Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2025	MATH4287-WE01

Title: High-Dimensional Statistics

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. Write your answer in the white-covered answer booklet with barcodes. Begin your answer to each question on a new page.

Revision:	

SECTION A

Q1 (a) Consider the clustering problem for a high dimensional data set with $n = 67$ observations and $p = 3000$ variables. Figure 1 shows the `scree plot` from the K-means clustering method with the number of clusters varying from 1 to 15. Based on this scree plot, what number of clusters is most appropriate for the K-means clustering of this data and why?

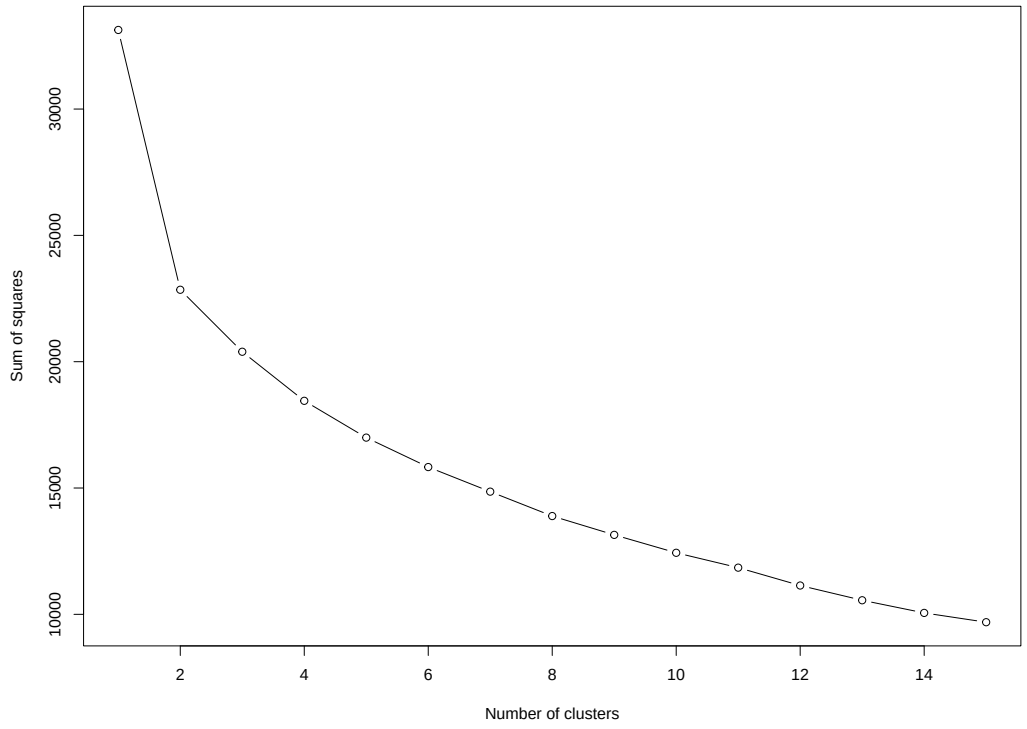

Figure 1: Scree plot for the K-means clustering in part (a) of Q1.

(b) Now consider hierarchical clustering for the data set in Part (a). Figure 2 shows the dendrogram from hierarchical clustering for this high dimensional data set. Based on this dendrogram, explain what type of hierarchical clustering (agglomerative or divisive) is used here.

(c) Using the dendrogram shown in Figure 2 on page 4, how many clusters do you suggest for this data? Explain your answer.

(d) The `R` output on the next page reports the proportion of variance and the cumulative proportion captured by the principal components for this data. Based on this output, how many principal components capture about 95% of the data variance?
Also, explain briefly why the total variance (100%) is fully captured by only 67 principal components while there are 3000 variables in this data set.

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     14.4311 7.9997 5.84184 4.51285 4.28174 3.95003 3.53042
Proportion of Variance  0.4149 0.1275 0.06798 0.04057 0.03652 0.03108 0.02483
Cumulative Proportion   0.4149 0.5424 0.61034 0.65091 0.68743 0.71852 0.74335
                          PC8    PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     3.3459 3.1447 3.11112 2.95285 2.61072 2.41062 2.27749
Proportion of Variance 0.0223 0.0197 0.01928 0.01737 0.01358 0.01158 0.01033
Cumulative Proportion  0.7657 0.7853 0.80463 0.82200 0.83558 0.84715 0.85749
                          PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     2.08230 2.02676 1.99309 1.87063 1.82158 1.76700 1.66751
Proportion of Variance 0.00864 0.00818 0.00791 0.00697 0.00661 0.00622 0.00554
Cumulative Proportion  0.86612 0.87431 0.88222 0.88919 0.89580 0.90202 0.90756
                          PC22    PC23    PC24    PC25   PC26    PC27    PC28
Standard deviation     1.61992 1.59664 1.53104 1.46407 1.3993 1.34242 1.30520
Proportion of Variance 0.00523 0.00508 0.00467 0.00427 0.0039 0.00359 0.00339
Cumulative Proportion  0.91279 0.91787 0.92254 0.92681 0.9307 0.93430 0.93769
                          PC29    PC30    PC31   PC32    PC33    PC34    PC35
Standard deviation     1.29655 1.26377 1.21520 1.1848 1.10976 1.10597 1.07783
Proportion of Variance 0.00335 0.00318 0.00294 0.0028 0.00245 0.00244 0.00231
Cumulative Proportion  0.94104 0.94422 0.94716 0.9500 0.95241 0.95485 0.95716
                          PC36    PC37    PC38    PC39    PC40    PC41   PC42
Standard deviation     1.06867 1.05511 1.01839 0.99151 0.97300 0.96405 0.9501
Proportion of Variance 0.00228 0.00222 0.00207 0.00196 0.00189 0.00185 0.0018
Cumulative Proportion  0.95944 0.96166 0.96372 0.96568 0.96757 0.96942 0.9712
                          PC43    PC44    PC45    PC46    PC47    PC48    PC49
Standard deviation     0.94500 0.93754 0.91271 0.88823 0.85382 0.83616 0.82573
Proportion of Variance 0.00178 0.00175 0.00166 0.00157 0.00145 0.00139 0.00136
Cumulative Proportion  0.97300 0.97475 0.97641 0.97798 0.97943 0.98082 0.98218
                          PC50    PC51    PC52    PC53    PC54    PC55    PC56
Standard deviation     0.81650 0.80496 0.78883 0.78213 0.77324 0.76232 0.75980
Proportion of Variance 0.00133 0.00129 0.00124 0.00122 0.00119 0.00116 0.00115
Cumulative Proportion  0.98351 0.98480 0.98604 0.98726 0.98845 0.98961 0.99076
                          PC57    PC58    PC59   PC60    PC61    PC62    PC63
Standard deviation     0.73549 0.72290 0.71643 0.7082 0.70021 0.69464 0.65413
Proportion of Variance 0.00108 0.00104 0.00102 0.0010 0.00098 0.00096 0.00085
Cumulative Proportion  0.99184 0.99288 0.99390 0.9949 0.99587 0.99684 0.99769
                          PC64    PC65    PC66      PC67
Standard deviation     0.65224 0.62365 0.58815 1.992e-14
Proportion of Variance 0.00085 0.00077 0.00069 0.000e+00
Cumulative Proportion  0.99854 0.99931 1.00000 1.000e+00
```

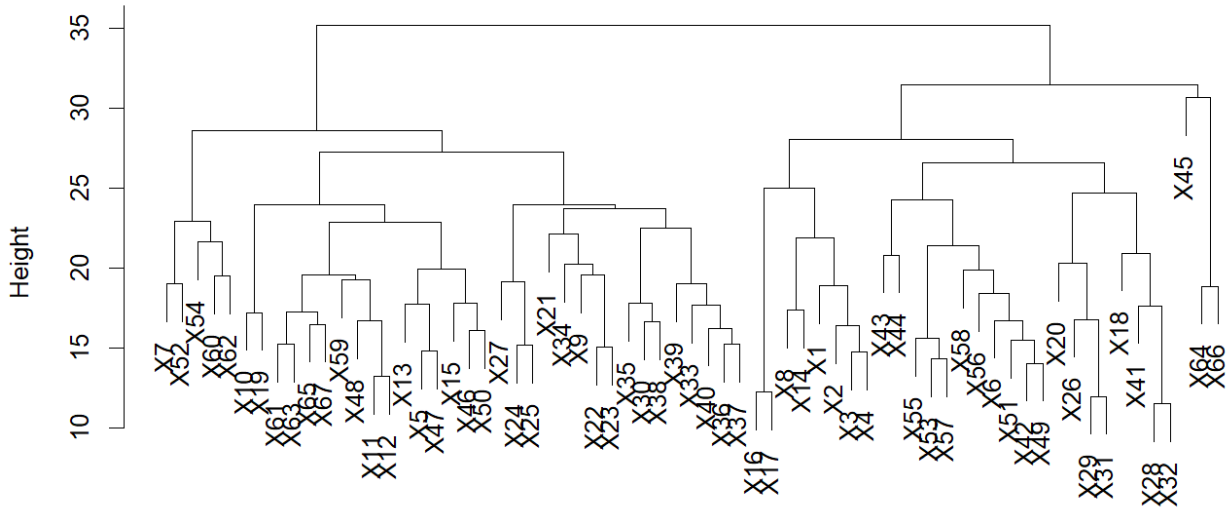

Figure 2: Dendrogram from the hierarchical clustering in part (b) of Q1.

Q2 Consider the linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The elastic net estimator of $\boldsymbol{\beta}$ is defined as follows

$$\hat{\boldsymbol{\beta}}_{\mathrm{EN}}(\lambda_1, \lambda_2) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\Big\{ \big\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\big\|_2^2 + \lambda_1 \big\|\boldsymbol{\beta}\big\|_1^1 + \lambda_2 \big\|\boldsymbol{\beta}\big\|_2^2 \Big\},$$

where $\lambda_1, \lambda_2 \geq 0$ are two separate regularisation parameters.

(a) Prove that the elastic net estimator with a fixed λ_2 can be computed using a lasso problem.

Hint: Make an appropriate enlargement of the design matrix $\boldsymbol{X}$ using some additional matrix and proceed from there.

(b) Consider a simple case where the columns of $\boldsymbol{X}$ are orthonormal, that is, we have $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_p$ with $\boldsymbol{I}_p$ being the $p \times p$ identity matrix, meaning that the columns of $\boldsymbol{X}$ are orthogonal with norm 1. Show that the elastic net estimator in this simple case can be obtained in closed form as follows

$$[\hat{\boldsymbol{\beta}}_{\mathrm{EN}}(\lambda_1, \lambda_2)]_j = \frac{\mathrm{sign}([\boldsymbol{X}^T\boldsymbol{Y}]_j)\big(\big|[\boldsymbol{X}^T\boldsymbol{Y}]_j\big| - \frac{\lambda_1}{2}\big)_+}{1+\lambda_2}, \quad j = 1, \ldots, p,$$

where $\mathrm{sign}(x) = 1(x > 0) - 1(x < 0)$ for $x \neq 0$, and also $(x)_+ = \max(x, 0)$.

Hint: For this question and throughout the paper you may use the following formulae on matrix differentiation. Let $\boldsymbol{x}$ be a vector, and further suppose that matrix $\boldsymbol{A}$ and vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are all not functions of $\boldsymbol{x}$. The following matrix differentiation results hold:

$$\begin{aligned}
\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}} &= \boldsymbol{0} \\
\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{x}} &= \boldsymbol{I} \\
\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} &= \boldsymbol{A} \\
\frac{\partial \boldsymbol{x}^T\boldsymbol{A}}{\partial \boldsymbol{x}} &= \boldsymbol{A}^T \\
\frac{\partial \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} &= \boldsymbol{x}^T\big(\boldsymbol{A} + \boldsymbol{A}^T\big) \qquad \text{the row representation} \\
\frac{\partial \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} &= \big(\boldsymbol{A} + \boldsymbol{A}^T\big)\boldsymbol{x} \qquad \text{the column representation} \\
\frac{\partial \boldsymbol{a}^T\boldsymbol{x}\boldsymbol{b}}{\partial \boldsymbol{x}} &= \boldsymbol{a}\boldsymbol{b}^T \\
\frac{\partial \boldsymbol{a}^T\boldsymbol{x}^T\boldsymbol{b}}{\partial \boldsymbol{x}} &= \boldsymbol{b}\boldsymbol{a}^T \\
\frac{\partial \boldsymbol{a}^T\boldsymbol{x}^T\boldsymbol{x}\boldsymbol{b}}{\partial \boldsymbol{x}} &= \boldsymbol{x}^T\big(\boldsymbol{a}\boldsymbol{b}^T + \boldsymbol{b}\boldsymbol{a}^T\big) \qquad \text{the row representation} \\
\frac{\partial \boldsymbol{a}^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{b}}{\partial \boldsymbol{x}} &= \big(\boldsymbol{a}\boldsymbol{b}^T + \boldsymbol{b}\boldsymbol{a}^T\big)\boldsymbol{x} \qquad \text{the column representation.}
\end{aligned}$$

SECTION B

Q3 (a) Recall the linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, here with a scaled matrix $\boldsymbol{X}$ (i.e., all columns of $\boldsymbol{X}$ have mean 0 and variance 1). Assume the random errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ 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., $\|\boldsymbol{a}\|_\infty = \max\left\{|a_1|, \ldots, |a_p|\right\}$). Prove that, on $\mathcal{F}$ with $\lambda \geq 2\lambda_0$, we have

$$\frac{2}{n}\left\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\right\|_2^2 + \lambda \left\|\hat{\boldsymbol{\beta}}_{S_0^c}\right\|_1^1 \leq 3\lambda \left\|\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^0_{S_0}\right\|_1^1,$$

where $\hat{\boldsymbol{\beta}}$ is the lasso estimator, $\boldsymbol{\beta}^0$ is the vector of unknown true parameter values, and S_0 denotes the active set.

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

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

(b) Suppose that the compatibility condition holds for S_0, that is, for some constant $\phi_0 > 0$ we can write

$$\left\|\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^0_{S_0}\right\|_2^2 \leq \left\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\right\|_2^2 / \left(n\phi_0^2\right).$$

Prove that, on $\mathcal{F}$ with $\lambda \geq 2\lambda_0$, we have

$$\frac{1}{n}\left\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\right\|_2^2 + \lambda \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_1^1 \leq 4\lambda^2 s_0 / \phi_0^2,$$

where $s_0 = |S_0|$ is the sparsity index which is the cardinality of active set.

(c) Assume $\lambda = 4\sigma\sqrt{\frac{t^2 + 2\log(p)}{n}}$, with a known σ (e.g., by prior knowledge or an appropriate estimate), for some constant $t > 0$. What would the result in part (b) claim about the optimality of lasso estimator $\hat{\boldsymbol{\beta}}$ with such choice of λ? Justify your claim mathematically.

Q4 (a) Consider the sparse PCA problem

$$\min_{\boldsymbol{v},\boldsymbol{\theta}} \sum_{i=1}^{n} \left\| \boldsymbol{X}_i - \boldsymbol{\theta}\boldsymbol{v}^T\boldsymbol{X}_i \right\|_2^2 + \lambda \left\| \boldsymbol{v} \right\|_2^2 + \lambda_1 \left\| \boldsymbol{v} \right\|_1^1, \quad \text{subject to } \boldsymbol{\theta}^T\boldsymbol{\theta} = 1.$$

For the case when both regularisation parameters λ and λ_1 are zero and $n > p$, prove that $\boldsymbol{v} = \boldsymbol{\theta}$ and is the leading principal component direction.

(b) Write down the sparse PCA problem for the general case of d components with $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$. Also, briefly explain how to carry out the computation for this general sparse PCA problem.

(c) We know that in K-means clustering, the (squared) Euclidean distance

$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) = \left\| \boldsymbol{X}_i - \boldsymbol{X}_j \right\|_2^2 = \sum_{l=1}^{p} \left(\boldsymbol{X}_{il} - \boldsymbol{X}_{jl} \right)^2$$

is often used as the dissimilarity measure for clustering the observations. Now suppose that we instead use the weighted Euclidean distance

$$d_W(\boldsymbol{X}_i, \boldsymbol{X}_j) = \frac{\sum_{l=1}^{p} w_l \left(\boldsymbol{X}_{il} - \boldsymbol{X}_{jl} \right)^2}{\sum_{l=1}^{p} w_l},$$

where the w_l are some non-negative weights for clustering. Show that the weighted Euclidean distance satisfies

$$d_W(\boldsymbol{X}_i, \boldsymbol{X}_j) = d(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = \sum_{l=1}^{p} \left(\boldsymbol{Z}_{il} - \boldsymbol{Z}_{jl} \right)^2,$$

where

$$\boldsymbol{Z}_{il} = \boldsymbol{X}_{il} \left(\frac{w_l}{\sum_{l=1}^{p} w_l} \right)^{1/2}.$$

What is the implication of this result when applying the K-means clustering method in this case? Explain your answer.