Vanderbilt University Biostatistics Comprehensive Examination

PhD Theory Exam Series 2

May 21–May 24, 2024

Instructions: Please adhere to the following guidelines:

- This exam is scheduled to be administered on Tuesday, May 21 at 9:00am, and will be due on Friday, May 24 at 5:00pm. This deadline is strict: late submissions will not be accepted.
- To turn in your exam, please use your assigned Box folder and e-mail your word-processed exam to Dr. Andrew Spieker by the deadline. This level of redundancy is designed to ensure that your exam is received by the deadline. If you would like to e-mail exam drafts along the way, that is perfectly acceptable—do not be concerned about spamming my inbox.
- There are four problems. Note that not all questions and their sub-questions are weighted equally. You are advised to pace yourself and to not spend too much time on any one problem.
- Answer each question clearly and to the best of your ability. Partial credit will be awarded for partially correct answers.
- Be as specific as possible in your responses.
- You may consult reference material (e.g., course notes, textbooks), though the work you turn in must be your own (this means no generative AI). This is an *individual effort.* Do not communicate about the exam with anyone. Vanderbilt University's academic honor code applies.
- Please direct clarifying questions by e-mail to Dr. Andrew Spieker, Dr. Bob Johnson, and Dr. Amir Asiaee.

---

1. $\boxed{\text{25 pts}}$ **Background**: A random process $\mathcal{C}(N) = \{N(t), t \in [0, \infty)\}$ is said to be a counting process if $N(t)$ is the number of events occurring from time 0 up to and including time t. For a counting process, we assume $N(0) = 0$. A counting process $\mathcal{C}(N)$ is called a Poisson process with rate $\lambda > 0$ (fixed) if all of the following conditions hold:

   - $N(0) = 0$,
   - $\mathcal{C}(N)$ has independent increments (times between sequential events), and
   - the number of events in any interval of length $\tau > 0$ has Poisson$(\lambda\tau)$ distribution.

   (a) Consider a Poisson process with rate λ. Let T_1 be the "arrival" time of the first event and T_n be the inter-arrival time between the $(n-1)^{\text{st}}$ and the n^{th} events. Show that $\{T_n : n = 1, 2, \ldots\}$ are independently and identically distributed exponential random variables with parameter λ.

   (b) Does a Poisson process have stationary increments? Explain your answer.

   (c) Let $Y_n \sim$ Binomial$(n, \lambda/n)$ where $\lambda > 0$. Show that $Y_n \xrightarrow{d} Y \sim$ Poisson(λ) using characteristic functions.

   (d) Argue that a counting process, $\mathcal{C}(M)$, with the following properties is a Poisson process.

   - $M(0) = 0$;
   - $\mathcal{C}(M)$ has independent and stationary increments; and
   - $\mathrm{P}\{M(\Delta) = 0\} = 1 - \lambda\Delta + o(\Delta)$,
$\mathrm{P}\{M(\Delta) = 1\} = \lambda\Delta + o(\Delta)$, and
$\mathrm{P}\{M(\Delta) \geq 2\} = o(\Delta)$

   for $\Delta > 0$ and fixed $\lambda > 0$. (Recall that the *little o* notation, $o(\Delta)$, may replace some $h(\Delta)$ if $h(\Delta)$ is negligible compared to Δ as $\Delta \to 0$; that is, $h(\Delta)/\Delta \to 0$ as $\Delta \to 0$).

   (e) Consider again the process defined in (a). Let $G_k = \sum_{i=1}^{k} T_i$, the time to the k^{th} event.

   [i] Plot the sequence $\{G_n\}$ up to $n = 1000$. Generate data using the following code:

```
1   n=1000; lambda = 1
2   set.seed(1395271)
3   G=c(0,cumsum(rexp(n,rate=lambda)))
```

   Discuss the plot. Is it helpful in viewing the properties of the sequence?

   [ii] Prove that $G_k \sim$ Gamma(k, λ). What are the mean and variance of G_k? Determine Cov$[G_k, G_m]$.

   [iii] Could we have just as well replaced the third line of the code in (e)[i] with the following code: `G=c(0,rgamma(1:n,1:n,rate=lambda))`? Explain your answer.

   [iv] We want to show in a figure where the sequence is potentially *out of control* by noting where G_n is above or below $\mathrm{E}[G_n] \pm 2\sqrt{\mathrm{Var}[G_n]}$. To simplify this, redraw the plot in (e)[i] after centering each G_n; that is, plot $G_n - \mathrm{E}[G_n]$. Include red curves (use `lwd=3`) that are ± 2 standard deviations from 0. Discuss the plot. Did the sequence remain in *control* up to $n = 1000$?

   [v] How does this stochastic sequence relate to the standard Brownian motion?

   [vi] What is the probability (or approximate probability) that the <u>centered</u> sequence first passes the horizontal line at 25 no later than the 750^{th} step in the sequence? Use the following to add the line to your last figure: `abline(h=25,lty=2,col="blue",lwd=2)`. You may use simulation to estimate and check your result, but you should provide an estimate using Brownian motion.

2. $\boxed{\text{25 pts}}$ Suppose $X_1, \ldots, X_n$ are i.i.d. random variables having the common distribution function F and density function f that you may assume in this problem to have a continuous first derivative. Let $\widehat{F}_n$ denote the empirical distribution function of the X_i's, and let $\{a_n\}_{n=1}^{\infty}$ denote some sequence of positive numbers. Consider the following estimator of f:

$$\widehat{f}_n(x) = \frac{\widehat{F}_n(x + a_n) - \widehat{F}_n(x - a_n)}{2a_n}.$$

(a) Argue that $Q_n(x) = 2na_n\widehat{f}_n(x) \sim \text{Binomial}(n, p_n(x))$, where $p_n(x) = F(x + a_n) - F(x - a_n)$.

(b) Determine $\mathrm{E}[\widehat{f}_n(x)]$, and show that $\mathrm{E}[\widehat{f}_n(x)] \longrightarrow f(x)$ if $a_n \longrightarrow 0$.

(c) Determine $\mathrm{Var}[\widehat{f}_n(x)]$, and show that $\mathrm{Var}[\widehat{f}_n(x)] \longrightarrow 0$ if $a_n \longrightarrow 0$ and $na_n \longrightarrow \infty$.

(d) Suppose again that $a_n \longrightarrow 0$ and $na_n \longrightarrow \infty$. Use the Lyapunov Central Limit Theorem to argue that:

$$\frac{2na_n\Big(\widehat{f}_n(x) - \mathrm{E}[\widehat{f}_n(x)]\Big)}{\sqrt{np_n(1 - p_n)}} \xrightarrow{d} \mathcal{N}(0, 1).$$

(e) Argue that if $n^{1/2}a_n^{3/2} \longrightarrow C \in [0, \infty)$, we can push the result of part (d) further as follows:

$$\sqrt{2na_n}\left(\frac{\widehat{f}_n(x) - f(x)}{\sqrt{\widehat{f}_n(x)}}\right) \xrightarrow{d} \mathcal{N}(0, 1).$$

Use this result to determine the form of a confidence interval for $f(x)$ that would be asymptotically valid for, e.g., sequences of the form $a_n = n^{-r}$, $1/3 < r < 1$.

(f) Suppose that $F(x) = 1 - \exp(-x)$, with $n = 100$. Below is sample code; run it line-by-line and be certain you understand each step. Present and comment on the graphical output.

```
set.seed(2024)
n <- 100
zz <- ecdf(rexp(n, rate = 1))
x <- seq(0,8,0.001)
an1 <- n^(-1/200)
an2 <- n^(-1/2)
fn1.hat <- (zz(x + an1) - zz(x - an1))/(2*an1)
fn2.hat <- (zz(x + an2) - zz(x - an2))/(2*an2)
plot(x, fn1.hat, type = "l", ylim = c(0, 1))
lines(x, dexp(x), col = "blue", lwd = 2)
plot(x, fn2.hat, type = "l", ylim = c(0, 1))
lines(x, dexp(x), col = "blue", lwd = 2)
```

(g) Again consider the case in which $F(x) = 1 - \exp(-x)$. Conduct a simulation study in which you vary the simulation parameters as follows:

- Sample sizes: $n = 10^2$, $n = 10^3$, and $n = 10^4$.
- Sequences: $a_n = n^{-3/4}$, $a_n = n^{-1/3}$, and $a_n = n^{-1/10}$.
- Values of x at which to estimate $f(x)$: $x = 0.25$, $x = 1$, and $x = 4$.

Present and compare the following finite-sample properties of $\widehat{f}_n(x)$, accounting for your findings:

- The average values of $\widehat{f}_n(x)$, $\sqrt{2na_n}(\widehat{f}_n(x) - \mathrm{E}[\widehat{f}_n(x)])$, and $\sqrt{2na_n}(\widehat{f}_n(x) - f(x))$ at each x.
- The empirical standard errors of $\widehat{f}(x)$ across simulation replicates.
- The coverage of a 95% confidence interval for $f(x)$, formed based on the result of part (e).

Please use a total of $M = 10{,}000$ simulation replicates per setting. You can use graphical and/or tabular methods to present your results; this problem is open-ended. Include your R code as an appendix.

---

3. $\boxed{\text{20 pts}}$ This problem aims to enrich your understanding about how the ridge penalty affects the leverage of individual observations in a simple linear regression model, and further seeks to elucidate what can go wrong if you fail to center a predictor prior to regularization. To that end, consider the setting in which you seek to estimate shrunken coefficients from the simple linear regression model $\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x$ via the ridge penalty. For simplicity, and without any serious loss to generality, consider X to be uniformly distributed between 0 and 1. Given a sample size of $n > 2$, define the leverage for an observation $\boldsymbol{x} = [1 \quad x]$ as:

$$P_\lambda(x) = \boldsymbol{x}^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{x},$$

where $\lambda \geq 0$ marks the penalty and $\mathbf{X}$ is the $n \times 2$ design matrix for the uncentered data. Throughout this problem, you may freely use without proof the following two facts:

- The graph of $y = ax^2 + bx + c$ $(a \neq 0)$ corresponds to a parabola with vertex occurring at $x = -b/2a$.
- The matrix products $\mathbf{AB}$ and $\mathbf{BA}$ have the same eigenvalues ($\mathbf{A}$ and $\mathbf{B}$ must clearly be square and of the same dimension for them to be conformable for multiplication in both directions).

(a) Determine the value of x, call it x_λ, at which $P_\lambda(x)$ is minimized. Conclude that $x_\lambda < x_0$ for $\lambda > 0$.

(b) Prove as a lemma to part (c) that if $\mathbf{A}$ and $\mathbf{B}$ are positive definite matrices of the same dimension, then $\mathbf{A} \succ \mathbf{B}$ implies that $\mathbf{B}^{-1} \succ \mathbf{A}^{-1}$. Please recall that the notation $\mathbf{A} \succ \mathbf{B}$ is a shorthand way to communicate that $\mathbf{A} - \mathbf{B}$ is a positive definite matrix.

(c) Show that for $x \in (0, 1)$, $P_\lambda(x) > P_{\lambda'}(x) > 0$ if $\lambda' > \lambda \geq 0$. Confirm this by running the following code (which might also help you with subsequent parts of this problem):

```
P <- function(x, X, lambda = 0) {
  x.t <- matrix(cbind(1, x), ncol = 2)
  p <- x.t %*% solve(t(X) %*% X + lambda * diag(2)) %*% t(x.t)
  return(diag(p))
}

set.seed(2024)
n <- 100
X <- cbind(1, runif(n, 0, 1))
x.p <- seq(0,1,0.01)

plot(x.p, P(x.p, X = X, lambda = 0), frame.plot = FALSE, xlab = "x",
     ylab = "Leverage", type = "l", lwd = 2, ylim = c(0, 0.04))
lines(x.p, P(x.p, X = X, lambda = 5))
lines(x.p, P(x.p, X = X, lambda = 10))
lines(x.p, P(x.p, X = X, lambda = 20))
```

(d) Argue that for $\lambda > 0$, $P_\lambda(x)$ is not a function of $P_0(x)$. A response relying on proper graphical reasoning will be considered sufficient for this problem (for instance, you may wish to include a graph and label it in a way that illustrates your point).

(e) Characterize the behavior of $P_\lambda(x)$ as $\lambda \nearrow \infty$ (i.e., for a fixed $n > 2$).

(f) Characterize the behavior of $P_\lambda(x)$ as $n \nearrow \infty$ (i.e., for a fixed $\lambda > 0$).

(g) Comment on the pragmatic implications of your findings in this problem; your answer can be heuristic and conceptual, but it should be thoughtful. If you need a starting point in crafting a response, re-read the first sentence of the problem description. A thoughtful response will consider how the answers to previous parts of the problem might change if the X's are centered in advance to have mean zero.

---

4. $\boxed{\text{30 pts}}$ It is often of interest to predict multiple outcomes from a common set of predictors. Though each outcome could be modeled as a distinct regression task, there may be between-outcome correlations. Consider a data set with N independent observations, each having D features and T outcomes. Let y_{nt} denote the t^{th} outcome for the n^{th} observation, and let x_{nd} represent the d^{th} feature for the n^{th} observation. Assuming the outcomes are linearly dependent on the features, the relationship can be modeled as:

$$y_{nt} = \sum_{d=1}^{D} x_{nd} b_{dt} + e_{nt} = \boldsymbol{x}_n^T \mathbf{b}_t + e_{nt},$$

where $\boldsymbol{x}_n, \mathbf{b}_t \in \mathbb{R}^D$, and e_{nt} is random noise. The data set comprises pairs of input-output vectors $\mathcal{D} = \{(\boldsymbol{x}_n, \mathbf{y}_n)\}_{n=1}^N$, with $\boldsymbol{x}_n \in \mathbb{R}^D$ and $\mathbf{y}_n \in \mathbb{R}^T$. The linear model in matrix form is expressed as:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \quad \mathbf{Y} \in \mathbb{R}^{N \times T}, \ \ \mathbf{X} \in \mathbb{R}^{N \times D}, \ \ \mathbf{E} \in \mathbb{R}^{N \times T}, \ \text{ and } \ \mathbf{B} \in \mathbb{R}^{D \times T}.$$

The noise vectors $\mathbf{e}_n$ are assumed to be multivariate normal with mean zero and a covariance matrix $\boldsymbol{\Sigma}$—that is, $\mathbf{e}_n \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ denote the precision matrix.

(a) Derive the negative log-likelihood $\mathrm{NNL}_{\mathcal{D}}(\mathbf{B}, \boldsymbol{\Omega}) \equiv -\log \mathcal{L}_{\mathcal{D}}(\mathbf{B}, \boldsymbol{\Omega})$, and simplify by removing non-essential terms.

(b) Treating the precision matrix, $\boldsymbol{\Omega}^*$, as known, derive the closed-form solution for $\widehat{\mathbf{B}}$, which minimizes $\mathrm{NNL}_{\mathcal{D}}(\mathbf{B}, \boldsymbol{\Omega}^*)$. Demonstrate that $\widehat{\mathbf{B}}$ does not depend upon $\boldsymbol{\Omega}^*$, effectively reducing the estimation to T independent ordinary least squares problems. *Hint*: Use the trace trick.

(c) Introduce a Frobenius-norm penalty of the matrix $\mathbf{B}$ to the negative log-likelihood as a way to mitigate overfitting. Call the objective function $\mathrm{PNLL}_{\mathcal{D}}(\mathbf{B}, \boldsymbol{\Omega}^*)$, for "penalized negative log-likelihood." Derive an equation that characterizes $\widehat{\mathbf{B}}$ under this regularization (a closed-form derivation is not necessary). Illustrate that the resulting penalized MLE solution is not equivalent to T independent ridge regressions. Specifically, demonstrate how the coefficients are coupled across tasks via $\boldsymbol{\Omega}^*$.

(d) Consider the scenario in which both $\boldsymbol{\Omega}$ and $\mathbf{B}$ are unknown. Is is known that $\mathrm{PNNL}_{\mathcal{D}}(\mathbf{B}, \boldsymbol{\Omega})$ is *not* jointly convex with respect to these variables.

   [i] Demonstrate that when $\boldsymbol{\Omega}$ is fixed, $\mathrm{PNLL}_{\mathcal{D}}$ is convex in $\mathbf{B}$. *Hint*: Note that the variables here are matrices and although the first derivative with respect to a matrix is easy, the second derivative required to show convexity is complicated. For that, you can vectorize the variables and use the Kronecker product identity: $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\mathrm{vec}(\mathbf{B})$, where $\otimes$ is the Kronecker product.

   [ii] When $\mathbf{B}$ is fixed, $\mathrm{PNLL}_{\mathcal{D}}$ is convex in $\boldsymbol{\Omega}$, a fact you are free to use without further proof. Based on these convexity properties, propose a gradient descent-based approach to find a local minimum for the penalized maximum likelihood estimation described in part (c). You should compute the gradients for the updates.

(e) Given the challenges of estimating $\boldsymbol{\Omega}$ in high-dimensional settings with limited samples, it becomes necessary to assume a simpler structure for $\boldsymbol{\Omega}$ using regularization norms.

   [i] Discuss and compare two regularization approaches: the nuclear norm, $\|\boldsymbol{\Omega}\|_{\text{nuc}} \equiv \sum_{i=1}^T \sigma_i(\boldsymbol{\Omega})$, and the ℓ_1-norm, $\|\boldsymbol{\Omega}\|_1 \equiv \sum_{i=1}^T \sum_{j=1}^T |\omega_{ij}|$. Based on their properties and their implications for the estimated precision matrix, argue which norm is more appropriate and why.

   [ii] Demonstrate that $f(\boldsymbol{\Omega}) \equiv \sum_{t=1}^T \|\boldsymbol{\Omega}_{t,:}\|_2$ qualifies as a norm (here, $\boldsymbol{\Omega}_{t,:}$ is the t^{th} row of the matrix $\boldsymbol{\Omega}$) and discuss what type of prior belief about the interrelationships between tasks is reflected by this norm.
