Department of Statistics

University of Wisconsin, Madison

PhD Qualifying Exam Option B

12:30-4:30pm, Room 133 SMI

- There are a total of FOUR (4) problems in this exam. Please do all FOUR (4) problems.
- Each problem must be done in a separate exam book.
- Please turn in FOUR (4) exam books.
- Please write your code name and **NOT** your real name on each exam book.

1. This problem investigates fixed design linear regressions in high-dimensional settings. Consider a study of human heritability where the goal is to estimate the contribution of many genetic variants (e.g., single nucleotide polymorphisms) to a quantitative phenotypic trait (e.g., height). Suppose that the study sample consists of n individuals. For each individual $i \in \{1, \dots, n\}$, we observe a pair of measurements $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the genotype vector across d genetic variants, and $y_i \in \mathbb{R}$ denotes the scalar-valued phenotype.

   Let $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$ denote the design matrix and $\mathbf{y} = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ be the response vector. Consider a linear model with i.i.d. mean-zero Gaussian noise

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n}), \tag{1}$$

   where $\boldsymbol{\beta} \in \mathbb{R}^d$ is the unknown coefficient and $\mathbf{I}_{n\times n}$ is an n-by-n identity matrix. Modern genomic dataset is often high-dimensional; that is, the number of features d is comparable, or even larger than, the sample size n. For simplicity, we assume $d = n$ and $\mathbf{X}$ has orthonormal columns. Consider the regularized estimator for $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{p}\|\boldsymbol{\beta}\|_p \right\}, \tag{2}$$

   where $\|\cdot\|_p$ denotes the vector p-norm; i.e., $\|\mathbf{a}\|_p = \left(\sum_{j=1}^d |a_j|^p\right)^{1/p}$ for a vector $\mathbf{a} = (a_1, \dots, a_d)^T \in \mathbb{R}^d$.

   The following questions consider $p = 1$ or 2 and $\lambda \geq 0$.

   (a) Let $\lambda = 0$.

      i. Give the distribution for $\hat{\boldsymbol{\beta}}$, the solution to the least-squares optimization (2).

      ii. Consider the prediction error for a new observation of the form $y_{\text{new}} = \mathbf{x}_{\text{new}}^T\boldsymbol{\beta} + \varepsilon$, for an arbitrary, fixed vector $\mathbf{x}_{\text{new}} \in \mathbb{R}^d$ and independent noise $\varepsilon \sim \mathcal{N}(0, 1)$. Find the expected squared prediction error, $\mathbb{E}(y_{\text{new}} - \mathbf{x}_{\text{new}}^T\hat{\boldsymbol{\beta}})^2$.

   Solution:

   i. The least-squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y}) = \mathbf{X}^T\mathbf{y}$, where we have used the fact that $\mathbf{X}^T\mathbf{X} = \mathbf{I}$ for orthogonal matrices. Plugging the model (1) into the estimator yields

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta},\ \mathbf{I}_{n\times n}).$$

ii. The expected squared prediction error is

$$\begin{aligned}\mathbb{E}(y_{\text{new}} - \mathbf{x}_{\text{new}}^T \hat{\boldsymbol{\beta}})^2 &= \mathbb{E}\left(\mathbf{x}_{\text{new}}^T \boldsymbol{\beta} + \varepsilon - \mathbf{x}_{\text{new}}^T \hat{\boldsymbol{\beta}}\right)^2 \\ &= \mathbb{E}\varepsilon^2 + \mathbb{E}\left(\mathbf{x}_{\text{new}}^T \hat{\boldsymbol{\beta}} - \mathbf{x}_{\text{new}}^T \boldsymbol{\beta}\right)^2 \\ &= 1 + \operatorname{Var}(\mathbf{x}_{\text{new}}^T \hat{\boldsymbol{\beta}}) \\ &= 1 + \|\mathbf{x}_{\text{new}}\|_2^2.\end{aligned}$$

(b) Let $p = 2$ and $\lambda > 0$.

i. Give an expression for $\hat{\boldsymbol{\beta}}^{\text{ridge}}$, the solution to the penalized least-squares optimization (2) in this case.

ii. Consider the prediction error for a new observation of the form $y_{\text{new}} = \mathbf{x}_{\text{new}}^T \boldsymbol{\beta} + \varepsilon$, for an arbitrary, fixed vector $\mathbf{x}_{\text{new}} \in \mathbb{R}^d$ and independent noise $\varepsilon \sim \mathcal{N}(0, 1)$. Find the expected squared prediction error, $\mathbb{E}(y_{\text{new}} - \mathbf{x}_{\text{new}}^T \hat{\boldsymbol{\beta}}^{\text{ridge}})^2$. Compare the result to part (a).

Solution:

i. The solution to the penalized least-squares optimization is

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \frac{1}{1+\lambda} \mathbf{X}^T \mathbf{y}. \tag{3}$$

In the special case when $\lambda = 0$, the solution reduces to the MLE in part (a).

ii. Following the calculation in part a(ii), we have

$$\mathbb{E}(y_{\text{new}} - \mathbf{x}_{\text{new}}^T \hat{\boldsymbol{\beta}}^{\text{ridge}})^2 = 1 + \mathbb{E}\left(\mathbf{x}_{\text{new}}^T \hat{\boldsymbol{\beta}}^{\text{ridge}} - \mathbf{x}_{\text{new}}^T \boldsymbol{\beta}\right)^2. \tag{4}$$

Now, the estimator (3) implies that

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} \sim \mathcal{N}\left(\frac{1}{1+\lambda}\boldsymbol{\beta},\ \frac{1}{(1+\lambda)^2}\mathbf{I}\right).$$

Hence,

$$\mathbf{x}_{\text{new}}^T \hat{\boldsymbol{\beta}}^{\text{ridge}} \sim \mathcal{N}\left(\frac{1}{1+\lambda}\mathbf{x}_{\text{new}}^T \boldsymbol{\beta},\ \frac{1}{(1+\lambda)^2}\|\mathbf{x}_{\text{new}}\|_2^2\right). \tag{5}$$

Plugging (5) into (4) gives

$$\begin{aligned}\mathbb{E}(y_{\text{new}} - \mathbf{x}_{\text{new}}^T \hat{\boldsymbol{\beta}}^{\text{ridge}})^2 &= 1 + \underbrace{\left[\mathbb{E}(\mathbf{x}_{\text{new}}^T \hat{\boldsymbol{\beta}}^{\text{ridge}}) - \mathbf{x}_{\text{new}}^T \boldsymbol{\beta}\right]^2}_{\text{Bias}^2} + \underbrace{\operatorname{Var}(\mathbf{x}_{\text{new}}^T \hat{\boldsymbol{\beta}}^{\text{ridge}})}_{\text{Variance}} \\ &= 1 + \left(\frac{\lambda}{1+\lambda}\right)^2 (\mathbf{x}_{\text{new}}^T \boldsymbol{\beta})^2 + \frac{1}{(1+\lambda)^2}\|\mathbf{x}_{\text{new}}\|_2^2.\end{aligned}$$

In the special case when $\lambda = 0$, the expected squared prediction error $\mathbb{E}(y_{\text{new}} - \mathbf{x}_{\text{new}}^T \hat{\boldsymbol{\beta}}^{\text{ridge}})^2$ reduces to $1 + \|\mathbf{x}_{\text{new}}\|_2^2$, the same expression as in part a(ii).

(c) This part does not rely on p or λ.

Suppose that a prior distribution $\boldsymbol{\beta}^{\text{prior}} \sim \mathcal{N}(\mathbf{0}, \sigma^2\Phi)$ is imposed to the model (1), where σ^2 is an unknown variance parameter, and Φ is a known positive definite matrix. Furthermore, assume $\boldsymbol{\beta}^{\text{prior}}$ and $\boldsymbol{\varepsilon}$ are independent.

i. Find the marginal distribution of $\mathbf{y}$.

ii. Propose an estimator for σ^2.

Solution: We compute the joint distribution of $(\mathbf{y}, \boldsymbol{\beta}^{\text{prior}})$. Note that

$$\begin{bmatrix} \mathbf{y} \\ \boldsymbol{\beta}^{\text{prior}} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{X} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}}_{\text{fixed}} \underbrace{\begin{bmatrix} \boldsymbol{\beta}^{\text{prior}} \\ \boldsymbol{\varepsilon} \end{bmatrix}}_{\text{random}}, \quad \text{where } \begin{bmatrix} \boldsymbol{\beta}^{\text{prior}} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma^2\Phi & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \right). \tag{6}$$

Since linear combinations of normal distribution are still normal distribution, (6) implies that

$$\begin{bmatrix} \mathbf{y} \\ \boldsymbol{\beta}^{\text{prior}} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma^2\mathbf{X}\Phi\mathbf{X}^T + \mathbf{I} & \sigma^2\mathbf{X}\Phi \\ \sigma^2\Phi\mathbf{X}^T & \sigma^2\Phi \end{bmatrix} \right).$$

Therefore,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0},\ \sigma^2\mathbf{X}\Phi\mathbf{X}^T + \mathbf{I}).$$

ii. Answer 1. Maximum-likelihood estimator:

$$\hat{\sigma}^2 = \arg\max_{\sigma^2} \left\{ -\frac{1}{2}\log\det(\sigma^2\mathbf{X}\Phi\mathbf{X}^T + \mathbf{I}) - \frac{1}{2}\mathbf{y}^T(\sigma^2\mathbf{X}\Phi\mathbf{X}^T + \mathbf{I})^{-1}\mathbf{y} \right\}.$$

Answer 2. Method-of-moment estimator:

$$\hat{\sigma}^2 \text{trace}(\mathbf{X}\Phi\mathbf{X}^T) + n = \sum_i y_i^2 \Rightarrow \hat{\sigma}^2 = \frac{\sum_i y_i^2 - n}{\text{trace}(\mathbf{X}\Phi\mathbf{X}^T)}.$$

(d) Let $p = 1$ and $\lambda > 0$. What value of λ would you suggest for this case and why?

Hint: you may use the following results.

i. The solution to the optimization (2) in this case is $\hat{\boldsymbol{\beta}}^{\text{lasso}} = (\hat{\beta}_1^{\text{lasso}}, \ldots, \hat{\beta}_d^{\text{lasso}})^T$ with

$$\hat{\beta}_j^{\text{lasso}} = \text{sign}(\hat{\beta}_j)\max(|\hat{\beta}_j| - \lambda,\ 0), \quad \text{for all } j = 1, \ldots, d,$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)^T$ denotes the least-squares estimator in Part (a), and $\text{sign}(x) = -1, 0$ or 1 according to $x < 0$, $x = 0$ or $x > 0$, respectively.

ii. Let $\{\varepsilon_i\}_{i=1,\ldots,n}$ be an i.i.d. sequence of $\mathcal{N}(0,1)$ noise. Then as $n \to \infty$, $\mathbb{P}(\max_{i=1,\ldots,n} \varepsilon_i \geq \sqrt{2.01 \times \log n}) \to 0$. Roughly speaking, the following approximation holds

$$\max_{i=1,\ldots,n} \varepsilon_i \approx \sqrt{2.01 \times \log n}, \quad \text{as} \quad n \to \infty.$$

Solution: I would suggest $\lambda = \sqrt{2.01 \times \log n}$. Under this choice of λ, we have vanishing family-wise type I error,

$$\begin{aligned}
\mathbb{P}\left(\|\hat{\boldsymbol{\beta}}^{\text{lasso}}\|_\infty > 0 \;\middle|\; \boldsymbol{\beta} = \mathbf{0}\right) &= \mathbb{P}\left(\|\hat{\boldsymbol{\beta}}\|_\infty > \lambda \;\middle|\; \boldsymbol{\beta} = \mathbf{0}\right) \\
&= \mathbb{P}\left(\|\mathbf{X}\boldsymbol{\varepsilon}\|_\infty > \lambda\right) \\
&= \mathbb{P}\left(\|\boldsymbol{\varepsilon}\|_\infty > \lambda\right) \to 0,
\end{aligned}$$

where the second line follows from the rotation invariance of multivariate normal distribution $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I})$. On the other hand, if all the features are important,

$$\begin{aligned}
\mathbb{P}\left(\|\hat{\boldsymbol{\beta}}^{\text{lasso}}\|_\infty > 0 \;\middle|\text{none of } \beta_j \text{ is zero}\right) &= \mathbb{P}\left(\|\hat{\boldsymbol{\beta}}\|_\infty > \lambda \;\middle|\; \text{none of } \beta_j \text{ is zero}\right) \\
&= \mathbb{P}\left(\|\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\varepsilon}\|_\infty > \lambda \;\middle|\; \text{none of } \beta_j \text{ is zero}\right) \\
&= \mathbb{P}\left(\|\boldsymbol{\beta} + \boldsymbol{\varepsilon}\|_\infty > \lambda \;\middle|\; \text{none of } \beta_j \text{ is zero}\right) \to 1.
\end{aligned}$$

Therefore, the lasso estimator achieves (weak) feature selection.