Final Examination: QUESTION BOOKLET

Prof. Will Fithian

Fall 2020

- The exam begins at 3:10pm and ends at 6:00pm. There is a grace period for turning in the exam until 6:10pm; after that, the exam accrues a 20-point penalty plus 20 points more for every additional 10 minutes of lateness. If you are unable to submit to Gradescope, take timestamped photos and send them to us by email as soon as you possibly can.
- Any communication with classmates or anyone else other than me during the exam, about any subject remotely related to statistics, is strictly forbidden. That includes statements like "Problem 2 is so hard!"
- The exam is open book, open notes, open lecture videos, and any general resources from the Internet (**not** any materials specifically related to this test, obviously). These are not standard problems so hunting around for the answers to them in textbooks is unlikely to be worth your time.
- **Some students are taking the exam later due to time zone issues. Do not post anything about the exam on Piazza until I post the solutions tomorrow afternoon.**
- All parts of all problems are worth 5 points. There are 20 total parts, for 100 total points.
- 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.
- For a multi-part problem, you may treat 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 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.
- You can ask questions by email to me, with [210A Exam] in the subject line, and I will respond as quickly as I can. But my answer to most questions is just "I am satisfied with the wording of the exam as written."
- Check your email every so often just in case I have to correct something.

Good luck!

1. One Poisson, two Poissons (30 points, 5 points / part).

Some useful facts / notation for this problem:

- For $\theta > 0$, the Poisson density for $X \sim \text{Pois}(\theta)$ is $\frac{\theta^x e^{-\theta}}{x!}$ on $x = 0, 1, \ldots$. The mean and variance are both θ.
- Let $P(n)$ denote the set of integers $0 \le i \le n$ with the same parity (odd/even) as n, i.e. for which $n - i$ is even:

$$P(n) = \{i \in 0, 1, \ldots, n : \; n - i \text{ is even}\},$$

so for example $P(10) = \{0, 2, 4, 6, 8, 10\}$ while $P(9) = \{1, 3, 5, 7, 9\}$.

Suppose we observe two independent random variables, with

$$X \sim \text{Pois}(\theta), \quad \text{and } Y \sim \text{Pois}(\theta^2),$$

where $\theta > 0$ is an unknown parameter.

(a) Show that the model is an exponential family and find its complete sufficient statistic.

(b) Give an explicit expression for the UMVU estimator of θ. Evaluate it when $X = Y = 2$ (give your answer as a fraction, or a decimal with at least 3 significant digits).

(c) Now suppose that you observe an i.i.d. sample of n pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ where each pair has the same distribution specified above. That is, $X_i \sim \text{Pois}(\theta)$ and $Y_i \sim \text{Pois}(\theta^2)$, independently. Give an explicit expression for the MLE $\hat{\theta}_n$ as a function of the data.

If $\sum_{i=1}^n X_i = \sum_{i=1}^n Y_i = 2n$, find the MLE for θ (give your answer as a fraction, or a decimal with at least 3 significant digits).

(d) Find the asymptotic distribution of $\hat{\theta}_n$ as $n \to \infty$. (Don't worry about checking any regularity conditions for this part).

(e) A simpler estimator for θ is

$$\tilde{\theta}_n = \frac{\overline{X}_n + \overline{Y}_n^{1/2}}{2},$$

where $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$ and $\overline{Y}_n = n^{-1} \sum_{i=1}^n Y_i$.

Find the asymptotic distribution of this estimator. Justify why it has the distribution you say and give its asymptotic relative efficiency.

(f) Now suppose we want to test our model against the alternative hypothesis that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are still i.i.d. pairs of independent Poisson random variables, but their means do not have the relationship we posited. In other words, in the expanded model

$$X_i \sim \text{Pois}(\theta), \quad Y_i \sim \text{Pois}(\lambda), \quad i = 1, \ldots, n,$$

Test $H_0 : \lambda = \theta^2$ against the alternative $H_1 : \lambda \neq \theta^2$, for large n. Suggest an asymptotic test from class or homework: give an explicit expression for the test statistic and an explicit rejection cutoff in terms of a quantile of a known distribution. (If you choose a well-known test that is appropriate for this kind of setting then you do **not** need to justify why your test has the correct null distribution in this case).

(**Hint:** there are at least three choices of asymptotic tests from class or homework; it might pay off to take a moment to consider which is easiest to carry out here).

1. Solution

(a) The likelihood is

$$\begin{aligned} p_\theta(x,y) &= \frac{\theta^x e^{-\theta}}{x!} \cdot \frac{\theta^{2y} e^{-\theta^2}}{y!} \\ &= \exp\{(x+2y)\log\theta - \theta - \theta^2\}\frac{1}{x!y!}, \end{aligned}$$

a one-parameter exponential family with natural parameter $\eta = \log\theta$, complete sufficient statistic $T = X + 2Y$, normalizing constant $B(\theta) = \theta + \theta^2$, and carrier density $\frac{1}{x!y!}$ (wrt the counting measure on pairs of non-negative integers). T is complete sufficient because $\log\theta$ varies over the entire real line, which is an open set.

(b) Because $\mathbb{E}_\theta X = \theta$, we can get a UMVUE by Rao-Blackwellizing it, to obtain the estimator

$$\delta(t) = \mathbb{E}[X \mid X + 2Y = t] = \left(\sum_{x \in P(t)} \frac{x}{x!\left(\frac{t-x}{2}\right)!}\right) \Bigg/ \left(\sum_{x \in P(t)} \frac{1}{x!\left(\frac{t-x}{2}\right)!}\right),$$

where $P(t)$ includes all possible values of X given $X + 2Y = t$, and the factors in the likelihood that involve θ cancel in the numerator and denominator.

If $X = Y = 2$ then $T = 6$, so

$$\delta(6) = \frac{\frac{0}{0!3!} + \frac{2}{2!2!} + \frac{4}{4!1!} + \frac{6}{6!0!}}{\frac{1}{0!3!} + \frac{1}{2!2!} + \frac{1}{4!1!} + \frac{1}{6!0!}} = \frac{486}{331} \approx 1.47$$

(c) An i.i.d. sample from an exponential family with sufficient statistic $X_i + 2Y_i$ is just another exponential family with sufficient statistic $T = \sum_i X_i + \sum_i 2Y_i$. The MLE sets the complete sufficient statistic equal to its expectation (in this case $n(\theta + 2\theta^2)$) and solves for θ:

$$2n\hat\theta_n^2 + n\hat\theta_n = T \iff \hat\theta_n = \frac{-n \pm \sqrt{n^2 + 8nT}}{4n} = \frac{-1 \pm \sqrt{1 + 8T/n}}{4}$$

Because $\hat\theta_n > 0$, we choose the positive root and the MLE is

$$\hat\theta_n = \frac{\sqrt{1 + 8T/n}}{4} - \frac{1}{4}$$

If $T = 6n$, we obtain $\hat\theta_n = \frac{3}{2} = 1.50$ (1.5 obviously acceptable too).

(d) As a function of θ, the log-likelihood and its derivatives are

$$\begin{aligned}
\ell_n(\theta) &= \log\theta \sum_i (X_i + 2Y_i) - n(\theta + \theta^2) - \sum_i \log(X_i!Y_i!) \\
\dot{\ell}_n(\theta) &= \frac{1}{\theta}\sum_i (X_i + 2Y_i) - n(1 + 2\theta) \\
\ddot{\ell}_n(\theta) &= -\frac{1}{\theta^2}\sum_i (X_i + 2Y_i) - 2n
\end{aligned}$$

We can calculate the Fisher information as either the variance of the score, in which case

$$J_1(\theta) = \mathrm{Var}_\theta(\dot{\ell}_1(\theta)) = \frac{1}{\theta^2}\mathrm{Var}_\theta(X_i + 2Y_i) = \frac{\theta + 4\theta^2}{\theta^2} = \frac{1 + 4\theta}{\theta},$$

or as minus the expectation of the second derivative, in which case

$$J_1(\theta) = -\mathbb{E}\,\ddot{\ell}_1(\theta) = \frac{1}{\theta^2}\mathbb{E}(X_i + 2Y_i) + 2 = \frac{\theta + 2\theta^2}{\theta^2} + 2 = \frac{1 + 4\theta}{\theta}.$$

Either way, we can apply our usual result about the asymptotic distribution of the MLE to obtain

$$\sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow N\left(0, J_1(\theta)^{-1}\right) = N\left(0, \frac{\theta}{1 + 4\theta}\right).$$

(e) This is a delta method problem for the differentiable function $g(x, y) = (x + \sqrt{y})/2$, $\nabla g(x, y) = \left(\frac{1}{2}, \frac{1}{4\sqrt{y}}\right)$. The argument to g is $(\overline{X}_n, \overline{Y}_n)$, whose distribution is given by

$$\sqrt{n}\left(\begin{pmatrix}\overline{X}_n \\ \overline{Y}_n\end{pmatrix} - \begin{pmatrix}\theta \\ \theta^2\end{pmatrix}\right) \Rightarrow N_2\left(0, \begin{pmatrix}\theta & 0 \\ 0 & \theta^2\end{pmatrix}\right),$$

from the CLT. The correct limiting variance is

$$\nabla g(\theta, \theta^2)' \begin{pmatrix}\theta & 0 \\ 0 & \theta^2\end{pmatrix} \nabla g(\theta, \theta^2) = \frac{\theta}{4} + \frac{1}{16} = \frac{4\theta + 1}{16},$$

and $g(\theta, \theta^2) = \theta$, so

$$\sqrt{n}\left(\tilde{\theta}_n - \theta\right) \Rightarrow N\left(0, \frac{4\theta + 1}{16}\right).$$

The asymptotic relative efficiency is atrocious:

$$\text{ARE}_\theta = \frac{\theta/(1+4\theta)}{(4\theta+1)/16} = \frac{16\theta}{16\theta^2+8\theta+1} = \frac{1}{\theta+1/2+1/\theta},$$

which is *maximized* at 40% when $\theta = 1$, but tends to 0 as θ becomes large *or* small (you did not need to analyze the ARE on your exam, the formula would be enough).

Essentially, this is because as $\theta \to \infty$, $\theta^2 \gg \theta$ so the estimator should be driven by the much more informative $\overline{Y}_n$, but as $\theta \to 0$, $\theta^2 \ll \theta$ so the estimator should be driven by the much more informative $\overline{X}_n$. Using the sufficient statistic $\overline{X}_n + 2\overline{Y}_n$ as our vehicle for estimation (as the UMVU and MLE both do) gets this right, because the one with larger mean will dominate the sum. By contrast $\tilde{\theta}_n$ does a very poor job because it gives both sources of information, $\overline{X}_n$ and $\overline{Y}_n$, an equal voice in determining the "ensemble" estimator.

This automatic, adaptive reweighting of evidence from $\overline{X}_n$ vs. $\overline{Y}_n$ is just the kind of "everyday miracle" that happens when we use the MLE (or even just when we make a sufficiency reduction). We'd have to think very hard to always get this kind of thing right if we had to design an estimator from scratch.

(f) The easiest choice here is the generalized likelihood ratio test. We have already calculated the MLE under the null in part (c), and the log-likelihood at the null MLE is

$$\max_{H_0} \ell_n = \ell_n(\hat{\theta}_n; X, Y) = n(\overline{X}_n + 2\overline{Y}_n)\log\hat{\theta}_n - n\hat{\theta}_n - n\hat{\theta}_n^2 - \sum_{i=1}^n \log(X_i!Y_i!).$$

Under the full model, the MLE for (θ, λ) is just $(\overline{X}_n, \overline{Y}_n)$, so the log-likelihood is

$$\max_{\theta,\lambda} \ell_n = n\overline{X}_n \log\overline{X}_n - n\overline{X}_n - \sum_{i=1}^n \log(X_i!) + n\overline{Y}_n\log\overline{Y}_n - n\overline{Y}_n - \sum_{i=1}^n \log(Y_i!).$$

The GLRT statistic is twice the difference, which we can simplify a bit to

$$G(X) = 2n\left\{\overline{X}_n\left(\log\frac{\overline{X}_n}{\hat{\theta}_n} - 1\right) + \overline{Y}_n\left(\log\frac{\overline{Y}_n}{\hat{\theta}_n^2} - 1\right) - \hat{\theta}_n(1+\hat{\theta}_n)\right\},$$

with $\hat{\theta}_n$ as defined in part (c). The null has one parameter and the alternative has two, so we should reject if $G(X)$ is above the upper α quantile of a χ_1^2 distribution.

2. A problem of limited means (20 points, 5 points / part).

Some useful facts for this problem:

- The uniform density Unif$[a, b]$ with parameters $a < b$ has density
$$\frac{1\{a \leq x \leq b\}}{b-a}, \quad \text{for } x \in \mathbb{R}.$$
Its mean and variance are $(a+b)/2$ and $(b-a)^2/12$, respectively.
- The exponential distribution Exp(λ) with scale parameter λ has density
$$\frac{1}{\lambda}e^{-x/\lambda}, \quad \text{for } x > 0.$$

The Gaussian density is printed in the preamble of Problem 2.

Assume that we are in the Gaussian sequence model with

$$X_i \overset{\text{ind.}}{\sim} N(\mu_i, 1), \qquad \text{for } i = 1, \ldots, d,$$

with the additional assumption that $|\mu_i| \leq \theta$ for some $\theta > 0$. Assume unless specified otherwise that θ is known.

(a) Give the MLE of $\mu_1, \ldots, \mu_d$ in this model.

(b) Give an unbiased estimator for the mean squared error of the MLE, as a function of $X_1, \ldots, X_d$ and θ.

(c) Now, suppose we introduce Bayesian assumptions: we assume additionally that $\mu_i \overset{\text{i.i.d.}}{\sim} \text{Unif}[-\theta, +\theta]$, still with θ known. Give an explicit expression for the Bayes estimator of $\mu_1, \ldots, \mu_d$ using squared error loss.

(d) Now, we relax the assumption that θ is known and introduce a hierarchical Bayesian model with an exponential hyperprior for θ:

$$\begin{aligned} \theta &\sim \text{Exp}(\lambda) \\ \mu_i \mid \theta &\overset{\text{i.i.d.}}{\sim} \text{Unif}[-\theta, +\theta], \quad i = 1, \ldots, d \\ X_i \mid \theta, \mu &\overset{\text{ind.}}{\sim} N(\mu_i, 1), \quad i = 1, \ldots, d. \end{aligned}$$

Suggest a Gibbs sampler algorithm to sample from the posterior distribution of $(\theta, \mu_1, \ldots, \mu_d)$. Give the update rules explicitly.

2. Solution

(a) The likelihood is

$$\ell(\mu; X) = \frac{1}{2}\sum_i (X_i - \mu_i)^2,$$

so the likelihood is maximized by taking $\hat{\mu}_i$ as close to X_i as possible, subject to the constraint that $|\hat{\mu}_i| \le \theta$. As a result,

$$\hat{\mu}_i = \begin{cases} \theta & X_i > \theta \\ X_i & -\theta \le X_i \le \theta \\ -\theta & X_i < -\theta \end{cases}.$$

(b) This is a SURE problem. Defining

$$h_i(x) = x_i - \hat{\mu}_i(x) = \begin{cases} x_i - \theta & x_i > \theta \\ 0 & -\theta \le x_i \le \theta \\ x_i + \theta & x_i < -\theta \end{cases},$$

we have

$$Dh(x)_{ii} = \frac{\partial h_i(x)}{\partial x_i} = \begin{cases} 0 & -\theta \le x_i \le \theta \\ 1 & \text{otherwise} \end{cases}.$$

Then Stein's unbiased risk estimator is

$$R(X) = d + \sum_{i=1}^d h_i(X)^2 - 2\sum_{i=1}^d Dh(X)_{ii}$$

$$= d + \sum_{i=1}^d (|X_i| - \theta)_+^2 - 2\sum_{i=1}^d 1\{|X_i| > \theta\}.$$

(c) Under this model, the pairs (μ_i, X_i) are i.i.d. so the posterior distribution only depends on X_i. The prior for μ_i is $\frac{1}{2\theta}1\{|\mu_i| \le \theta\}$ and the likelihood is $\phi(X_i - \mu_i)$ where $\phi(z) = (2\pi)^{-1/2}e^{-z^2/2}$, so the posterior density is

$$p(\mu_i \mid X) = \frac{\phi(X_i - \mu_i)}{\int_{-\theta}^{\theta}\phi(X_i - u)\,du} \cdot 1\{|\mu_i| \le \theta\},$$

where the $1/2\theta$ term cancels on the top and the bottom. This is the distribution of a $N(X_i, 1)$ random variable truncated to the interval $[-\theta, \theta]$, which we can write as $N(X_i, 1)1\{|\mu_i| \le \theta\}$.

The Bayes estimator is the posterior expectation,

$$\mathbb{E}_\theta[\mu_i \mid X] = \frac{\int_{-\theta}^{\theta} u\,\phi(X_i - u)\,du}{\int_{-\theta}^{\theta} \phi(X_i - u)\,du}.$$

(d) We have just shown that, conditional on (θ, X), the coordinates of μ are truncated normal random variables:

$$\mu_i \mid X, \theta \overset{\text{ind.}}{\sim} N(X_i, 1)1\{|\mu_i| \leq \theta\}.$$

They are conditionally independent because the pairs (μ_i, X_i) are i.i.d. given θ (note they are not marginally independent if θ is random). The conditional density of θ given (μ, X) is proportional to

$$\begin{aligned} p(\theta \mid \mu, X) &\propto \frac{1}{\lambda} e^{-\theta/\lambda} \prod_{i=1}^{d} \frac{1}{2\theta} \cdot 1\{|\mu_i| \leq \theta\} \cdot \phi(X_i - \mu_i) \\ &\propto \theta^{-d}\, e^{-\theta/\lambda} \cdot 1\left\{\theta \geq \max_i |\mu_i|\right\}, \end{aligned}$$

so normalizing it gives

$$p(\theta \mid \mu, X) = \frac{\theta^{-d}\, e^{-\theta/\lambda}}{\int_{\max_i |\mu_i|}^{\infty} u^{-d}\, e^{-u/\lambda}\, du} \cdot 1\left\{\theta \geq \max_i |\mu_i|\right\}. \qquad (1)$$

We could sort of call this a truncated Gamma$(-d+1, \lambda)$ distribution, but the “shape parameter” is negative (note the density wouldn't be normalizable if the lower bound of the support were at 0). Anyway the density is given explicitly above, so we can sample from it by plugging a uniform random variable into its inverse CDF.

So the Gibbs sampler iterates between:

Step 1. Sample $\mu_1^{(t+1)}, \ldots, \mu_n^{(t+1)} \overset{\text{ind.}}{\sim} N(X_i, 1)1\{|\mu_i| \leq \theta^{(t)}\}$.

Step 2. Sample $\theta^{(t+1)}$ from $p(\theta \mid \mu^{(t+1)}, X)$ as given in (1).

I kind of wish I'd told you to use the improper “flat” prior on θ, with $p(\theta) \equiv 1$. Then the Gibbs update for θ would have been exactly a Pareto distribution, which would have been kind of cool. Oh, well!

3. Gamma palooza (25 points, 5 points / part).

Some useful facts for this problem:

- For shape parameter $k > 0$ (not necessarily an integer) and scale parameter $\sigma > 0$, the Gamma(k, σ) distribution has density

$$\frac{1}{\sigma^k \Gamma(k)} x^{k-1} e^{-x/\sigma}, \qquad \text{for } x > 0.$$

  Its mean and variance are $k\sigma$ and $k\sigma^2$, respectively.

- The χ^2_d distribution is Gamma$(d/2, 2)$. It is usually defined when d is an integer, but the density is still a proper density for any $d > 0$. The same is true for distributions derived from the χ^2 like t or F whose "degrees of freedom" argument(s) can take on any positive real value.

Assume that we observe independent random variables X_{ij} with

$$X_{ij} \overset{\text{ind.}}{\sim} \text{Gamma}(k_i, \sigma_j), \qquad \text{for } i = 1, \ldots, n \geq 2, \;\; \text{and } j = 1, 2.$$

Unless otherwise specified, assume all k_i and σ_j are unknown and strictly positive (different parts of the problem will consider simpler submodels). Let $S_j = \sum_{i=1}^n X_{ij}$ and $M_i = X_{i1}X_{i2}$.

(a) Show that $T(X) = (S_1, S_2, M_1, \ldots, M_n)$ is a complete sufficient statistic for this model.

(b) Assume (for this part **only**) that $k_1, \ldots, k_n$ are known. Give an explicit formula for an exact equal-tailed confidence interval for σ_2/σ_1, in terms of the sufficient statistics described above and quantiles for one or more known distributions from class.

(c) Assume instead (for this part **only**) that σ_1 and σ_2 are known, and also it is known that $k_1 = k_2 = \cdots = k_n = k$, but the common value k is unknown. Suggest a UMP test of the hypothesis $H_0 : \; k = k_0$ against the alternative $H_1 : \; k > k_0$, where k_0 is generic. Give the test statistic and explain how to calculate the rejection cutoff (give an explicit recipe that anyone can follow).

(d) Suppose (for this part **only**) that $n = 2$ with all of $k_1, k_2, \sigma_1, \sigma_2$ unknown. Suggest an exact UMPU test of $H_0 : \; k_1 = k_2$ against $H_1 : \; k_1 > k_2$. Say what test statistic you would use and give a precise mathematical description of the rejection cutoff, but you do **not** need to give an explicit expression or recipe for how to calculate it.

(e) (*) Drop all assumptions from previous parts, so n is arbitrary and no parameters of the model are known.

Suppose that we begin doubting the validity of our Gamma model, and we want to generalize it to replace the Gamma family with a generic scale family:

$$X_{ij} \overset{\text{i.i.d.}}{\sim} G_i(x/\sigma_j),$$

for a generic, unknown, continuous distribution function G_i that puts all its mass on positive values of x (i.e., $G_i(0) = 0$). We want to guarantee Type I error control no matter what $G_1, \ldots, G_n$ are.

Explain how to calculate an exact 95% confidence interval for σ_2/σ_1. Your interval must be nontrivial; we will not award any points for answers like "flip a coin and cover the entire parameter space with probability 95%."

(**Hint:** This problem is closely related to testing $H_0 : \sigma_1 = \sigma_2$ against $H_1 : \sigma_1 > \sigma_2$. The testing problem might be easier to think about at first, and partial credit will be awarded for making progress on it.)

3. Solution

(a) The likelihood is

$$\begin{aligned} p_{k,\sigma}(x) &= \prod_{i,j} \frac{1}{\sigma_j^{k_i}\Gamma(k_i)} x_{ij}^{k_i-1} e^{-x_{ij}/\sigma_j} \\ &= \exp\left\{ \sum_{i,j} k_i \log x_{ij} - \sum_{i,j} x_{ij}/\sigma_j - \sum_{i,j} k_i \log \sigma_j - \log\Gamma(k_i) \right\} \frac{1}{\prod_{i,j} x_{ij}} \\ &= \exp\left\{ \sum_{i=1}^n k_i \log M_i + \sum_{j=1}^2 \frac{1}{\sigma_j} S_j - \sum_{i,j} k_i \log \sigma_j - \log\Gamma(k_i) \right\} \frac{1}{\prod_{i,j} x_{ij}}, \end{aligned}$$

and exponential family with natural parameter $\eta = \left(k_1, \dots, k_n, \frac{1}{\sigma_1}, \frac{1}{\sigma_2}\right)$ and sufficient statistic $T(X) = (\log M_1, \dots, \log M_n, S_1, S_2)$. Equivalently, $(M_1, \dots, M_n, S_1, S_2)$ is complete sufficient as well since it is in one-to-one correspondence with $T(X)$. Furthermore, the natural parameter ranges over $\mathbb{R}_+^{n+2}$, which contains an open set, so the model is full rank.

(b) If $k_1, \dots, k_n$ are known then (S_1, S_2) is complete sufficient. We can make a sufficiency reduction, after which we have

$$S_j \overset{\text{ind.}}{\sim} \text{Gamma}(k_+, \sigma_j) = \frac{\sigma_j}{2}\chi^2_{2k_+}, \qquad \text{where } k_+ = \sum_i k_i.$$

Let $\rho = \sigma_2/\sigma_1$ and $R = S_2/S_1$, then

$$R = \rho \frac{S_2/\sigma_2}{S_1/\sigma_1} \sim \rho F_{2k_+, 2k_+}.$$

As a result, if c_1 and c_2 are respectively the lower and upper $\alpha/2$ quantiles of $F_{2k_+,2k_+}$, then

$$1 - \alpha = \mathbb{P}_\rho\left[c_1 \le R/\rho \le c_2\right] = \mathbb{P}_\rho\left[\frac{R}{c_2} \le \rho \le \frac{R}{c_1}\right].$$

So the CI is $\left[\frac{R}{c_2}, \frac{R}{c_1}\right]$ (note the ordering).

(c) If σ_1 and σ_2 are known, and $k_1 = \cdots = k_n = k$, then the likelihood reduces to the one-parameter exponential family,

$$p_k(x) = \exp\left\{ k \sum_i \log M_i - n\log(\sigma_1\sigma_2) - 2n\log\Gamma(k) \right\} \cdot \frac{e^{-S_1/\sigma_1 - S_2/\sigma_2}}{\prod_{i,j} x_{ij}},$$

so we should reject for large values of $P = \prod_{i=1}^n M_i = \prod_{i,j} X_{ij}$, or equivalently for large values of $P_1 = \frac{1}{\sigma_1^n \sigma_2^n} \prod_{i,j} X_{ij}$, whose distribution doesn't depend on σ_1, σ_2. To find the rejection threshold for P_1, which is a product of $2n$ independent Gamma$(k_0, 1)$ random variables under the null, we can repeatedly sample from this distribution and take the upper α quantile of the empirical distribution. This is exact up to Monte Carlo error, which we can make as small as we want.

(d) We can rewrite the likelihood again as

$$\exp\left\{\frac{k_1 - k_2}{2}(\log M_1 - \log M_2) + \frac{k_1 + k_2}{2}(\log M_1 + \log M_2)\right.$$
$$\left. + \frac{1}{\sigma_1} S_1 + \frac{1}{\sigma_2} S_2 - \sum_{i,j} k_i \log \sigma_j - \log \Gamma(k_i)\right\} \frac{1}{\prod_{i,j} x_{ij}},$$

so that the first term in the exponent represents the parameter of interest and the other three represent the nuisance parameters $(k_1 + k_2, \sigma_1, \sigma_2)$. The UMPU test, then, rejects for large values of M_1/M_2, conditional on $(M_1 M_2, S_1, S_2)$.

(e) If $\rho = \sigma_2/\sigma_1$, then

$$X_{i1},\ X_{i2}/\rho \overset{\text{i.i.d.}}{\sim} G_i(x/\sigma_1), \quad \text{independently for } i = 1, \ldots, n.$$

We could use this to construct a permutation test of $H_0 : \ \rho = \rho_0$. A sufficient statistic for the null model consists of the unordered sets $\{X_{i1}, X_{i2}/\rho_0\}$ for $i = 1, \ldots, n$. Write that statistic as $U(X)$; then we can sample from the distribution of X given $U(X)$ by randomly permuting within each pair $(X_{i1}, X_{i2}/\rho_0)$ independently for each i. Under alternative values of $\rho > \rho_0$, we will tend to have $X_{i2}/\rho_0 > X_{i1}$, so we can pick any test statistic that will tend to be large when that is the case. (Getting everything right up to here would be enough for 4 points out of the possible 5).

If we pick a test statistic that we can easily evaluate for all the different candidate values of ρ_0, then we can get a confidence interval too. There are a lot of possible choices but a robust and convenient one would be

$$\begin{aligned} B(X; \rho_0) &= \sum_{i=1}^{n} 1\{X_{i2}/\rho_0 > X_{i1}\} \\ &= \sum_{i=1} 1\{X_{i2}/X_{i1} > \rho_0\} \\ &\overset{H_0}{\sim} \text{Binom}\,(n, 1/2)\,, \end{aligned}$$

because the exchangeability of $(X_{i1}, X_{i2}/\rho_0)$ under the null means each indicator has an equal 50% chance to be 1 or 0. Let $R_i = X_{i2}/X_{i1}$, then this is saying if $\rho = \rho_0$ then the sample median value of R_i should be about ρ_0, and we reject if too many R_i values are above ρ_0. If we want to make the test two-sided, then we should reject if too many *or* too few R_i values are above ρ_0. This test statistic is convenient in part because we don't have to do Monte Carlo sampling; we just know the null distribution of B and it doesn't even depend on U.

Ignoring randomizing at the boundary, assume $b_1 \geq 1$ and $b_2 = n - b_1$ are respectively the lower and upper $\alpha/2$ quantiles of Binom$(n, 1/2)$ (if $b_1 = 0$ then n is too small for the binomial test to ever reject, so we have to randomize at the boundary or use a different approach). Then the two-sided permutation test *fails* to reject at level α if

$$b_1 \leq \sum_{i=1} 1\{R_i > \rho_0\} \leq n - b_1 \iff R_{(n+1-b_1)} \leq \rho_0 < R_{(b_1)},$$

where $R_{(1)} > R_{(2)} > \cdots > R_{(n)}$ are the order statistics of $R_1, \ldots, R_n$. As a result, $[R_{(n+1-b_1)}, R_{(b_1)}]$ is a valid confidence interval for ρ (note $\mathbb{P}_\rho(R_{(b_1)} = \rho) = 0$ so we can use the closed interval without affecting the coverage).

4. Apocalypse τ (25 points, 5 points / part).

Some useful facts for this problem:

- For $\sigma^2 > 0$ and $\mu \in \mathbb{R}$, the Gaussian density for $X \sim N(\mu, \sigma^2)$ is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad \text{for } x \in \mathbb{R}.$$

  Its mean and variance are μ and σ^2.

Assume we observe i.i.d. pairs (X_i, Y_i) for $i = 1, \ldots, n$, where $X_1, \ldots, X_n \in \mathbb{R}^k$ are sampled from a known density $q(x)$ and Y_i are real numbers with

$$Y_i = f_\tau(X_i) + \varepsilon_i, \qquad \text{where } \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

Assume the errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent of $X_1, \ldots, X_n$.

The parameters $\tau \in [-1, 1]$ and $\sigma^2 > 0$ are fixed and unknown, but the real-valued function $f_\tau(x)$ is known up to its parameter τ.

Assume that

- $f_\tau(x)$ is infinitely differentiable *with respect to* τ, with first and second derivatives

$$g_\tau(x) = \frac{\partial f}{\partial \tau}(x), \qquad \text{and } h_\tau(x) = \frac{\partial^2 f}{\partial \tau^2}(x).$$

- $g_\tau(x) > 0$ for all τ and x.
- $|g_\tau(x)|, |h_\tau(x)| \leq 1$ for all τ and x.

(a) Assume (for this part **only**) that X_i are fixed instead of random, while the errors still have the same distribution, $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$.

Consider testing $H_0 : \tau = 0$ against the alternative $H_1 : \tau \neq 0$ using the test statistic

$$T = \frac{\sum_{i=1}^n g_0(X_i)(Y_i - f_0(X_i))}{\hat{\sigma}\left(\sum_{i=1}^n g_0(X_i)^2\right)^{1/2}},$$

where

$$\hat{\sigma}^2 = \frac{1}{d}\left[\sum_{i=1}^n (Y_i - f_0(X_i))^2 - \frac{\left[\sum_{i=1}^n g_0(X_i)(Y_i - f_0(X_i))\right]^2}{\sum_{i=1}^n g_0(X_i)^2}\right]$$

What number should we plug in for d? Give the distribution of T under the null, and justify your answer.

(b) Now go back to assuming that X_i are random, sampled i.i.d. from an unknown distribution. Show that the test from part (a) still works; i.e. its distribution under the null is independent of $X_1, \ldots, X_n$.

(c) Assume (for this part **only**) that σ^2 is known. Show that the MLE $\hat{\tau}_n$ is consistent for τ as $n \to \infty$. (For full credit, please check appropriate conditions).

(d) Continue to assume (for this part **only**) that σ^2 is known. Assuming the MLE is consistent, and $\tau \in (-1, 1)$ (i.e. not at the boundary of the parameter space), find its asymptotic distribution as $n \to \infty$. (You do not need to check conditions for this).

(e) (*) Suppose we add an intercept to the model, so

$$Y_i \mid X_1, \ldots, X_n \overset{\text{ind.}}{\sim} N(\alpha + f_\tau(X_i), \sigma^2).$$

Can we still estimate τ consistently as $n \to \infty$ using maximum likelihood? Prove or give a counterexample.

4. Solution

(a) Let $z = (g_0(X_1), \ldots, g_0(X_n)) \in \mathbb{R}^n$, which is a fixed nonzero vector, and define the unit vector $q_1 = z/\|z\|$. Under H_0, $Y_i - f_0(X_i) = \varepsilon_i$, so the test statistic is

$$T = \frac{z'\varepsilon}{\hat{\sigma}\|z\|} = \frac{q_1'\varepsilon}{\sqrt{\hat{\sigma}^2}},$$

and the variance estimator is

$$\hat{\sigma}^2 = \frac{1}{d}\left[\|\varepsilon\|^2 - (q_1'\varepsilon)^2\right] = \frac{1}{d}\left[\varepsilon'(I_n - q_1 q_1')\varepsilon\right] = \frac{1}{d}\|Q_r'\varepsilon\|^2,$$

where $Q_r \in \mathbb{R}^{n\times(n-1)}$ is chosen so $Q_r Q_r' = I_n - q_1 q_1'$, an $(n-1)$-dimensional projection matrix. Furthermore, $Q_r' q_1 = 0$ so $Q_r'\varepsilon$ and $q_1'\varepsilon$ are independent, with $q_1'\varepsilon \sim N(0, \sigma^2)$ and $\|Q_r'\varepsilon\|^2 \sim \sigma^2\chi^2_{n-1}$, under the null, so we should take $d = n-1$ and

$$T = \frac{N(0,\sigma^2)}{\sqrt{\frac{\sigma^2}{n-1}\chi^2_{n-1}}} \sim t_{n-1}.$$

(b) If we condition on $X_1, \ldots, X_n$, we are back in the same situation as in part (a), and we have just derived that T is conditionally t_{n-1} distributed, given $X_1, \ldots, X_n$. If the conditional distribution of T given X doesn't depend on X, then T is independent of X.

(c) Because $[-1, 1]$ is a compact parameter space, to apply our theorem from class we only need to establish that the model is identifiable, and

$$\mathbb{E}_{\tau_0}\left[\sup_{\tau\in[-1,1]} |\ell_1(\tau) - \ell_1(\tau_0)|\right] < \infty.$$

The log-likelihood and its derivative for a single pair (X_i, Y_i) is

$$\begin{aligned}\ell_1(\tau; X_i, Y_i) &= \frac{1}{2\sigma^2}(Y_i - f_\tau(X_i))^2 - \frac{1}{2}\log(2\pi\sigma^2) + q(X_i)\\ \dot{\ell}_1(\tau; X_i, Y_i) &= \frac{1}{\sigma^2} g_\tau(X_i)(Y_i - f_\tau(X_i))\end{aligned}$$

Because $|g_\tau(X_i)| \leq 1$, the first derivative is bounded by

$$\begin{aligned}|\dot{\ell}_1(\tau; X_i, Y_i)| &\leq \frac{|Y_i - f_\tau(X_i)|}{\sigma^2}\\ &\leq \frac{|Y_i - f_{\tau_0}(X_i)|}{\sigma^2} + \frac{|f_\tau(X_i) - f_{\tau_0}(X_i)|}{\sigma^2}\\ &\leq \frac{|\varepsilon_i| + 2}{\sigma^2},\end{aligned}$$

where the last step uses the fact that

$$|f_{\tau_2}(x) - f_{\tau_1}(x)| \leq |\tau_2 - \tau_1| \sup_{\tau \in [-1,1]} |g_\tau(x)| \leq 2,$$

for any x and $\tau_1, \tau_2 \in [-1, 1]$. As a result, for any $\tau_0 \in [-1, 1]$, we have

$$\begin{aligned}\mathbb{E}_{\tau_0}\left[\sup_{\tau \in [-1,1]} |\ell_1(\tau) - \ell_1(\tau_0)|\right] &\leq \mathbb{E}_{\tau_0}\left[\sup_{\tau \in [-1,1]} |\tau - \tau_0| \cdot \frac{|\varepsilon_i| + 2}{\sigma^2}\right] \\ &\leq \mathbb{E}_{\tau_0}\left[\frac{2|\varepsilon_i| + 4}{\sigma^2}\right],\end{aligned}$$

which is certainly finite.

As for identifiability, consider $\tau_2 > \tau_1$ and note

$$f_{\tau_2}(x) - f_{\tau_1}(x) = (\tau_2 - \tau_1) g_{\tilde{\tau}(x)}(x),$$

where $\tilde{\tau}(x) \in [\tau_1, \tau_2]$ is defined implicity by the mean value theorem. Then

$$\mathbb{E}_{\tau_2} Y_i - \mathbb{E}_{\tau_1} Y_i = \mathbb{E}\left[f_{\tau_2}(X_i) - f_{\tau_1}(X_i)\right] = (\tau_2 - \tau_1)\mathbb{E}[g_{\tilde{\tau}(X_i)}(X_i)] > 0,$$

since $g_{\tilde{\tau}(X_i)}(X_i) > 0$ almost surely (we will accept more informal arguments along the same lines).

(d) We have assumed consistency, and that τ is in the interior of the parameter space. The other conditions can be checked using similar methods as in the previous part, but you did not need to check them. Continuing our calculation from part (c), the second derivative of ℓ_1 is

$$\ddot{\ell}_1(\tau; X_i, Y_i) = \frac{1}{\sigma^2}\left\{h_\tau(X_i)(Y_i - f_\tau(X_i)) - g_\tau(X_i)^2\right\},$$

and its expectation is

$$J_1(\tau) = \mathbb{E}_\tau\left[\mathbb{E}_\tau[\frac{1}{\sigma^2}\left\{h_\tau(X_i)(Y_i - f_\tau(X_i)) - g_\tau(X_i)^2\right\} \mid X_i]\right] = \frac{1}{\sigma^2}\mathbb{E}_\tau g_\tau(X_i)^2.$$

As a result, the asymptotic distribution of the MLE is

$$\sqrt{n}\,(\hat{\tau}_n - \tau) \Rightarrow N\left(0,\ \frac{\sigma^2}{\mathbb{E}_\tau g_\tau(X_i)^2}\right)$$

(e) No, we can't (necessarily) estimate τ consistently anymore. The problem is that the intercept breaks our proof of identifiability. Indeed, let $f_\tau(x) \equiv \tau$, which satisfies all of the conditions in the preamble since $g_\tau(x) \equiv 1$ and $h_\tau(x) \equiv 0$; then τ is unidentifiable in the model so there is no way we can hope to estimate it using the MLE or any other method.

To get more specific in terms of the MLE, the likelihood function $\ell_n(\alpha - \tau, \tau; X_i, Y_i)$ is constant for $\tau \in [-1, +1]$ and any (α, τ) with $\alpha + \tau = \overline{Y}_n$ is a valid MLE, so clearly the second coordinate can't be converging to the correct value of τ for any sequence of MLEs.