

Homework 3

DUE: SATURDAY, FEBRUARY 8, 11:59PM

For all the problems below, assume $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

Problem 3.1 (Warmup: eigenvalues and eigenvectors of $A + \lambda I$). Suppose $A = X\Lambda X^{-1} \in \mathcal{M}_{n \times n}(\mathbb{C})$ where $X \in \mathcal{M}_{n \times n}(\mathbb{C})$ is an invertible matrix having the eigenvectors of A as its columns, and $\Lambda \in \mathcal{M}_{n \times n}(\mathbb{C})$ is a diagonal matrix having the eigenvalues of A on its main diagonal.

- a) Show that we can write the identity matrix $I_{n \times n}$ as

$$I_{n \times n} = X I_{n \times n} X^{-1}. \quad (3.1)$$

Conclude that for any $\lambda \in \mathbb{C}$,

$$\lambda I_{n \times n} = X(\lambda I_{n \times n})X^{-1}. \quad (3.2)$$

- b) For any $\lambda \in \mathbb{C}$, show that

$$A + \lambda I_{n \times n} = X(\Lambda + \lambda I_{n \times n})X^{-1} \quad (3.3)$$

- c) Use part b) to show that

$$(A + \lambda I_{n \times n})X = X(\Lambda + \lambda I_{n \times n}). \quad (3.4)$$

Use this identity to identify the eigenvectors and eigenvalues of $A + \lambda I_{n \times n}$.

- d) Suppose A is not an invertible matrix but its eigenvalues all have non-negative real parts (in particular we're assuming that A does not have any negative eigenvalues). Use part c) to explain why $A + \lambda I_{n \times n}$ will always be invertible for any $\mathbb{R} \ni \lambda > 0$.

Problem 3.2 (Ridge regression part II). Recall that if an $m \times n$ matrix A has independent columns, then $A^T A$ is an invertible $n \times n$ square matrix. Thus, the least squares solution $\hat{\mathbf{x}}$ satisfying $A^T A \hat{\mathbf{x}} = A^T \mathbf{b}$ can be written as

$$\hat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b}. \quad (3.5)$$

However, if A represents a data matrix where its rows represent samples and its columns represents features, it is very common for the columns of A to be dependent (or very close to being dependent) if there is high correlation between the feature variables. So in practice, the matrix $A^T A$ is usually very close to being non-invertible. In statistics, this is referred to as the phenomenon of *multicollinearity*.

One way to combat multicollinearity is by replacing the standard least squares estimator with the ridge estimator

$$\hat{\mathbf{x}}_{\text{ridge}} = (A^T A + \lambda I_{n \times n})^{-1} A^T \mathbf{b}, \quad (3.6)$$

where $\mathbb{R} \ni \lambda > 0$ is a positive real number. Our goal in this problem is to show that for any $\lambda > 0$, the ridge estimator is well-defined, by showing that the matrix $A^T A + \lambda I_{n \times n}$ is invertible.

- a) Use the spectral theorem to show that there exists an orthogonal matrix Q and a diagonal matrix Λ such that

$$A^T A = Q \Lambda Q^T. \quad (3.7)$$

- b) Show that one can write

$$A^T A + \lambda I_{n \times n} = Q(\Lambda + \lambda I_{n \times n})Q^T. \quad (3.8)$$

- c) Using the previous problem, how are the eigenvalues of $A^T A + \lambda I_{n \times n}$ related to eigenvalues of $A^T A$?
d) Conclude that for any $\lambda > 0$, the $n \times n$ matrix $A^T A + \lambda I_{n \times n}$ is invertible. This shows that the ridge coefficient $\hat{\mathbf{x}}_{\text{ridge}}$ is well-defined for any $\lambda > 0$.

Problem 3.3 (Another problem on correlation). Suppose $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$ are random variables for which the pairwise correlation coefficients are all equal to $\rho \in \mathbb{R}$. In other words, for all $i, j \in \{1, \dots, n\}$, we have $\text{Corr}(X_i, X_j) = \rho$ for $i \neq j$. In this problem our goal is to find the range of possible values of ρ .

- a) Let $A \in \mathcal{M}_{n \times n}(\mathbb{R})$ be the correlation matrix of the random vector $\mathbf{X} : \Omega \rightarrow \mathbb{R}^n$ defined via $\mathbf{X} = (X_1 \dots X_n)^T$. Show that

$$A = (1 - \rho)I_{n \times n} + \begin{pmatrix} \rho & \rho & \cdots & \rho \\ \rho & \rho & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & \rho \end{pmatrix} = (1 - \rho)I_{n \times n} + \rho B, \quad (3.9)$$

where $B \in \mathcal{M}_{n \times n}(\mathbb{R})$ is a matrix with all entries equal to 1.

- b) Show that if $\lambda \in \mathbb{C}$ is an eigenvalue of B , then either $\lambda = 0$ or $\lambda = n$.
c) Use (3.9) and part b) to show that if $\lambda \in \mathbb{C}$ is an eigenvalue of A , then either $\lambda = 1 - \rho$ or $\lambda = 1 + (n - 1)\rho$.
d) Use part c) and the fact that A is positive semi-definite to show that

$$-\frac{1}{n-1} \leq \rho \leq 1. \quad (3.10)$$

- e) Let $n = 3$. Construct explicit examples of random variables $X_1, X_2, X_3 : \Omega \rightarrow \mathbb{R}$ for which the minimum and maximum values of ρ in (3.10) are achieved. (Hint: use Homework 2 Problem 5)

Problem 3.4 (Row stochastic matrices). Later we will encounter a class of matrices referred to as *stochastic matrices* or sometimes *Markov matrices* when we discuss finite-state Markov chains. Stochastic matrices in this context are used to model the transition probabilities of a discrete dynamical system.

A matrix $M \in \mathcal{M}_{n \times n}(\mathbb{R})$ is said to be a row stochastic matrix if all of its entries are non-negative and the sum of the entries in each row is equal to 1.

- a) Translate the definition above to the following: show that $M \in \mathcal{M}_{n \times n}(\mathbb{R})$ is row stochastic iff. all the entries of M are non-negative and $M\mathbf{1}_{n \times 1} = \mathbf{1}_{n \times 1}$, where $\mathbf{1}_{n \times 1} \in \mathbb{R}^n$ is the column vector with all entries equal to 1.
- b) Show that if $M_1, \dots, M_k \in \mathcal{M}_{n \times n}(\mathbb{R})$ are row stochastic matrices, then the product $\prod_{i=1}^k M_i = M_1 \cdots M_k \in \mathcal{M}_{n \times n}(\mathbb{R})$ is also a row stochastic matrix.

Problem 3.5 (Linear regression and cloning datasets).

Suppose one is working on a dataset with m samples, p features, and 1 target and sets up a linear regression model with a design matrix $X \in \mathcal{M}_{m \times (p+1)}(\mathbb{R})$, a target variable $\mathbf{y} \in \mathbb{R}^m$ and tries to solve for the least squares regressor $\hat{\beta} \in \mathbb{R}^{p+1}$. After solving for the least squares regressor $\hat{\beta}$, they then decided to “clone the data” and run the regression again to see if anything changes. For example, if the original dataset had 3 samples with one target and one predictor, then the cloned dataset would have 6 samples:

x	y
0	2
1	2
2	8

x	y
0	2
1	2
2	8
0	2
1	2
2	8

FIGURE 1. Original dataset on the left vs doubled dataset on the right

In general with $X \in \mathcal{M}_{m \times (p+1)}(\mathbb{R})$, $\mathbf{y} \in \mathbb{R}^m$, this means that instead of looking for the least squares estimator $\hat{\mathbf{x}} \in \mathbb{R}^{p+1}$ to $X\beta = \mathbf{y}$, they instead try to look for the least squares estimator $\hat{\beta}_2 \in \mathbb{R}^{p+1}$ for $X_2\beta = \mathbf{y}_2$, where

$$X = \begin{pmatrix} X \end{pmatrix} \in \mathcal{M}_{m \times (p+1)}(\mathbb{R}), \quad X_2 = \begin{pmatrix} X \\ X \end{pmatrix} \in \mathcal{M}_{2m \times (p+1)}(\mathbb{R}), \quad \mathbf{y} = \begin{pmatrix} \mathbf{y} \end{pmatrix} \in \mathbb{R}^m, \quad \mathbf{y}_2 = \begin{pmatrix} \mathbf{y} \\ \mathbf{y} \end{pmatrix} \in \mathbb{R}^{2m}. \quad (3.11)$$

For example, for the dataset in Figure 1, under the standard simple linear regression model $\mathbf{y} = X\beta + \varepsilon$ where we assume the random vector $\varepsilon \sim N(0, \sigma^2 I)$, we would set up

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 2 \\ 2 \\ 8 \end{pmatrix}, \quad \mathbf{y}_2 = \begin{pmatrix} 2 \\ 2 \\ 8 \\ 2 \\ 2 \\ 8 \end{pmatrix}. \quad (3.12)$$

Intuitively, nothing should really change because no new information has been added to the original dataset. Is this true? Let's investigate.

- Explain briefly why $\text{Col}(X) \neq \text{Col}(X_2)$ yet the dimensions of $\text{Col}(X)$ and $\text{Col}(X_2)$ are the same.
- Show that $\text{Null}(X) = \text{Null}(X_2)$.
- Assuming that the original X matrix has independent columns, show that the unique least squares solution $\hat{\beta}$ solving $X^T X \hat{\beta} = X^T \mathbf{y}$ is the same as the unique least squares regressor solving $(X_2)^T X_2 \hat{\beta} = (X_2)^T \mathbf{y}_2$, however in the cloned system the $\text{RSS}(\hat{\beta})$ is larger by a factor of 2. What about the mean squared error (MSE) $\frac{\text{RSS}(\hat{\beta})}{n}$ in the original and cloned systems? Use this observation to explain why the MSE might be preferred over RSS as a more reliable metric in practice.
- What about the coefficient of determination R^2 , the unbiased estimator $\hat{\sigma}^2 = \frac{\text{RSS}(\hat{\beta})}{n-p-1}$ for σ^2 , the standard errors $\text{SE}(\hat{\beta}_i)$, and also the t -statistics $t_i = \frac{\hat{\beta}_i - \beta_i}{\text{SE}(\hat{\beta}_i)}$ for $1 \leq i \leq 2$? What would happen to the confidence intervals for β_i if we were to use the t -statistics from the cloned system?

Problem 3.6 (Characteristic functions and affine transformations of Gaussian random variables). Let $\mathbf{X} : \Omega \rightarrow \mathbb{R}^n$ be a random vector. The *characteristic function* of \mathbf{X} is defined as the function $\phi_{\mathbf{X}} : \mathbb{R}^n \rightarrow \mathbb{C}$ defined via

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}[e^{i\mathbf{t} \cdot \mathbf{X}}] = \int_{\mathbb{R}^n} e^{i\mathbf{t} \cdot \mathbf{x}} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}. \quad (3.13)$$

For example, if $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, then the characteristic function $\phi_{\mathbf{X}} : \mathbb{R}^n \rightarrow \mathbb{C}$ of \mathbf{X} is given by

$$\phi_{\mathbf{X}}(\mathbf{t}) = \exp \left(i\boldsymbol{\mu} \cdot \mathbf{t} - \frac{1}{2} \mathbf{t}^T \Sigma \mathbf{t} \right). \quad (3.14)$$

If $n = 1$, then this says that if $X \sim N(\mu, \sigma^2)$, then the characteristic function $\phi_X : \mathbb{R} \rightarrow \mathbb{C}$ of X is given by

$$\phi_X(t) = e^{i\mu t - \frac{1}{2}\sigma^2 t^2}. \quad (3.15)$$

The characteristic function is a powerful tool as it allows one to prove many results with ease. Below are two properties of the characteristic function that we will use in this problem.

- If $\mathbf{X}_1, \dots, \mathbf{X}_n : \Omega \rightarrow \mathbb{R}^n$ are independent random vectors, then the characteristic function of their sum is the product of their individual characteristic functions.
- The characteristic function of a random vector uniquely determines its distribution: if \mathbf{X} and \mathbf{Y} are random vectors such that $\phi_{\mathbf{X}} = \phi_{\mathbf{Y}}$, then $F_{\mathbf{X}} = F_{\mathbf{Y}}$, where $F_{\mathbf{X}}$ and $F_{\mathbf{Y}}$ are the cumulative distribution functions of \mathbf{X} and \mathbf{Y} , respectively.

These essentially follow from properties of the *Fourier transform*, which we will explore later in the course.

- Show that if X_1, \dots, X_n are independent random variables and $X_i \sim N(\mu_i, \sigma_i^2)$, then their sum $X = X_1 + \dots + X_n$ is also a Gaussian random variable with mean $\mu = \mu_1 + \dots + \mu_n$ and variance $\sigma^2 = \sigma_1^2 + \dots + \sigma_n^2$. What is the exact distribution of their mean $\bar{X} = \frac{X}{n}$?
- Suppose $\mathbf{X} : \Omega \rightarrow \mathbb{R}^n$ is a random vector and $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$. Let $A \in \mathcal{M}_{n \times n}(\mathbb{R})$ and $\mathbf{b} \in \mathbb{R}^n$. Show that the random vector $A\mathbf{X} + \mathbf{b}$ is also Gaussian with mean $A\boldsymbol{\mu} + \mathbf{b}$ and covariance matrix $A\Sigma A^T$.