STAT3006 Assignment 4—High-Dimensional Inference

Due Date: 15th November 2021
Weighting: 25%

Instructions

- The assignment consists of **three (3) problems; Problems 1 and 2 are worth 10 Marks each, and Problem 3 is worth 5 Mark.** Each Mark is equally weighted and is worth 1% of the overall course grade.
- The mathematical elements of the assignment can be completed by hand, **in LaTeX (preferably)**, or in Word (or other typesetting software). The mathematical derivations and manipulations should be accompanied by clear explanations in English regarding necessary information required to interpret the mathematical exposition.
- Computation problems can be answered using your programming language of choice, although R is generally recommended, or Python if you are uncomfortable with R. As with the mathematical exposition, you may choose to typeset your answers to the problems in whatever authoring or word processing software that you wish. You should also maintain a copy of any codes that you have produced.
- Computer generated plots and hand drawn graphs should be included together with the text where problems are answered.
- The assignment will require four (4) files containing data, that you can can download from the Assignment 4 section on Blackboard. These files are: `zip.txt`, `golub_genes.csv`, `golub_labels.csv`, and `prostate.csv`.
- Submission files should include the following (which ever applies to you):
  - Scans of handwritten mathematical exposition.
  - Typeset mathematical exposition, outputted as a `pdf` file.
  - Typeset answers to computational problems, outputted as a `pdf` file.

– Program code/scripts that you wish to submit, outputted as a `txt` file.

- **All submission files should be labeled with your name and student number and archived together in a `zip` file and submitted at the TurnItIn link on Blackboard.** We suggest naming using the convention:

  `FirstName_LastName_STAT3006A4_[Problem_XX/Problem_XX_Part_YY].[FileExtension]`.

- As per `my.uq.edu.au/information-and-services/manage-my-program/student-integrityand-conduct/academic-integrity-and-student-conduct`, what you submit should be your own work. Even where working from sources, you should endeavour to write in your own words. You should use consistent notation throughout your assignment and define whatever is required.

Problem 1 [10 Marks]

Consider the data set `zip.txt`, which contains $n = 7291$ rows of data, where each row is an observation

$$\boldsymbol{Z}_i^\top = \left(Y_i, \boldsymbol{X}_i^\top\right) \in \mathbb{R}^{1+q},$$

with $q = 256$. Here $Y_i \in \{0, 1, \ldots, 9\}$ is a **label** indicating the **digit** that is represented by the **vectorized** 16×16 **matrix (image)** $\boldsymbol{X}_i \in \mathbb{R}^q$.

Part 1 [1 Mark]

Select one of the digits $y \in \{0, 1, \ldots, 9\}$ and **plot** $m = 9$ **unique images with** $Y_i = y$**, in the same plot, as characterized by** $\boldsymbol{X}_i$ **.**

Part 2 [2 Marks]

Using all $n = 7291$ observations of the sample $\tilde{\mathbf{X}}_n = \left(\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_n\right)$, where $\tilde{\boldsymbol{X}}_i = \boldsymbol{X}_i - \bar{\boldsymbol{X}}_n$ $(i \in [n])$ and $\bar{\boldsymbol{X}}_n = n^{-1} \sum_{i=1}^n \boldsymbol{X}_i$, **obtain a solution to the optimization problem:**

$$\hat{\mathbf{F}}, \hat{\mathbf{R}} = \underset{\mathbf{F} \in \mathbb{R}^{s \times q}, \mathbf{R} \in \mathbb{R}^{q \times s}}{\arg\min} \sum_{i=1}^n \left\| \tilde{\boldsymbol{X}}_i - \mathbf{R}\mathbf{F}\tilde{\boldsymbol{X}}_i \right\|_2^2,$$

for $s = 4$ **and report the minimum value:**

$$\sum_{i=1}^n \left\| \tilde{\boldsymbol{X}}_i - \hat{\mathbf{R}}\hat{\mathbf{F}}\tilde{\boldsymbol{X}}_i \right\|_2^2.$$

Part 3 [2 Marks]

Using the results from Part 2, plot the forward mappings $\hat{\boldsymbol{W}}_1, \ldots, \hat{\boldsymbol{W}}_n$, where $\hat{\boldsymbol{W}}_i = \hat{\mathbf{F}}\tilde{\boldsymbol{X}}_i$, for each $i \in [n]$, colored by the labels Y_i. Discuss whether there appears to be differences in the distributions of the forward mapped observations $\hat{\boldsymbol{W}}_1, \ldots, \hat{\boldsymbol{W}}_n$, corresponding to different values of the labels Y_i.

Part 4 [1 Mark]

Via a spectral decomposition of the Grammian

$$\tilde{\mathbf{G}} = \sum_{i=1}^{n} \tilde{\boldsymbol{X}}_i \tilde{\boldsymbol{X}}_i^\top$$

report the proportion of total variance that is explained by $s = 4$ eigenvectors corresponding to the first s largest eigenvalues.

Part 5 [2 Marks]

Using all $n = 7291$ observations of the sample $\mathbf{X}_n = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, **obtain a solution** $\hat{\boldsymbol{\theta}} = \left(\hat{\sigma}^2, \hat{\boldsymbol{\mu}}, \hat{\mathbf{R}}\right)$ **to the optimization problem**

$$\arg\max_{\boldsymbol{\theta}} \ \log L_n(\boldsymbol{\theta}),$$

where

$$\begin{aligned}\log L_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log \phi\left(\boldsymbol{X}_i; \boldsymbol{\mu}, \mathbf{R}\mathbf{R}^\top + \sigma^2 \mathbf{I}_q\right) \\ &= -\frac{n}{2} q \log(2\pi) - \frac{n}{2} \log\left|\mathbf{R}\mathbf{R}^\top + \sigma^2 \mathbf{I}_q\right| - -\frac{1}{2} \text{trace}\left(\left[\mathbf{R}\mathbf{R}^\top + \sigma^2 \mathbf{I}_q\right]^{-1} \mathbf{S}(\boldsymbol{\mu})\right),\end{aligned}$$

$\mathbf{S}(\boldsymbol{\mu}) = n^{-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^\top$, and $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\mu}, \mathbf{R})$, with $\sigma^2 > 0$, $\boldsymbol{\mu} \in \mathbb{R}^q$, and $\mathbf{R} \in \mathbb{R}^{q \times s}$, for $s = 4$.

Part 6 [1 Marks]

Using the results from Part 5, **estimate the posterior expectations of the latent variables** $\boldsymbol{W}_i$**:**

$$\mathbf{E}(\boldsymbol{W}_i | \boldsymbol{X}_i) = \left(\mathbf{R}^\top \mathbf{R} + \sigma^2 \mathbf{I}_s\right)^{-1} \mathbf{R}^\top (\boldsymbol{X}_i - \boldsymbol{\mu})$$

for each $i \in [n]$**, and plot the estimated posterior expectations, colored by the labels** Y_i.

Part 7 [1 Marks]

Using all $n = 7291$ observations of the sample $\mathbf{X}_n = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, **use 3-layer autoencoder with some activation function $a : \mathbb{R} \to \mathbb{R}$ (of your choice) to obtain an s-dimensional nonlinear dimensionality reduction**

$$\boldsymbol{W}_i = \mathscr{F}(\boldsymbol{X}_i), \text{ for each } i \in [n]$$

and for $s = 4$, where

$$\mathscr{F} : \mathbb{R}^q \to \mathbb{R}^s, \boldsymbol{x} \mapsto (a(\mathbf{f}_1 \boldsymbol{x} + c_1), \ldots, a(\mathbf{f}_s \boldsymbol{x} + c_s)),$$

for some vectors $\mathbf{f}_1, \ldots, \mathbf{f}_s \in \mathbb{R}^q$ and scalars $c_1, \ldots, c_s \in \mathbb{R}$. Then, **plot $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$, colored by the labels Y_i.**

Problem 2 [10 Marks]

Consider the data set `golub_genes.csv` , which contains $q = 3571$ rows of data corresponding observations $\boldsymbol{X}_i \in \mathbb{R}^q$ ($i \in [n]$), in each of the $n = 72$ columns. Here, each row corresponds to the expression levels of a gene j across the n cells, corresponding to the columns. The data set `golub_labels.csv` then contains the corresponding label corresponding to the cell type of each of the n columns of `golub_genes.csv`, where the cells are either labeled as "ALL" or "AML", where ALL stands for Acute Lymphoblastic Leukemia, and AML stands for Acute Myeloid Leukemia. For each $i \in [n]$, we will write $Y_i = 1$ if cell i is ALL and $Y_i = 2$ if cell i is AML.

Part 1 [2 Mark]

Let $f_1(\boldsymbol{x}) = f(\boldsymbol{x}|Y = 1)$ and $f_2(\boldsymbol{x}) = f(\boldsymbol{x}|Y = 2)$ be the probability density functions of the gene expression levels for ALL and AML cells, respectively. Using an maximum mean discrepancy statistic with kernel of the form:

$$\kappa_g(\boldsymbol{x}, \boldsymbol{y}) = g\left(\|\boldsymbol{x} - \boldsymbol{y}\|_2^2\right),$$

where $g(t) = \exp\{-\beta t\}$ for $\beta = 2^{-28}$, and using the data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and $Y_1, \ldots, Y_n$, (assumed to be independent and identically distributed), **test the hypotheses**

$$\mathbf{H}_0 : f_1 = f_2 \text{ versus } \mathbf{H}_0 : f_1 \neq f_2,$$

at the $\alpha = 0.1$ significance level. That is, report the test statistic, critical value, and decision that is made. If your decisions is to not reject the null hypothesis H_0, then **comment on whether**

or not you believe that the test was powerful enough to reject the null hypothesis, based on the sample size.

Part 2 [2 Mark]

Let $f_{1j}(x) = f(x_1|Y=1)$ and $f_{2j}(x) = f(x_2|Y=2)$ be the marginal probability density functions of the gene expression levels of the jth gene for ALL and AML cells, respectively. Let

$$\mathscr{P}(f) = \int_{\mathbb{R}} x f(x)\,\mathrm{d}x$$

be the mean of univariate probability density function $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$. For each $j \in [q]$, **compute a p-value P_j for a test of the hypotheses**

$$\mathbf{H}_0 : \mathscr{P}(f_{1j}) = \mathscr{P}(f_{2j}) \text{ versus } \mathbf{H}_0 : \mathscr{P}(f_{1j}) \neq \mathscr{P}(f_{2j}),$$

using the data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and $Y_1, \ldots, Y_n$. **Plot the p-values $\mathbf{P}_q = (P_1, \ldots, P_q)$ using a histogram.**

Part 3 [2 Mark]

Following from Part 2, **plot the empirical cumulative distribution function (ECDF) for the sample of p-values:**

$$F(p; \mathbf{P}_q) = \frac{1}{q}\sum_{j=1}^{q} [\![p \leq P_j]\!],$$

along with the cumulative distribution function of the uniform distribution on the domain $[0, 1]$. Comment on whether the distribution of p-values is sub-uniform or not, and whether or not this observation conforms with the conclusions made in Part 1.

Part 4 [2 Mark]

Using the Benjamini–Hochberg and Benjamini–Yekutieli methods, **identify sets of genes $j \in [q]$ that are significant at the false discovery rate controlled level of $\alpha_{\mathbf{FD}} = 0.05$. Report how many of the genes are significant under each method, and report the largest p-value that was rejected under each method.**

Part 5 [1 Mark]

Using exploratory techniques and the samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and $Y_1, \ldots, Y_n$, **explain whether you believe that the outcomes from either of the methods applied in Part 4 are valid by way of a discussion of the required assumptions.**

Part 6 [1 Mark]

Prove that

$$\sum_{j=1}^{\infty} \frac{\delta\left(\min\left\{j, m\right\}\right)}{j\left(j+1\right)} \leq 1,$$

for $\delta\left(k\right) = \left(2m\right)^{-1} k\left(k+1\right)$, and provide a false discovery rate step-up rejection procedure based on this observation.

Problem 3 [5 Marks]

Let $\left(\boldsymbol{X}_1, Y_1\right), \ldots, \left(\boldsymbol{X}_n, Y_n\right)$ be an independent and identically distributed sample of pairs of covariates and responses, where $\boldsymbol{X}_i \in \mathbb{R}^q$ and $Y_i \in \mathbb{R}$, where q may be larger than n.

Suppose that

$$\mathrm{E}\left[Y_i | \boldsymbol{X}_i = \boldsymbol{x}_i\right] = \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}_i, \tag{1}$$

for some $\alpha \in \mathbb{R}$ and potentially sparse $\boldsymbol{\beta} \in \mathbb{R}^q$ (here, we take sparse to mean that many of the coordinates $\beta_j = 0$ for many $j \in [q]$), where $\boldsymbol{\beta}^\top = \left(\beta_1, \ldots, \beta_q\right)$. We wish to estimate $\boldsymbol{\beta}$ via the so-called **elastic net-penalized least squares estimator**:

$$\hat{\alpha}, \hat{\boldsymbol{\beta}} = \underset{\alpha, \boldsymbol{\beta} \in \mathbb{R}^q}{\arg\min} \ \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - \alpha - \boldsymbol{\beta}^\top \boldsymbol{X}_i\right)^2 + \lambda \left\{ \left\|\boldsymbol{\beta}\right\|_1 + \left\|\boldsymbol{\beta}\right\|_2^2 \right\}, \tag{2}$$

for some $\lambda \geq 0$.

Part 1 [1 Mark]

Argue that Problem (2) is equivalent to the problem:

$$\hat{\alpha}, \hat{\boldsymbol{\beta}} = \underset{\alpha, \boldsymbol{\beta} \in \mathbb{B}(\gamma)}{\arg\min} \ \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - \alpha - \boldsymbol{\beta}^\top \boldsymbol{X}_i\right)^2, \tag{3}$$

where $\mathbb{B}\left(\gamma\right) = \left\{\boldsymbol{\beta} \in \mathbb{R}^q : \left\|\boldsymbol{\beta}\right\|_1 + \left\|\boldsymbol{\beta}\right\|_2^2 \leq \gamma\right\}$, for some $\gamma > 0$.

Part 2 [1 Mark]

Plot the set $\mathbb{B}\left(\gamma\right) = \left\{\boldsymbol{\beta} \in \mathbb{R}^q : \left\|\boldsymbol{\beta}\right\|_1 + \left\|\boldsymbol{\beta}\right\|_2^2 \leq \gamma\right\}$ for some value of γ and discuss whether you believe that elastic net-penalized least squares estimator can be sparse.

Part 3 [2 Mark]

Devise an algorithm for solving either Problem 2 or Problem 3.

Part 4 [1 Mark]

The rows of the file `prostate.csv` contains $n = 98$ pairs of covariates $\boldsymbol{X}_i \in \mathbb{R}^q$ ($q = 8$; in the first 8 columns) and response Y_i (in the 9th column), for $i \in [n]$. Assuming that these data admit the relationship (2), **compute estimates of α and $\boldsymbol{\beta}$ for these data using the elastic net-penalized least squares estimator for some increasing sequence of γ (or decreasing sequence of $\lambda \geq 0$). Plot the trajectory of the sequence of estimates as γ increases (or λ decreases).**