Student ID:

Final Examination: QUESTION BOOKLET

Prof. William Fithian

Fall 2018

- Do *NOT* open this question booklet until you are told to do so.
- Write your Student ID number at the top of this page.
- Write your solutions in this booklet.
- No electronic devices are allowed during the exam.
- Be neat! If we can't read it, we can't grade it.
- You can treat any results from lecture or homework as "known," and use them in your work without rederiving them, but do make clear what result you're using. You do not need to explicitly check regularity conditions for the theorems from class that required them.
- For a multi-part problem, you may treat the results of previous parts as given (if you don't prove the result for part (a), you can still use it to solve part (b)).
- I have starred some parts which I believe are the most difficult, and which I expect most students won't necessarily be able to solve in the time allotted. They are generally not worth more points than the less difficult parts, so don't waste too much time on them until you're happy with your answers to the latter.
- Be careful to justify your reasoning and answers. We are primarily interested in your understanding of concepts, so show us what you know.
- Good luck!

1. A curved Gaussian family (20 points, 4 points / part). Some useful facts for this problem:

- Recall that the Gaussian density function for $Z \sim N(\mu, \sigma^2)$ is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Suppose that

$$X_1, \ldots, X_n = \begin{pmatrix} X_{1,1} \\ X_{1,2} \end{pmatrix}, \ldots, \begin{pmatrix} X_{n,1} \\ X_{n,2} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} N_2(\mu(\theta), I_2),$$

for $\theta \in \mathbb{R}$ and $\mu(\theta) = \binom{\theta}{\theta^2}$.

(a) Show that $T(X) = \sum_i X_i \in \mathbb{R}^2$ is a minimal sufficient statistic but is not complete sufficient.

(b) Find the Fisher information $J_n(\theta)$, i.e. the information about θ in the complete sample.

(c) Consider the score test of $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$. Give an explicit expression for both the test statistic and its rejection threshold, and show that the test achieves finite-sample control of the Type I error rate.

(d) Find the asymptotic distribution of $\hat{\mu}_2 = \hat{\theta}^2$, the MLE for the expectation of $X_{i,2}$, when $\theta \neq 0$. Compare its asymptotic relative efficiency to the "obvious" estimator $\frac{1}{n}\sum_{i=1}^n X_{i,2}$.

(e) (*) If $\theta = 0$, find the asymptotic distribution of $\hat{\theta}^2$, appropriately centered and scaled (feel free to use heuristic arguments).

Problem 1 answers continued (1):

Problem 1 answers continued (2):

Problem 1 answers continued (3):

2. Species abundance (20 points, 5 points / part). Some useful facts for this problem:

- Recall that the Poisson distribution $X \sim \text{Pois}(\lambda)$ has probability mass function
$$\frac{\lambda^x e^{-\lambda}}{x!},$$
on $x = 0, 1, 2, \ldots$. X has mean λ and variance λ.

- The multinomial distribution $(X_1, \ldots, X_d) \sim \text{Multinom}(n, \pi)$ has probability mass function
$$\frac{n!}{\prod_i x_i!} \prod_i \pi_i^{x_i}$$

- Suppose $X_i \overset{\text{ind.}}{\sim} \text{Pois}(\lambda_i)$ for $i = 1, \ldots, d$, and let $X_+ = \sum_i X_i$ and $\lambda_+ = \sum_i \lambda_i$. Then conditional on $X_+ = x_+$, we have
$$(X_1, \ldots, X_d) \sim \text{Multinomial}(x_+, (\lambda_1, \ldots, \lambda_d)/\lambda_+).$$

Consider an ecological sampling problem where we visit m sites and for each of s species, we count the total number of individuals at each site. Let $N_j^{(i)}$ denote the number of individuals of species j at site i. Hence we observe a table of counts of the form

	Species		
Sites	1	$\cdots$	s
1	$N_1^{(1)}$	$\cdots$	$N_s^{(1)}$
$\vdots$	$\vdots$	$N_j^{(i)}$	$\vdots$
m	$N_1^{(m)}$	$\cdots$	$N_s^{(m)}$

We will assume throughout that the rows $N^{(1)}, \ldots, N^{(m)}$ are i.i.d. random vectors in $\mathbb{R}^s$ (but the coordinates $N_1^{(i)}, \ldots, N_s^{(i)}$ within a single site are *not* i.i.d.).

(a) First assume $N_j^{(i)} \overset{\text{ind.}}{\sim} \text{Pois}(\lambda_j)$ for $j = 1, \ldots, s$, and that at each site the species counts are independent, i.e.
$$N^{(i)} \overset{\text{i.i.d.}}{\sim} p_\lambda(n) = \prod_{j=1}^{s} \frac{\lambda_j^{n_j} e^{-\lambda_j}}{n_j!} \tag{1}$$

Find a complete sufficient statistic for the entire data table and give a UMVU estimator for λ_j, the average abundance of species j, explaining why it is UMVU.

(b) Next, assume we use outside data to compute a dissimilarity measure $d(j,k) \in [0,\infty)$ for each pair of species $1 \le j < k \le s$; for example $d(j,k)$ could denote how long ago the species diverged in their evolution. Take the $d(j,k)$ as fixed and known.

We might expect that some latent characteristics of the habitat at site i cause similar species to be more or less common together, and we can test this hypothesis by modifying our model:

$$N^{(i)} \overset{\text{i.i.d.}}{\sim} p_{\lambda,\beta}(n) \;\propto\; \prod_{j=1}^{s} \frac{\lambda_j^{n_j} e^{-\lambda_j}}{n_j!} \;\times \prod_{1 \le j < k \le s} \exp\{\beta e^{-d(j,k)} n_j n_k\}. \quad (2)$$

Show that this model is an exponential family with $s+1$ sufficient statistics, and find the natural parameter corresponding to each (as always, there are multiple ways to write these). You do *not* need to find the normalizing constant.

(c) Find a UMPU test of $H_0 : \beta = 0$ (independence) vs. $H_1 : \beta > 0$ (positive correlation between similar species) and explain how to find its critical value.

(d) (*) Now suppose we want to make our test more robust by dropping the Poisson assumption: under the null hypothesis the species counts are still independent, but now with unknown distributions (still supported on the non-negative integers):

$$N^{(i)} \overset{\text{i.i.d.}}{\sim} \prod_{j=1}^{s} F_j(n_j) \quad (\text{under } H_0),$$

and under the alternative the counts of similar species are still more correlated. Modify your test from part (b) so that it controls finite-sample Type I error, in this nonparametric model.

Problem 2 answers continued (1):

Problem 2 answers continued (2):

Problem 2 answers continued (3):

3. Inverse gamma prior (20 points, 4 points / part). Some useful facts for this problem:

- Recall that the Gaussian density function for $Z \sim N(\mu, \sigma^2)$ is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- A χ^2_d random variable has mean d and variance $2d$.

- If Y is a Gamma(α, β) random variable (in its "rate parameterization") then it has density

$$\frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp\{-\beta y\},$$

  on $(0, \infty)$. Y has mean α/β and variance α/β^2. This distribution is defined for $\alpha, \beta > 0$.

- The inverse-gamma distribution (denoted $IG(\alpha, \beta)$) is the distribution of $W = 1/Y$ where $Y \sim$ Gamma(α, β). Then $W \in (0, \infty)$ has the density

$$\frac{\beta^\alpha}{\Gamma(\alpha)} w^{-\alpha-1} \exp\{-\beta/w\}.$$

  Note that β is a scale parameter for W. W has mean $\frac{\beta}{\alpha-1}$ provided $\alpha > 1$, and variance $\frac{\beta-1}{(\alpha-1)^2(\alpha-2)}$ provided $\alpha > 2$. This distribution is likewise defined for $\alpha, \beta > 0$.

- Define the *squared relative error* loss function

$$L_{\text{rel}}(d, \theta) = \left(\frac{d-\theta}{\theta}\right)^2 = \left(\frac{d}{\theta} - 1\right)^2,$$

  and define the corresponding risk function $R_{\text{rel}}(\delta(\cdot), \theta) = \mathbb{E}_\theta[L_{\text{rel}}(\delta(X), \theta)]$.

Consider the Bayesian model with

$$\theta \sim IG(\alpha, \beta),$$
$$X_1, \ldots, X_n \mid \theta \overset{\text{i.i.d.}}{\sim} N(0, \theta)$$

Note that the variance is θ, not θ^2, and assume $n \geq 2$.

(a) Find the posterior distribution of θ given $X = (X_1, \ldots, X_n)$ and the Bayes estimator for θ under the usual squared error loss.

(b) Give the mean squared error of the Bayes estimator from part (a), as a function of θ (you don't need to try too hard to simplify it).

(c) Find the Bayes estimator for θ under the squared relative error loss L_{rel}.

(d) (*) For the estimator in part (c), find the risk function $R_{\text{rel}}(\delta(\cdot), \theta)$ as a function of θ and show that the Bayes risk is $\frac{2}{n+2(\alpha+1)}$.

(e) For the relative squared error risk, find a linear estimator of the form $\delta(X) = a\sum_{i=1}^{n} X_i^2$ that is minimax, and prove it is minimax.

Problem 3 answers continued (1):

Problem 3 answers continued (2):

Problem 3 answers continued (3):

4. Inference in the Laplace family (15 points, 5 points / part). Some useful facts for this problem:

- The Laplace location family with location parameter θ is given by the density $p_\theta(x) = f(x-\theta)$, where $f(x) = \frac{1}{2}e^{-|x|}$.
- The sign function is defined as

$$\text{sign}(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}.$$

Assume we observe $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Laplace}(\theta)$.

(a) Find the score test of $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$. Give the test statistic and threshold value in terms of a quantile of a binomial distribution (for simplicity you may assume α is chosen so the binomial distribution has an exact α quantile, so the test need not be randomized. Also note $\mathbb{P}_\theta(X = 0) = 0$ for all θ, so you don't need to worry about what happens there).

(b) Suppose we are not so sure about the Laplace assumption, but we do believe the data come from a symmetric location family, meaning $p_\theta(x) = f(x-\theta)$ for some unknown $f : \mathbb{R} \to [0, \infty)$ that integrates to 1 and is symmetric about the origin (we can think of the nonparametric family as being parameterized by (θ, f)). Show that the test from part (a) is still a valid, finite-sample test of $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$ in this larger family.

(c) Now consider testing $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$ (note the null hypothesis now includes negative values of θ). Show that the test from part (a) is a valid and unbiased level-α test for the nonparametric family from part (b).

Problem 4 answers continued (1):

Problem 4 answers continued (2):

Problem 4 answers continued (3):