Durham University

# EXAMINATION PAPER

| Examination Session: | Year: | Exam Code: |
|---|---|---|
| May/June | 2026 | MATH4287-WE01 |

| Title: |
|---|
| High-Dimensional Statistics |

| | | |
|---|---|---|
| Time: | 2 hours | |
| Additional Material provided: | None | |
| Materials Permitted: | None | |
| Calculators Permitted: | No | Models Permitted: Use of electronic calculators is forbidden. |

| | |
|---|---|
| Instructions to Candidates: | Answer all questions.<br>The indicative marks shown in brackets for the main parts of each question are given as a guide to the weighting the markers expect to apply.<br>Write your answer in the white-covered answer booklet with barcodes.<br>Begin your answer to each question on a new page. |

| Revision: | |
|---|---|

ED01/2026

## SECTION A

**1.** (a) Consider the clustering problem for a high dimensional data set with $n = 70$ observations and $p = 496$ variables. Figure 1 shows the `scree plot` from the $K$-means clustering method with the number of clusters varying from 1 to 10. Based on this scree plot, what number of clusters is most appropriate for the $K$-means clustering of this data and why? **[2]**

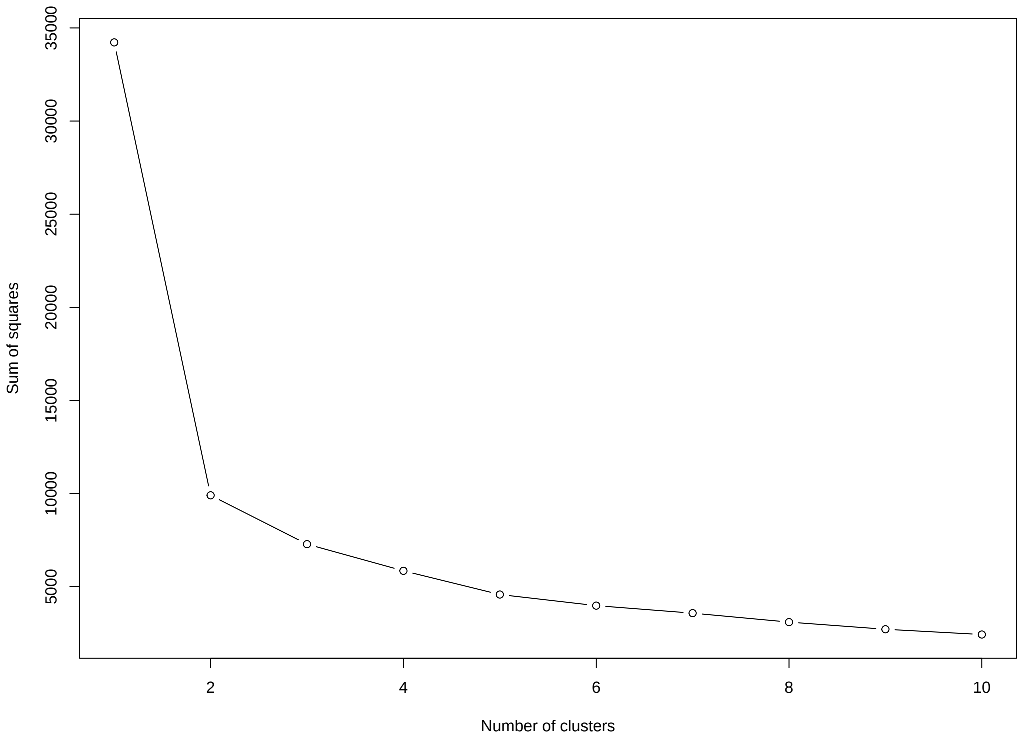


Figure 1: Scree plot for the $K$-means clustering in part (a) of Q1.

(b) Now consider hierarchical clustering for the data set in part (a). Figure 2 shows the dendrogram for this high dimensional data set. Based on this dendrogram, explain what type of hierarchical clustering (agglomerative or divisive) is used here. **[3]**

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

(d) The R output below 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 70 principal components while there are 496 variables in this data set. [3]

```
Importance of components:
                           PC1     PC2     PC3     PC4     PC5     PC6
Standard deviation     20.0990 5.84117 4.05048 3.07309 2.60324 2.24181
Proportion of Variance  0.8145 0.06879 0.03308 0.01904 0.01366 0.01013
Cumulative Proportion   0.8145 0.88324 0.91632 0.93536 0.94902 0.95916
                           PC7     PC8     PC9    PC10   PC11    PC12    PC13
Standard deviation     1.89518 1.56944 1.29460 1.20682 1.1131 1.03767 0.93170
Proportion of Variance 0.00724 0.00497 0.00338 0.00294 0.0025 0.00217 0.00175
Cumulative Proportion  0.96640 0.97136 0.97474 0.97768 0.9802 0.98235 0.98410
                          PC14    PC15   PC16    PC17    PC18    PC19    PC20
Standard deviation     0.90063 0.79223 0.7707 0.69298 0.67716 0.59273 0.57246
Proportion of Variance 0.00164 0.00127 0.0012 0.00097 0.00092 0.00071 0.00066
Cumulative Proportion  0.98573 0.98700 0.9882 0.98917 0.99009 0.99080 0.99146
                         PC21    PC22    PC23    PC24    PC25   PC26    PC27
Standard deviation     0.5451 0.53439 0.50080 0.48986 0.47195 0.4429 0.42933
Proportion of Variance 0.0006 0.00058 0.00051 0.00048 0.00045 0.0004 0.00037
Cumulative Proportion  0.9921 0.99263 0.99314 0.99362 0.99407 0.9945 0.99484
                          PC28    PC29   PC30    PC31    PC32    PC33    PC34
Standard deviation     0.41866 0.39469 0.3878 0.36977 0.35750 0.35265 0.33864
Proportion of Variance 0.00035 0.00031 0.0003 0.00028 0.00026 0.00025 0.00023
Cumulative Proportion  0.99519 0.99551 0.9958 0.99609 0.99634 0.99659 0.99683
                           PC35    PC36   PC37    PC38    PC39    PC40    PC41
Standard deviation     0.32384 0.31944 0.3114 0.29987 0.29283 0.27729 0.26857
Proportion of Variance 0.00021 0.00021 0.0002 0.00018 0.00017 0.00016 0.00015
Cumulative Proportion  0.99704 0.99724 0.9974 0.99762 0.99779 0.99795 0.99809
                            PC42    PC43    PC44    PC45    PC46   PC47   PC48
Standard deviation     0.26227 0.25473 0.24428 0.23801 0.23340 0.2271 0.2220
Proportion of Variance 0.00014 0.00013 0.00012 0.00011 0.00011 0.0001 0.0001
Cumulative Proportion  0.99823 0.99836 0.99848 0.99860 0.99871 0.9988 0.9989
                          PC49    PC50    PC51    PC52    PC53    PC54
Standard deviation     0.21648 0.20925 0.20450 0.19436 0.19295 0.18783
Proportion of Variance 0.00009 0.00009 0.00008 0.00008 0.00008 0.00007
Cumulative Proportion  0.99900 0.99909 0.99918 0.99925 0.99933 0.99940
                          PC55    PC56    PC57    PC58    PC59    PC60
Standard deviation     0.17912 0.17626 0.17299 0.16042 0.15796 0.15079
Proportion of Variance 0.00006 0.00006 0.00006 0.00005 0.00005 0.00005
Cumulative Proportion  0.99946 0.99953 0.99959 0.99964 0.99969 0.99974
                          PC61    PC62    PC63    PC64    PC65    PC66
Standard deviation     0.14547 0.13486 0.13174 0.12291 0.12257 0.11675
Proportion of Variance 0.00004 0.00004 0.00003 0.00003 0.00003 0.00003
Cumulative Proportion  0.99978 0.99981 0.99985 0.99988 0.99991 0.99994
                          PC67    PC68    PC69      PC70
Standard deviation     0.11216 0.09644 0.09490 8.926e-15
Proportion of Variance 0.00003 0.00002 0.00002 0.000e+00
Cumulative Proportion  0.99996 0.99998 1.00000 1.000e+00
```

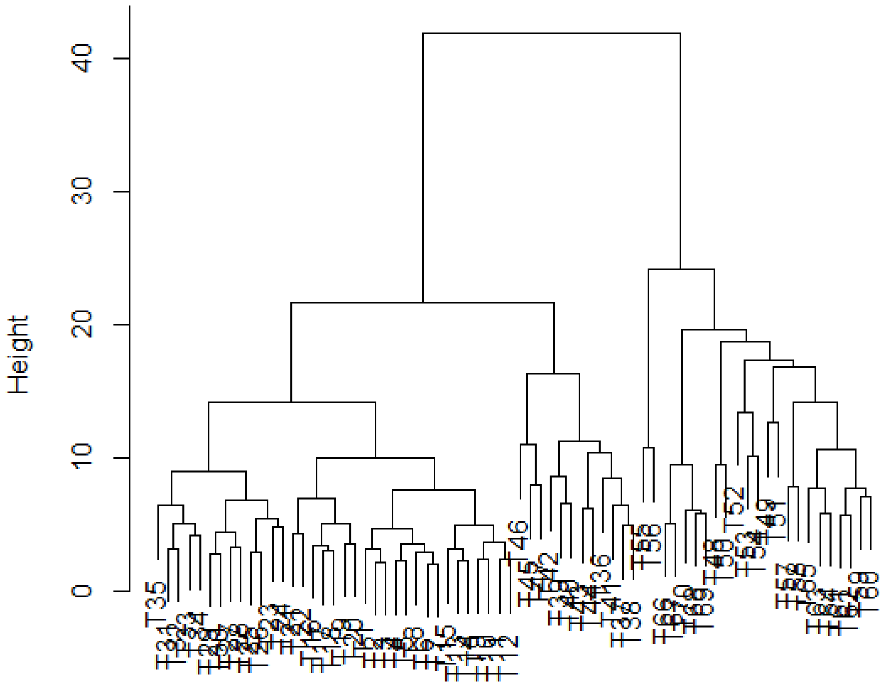


Figure 2: Dendrogram from the hierarchical clustering in part (b) of Q1.

---

**2.** (a) Consider the high dimensional linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the sample size $n$ is smaller than the dimension $p$ (i.e., $n \ll p$). Let $\hat{\boldsymbol{\beta}}_{\text{Ridge}} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ be the ridge regression solution, where $\lambda$ is the corresponding regularisation parameter. Prove that $\left\|\hat{\boldsymbol{\beta}}_{\text{Ridge}}\right\|_2^2$ increases as the tuning parameter $\lambda \to 0$. **[5]**

Hint: Use the full SVD of $\boldsymbol{X}$ as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are two orthogonal matrices (i.e., $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}_n$ and $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}_p$), and $\boldsymbol{D}$ is an $n \times p$ matrix of singular values.

(b) Suppose we run a ridge regression with parameter $\lambda$ on a single covariate $\boldsymbol{X}_1$, and get coefficient $\alpha$. We now include an exact copy $\boldsymbol{X}_1^* = \boldsymbol{X}_1$, and refit our ridge regression. Show that both coefficients are identical, and derive their value. Show in general that if $m$ copies of a covariate $\boldsymbol{X}_j$ are included in a ridge regression, their coefficients are all the same. **[5]**

---

## SECTION B

**3.** For high dimensional settings $n \ll p$, consider the following linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}_I\boldsymbol{\beta}_I + \boldsymbol{X}_N\boldsymbol{\beta}_N + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta}_I \in \mathbb{R}^q$ is a vector of parameters of interest and $\boldsymbol{\beta}_N \in \mathbb{R}^{p-q}$ is a vector of nuisance parameters, and $\boldsymbol{X}_I \in \mathbb{R}^{n\times q}$ and $\boldsymbol{X}_N \in \mathbb{R}^{n\times(p-q)}$ denote the two matrices of covariates associated with the parameters of interest and the nuisance parameters, respectively. Suppose that we want to estimate parameters $\boldsymbol{\beta}_I$ and $\boldsymbol{\beta}_N$ by solving the following optimisation problem:

$$\min_{\boldsymbol{\beta}_I\in\mathbb{R}^q,\boldsymbol{\beta}_N\in\mathbb{R}^{p-q}} \left\{\frac{1}{n}\big\|\boldsymbol{Y} - \boldsymbol{X}_I\boldsymbol{\beta}_I - \boldsymbol{X}_N\boldsymbol{\beta}_N\big\|_2^2 + \lambda_1\big\|\boldsymbol{\beta}_I\big\|_2^2 + \lambda_2\big\|\boldsymbol{\beta}_N\big\|_1^1\right\},$$

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

(a) Briefly explain whether or not this regularisation problem is different than the elastic net. **[3]**

(b) For a fixed given $\boldsymbol{\beta}_N$, find the estimate of $\boldsymbol{\beta}_I$ and denote it by $\hat{\boldsymbol{\beta}}_I$. **[5]**

(c) Now suppose $\boldsymbol{\beta}_N$ is no longer fixed. Let $\boldsymbol{P}_I = \boldsymbol{X}_I(\boldsymbol{X}_I^T\boldsymbol{X}_I + n\lambda_1\boldsymbol{I}_q)^{-1}\boldsymbol{X}_I^T$ for $\lambda_1 > 0$. By substituting $\boldsymbol{\beta}_I$ in the above optimisation problem with its estimate $\hat{\boldsymbol{\beta}}_I$ you obtained in part (b), prove that finding the estimate of $\boldsymbol{\beta}_N$ leads to solving

$$\min_{\boldsymbol{\beta}_N\in\mathbb{R}^{p-q}} \left\{\frac{1}{n}\big\|(\boldsymbol{I}_n - \boldsymbol{P}_I)^{1/2}(\boldsymbol{Y} - \boldsymbol{X}_N\boldsymbol{\beta}_N)\big\|_2^2 + \lambda_2\big\|\boldsymbol{\beta}_N\big\|_1^1\right\},$$

which is a lasso problem with the modified response variable $(\boldsymbol{I}_n - \boldsymbol{P}_I)^{1/2}\boldsymbol{Y}$ and the modified covariates $(\boldsymbol{I}_n - \boldsymbol{P}_I)^{1/2}\boldsymbol{X}_N$. **[7]**

Hint: In your calculations, you can define $\boldsymbol{H} = \boldsymbol{X}_I^T\boldsymbol{X}_I + n\lambda_1\boldsymbol{I}_q$ to simplify expressions such as $(\boldsymbol{I}_n - \boldsymbol{P}_I)^2 = \big(\boldsymbol{I}_n - \boldsymbol{X}_I\boldsymbol{H}^{-1}\boldsymbol{X}_I^T\big)\big(\boldsymbol{I}_n - \boldsymbol{X}_I\boldsymbol{H}^{-1}\boldsymbol{X}_I^T\big)$ for subsequent calculations.

---

**4.** (a) Consider high dimensional PCA with an $n \times p$ data matrix $\boldsymbol{X}$ where $n \ll p$. The principal components are calculated based on eigenvectors of the sample covariance $\boldsymbol{\Sigma}_n := \frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{X}$, which can be efficiently computed using the SVD of $\boldsymbol{X}$ in high dimensional settings with $n \ll p$. Another approach for computing the eigenvectors of $\boldsymbol{\Sigma}_n = \frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{X}$ is through computing the eigenvectors of the much smaller $n \times n$ matrix $\boldsymbol{\Sigma}_n^* := \frac{1}{n-1}\boldsymbol{X}\boldsymbol{X}^T$, since there is a relationship between these two. Show how the eigenvectors of $\boldsymbol{\Sigma}_n$ can be calculated from the eigenvectors of $\boldsymbol{\Sigma}_n^*$. Also, explain briefly how this relationship can be used for PCA on $\boldsymbol{X}$ when $n \ll p$. **[6]**

(b) Suppose that the eigenvalues of the scaled sample covariance matrix $\boldsymbol{S}_n := \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$ are denoted by $\lambda_j$, $j = 1, \ldots, p$. As required in the so-called moment method for the proof of the Marchenko-Pastur distribution of eigenvalues $\lambda_j$, prove that for any integer $k \geq 1$

$$E\Big(\mathrm{tr}\big[(\boldsymbol{S}_n)^k\big]\Big) = E\Big(\sum_{j=1}^{p} \lambda_j^k\Big).$$

**[6]**

(c) Consider the spike model for high dimensional PCA with a rank 1 perturbed covariance matrix such as $\boldsymbol{\Sigma} = \boldsymbol{I}_p + \beta\boldsymbol{v}\boldsymbol{v}^T$, where $\boldsymbol{v}$ is a unit norm vector and $\beta \geq 0$. Show that the largest eigenvalue of the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{I}_p + \beta\boldsymbol{v}\boldsymbol{v}^T$ in the case of $\boldsymbol{v} = \boldsymbol{e}_1$, where $\boldsymbol{e}_1$ is the first element of the canonical or standard basis, is given by $1 + \beta$. **[3]**

---