EXAMINATION PAPER

Examination Session: May/June	**Year:** 2024	**Exam Code:** MATH44220-WE01

Title:

Deep Learning, AI and Data Analytics V: Paper 1

Time:	2 hours	
Additional Material provided:		
Materials Permitted:		
Calculators Permitted:	No	Models Permitted: Use of electronic calculators is forbidden.

Instructions to Candidates:	Answer all questions. Section A is worth 40% and Section B is worth 60%. Within each section, all questions carry equal marks. Students must use the mathematics specific answer book.

Revision:	

SECTION A

Q1 (a) Consider the high dimensional PCA for a data set with $n = 125$ observations and $p = 496$ variables. Figure 1, which is calculated using `R`, shows the percentage of variance captured by the first 20 principal components. Based on this plot, what is an optimal number of principal components explaining a reasonably high amount of variance? Justify your answer.

(b) Figure 2 shows histograms of the loadings for the first 3 principal components, where the vertical lines highlight the locations of the 5^{th} and 95^{th} percentiles. Using these histograms, comment on the loadings of these 3 leading principal components. Also, explain whether the loadings suggest a more efficient approach for PCA on this data.

(c) Figure 3 shows the distribution of the eigenvalues from the PCA for this data. Comment on the eigenvalues and explain why some eigenvalues are very big.

Figure 1: Percentage of variance explained by the first 20 principal components.

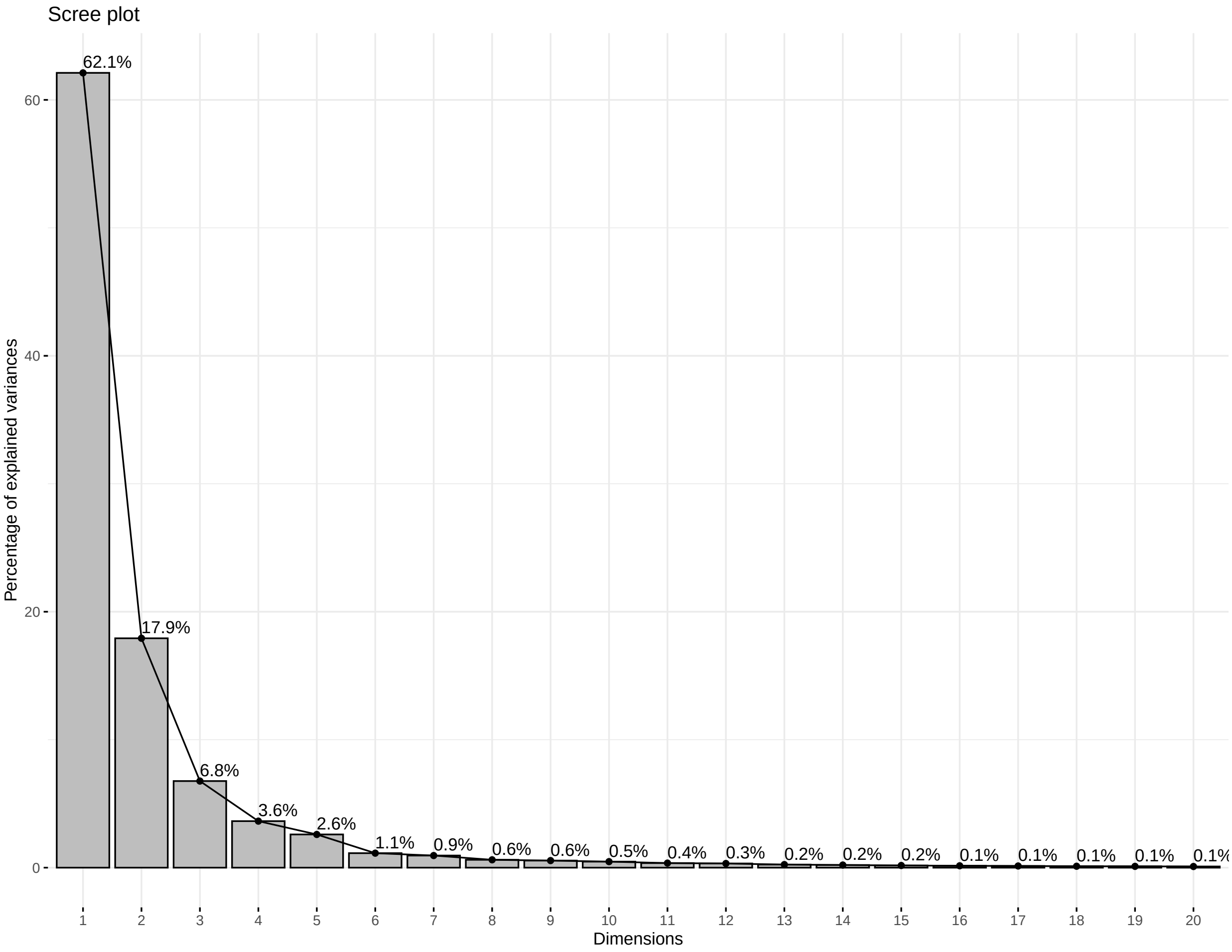


Figure 2: Histograms of the loadings for the first 3 principal components.

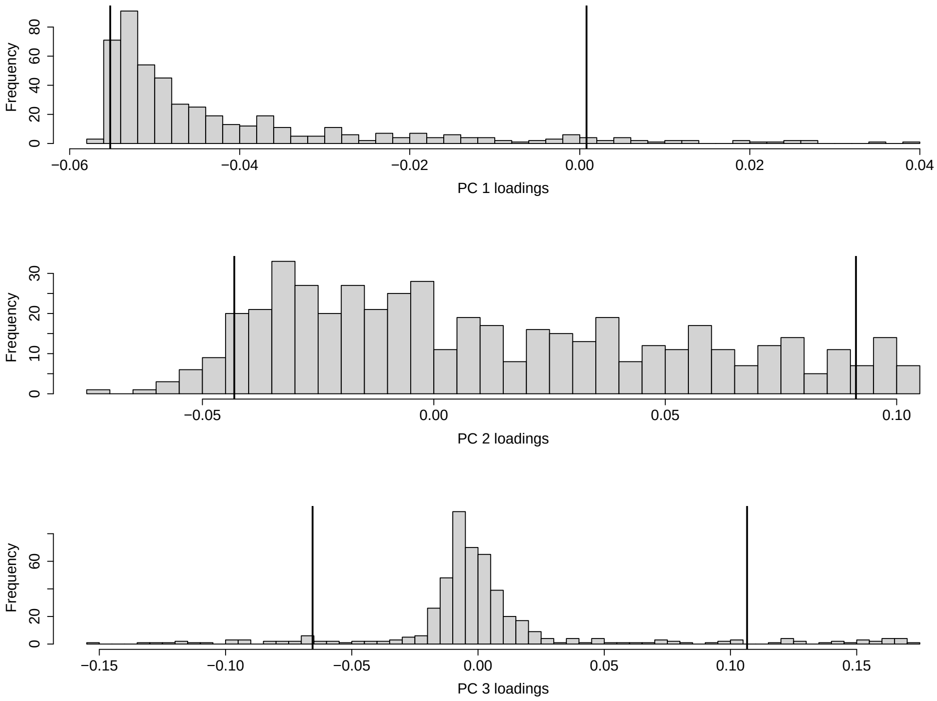
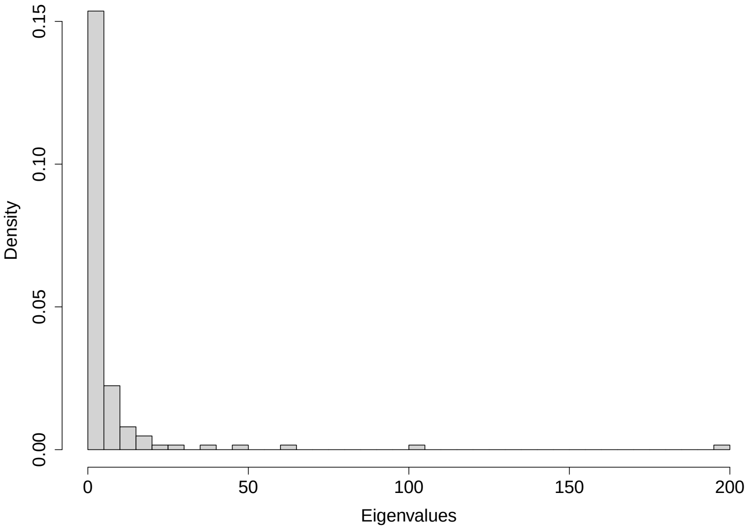


Figure 3: Distribution of the eigenvalues from the PCA result.



Q2 (a) Consider the linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with a scaled matrix $\boldsymbol{X}$ (i.e., all columns of $\boldsymbol{X}$ have mean 0 and variance 1). Assume the random errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$ are Gaussian and all independent with mean 0 and variance σ^2. Let $\mathcal{F} := \left\{\frac{2}{n}\left\|\boldsymbol{\varepsilon}^T\boldsymbol{X}\right\|_\infty \leq \lambda_0\right\}$, where $\|\cdot\|_\infty$ denotes the L_∞-norm of a vector (e.g., $\|\boldsymbol{a}\|_\infty = \max\left\{|a_1|, \dots, |a_p|\right\}$). Prove that, for all $t > 0$ and $\lambda_0 = 2\sigma\sqrt{\frac{t^2+2\log(p)}{n}}$,

$$P(\mathcal{F}) \geq 1 - 2\exp(-t^2/2).$$

Hint: You may use the following inequality for a standard normal random variable Z

$$P(|Z| > c) \leq 2\exp(-c^2/2), \quad \forall\ c > 0.$$

(b) Let $\hat{\boldsymbol{\beta}}$ denote the lasso estimator for $\boldsymbol{\beta}$. Under the conditions of part (a) and with $\lambda \geq 2\lambda_0 = 4\sigma\sqrt{\frac{t^2+2\log(p)}{n}}$, show that

$$\frac{2}{n}\left\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\right\|_2^2 \leq 3\lambda\left\|\boldsymbol{\beta}^0\right\|_1^1,$$

where $\boldsymbol{\beta}^0$ denotes the vector of (unknown) true parameter values.

Hint: Use the basic inequality for $\hat{\boldsymbol{\beta}}$, that is,

$$\frac{1}{n}\left\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\right\|_2^2 + \lambda\left\|\hat{\boldsymbol{\beta}}\right\|_1^1 \leq \frac{2}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \lambda\left\|\boldsymbol{\beta}^0\right\|_1^1,$$

and the Holder's inequality for two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ which is

$$\boldsymbol{u}^T\boldsymbol{v} \leq \left\|\boldsymbol{u}\right\|_q^1\left\|\boldsymbol{v}\right\|_r^1, \qquad \frac{1}{q} + \frac{1}{r} = 1.$$

SECTION B

Q3 Consider the L_1-ball of radius $t > 0$ defined as $B_{L_1}(t) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_1^1 \leq t\}$. Suppose we want to project a vector $\boldsymbol{\beta} \in \mathbb{R}^p$ on the L_1-ball of radius t. When $\|\boldsymbol{\beta}\|_1^1 \leq t$, the projection is simply $\boldsymbol{\beta}$. We assume in the following that $\|\boldsymbol{\beta}\|_1^1 > t$. For $\lambda > 0$, define

$$\boldsymbol{S}_\lambda(\boldsymbol{\beta}) = \left[\boldsymbol{\beta}_j\left(1 - \frac{\lambda}{|\boldsymbol{\beta}_j|}\right)_+\right]_{j=1,\dots,p}, \qquad \text{with } (x)_+ = \max(x, 0).$$

Note that $\boldsymbol{S}_\lambda(\boldsymbol{\beta})$ is a p-dimensional vector.

(a) Prove that $\boldsymbol{S}_\lambda(\boldsymbol{\beta}) \in \underset{\boldsymbol{\alpha}\in\mathbb{R}^p}{\arg\min}\left\{\left\|\boldsymbol{\beta} - \boldsymbol{\alpha}\right\|_2^2 + 2\lambda\left\|\boldsymbol{\alpha}\right\|_1^1\right\}$.

(b) Show that the projection of $\boldsymbol{\beta}$ on the L_1-ball of radius t is given by $\boldsymbol{S}_{\hat{\lambda}}(\boldsymbol{\beta})$, where $\hat{\lambda} > 0$ is such that $\|\boldsymbol{S}_{\hat{\lambda}}(\boldsymbol{\beta})\|_1^1 = t$.

Q4 In cluster analysis, the heterogeneity of a group of observations or cluster G can be measured by the inertia inside the group or cluster, which is defined as

$$I_G = \sum_{i=1}^{N_G} \left\| \boldsymbol{X}_i - \bar{\boldsymbol{X}}_G \right\|_2^2,$$

where N_G and $\bar{\boldsymbol{X}}_G$ are, respectively, the number of observations and the mean of observations in cluster G.

(a) Show that $I_G = N_G \mathrm{tr}\big(S_G\big)$ where S_G denotes the sample covariance matrix of the observations in cluster G.

(b) In hierarchical clustering, when two clusters G and H are merged, we are interested to calculate the inertia of the merged cluster $G \cup H$. Prove that the corresponding inertia, denoted by $I_{G\cup H}$, is given by

$$I_{G\cup H} = I_G + I_H + \sum_{j=1}^{p} \left\{ N_G\big(\bar{\boldsymbol{X}}_{G,j} - \bar{\boldsymbol{X}}_{G\cup H,j}\big)^2 + N_H\big(\bar{\boldsymbol{X}}_{H,j} - \bar{\boldsymbol{X}}_{G\cup H,j}\big)^2 \right\},$$

in which $\bar{\boldsymbol{X}}_{G,j}$ denotes the j-th element of $\bar{\boldsymbol{X}}_G$, and $\bar{\boldsymbol{X}}_{H,j}$ denotes the j-th element of $\bar{\boldsymbol{X}}_H$. Similarly, $\bar{\boldsymbol{X}}_{G\cup H,j}$ denotes the j-th element of $\bar{\boldsymbol{X}}_{G\cup H}$.

(c) When merging two clusters G and H as in part (b), show that the increase of inertia, defined as $\Delta(G,H) := I_{G\cup H} - (I_G + I_H)$, can be written as

$$\Delta(G,H) = \frac{N_G N_H}{N_G + N_H} \sum_{j=1}^{p} \big(\bar{\boldsymbol{X}}_{G,j} - \bar{\boldsymbol{X}}_{H,j}\big)^2,$$

and comment on this result.

END