Homework 3 solutions

DUE: SATURDAY, FEBRUARY 8, 11:59PM

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

Department of Mathematical Sciences, Carnegie Mellon University
E-mail address: jkoganem@andrew.cmu.edu.
Date: February 12, 2025.

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
$$I_{n\times n} = XI_{n\times n}X^{-1}. \tag{3.1}$$
Conclude that for any $\lambda \in \mathbb{C}$,
$$\lambda I_{n\times n} = X(\lambda I_{n\times n})X^{-1}. \tag{3.2}$$

b) For any $\lambda \in \mathbb{C}$, show that
$$A + \lambda I_{n\times n} = X(\Lambda + \lambda I_{n\times n})X^{-1} \tag{3.3}$$

c) Use part b) to show that
$$(A + \lambda I_{n\times n})X = X(\Lambda + \lambda I_{n\times n}). \tag{3.4}$$
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 $\mathbb{R} \ni \lambda > 0$.

Solution.

a) We note that
$$XI_{n\times n}X^{-1} = XX^{-1} = I_{n\times n}, \tag{3.5}$$
therefore
$$\lambda I_{n\times n} = \lambda(XI_{n\times n}X^{-1}) = X(\lambda I_{n\times n})X^{-1}. \tag{3.6}$$

b) We note that by assumption, $A = X\Lambda X^{-1}$, therefore
$$A + \lambda I_{n\times n} = X\Lambda X^{-1} + X(\lambda I_{n\times n})X^{-1} = X(\Lambda + \lambda I_{n\times n})X^{-1}. \tag{3.7}$$

c) Since X is invertible, this shows that

d)
$$(A + \lambda I_{n\times n})X = X(\Lambda + \lambda I_{n\times n})X^{-1}X = X(\Lambda + \lambda I_{n\times n}). \tag{3.8}$$
By examining the columns of the LHS and RHS we see that the columns of X are eigenvectors of $A+\lambda I_{n\times n}$ and the diagonal elements of $\Lambda + \lambda I_{n\times n}$ are eigenvalues of A. Since X is invertible, we also know that $A + \lambda I_{n\times n}$ and $\Lambda + \lambda I_{n\times n}$ have the same characteristic polynomial. This follows from $\det X \det X^{-1} = 1$, which implies

$$\begin{aligned} p_{A+\lambda I_{n\times n}}(\alpha) = \det(A + \lambda I_{n\times n} - \alpha I_{n\times n}) = \det(X(\Lambda + \lambda I_{n\times n} - \alpha I_{n\times n})X^{-1}) \\ = \det(\Lambda + \lambda I_{n\times n} - \alpha I_{n\times n}) = p_{\Lambda+\lambda I_{n\times n}}(\alpha),\ \alpha \in \mathbb{C}. \end{aligned} \tag{3.9}$$
This shows that $A + \lambda I_{n\times n}$ and $\Lambda + \lambda I_{n\times n}$ have the same eigenvalues. Since the eigenvalues of $\Lambda + \lambda I_{n\times n}$ are the entries on its main diagonal, as it is a diagonal matrix, we may conclude that the eigenvalues of $A + \lambda I_{n\times n}$ are the eigenvalues of A shifted uniformly by a factor of λ.

e) We note that a square matrix is invertible if and only if it does not have zero as an eigenvalue. By assumption, since all of the eigenvalues of A have non-negative real parts and $\lambda > 0$, by the result from the previous part we may conclude that all the eigenvalues of $A + \lambda I_{n\times n}$ have strictly positive real part, therefore $A + \lambda I_{n\times n}$ cannot have zero as an eigenvalue and is thus invertible.

□

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

$$\hat{\boldsymbol{x}} = (A^TA)^{-1}A^T\boldsymbol{b}. \tag{3.10}$$

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^TA 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

$$\hat{\boldsymbol{x}}_{\text{ridge}} = (A^TA + \lambda I_{n\times n})^{-1}A^T\boldsymbol{b}, \tag{3.11}$$

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 $A^TA + \lambda I_{n\times n}$ is invertible.

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

$$A^TA = Q\Lambda Q^T. \tag{3.12}$$

b) Show that one can write

$$A^TA + \lambda I_{n\times n} = Q(\Lambda + \lambda I_{n\times n})Q^T. \tag{3.13}$$

c) Using the previous problem, how are the eigenvalues of $A^TA + \lambda I_{n\times n}$ related to eigenvalues of A^TA?

d) Conclude that for any $\lambda > 0$, the $n \times n$ matrix $A^TA + \lambda I_{n\times n}$ is invertible. This shows that the ridge coefficient $\hat{\boldsymbol{x}}_{\text{ridge}}$ is well-defined for any $\lambda > 0$.

Solution.

a) We note that A^TA is an $n \times n$ real symmetric matrix since A is real and

$$(A^TA)^T = A^T(A^T)^T = A^TA. \tag{3.14}$$

So by the spectral theorem, there exists an orthogonal matrix Q and a diagonal matrix Λ for which

$$A^TA = Q\Lambda Q^T. \tag{3.15}$$

b) Note that since Q is orthogonal, we have $Q^{-1} = Q^T$. Therefore $QQ^T = I$. So $\lambda I = Q(\lambda I)Q^T$, and

$$A^TA + \lambda I_{n\times n} = Q\Lambda Q^T + Q(\lambda I_{n\times n})Q^T = Q(\Lambda + \lambda I_{n\times n})Q^T. \tag{3.16}$$

c) We learned from last week's homework that the eigenvalues of $A^TA + \lambda I_{n\times n}$ are the eigenvalues of A^TA shifted by a factor of λ.

d) We've shown in lecture that the eigenvalues of A^TA are non-negative. Since $\lambda > 0$, we have $\lambda_i + \lambda > 0$ for each eigenvalue λ_i of A, so the eigenvalues of $A^TA + \lambda I_{n\times n}$ are strictly positive. Since $A^TA + \lambda I_{n\times n}$ does not have 0 as an eigenvalue, this implies that $A^TA + \lambda I_{n\times n}$ is invertible.

□

Problem 3.3 (Another problem on correlation)**.** Let $\mathbb{Z} \ni n \geq 2$. Suppose $X_1, \dots, X_n : \Omega \to \mathbb{R}$ 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 $\mathrm{Corr}(X_i, X_j) = \rho$ 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 $\boldsymbol{X} : \Omega \to \mathbb{R}^n$ defined via $\boldsymbol{X} = (X_1 \ \dots \ X_n)^T$. Show that

$$A = (1-\rho)I_{n\times n} + \begin{pmatrix} \rho & \rho & \cdots & \rho \\ \rho & \rho & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & \rho \end{pmatrix} = (1-\rho)I_{n\times n} + \rho B, \tag{3.17}$$

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 (3.17) 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

$$-\frac{1}{n-1} \leq \rho \leq 1. \tag{3.18}$$

e) Let $n = 3$. Construct explicit examples of random variables $X_1, X_2, X_3 : \Omega \to \mathbb{R}$ for which the minimum and maximum values of ρ in (3.18) are achieved. (Hint: use Homework 2 Problem 5)

Solution. For part a), we note that

$$A = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} = \begin{pmatrix} (1-\rho)+\rho & \rho & \cdots & \rho \\ \rho & (1-\rho)+\rho & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & (1-\rho)+\rho \end{pmatrix} = (1-\rho)I_{n\times n} + \rho B. \tag{3.19}$$

For part b), suppose

$$\boldsymbol{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{C}^n \setminus \{\boldsymbol{0}\} \tag{3.20}$$

is an eigenvalue of B with eigenvalue $\lambda \in \mathbb{C}$. Then we have

$$v_1 + v_2 + \cdots + v_n = \lambda v_1 \tag{3.21}$$
$$v_1 + v_2 + \cdots + v_n = \lambda v_2, \tag{3.22}$$
$$\vdots \tag{3.23}$$
$$v_1 + v_2 + \cdots + v_n = \lambda v_n. \tag{3.24}$$

This implies

$$\lambda v_1 = \lambda v_2 = \cdots = \lambda v_n, \tag{3.25}$$

so either $\lambda = 0$ or $v_1 = v_2 = \cdots = v_n$. In the latter case, we have $v_1 + v_2 + \cdots + v_n = nv_i = \lambda v_i$ for all $1 \leq i \leq n$. Since at least one of the v_i's is non-zero, we must have $\lambda = n$.

For part c), we note that if $\boldsymbol{v} \in \mathbb{C}^n \setminus \{\boldsymbol{0}\}$ is an eigenvector of B with eigenvalue λ, then we have

$$A\boldsymbol{v} = (1-\rho)I_{n\times n}\boldsymbol{v} + \rho B\boldsymbol{v} = (1-\rho)\boldsymbol{v} + \rho\lambda\boldsymbol{v} = (1-\rho+\lambda\rho)\boldsymbol{v}. \tag{3.26}$$

This shows that $\boldsymbol{v}$ is an eigenvector of A with eigenvalue $1-\rho+\lambda\rho$. Conversely, if $\boldsymbol{v}$ is an eigenvector of A with eigenvalue λ, then we have

$$\rho B\boldsymbol{v} = A\boldsymbol{v} - (1-\rho)\boldsymbol{v} = \lambda\boldsymbol{v} - (1-\rho)\boldsymbol{v} = (\lambda - 1 + \rho)\boldsymbol{v}. \tag{3.27}$$

If $\rho \neq 0$, then the above calculations show that A, B share the same eigenvectors and the eigenvalues of A are $1-\rho$ and $1+(n-1)\rho$. If $\rho = 0$, then $A = I_{n\times n}$, which implies that its eigenvalues are all equal to $1 = 1-\rho = 1+(n-1)\rho$. Since A is positive semi-definite, we must have $1-\rho \geq 0$ and $1+(n-1)\rho \geq 0$, which implies that $-\frac{1}{n-1} \leq \rho \leq 1$. For the last part, we note that if $\rho = 1$, we can simply consider random variables X_1, X_2, X_3 for which $X_1 = X_2 = X_3$. If $\rho = -\frac{1}{3-1} = -\frac{1}{2}$, we learned from Homework 2 Problem 5 that every positive semi-definite matrix is a covariance

matrix, in the sense that we can construct a random vector $(X_1, X_2, X_3) : \Omega \to \mathbb{R}^3$ for which the covariance matrix is A. Following the construction from Homework 2 Problem 5, we first orthogonally diagonalize the matrix

$$A = \begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix}. \tag{3.28}$$

From part c) we learned that the eigenvalues of A are either $\frac{3}{2}$ or 0. To find eigenvectors corresponding to $\lambda = 0$, we note that

$$A \sim \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix} \sim \begin{pmatrix} 1 & -2 & 1 \\ 0 & -3 & 3 \\ 0 & 3 & -3 \end{pmatrix} \sim \begin{pmatrix} 1 & -2 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix} \tag{3.29}$$

Therefore a basis for $\text{Null}(A)$ is

$$\mathcal{B}_{\text{Null}(A)} = \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}. \tag{3.30}$$

Note that since the row sums of A are all 0, the vector $\boldsymbol{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ is guaranteed to be a corresponding eigenvector; row reducing A shows that this is the only eigenvector corresponding to $\lambda = 0$.

Next we note that

$$A - \frac{3}{2} I_{3\times 3} = \begin{pmatrix} -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{3.31}$$

This implies that a basis for $\text{Null}(A - \frac{3}{2} I_{3\times 3})$ is given by

$$\mathcal{B}_{\text{Null}(A - \frac{3}{2} I_{3\times 3})} = \left\{ \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \right\}. \tag{3.32}$$

To construct an orthogonal basis for this subspace, we perform Gram-Schmidt:

$$\boldsymbol{v}_2 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \qquad \boldsymbol{q}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \tag{3.33}$$

$$\boldsymbol{v}_3 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} - \frac{1}{2} \left(\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \right) \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \tag{3.34}$$

$$= \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}, \qquad \boldsymbol{q}_3 = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}. \tag{3.35}$$

This shows that $A = Q\Lambda Q^T$ for

$$Q = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \frac{3}{2} & 0 & 0 \\ 0 & \frac{3}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{3.36}$$

Let $\boldsymbol{Z} : \Omega \to \mathbb{R}^3$ be a random vector with mean $\mathbf{0}$ and covariance matrix I. Then from Homework 2 Problem 5, it follows that the random vector $\boldsymbol{X} = Q\Lambda^{1/2}\boldsymbol{Z}$ has mean $\mathbf{0}$ and covariance matrix A. □

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 $M_1, \ldots, M_k \in \mathcal{M}_{n\times n}(\mathbb{R})$ 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.

Solution. We note that for any matrix $M \in \mathcal{M}_{n\times n}(\mathbb{R})$, if we write M in terms of its rows as

$$M = \begin{pmatrix} \boldsymbol{m}_1^T \\ \vdots \\ \boldsymbol{m}_n^T \end{pmatrix}, \tag{3.37}$$

then

$$M\mathbf{1}_{n\times 1} = \begin{pmatrix} \boldsymbol{m}_1^T\mathbf{1}_{n\times 1} \\ \vdots \\ \boldsymbol{m}_n^T\mathbf{1}_{n\times 1} \end{pmatrix}, \tag{3.38}$$

where $\boldsymbol{m}_i^T\mathbf{1}_{n\times 1}$ is the sum of the entries in the ith row of M. Therefore it follows that M is row stochastic iff. all the entries of M are non-negative and $M\mathbf{1}_{n\times 1} = \mathbf{1}_{n\times 1}$.

For part b), we note that it suffices to prove this for $k = 2$, the general case follows from a simple induction argument. Suppose $M_1, M_2 \in \mathcal{M}_{n\times n}(\mathbb{R})$ are row stochastic matrices. Then we have

$$(M_1M_2)\mathbf{1}_{n\times 1} = M_1(M_2\mathbf{1}_{n\times 1}) = M_1\mathbf{1}_{n\times 1} = \mathbf{1}_{n\times 1}. \tag{3.39}$$

This shows that M_1M_2 is row stochastic. □

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 $\boldsymbol{y} \in \mathbb{R}^m$ and tries to solve for the least squares regressor $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$. After solving for the least squares regressor $\hat{\boldsymbol{\beta}}$, 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 $X \in \mathcal{M}_{m\times(p+1)}(\mathbb{R}), \boldsymbol{y} \in \mathbb{R}^m$, this means that instead of looking for the least squares estimator $\hat{\boldsymbol{x}} \in \mathbb{R}^{p+1}$ to $X\boldsymbol{\beta} = \boldsymbol{y}$, they instead try to look for the least squares estimator $\hat{\boldsymbol{\beta}}_2 \in \mathbb{R}^{p+1}$ for $X_2\boldsymbol{\beta} = \boldsymbol{y}_2$, where

$$X = \begin{pmatrix} X \end{pmatrix} \in \mathcal{M}_{m\times(p+1)}(\mathbb{R}),\ X_2 = \begin{pmatrix} X \\ X \end{pmatrix} \in \mathcal{M}_{2m\times(p+1)}(\mathbb{R}),\ \boldsymbol{y} = \begin{pmatrix} \boldsymbol{y} \end{pmatrix} \in \mathbb{R}^m,\ \boldsymbol{y}_2 = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{y} \end{pmatrix} \in \mathbb{R}^{2m}. \tag{3.40}$$

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

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} 2 \\ 2 \\ 8 \end{pmatrix}, \quad \boldsymbol{y}_2 = \begin{pmatrix} 2 \\ 2 \\ 8 \\ 2 \\ 2 \\ 8 \end{pmatrix}. \tag{3.41}$$

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

a) Explain briefly why $\mathrm{Col}(X) \neq \mathrm{Col}(X_2)$ yet the dimensions of $\mathrm{Col}(X)$ and $\mathrm{Col}(X_2)$ are the same.
b) Show that $\mathrm{Null}(X) = \mathrm{Null}(X_2)$.
c) Assuming that the original X matrix has independent columns, show that the unique least squares solution $\hat{\boldsymbol{\beta}}$ solving $X^TX\hat{\boldsymbol{\beta}} = X^T\boldsymbol{y}$ is the same as the unique least squares regressor solving $(X_2)^TX_2\hat{\boldsymbol{\beta}} = (X_2)^T\boldsymbol{y}_2$, however in the cloned system the $\mathrm{RSS}(\hat{\boldsymbol{\beta}})$ is larger by a factor of 2. What about the mean squared error (MSE) $\frac{\mathrm{RSS}(\hat{\boldsymbol{\beta}})}{m}$ 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.
d) What about the coefficient of determination R^2, the unbiased estimator $\hat{\sigma}^2 = \frac{\mathrm{RSS}(\hat{\boldsymbol{\beta}})}{m-p-1}$ for σ^2, the standard errors $\mathrm{SE}(\hat{\beta}_i)$, and also the t-statistics $t_i = \frac{\hat{\beta}_i - \beta_i}{\mathrm{SE}(\hat{\beta}_i)}$ 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?

Solution.

a) We note that $\text{Col}(X) \subseteq \mathbb{R}^m$ whereas $\text{Col}(X_2) \subseteq \mathbb{R}^{2m}$, therefore $\text{Col}(X) \neq \text{Col}(X_2)$. However,

$$\text{rref}(X_2) = \begin{pmatrix} \boxed{\text{rref}(X)} \\ \boxed{0_{m\times n}} \end{pmatrix}, \tag{3.42}$$

so X and X_2 admit the same number of pivots, therefore $\dim \text{Col}(X) = \dim \text{Col}(X_2)$.

b) We note that by (3.42), $\text{rref}(X)\boldsymbol{x} = \mathbf{0}_m$ if and only if $\text{rref}(X_2)\boldsymbol{x} = \mathbf{0}_{2m}$ since adding an equation of the form $0x_1 + \ldots + 0x_n = 0$ does not change the underlying solution set. Therefore $\text{Null}(X) = \text{Null}(\text{rref}(X)) = \text{Null}(\text{rref}(X_2)) = \text{Null}(X_2)$.

c) We note that by block multiplication,

$$(X_2)^T X_2 = \begin{pmatrix} \boxed{X^T} & \boxed{X^T} \end{pmatrix} \begin{pmatrix} \boxed{X} \\ \boxed{X} \end{pmatrix} = X^T X + X^T X = 2X^T X, \tag{3.43}$$

and

$$(X_2)^T \boldsymbol{y}_2 = \begin{pmatrix} \boxed{X^T} & \boxed{X^T} \end{pmatrix} \begin{pmatrix} \boxed{\boldsymbol{y}} \\ \boxed{\boldsymbol{y}} \end{pmatrix} = X^T \boldsymbol{y} + X^T \boldsymbol{y} = 2X^T \boldsymbol{y}. \tag{3.44}$$

Therefore $X^T X\hat{\boldsymbol{x}} = X^T \boldsymbol{y} \iff (X_2)^T X_2 \hat{\boldsymbol{x}} = (X_2)^T \boldsymbol{y}_2$. However,

$$\boldsymbol{y}_2 - X_2\hat{\boldsymbol{x}} = \begin{pmatrix} \boxed{\boldsymbol{y}} \\ \boxed{\boldsymbol{y}} \end{pmatrix} - \begin{pmatrix} \boxed{X} \\ \boxed{X} \end{pmatrix} \hat{\boldsymbol{x}} = \begin{pmatrix} \boxed{\boldsymbol{y} - X\hat{\boldsymbol{x}}} \\ \boxed{\boldsymbol{y} - X\hat{\boldsymbol{x}}} \end{pmatrix} \implies \|\boldsymbol{y}_2 - X_2\hat{\boldsymbol{x}}\|^2 = 2\|\boldsymbol{y} - X\hat{\boldsymbol{x}}\|^2, \tag{3.45}$$

so the error increases by a factor of 2 in the cloned system. However, if we consider the mean squared error, since there are $2m$ samples in the cloned system and m samples in the original system, we have

$$\frac{1}{2m}\|\boldsymbol{y}_2 - X_2\hat{\boldsymbol{x}}\|^2 = \frac{2}{2m}\|\boldsymbol{y} - X\hat{\boldsymbol{x}}\| = \frac{1}{m}\|\boldsymbol{y} - X\hat{\boldsymbol{x}}\|, \tag{3.46}$$

therefore the mean squared error does not change. In practice, this means that the MSE is a more robust measure of error, as it is not affected by artificial duplication of data.

d) We note that the coefficient of determination R^2 is defined via

$$R^2 = 1 - \frac{2\text{RSS}_{\text{original}}(\hat{\boldsymbol{\beta}})}{2\text{TSS}_{\text{original}}}, \tag{3.47}$$

Since the scaling factor is the same in both the RSS and TSS, the coefficient of determination does not change. The unbiased estimator for σ^2 in the cloned system is $\frac{2\text{RSS}_{\text{original}}}{2m-p-1}$, which means that it is roughly unchanged if $p-1$ is small.

For the standard errors, recall that $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1}) \approx N(\boldsymbol{\beta}, \hat{\sigma}^2(X^TX)^{-1})$. For $p = 1$, we have

$$\hat{\sigma}^2(X^TX)^{-1} = \frac{\hat{\sigma}^2}{(m-1)s_x^2} \begin{pmatrix} \frac{1}{m}\sum_{i=1}^m x_i^2 & -\frac{1}{m}\sum_{i=1}^m x_i \\ -\frac{1}{m}\sum_{i=1}^m x_i & 1 \end{pmatrix} \tag{3.48}$$

In the cloned system, s_x^2 and $\hat{\sigma}^2$ are roughly the same, however the scaling factor $\frac{1}{m-1}$ is replaced by $\frac{1}{2m-1}$, which means the standard errors in the cloned system get scaled down by roughly a factor of $\sqrt{2}$. As a result, the t-statistics in the cloned system scales up by roughly a factor of $\sqrt{2}$, and the confidence intervals for β_i in the cloned system shrinks by roughly a factor of $\sqrt{2}$.

For general p, the scaling factors for $\hat{\sigma}^2$, the t-statistics, and the standard errors will also be affected by the value of p, however tracking this explicitly is more difficult as it depends on $(X^TX)^{-1}$.

□

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

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}[e^{i\boldsymbol{t}\cdot\boldsymbol{X}}] = \int_{\mathbb{R}^n} e^{i\boldsymbol{t}\cdot\boldsymbol{x}} f_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x}. \tag{3.49}$$

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

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(i\boldsymbol{\mu}\cdot\boldsymbol{t} - \frac{1}{2}\boldsymbol{t}^T\Sigma\boldsymbol{t}\right). \tag{3.50}$$

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

$$\phi_X(t) = e^{i\mu t - \frac{1}{2}\sigma^2 t^2}. \tag{3.51}$$

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 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n : \Omega \to \mathbb{R}^n$ 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 $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random vectors such that $\phi_{\boldsymbol{X}} = \phi_{\boldsymbol{Y}}$, then $F_{\boldsymbol{X}} = F_{\boldsymbol{Y}}$, where $F_{\boldsymbol{X}}$ and $F_{\boldsymbol{Y}}$ are the cumulative distribution functions of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

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

a) Show that if $X_1, \ldots, X_n$ are independent random variables and $X_i \sim N(\mu_i, \sigma_i^2)$, then their sum $X = X_1 + \cdots + X_n$ is also a Gaussian random variable with mean $\mu = \mu_1 + \cdots + \mu_n$ and variance $\sigma^2 = \sigma_1^2 + \cdots + \sigma_n^2$. What is the exact distribution of their mean $\bar{X} = \frac{X}{n}$?

b) Suppose $\boldsymbol{X} : \Omega \to \mathbb{R}^n$ is a random vector and $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$. Let $A \in \mathcal{M}_{n\times n}(\mathbb{R})$ and $\boldsymbol{b} \in \mathbb{R}^n$. Show that the random vector $A\boldsymbol{X} + \boldsymbol{b}$ is also Gaussian with mean $A\boldsymbol{\mu} + \boldsymbol{b}$ and covariance matrix $A\Sigma A^T$.

Solution. We note that the characteristic function of X is $\phi_X : \mathbb{R} \to \mathbb{C}$ given by

$$\begin{aligned}\phi_X(t) = \mathbb{E}[e^{itX}] = \mathbb{E}[e^{it(X_1+\cdots+X_n)}] = \mathbb{E}[e^{itX_1}\cdots e^{itX_n}] = \mathbb{E}[e^{itX_1}]\cdots\mathbb{E}[e^{itX_n}] \\ = \prod_{i=1}^n \phi_{X_i}(t) = \prod_{i=1}^n e^{i\mu_i t - \frac{1}{2}\sigma_i^2 t^2} = e^{i\mu t - \frac{1}{2}\sigma^2 t^2},\end{aligned} \tag{3.52}$$

where

$$\mu = \mu_1 + \cdots + \mu_n, \quad \sigma^2 = \sigma_1^2 + \cdots + \sigma_n^2. \tag{3.53}$$

Since the characteristic function determines the distribution, we have that $X \sim N(\mu, \sigma^2)$. The distribution of the mean $\bar{X} = \frac{X}{n}$ is then $N\left(\frac{\mu}{n}, \frac{\sigma^2}{n^2}\right)$.

For part b), we note that the characteristic function of $A\boldsymbol{X} + \boldsymbol{b}$ is $\phi_{A\boldsymbol{X}+\boldsymbol{b}} : \mathbb{R}^n \to \mathbb{C}$ given by

$$\begin{aligned}\phi_{A\boldsymbol{X}+\boldsymbol{b}}(\boldsymbol{t}) &= \mathbb{E}[\exp(i\boldsymbol{t}\cdot(A\boldsymbol{X}+\boldsymbol{b}))] = \mathbb{E}[\exp(i\boldsymbol{t}\cdot A\boldsymbol{X})\exp(i\boldsymbol{t}\cdot\boldsymbol{b})] \\ &= \exp(i\boldsymbol{t}\cdot\boldsymbol{b})\mathbb{E}[\exp(i\boldsymbol{t}\cdot A\boldsymbol{X})] = \exp(i\boldsymbol{t}\cdot\boldsymbol{b})\mathbb{E}[\exp(iA^T\boldsymbol{t}\cdot\boldsymbol{X})] = \exp(i\boldsymbol{t}\cdot\boldsymbol{b})\phi_{\boldsymbol{X}}(A^T\boldsymbol{t}) \\ &= \exp(i\boldsymbol{t}\cdot\boldsymbol{b})\exp\left(i\boldsymbol{\mu}\cdot A^T\boldsymbol{t} - \frac{1}{2}(A^T\boldsymbol{t})^T\Sigma(A^T\boldsymbol{t})\right) = \exp\left(i(A\boldsymbol{\mu}+\boldsymbol{b}) - \frac{1}{2}\boldsymbol{t}^T(A\Sigma A^T)\boldsymbol{t}\right).\end{aligned} \tag{3.54}$$

Since the characteristic function determines the distribution, it follows that $A\boldsymbol{X} + \boldsymbol{b} \sim N(A\boldsymbol{\mu} + \boldsymbol{b}, A\Sigma A^T)$. □