Ph.D. Qualifying Exam
Thursday–Friday, January 2–3, 2020

- **Begin your solution to each problem on a new sheet of paper.**
- **Statistics PhD students should do the 5106 problems.**
- **Biostatistics PhD students should do the 5198 problems.**
- **All students should do the 5166 and 5167 problems.**

Problem 1. (5106) Let Y be a discrete random variable with nonnegative integer values. For any $y \in \{0, 1, 2, \cdots\}$,

$$Prob(Y = y) = \alpha_1 P_1(y; \lambda_1) + \alpha_2 P_2(y; \lambda_2),$$

where P_1 and P_2 are two Poisson probability mass functions with means λ_1 and λ_2, respectively. Also, $0 \leq \alpha_1, \alpha_2 \leq 1$, such that $\alpha_1 + \alpha_2 = 1$. Given n i.i.d. observations $\{Y_i\}_{i=1}^n$, our goal is to find the maximum likelihood estimate of

$$\theta = (\alpha_1, \lambda_1, \alpha_2, \lambda_2).$$

Use the EM algorithm for iteratively estimating θ. Let $\theta^{(m)}$ be the estimated value at the m-step. Derive the mathematical formula to update it for $\theta^{(m+1)}$.

Problem 2. (5106) Consider a data set of observations $\{x_n\}$ where $n = 1, \cdots, N$, and x_n is a Euclidean variable with dimensionality D. Our goal is to project the data onto a space having dimensionality $M < D$ while maximizing the variance of the projected data. Prove that the optimal linear projection for which the variance of the projected data is maximized is defined by the M eigenvectors $U_1, \cdots, U_M$ of the data covariance matrix S corresponding to the M largest eigenvalues $\lambda_1, \cdots, \lambda_M$.

Problem 3. (5166) Consider the following linear model for a randomized block design:

$$y_{ti} = \mu + \beta_i + \tau_t + \epsilon_{ti}, t = 1, \ldots, k; i = 1, \ldots, n,$$

where μ is an overall mean, τ_t is the effect of tth treatment, β_i is the effect of ith block, $\{\epsilon_{ti} : t = 1, \ldots, k; i = 1, \ldots, n\}$ are assumed to be i.i.d. $N(0, \sigma^2)$.

(a) The least squares estimate of τ_t is $\hat{\tau}_t = \bar{y}_{t\cdot} - \bar{y}_{\cdot\cdot}$. Find the expectation and variance of $\hat{\tau}_t$.

(b) Show the decomposition of variation for the experiment: $S_D = S_B + S_T + S_R$ where

- S_D: Total Variation of the observations,
- S_B: Sum of Squares for Blocks,
- S_T: Sum of Squares for Treatments,
- S_R: Sum of Squares for Experimental Errors.

(c) Show that under the Null Hypothesis $H_0 : \tau_t = 0$ for $t = 1, \ldots, k$, S_T and S_R are independent.

Problem 4. (5166) Two treatments, A and B, are compared in an experimental design. For each treatment, measurements are recorded successively in time. Suppose that the measurements follow the models:

$$X_i(t) = \mu_i + \xi_i(t), \quad i = 1, 2;$$

where the processes $\{\xi_i(t), i = 1, 2; t = -1, 0, 1, \ldots, \}$ are generated by the models:

$$\xi_1(t) - 0.8\xi_1(t-1) = a_1(t), \quad \xi_2(t) = a_2(t) + 2.5a_2(t-1),$$

where $\{a_i(t), i = 1, 2; t = 0, \pm 1, \pm 2 \ldots, \}$ are i.i.d. $N(0, \sigma^2)$ variables.

Let $\bar{X}_i = \sum_{t=1}^{n} X_i(t)/n$ for $i = 1, 2$.

(a) Show that $\xi_1(t) = \sum_{j=0}^{\infty} (0.8)^j a_1(t-j)$ is a solution of the equation

$$\xi_1(t) - 0.8\xi_1(t-1) = a_1(t).$$

Is the process $\{\xi_1(t), t = -1, 0, 1, \ldots, \}$ stationary?

(b) Based on the expression in (a) for $\xi_1(t)$, calculate autocorrelations

$$\gamma_k = \mathrm{Cov}(\xi_1(t), \xi_1(t+k))$$

for $k \geq 0$. Show that $\gamma_k = 0.8\gamma_{k-1}$ for any $k \geq 0$.

(c) What is the distribution of $\bar{X}_2$?

Problem 5. (5167) Consider the following simple linear regression model of Y on X_1:

$$Y = \alpha + \beta X_1 + e. \tag{1}$$

Suppose $n = 60$, $\bar{X}_1 = -3.3$, $\bar{Y} = 5.5$, $SXX = 100.0$, $SYY = 123.4$, and $SXY = 88.8$.

(a) Find the following: RSS, R^2, $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2 \equiv \widehat{var}(e)$.

(b) Test the hypothesis (at level 0.05) that $\beta = 1$ versus $\beta > 1$.

(c) What assumptions are required for the test in part (b) (in addition to (1))?

(d) Find the 95% prediction interval for a new observation $X_1^\star = -2$.

(e) Consider the following regression model,

$$Y_i = \mu + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\text{var}(\epsilon_i) = \sigma^2$ for all i, and $\text{Corr}(\epsilon_i, \epsilon_j) = \rho > 0$ for $i \neq j$. Is the sample mean $\overline{Y} = \sum_{i=1}^n Y_i/n$ still the MLE for μ? What is the variance of $\overline{Y}$? Does the variance go to zero when $n \to \infty$ in this case, and why?

Some R output that might be helpful.

```
> qt(c(0.95,0.975,0.99,0.995),57)
[1] 1.672029 2.002465 2.393568 2.664870
> qt(c(0.95,0.975,0.99,0.995),58)
[1] 1.671553 2.001717 2.392377 2.663287
> qt(c(0.95,0.975,0.99,0.995),59)
[1] 1.671093 2.000995 2.391229 2.661759
```

Problem 6. (5167) Suppose we have univariate response Y and multivariate predictor $X \in \mathbb{R}^p$. We are interested in the following regression model,

$$Y = \beta_0 + \beta^T X + \varepsilon = \widetilde{\beta}^T \widetilde{X} + \varepsilon,$$

where $\widetilde{\beta} = (\beta_0, \beta^T)^T \in \mathbb{R}^{p+1}$ and $\widetilde{X} = (1, X^T)^T \in \mathbb{R}^{p+1}$. Let $Y_n \in \mathbb{R}^{n\times 1}$, $X_n \in \mathbb{R}^{n\times p}$ and $\widetilde{X}_n = (1_n, X_n) \in \mathbb{R}^{n\times(p+1)}$ be the data matrices, where 1_n is a vector of ones. You may use this result to answer the questions: Suppose $M = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$ is a $m \times m$ symmetric positive definite matrix, then its inversion can be written as

$$M^{-1} = \begin{pmatrix} A^{-1} - A^{-1}BDB^TA^{-1} & -A^{-1}BD \\ -DB^TA^{-1} & D \end{pmatrix},$$

where $D = (C - B^TA^{-1}B)^{-1}$.

(a) Write down the ordinary least squares estimates for $\widetilde{\beta}$, β_0 and β using the given notation.

(b) Let $\overline{X}$ and S be the sample mean and sample covariance of X. Show that

$$(\widetilde{X}_n^T\widetilde{X}_n)^{-1} = \begin{pmatrix} n^{-1} + (n-1)^{-1}\overline{X}^TS^{-1}\overline{X} & -(n-1)^{-1}\overline{X}^TS^{-1} \\ -(n-1)^{-1}S^{-1}\overline{X} & (n-1)^{-1}S^{-1} \end{pmatrix}$$

(c) Suppose we want to estimate the inverse of covariance matrix denoted as $\Theta \equiv \{\mathrm{Cov}(X)\}^{-1} \in \mathbb{R}^{p\times p}$. When p is very large, directly computing the inversion of the sample covariance matrix S is not practical. When Θ is sparse, the graphical lasso method uses a regression "trick" to obtain the sparse estimator of the inverse covariance. The algorithm updates one column at a time for Θ. To see this, we decompose Θ as

$$\Theta = \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^T & \theta_{22} \end{pmatrix},$$

where $\theta_{12} \in \mathbb{R}^{(p-1)\times 1}$. Suppose $\widehat{\Theta} = S^{-1}$, show that $\widehat{\theta_{12}}/\widehat{\theta_{22}}$ is the least squares regression coefficient estimator of X_p regressed on $(X_1, \ldots, X_{p-1})$. Discuss how to obtain $\widehat{\Theta}$ when the sample covariance is not positive definite due to small-n-large-p (hint: use lasso regression).

Begin your solution to each problem on a new sheet of paper.

Problem 9. (5326) Answer the following. The parts are not related.

(a) Define $Y = \tan X$ where X is uniformly distributed on the interval $(-\pi/2, \pi/2)$. Find the cumulative distribution function (cdf) and the density (pdf) of Y.

Note: $\dfrac{d}{du}\tan^{-1}(u) = \dfrac{1}{1+u^2}$

(b) Let Z have the Cauchy density given by

$$f_Z(z) = \frac{1}{\pi(1+z^2)}, \quad -\infty < z < \infty.$$

Show that $Z \stackrel{d}{=} 1/Z$, that is, Z and $1/Z$ have the same distribution.

(c) Find the density (pdf) of the ratio S/T where S, T are i.i.d. $N(0,1)$ (standard normal).

Problem 10. (5326) An urn contains R red balls and G green balls. Balls are drawn at random from this urn until a red ball is drawn. What is the probability that **more** than k draws are required? (Assume $k \geq 1$ is some given integer.) Answer this question in each of the following situations. Simplify your answers.

(a) Draws are done **WITH** replacement.

(b) Draws are done with**OUT** replacement. (In this case assume that $1 \leq k \leq G$.)

(c) The urn is a Polya urn: after a ball is drawn, it is returned to the urn and another ball of the same color is added to the urn.

The following table will be useful for solving the STA 5327 problems.

Name	**Notation**	density $f(x)$	$\mu = E(X)$	Var(X)
Normal	$\mathrm{N}(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$	μ	σ^2

Problem 11. (5327) Consider independent observations $(X_i, Y_i), i = 1, \ldots, n$, where

$$P(Y_i = k) = \pi_k\,, \qquad X_i \mid Y_i = k \sim N(\mu_k, \sigma^2),$$

with $Y_i \in \{1, 2\}$, $0 < \pi_1, \pi_2 < 1$, $\pi_1 + \pi_2 = 1$, $\mu_k \in \mathbb{R}$, $\sigma^2 > 0$.

(a) Construct consistent estimators for $\pi_1, \pi_2, \mu_1, \mu_2$ and σ^2. Which of them (if any) are biased?

(b) Verify that (X_i, Y_i) satisfies the following equation:

$$\log \frac{P(Y_i = 2 \mid X_i = x)}{P(Y_i = 1 \mid X_i = x)} = \beta_0 + \beta x.$$

Derive the explicit expressions of β_0 and β.

(c) Find a consistent estimate for β_0. Prove the consistency.

Problem 12. (5327) Consider independent observations $X_i, i = 1, \ldots, n$, where $X_i \sim N(\mu, 1)$ with $|\mu| \leq 1$.

(a) Let $\bar{X}$ be the sample mean. Is $\bar{X}$ admissible? Give a rigorous proof to your conclusion.

(b) Find the likelihood ratio test for

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu \neq 0, |\mu| \leq 1.$$

(c) If n is very large, is the knowledge of $|\mu| \leq 1$ important for testing the hypothesis in part (b)?

Problem 13. (6346) Answer the following.

(a) Give the definition of a submartingale.

(b) State Doob's decomposition theorem and prove that the decomposition is unique.

(c) Suppose X_n is a submartingale and its Doob decomposition is $X_n = X_0 + Y_n + A_n$ where Y_n is the martingale. Show $E(A_n) < \infty$.

Problem 14. (6346) Let $E_1, E_2, \ldots$ be a sequence of events and μ a probability measure.

(a) Show that $\mu\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} \mu(E_i)$.

(b) If $\mu(E_i) = 0$ for all i, prove that $\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = 0$.

(c) If $\mu(E_i) = 1$ for all i, prove that $\mu\left(\bigcap_{i=1}^{\infty} E_i\right) = 1$.

(d) If $E_i \subset E_{i+1}$, prove that $\mu(\lim_{i\to\infty} E_i) = \lim_{i\to\infty} \mu(E_i)$.