Consider estimating the failure rates $\lambda_1, \cdots, \lambda_p$ of independent exponential random variables $X_1, \cdots, X_p$. So X_i has density $\lambda_i e^{-\lambda_i x}, x > 0$.

(a) (3 pts) Following a Bayesian approach, suppose the unknown parameters are modeled as random variables $\Lambda_1, \cdots, \Lambda_p$. For a prior distribution, assume these variables are i.i.d. from a gamma distribution with shape parameter α and unit scale parameter, so Λ_i has density $\lambda^{\alpha-1}e^{-\lambda}/\Gamma(\alpha)$, $\lambda > 0$. Determine the marginal density of X_i in this model.

(b) (3 pts) Find the Bayes estimate of Λ_i in the Bayesian model with squared error loss.

(c) (3 pts) The Bayesian model gives a family of joint distributions for $X_1, \cdots, X_p$ indexed solely by the parameter α. Determine the maximum likelihood estimate of α for this family.

5. Consider an autoregressive model in which $X_1 \sim N(\theta, \sigma^2/(1-\rho^2))$ and the conditional distribution of X_{j+1} given $X_1 = x_1, \cdots, X_j = x_j$, is $N(\theta + \rho(x_j - \theta), \sigma^2)$, $j = 1, \cdots, n+1$. Assume ρ to be known.

(a) (5 pts) Find the Fisher information matrix, $I(\theta, \sigma)$.

(b) (2 pts) Give a lower bound for the variance of an unbiased estimator of θ.

Hint: Define $\epsilon_j = X_j - \theta$ and $\eta_{j+1} = \epsilon_{j+1} - \rho\epsilon_j$. You may use the fact that $\eta_2, \cdots, \eta_n$ are i.i.d. $N(0, \sigma^2)$ and are independent of ϵ_1.

6. (4 pts) Let X be a single sample from the geometric distribution $P(X = x) = p(1-p)^{x-1}$, $x = 1, \cdots$, with an unknown $p \in (0, 1)$. Show that $I(X = 1)$ is a minimax estimator of p under the loss function $(a-p)^2/(p(1-p))$.

7. (4 pts) Let $X = (X_1, \cdots, X_n)$ be a sample from the uniform distribution $U(\theta, \theta+1)$. Find the UMP test for testing $H : \theta \leq \theta_0$ against $K : \theta > \theta_0$ at level $\alpha \in (0, 1)$.

8. (5 pts) Suppose we observe a single observation X from the density

$$f_\theta(x) = c(\theta)|x|e^{-(x-\theta)^2/2}$$

Find the uniformly most powerful unbiased test of $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$.

9. Suppose at each covariate x_i, two correlated responses $(y_i, \tilde{y}_i)$ are observed, following the **full-rank** linear regression model:

$$y_i = x_i'\beta + \epsilon_i,\ \tilde{y}_i = x_i'\beta + \tilde{\epsilon}_i, \quad i = 1, 2, \ldots, n,$$

where $x_i \in \mathbb{R}^p$, and each pair $(\epsilon_i, \tilde{\epsilon}_i)$ is independently sampled from $\mathcal{N}(0, \Sigma)$ with the unknown covariance matrix $\Sigma = \begin{pmatrix} \sigma^2 & \tau \\ \tau & \sigma^2 \end{pmatrix} \in \mathbb{R}^{2\times 2}$. Define the following notation:

$$X = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} \in \mathbb{R}^{n\times p}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n, \quad \tilde{Y} = \begin{pmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_n \end{pmatrix} \in \mathbb{R}^n.$$

(a) (5 pts) Show that the least squares estimator for β equals $\frac{1}{2}(X'X)^{-1}X'(Y+\tilde{Y})$. Further prove that it is equivalent to the maximum likelihood estimator.

(b) (5 pts) Compute the expectation of the SSE given by $\sum_{i=1}^n [(y_i - x_i'\hat{\beta})^2 + (\tilde{y}_i - x_i'\hat{\beta})^2]$, where $\hat{\beta}$ is least squares estimator in (a). Based on this result, obtain an unbiased estimator for τ.

(c) (5 pts) Define the estimator $\bar{\beta} = (X'X)^{-1}X'(\lambda Y + (1-\lambda)\tilde{Y})$, where $\lambda \in (0, 1)$ is a given constant. Base on $\bar{\beta}$ to construct a $100(1-\alpha)\%$ confidence region for β.

10. Consider the model,

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where $x_i \in \mathbb{R}^p, \epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, and $f(\cdot)$ can be a non-linear function.

(a) (5 pts) Consider the local regression prediction method: for any given covariate $x \in \mathbb{R}^p$, the prediction is $\hat{f}(x) = x'\hat{\beta}_x$ with $\hat{\beta}_x$ given by

$$\hat{\beta}_x = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^n \left[e^{\frac{-1}{2h}\|x-x_i\|_2^2} \cdot (y_i - x_i'\beta)^2 \right],$$

where $h > 0$ is a constant. Compute its expected in-sample prediction error, and further give an unbiased estimator for the error (suppose σ^2 is known).

(b) (5 pts) Consider the case when the model is linear with orthogonal design: $f(x_i) = x_i'\beta, 1 \leq i \leq n$ and $\frac{1}{n}X'X = I_p$. Suppose we would like to select among two model candidates: one including the first p_1 predictors ($p_1 < p$) and one having all the p predictors. How do you select between them using leave-one-out cross validation (LOCV) ? Provide arguments to show that when the first model is the true model, LOCV has non-vanishing probability of selecting the wrong model, as $n \to \infty$ and $\max_{1\leq i\leq n} \|x_i\|_2^2/n \to 0$.

11. Consider the international trade data where y_{ij} denotes the export volume (in log billions of dollars) from country i to country j for $1 \leq i \neq j \leq n$. We model the data by the following two-way random effect model:

$$y_{ij} = \mu + a_i + b_j + \epsilon_{ij}, \quad 1 \leq i \neq j \leq n, \tag{1}$$

where $\mu \in \mathbb{R}$ is the mean parameter, $\{(a_i, b_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$, $\{(\epsilon_{ij}, \epsilon_{ji})\}_{1\leq i<j\leq n} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \Omega)$, with $\Sigma = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}, \Omega = \begin{pmatrix} \sigma^2 & \tau \\ \tau & \sigma^2 \end{pmatrix}$, and ϵ_{ij}'s are independent from the a_i's, b_j's. The parameters in the model are μ, Σ, Ω. Denote the whole data $\{y_{ij} : 1 \leq i \neq j \leq n\}$ by the vector $\mathbf{Y} \in \mathbb{R}^{n(n-1)}$.

(a) (5 pts) Compute $\mathbb{E}(\mathbf{Y})$ and $\mathrm{Cov}(\mathbf{Y})$. Base on these moment calculations to prove the identifiability of the model.

(b) (5 pts) Describe a method to estimate the parameter σ_{ab}.

(c) (5 pts) For a given triplet (i, j, k), we may use $\theta_{ijk} = \mathbb{E}[(y_{ij} - \mathbb{E}y_{ij})(y_{jk} - \mathbb{E}y_{jk})(y_{ki} - \mathbb{E}y_{ki})]$ to measure the third-order dependence among the data. Suppose we add a multiplicative effect to obtain a new model:

$$y_{ij} = \mu + a_i + b_j + u_i^T v_j + \epsilon_{ij}, \quad 1 \leq i \neq j \leq n, \tag{2}$$

where $u_i, v_j \in \mathbb{R}^2$, $\{(u_i, v_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \Theta)$ for some $\Theta \in \mathbb{R}^{4\times 4}$, and (u_i, v_i)'s are independent from all the other random variables in the model. Compute θ_{ijk} under both Models (1) and (2) and explain the difference.