Australian National University

Student Number u 7192786

Research School of Finance, Actuarial Studies and Statistics

FINAL EXAMINATION

Semester 2, 2022

Final Exam for STAT3050/4050/7050 Advanced Statistical Learning

Writing Time (including Scanning and Submission) 240 minutes

Exam Conditions:

Centrally scheduled examination

Permitted Materials:

Any materials are permitted.

Instructions to Students:

1. This exam paper consists of a total of 12 pages. Please ensure your paper has the correct number of pages.
2. This exam includes a total of 10 questions. The questions are of unequal value, with marks indicated for each question. You must attempt to answer all questions unless directed otherwise.
3. Please include all workings for each question as marks will not be awarded for answers that do not include workings.

Total Marks of Final Exam for STAT3050 = 92 marks
Total Marks of Final Exam for STAT4050/7050 = 100 marks

This exam will count towards 55% of your final grade for the course.

Question 1 True or False [24 marks]

Determine whether the following statements are true or false. For each answer, please provide your detailed reasoning.

(1). **[4 marks]** When the covariates observations $\mathbf{X}$ satisfies the orthonormal condition ($\mathbf{X}\mathbf{X}^\top = \mathbf{I}$), elastic net estimation is equal to lasso estimation. In view of this, elastic net estimation outperforms lasso estimation only when covariates are correlated.

(2). **[4 marks]** Starting from any weak classifier, Adaboost algorithm will result in zero training error with a large number of iterations.

(3). **[4 marks]** When the sample size $n = 60$ and the number of covariates p increases from 1 to 100, the test MSE of the least squares estimator will exhibit the double-descent phenomenon.

(4). **[4 marks]** For clustering problems that can be solved by K-means clustering algorithm, spectral clustering is also applicable.

(5). **[4 marks]** Random forests always performs better than bagging, in the sense of producing estimation of smaller variance.

(6). **[4 marks]** Kernel smooth estimation has bad performances for multivariate regression (the number of covariates $p > 3$) no matter if the sample size is large or small.

(1) FALSE

- Under orthonormal condition, elastic net $= \frac{\hat{\theta}_i(\text{lasso})}{1+\lambda_2}$ so it is a double shrinkage, and no covariate can be correlated ($X^TX = I$)
- elastic net outperform lasso if the covariates are highly correlated as elastic net expriences the grouping effect where lasso only pick a one from highly correlated group

(2) FALSE

- The weak learner should be better than the random guess to reach 0 traning error after large iteration.

(3) FALSE

- The least square estimator will encounter the curse of dimensionality when $p > n$, the (X^TX) will not be invertible. Unless we define a generalized inverse of X^TX as $(X^TX)^+$ then the least square estimate will experience the double descent on test MSE.

(4) True. The spectral clustering can also do a k-mean clustering by computing the unormalized graph Laplacian. If we choise $w_{ii'} = s_{ii'} = \exp(-d_{ii'}^2/c)$ then spectral can also do k-means, clustiy based on ~~Existence~~ distance.

(5) FALSE. Random forest only outperforms bagging when the correlation between trees is greater than 0. If correlation is 0, then we have

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 = \frac{\sigma^2}{B} = \text{Bagging result}$$

In this sense, random forest will not outperform bagging.

(6) True, when $p > 3$, it is impossible to maintain localness and sizeable sample in the neighborhood. Increase the sample size only make worse of the localness and sizeable sample in the neighborhood by introducing more bias and variance.

Question 2 [12 marks]

Consider the penalized least squares estimator $\widehat{\boldsymbol{\theta}} = \left(\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_p\right)$, which is the minimizer of the following objective function

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ji}\theta_j\right)^2 + \sum_{j=1}^{p} \lambda \cdot g(|\theta_j|),$$

where the penalty function $\lambda \cdot g(|z|)$ takes the following form

$$\lambda \cdot g(|z|) = \begin{cases} \lambda \cdot |z|, & |z| \leq \lambda \\ \frac{2\gamma\lambda|z| - z^2 - \lambda^2}{2(\gamma - 1)}, & \lambda < |z| < \gamma\lambda \\ \frac{\lambda^2(\gamma+1)}{2}, & |z| \geq \gamma\lambda. \end{cases}$$

(1). **[3 marks]** Is the estimator $\widehat{\boldsymbol{\theta}}$ unbiased? Please justify your conclusion. ✓

(2). **[3 marks]** Does the estimator $\widehat{\boldsymbol{\theta}}$ satisfy sparsity? Please justify your conclusion. ✓

(3). **[3 marks]** Does the estimator $\widehat{\boldsymbol{\theta}}$ satisfy continuity? Please justify your conclusion. ✗

(4). **[3 marks]** If the penality function $\lambda \cdot g(|z|)$ is changed into the lasso penalty function $\lambda \cdot |z|$, we obtain the lasso estimator $\tilde{\boldsymbol{\theta}}$. Please state one property (among unbiasedness, sparsity, continuity) that $\widehat{\boldsymbol{\theta}}$ satisfies but $\tilde{\boldsymbol{\theta}}$ does not. Please justify your conclusion.

(1) Yes, For large ture parameter $|\theta_j| \geq \gamma\lambda$

$$\lambda\, g(|\theta_j|) = \frac{\lambda^2(\gamma+1)}{2} = g_\lambda(|\theta_j|)$$

$$\Rightarrow g_\lambda'(|\theta_j|) = 0 \quad , \quad |\theta_j| \geq \gamma\lambda$$

$\Rightarrow$ This penality has unbiasedness property.

(2) Sparsity means a thresholding rule, which set small covariates to 0 to reduce model complexity. Then, $|z| < \lambda$, $\lambda \cdot g(|z|)$ is equal to $\lambda \cdot |z|$, This is a lasso (soft thresholding) which shrinks covariates towards 0 and reduce the model complexity.

$$g_\lambda'(|\theta_j|) = \begin{cases} \lambda, & |\theta_j| \le \lambda \\ \gamma\lambda / (\gamma - 1), & \lambda \le |\theta_j| < \gamma\lambda \\ 0, & |\theta_j| \ge \gamma\lambda \end{cases}$$

We can see that no matter where the minimum of function $|\theta_j| + g_\lambda'(|\theta_j|)$ is, the function is always positive as $\lambda, \gamma > 0$

(3)

$$|\theta_j| + g_\lambda'(|\theta_j|) = \begin{cases} \lambda + |\theta_j|, & |\theta_j| \le \lambda \\ r\lambda/r+1 + |\theta_j|, & \lambda \le |\theta_j| < r\lambda \\ |\theta_j|, & |\theta_j| \ge r\lambda \end{cases}$$

As $\lambda, r > 0$, the minimum of the function attains at $|\theta_j|$, However, in the case $|\theta_j| \ge r\lambda > 0$. The minimum is not attained at 0, so the continuity property does not held.

(4) Unbiasedness. The lasso is biased.

the lasso $g(\cdot)$ can be written as:

$$g_\lambda(|\theta_j|) = \lambda|\theta_j|$$

$$g'_\lambda(|\theta_j|) = \lambda \neq 0$$

Then lasso is biased, but SCAD is not.

Question 3 [14 marks]

Consider a clustering problem with two features x_1 and x_2. In Figure 1, we have four observations ①, ②, ③ and ④, which are denoted by (x_{11}, x_{21}), (x_{12}, x_{22}), (x_{13}, x_{23}) and (x_{14}, x_{24}), respectively.

(a). **[2 marks]** Define the dissimilarity measure as

$$D(①, ⓚ) = (x_{1i} - x_{1k})^2 + (x_{2i} - x_{2k})^2, \quad i, k = 1, 2, 3, 4.$$

Based on the proximity matrix $\mathbf{D} = (D(①, ⓚ))_{4\times 4}$, please cluster these four observations into two clusters. Please provide your reasoning.

(b). **[4 marks]** One researcher aims to apply PCA on these four observations ①, ②, ③ and ④, and keep the first principal component scores z_1, z_2, z_3 and z_4, respectively. Then this researcher applies the following dissimilarity measure

$$\bar{D}(①, ⓚ) = (z_i - z_k)^2, \quad i, k = 1, 2, 3, 4.$$

Please help this researcher obtain the clustering result based on the proximity matrix $\bar{\mathbf{D}} = \left(\bar{D}(①, ⓚ)\right)_{4\times 4}$. Please provide your reasoning.

(c). **[4 marks]** Consider another dissimilarity measure

$$\widetilde{D}(①, ⓚ) = w_1 (x_{1i} - x_{1k})^2 + w_2 (x_{2i} - x_{2k})^2, \quad i, k = 1, 2, 3, 4. \tag{0.1}$$

where

$$w_1 = \frac{1}{2\widehat{var}(x_1)}, \quad w_2 = \frac{1}{2\widehat{var}(x_2)},$$

with $\widehat{var}(x_1)$ and $\widehat{var}(x_2)$ are sample variance of x_1 and x_2, respectively.

In terms of the proximity matrix $\widetilde{\mathbf{D}} = \left(\widetilde{D}(①, ⓚ)\right)_{4\times 4}$, please cluster these four observations into two clusters. Please provide your reasoning.

(d). **[4 marks]** Please create a new proximity matrix $\widehat{\mathbf{D}} = \left(\widehat{D}(①, ⓚ)\right)_{4\times 4}$, under which the clustering result is

Cluster 1: ①, ④.

Cluster 2: ②, ③.

Please justify your results.

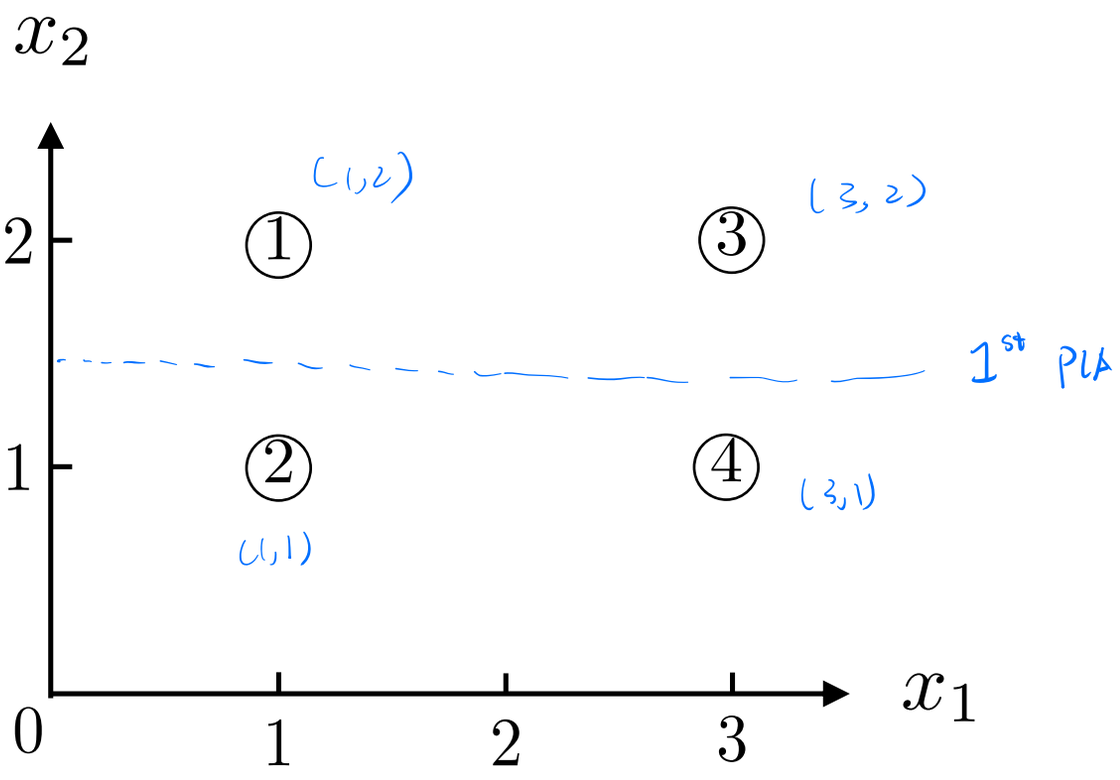


Figure 1: Question 3

(a)

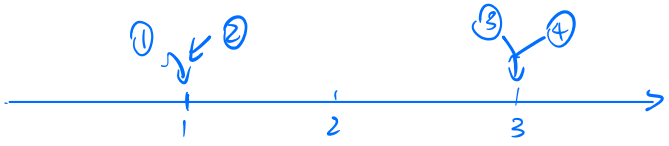
$$D = \begin{array}{c} i \backslash k \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & D(①,②) & D(①,③) & D(①,④) \\ D(②,①) & 0 & \cdots & \\ \vdots & \ddots & 0 & \vdots \\ & & & 0 \end{pmatrix} \end{array}$$

The proximity has each element as the l2 distance. Then, the point with smallest distance (dissimilarity) will be grouped into one cluster.

Cluster 1: ① ②

Cluster 2: ③ ④

(b) The first PCA score is shown above, then we have



then

$$D = \begin{pmatrix} 0 & 0 & 4 & 4 \\ 0 & 0 & 4 & 4 \\ 4 & 4 & 0 & 0 \\ 4 & 4 & 0 & 0 \end{pmatrix}$$

then, cluster 1: ① ②

cluster 2: ③ ④

(c) $\widehat{var}(x_1) = (1-2)^2 + (3-2)^2 = 2$

$\widehat{var}(x_2) = (1-1.5)^2 + (2-1.5)^2 = 0.5$

$\tilde{D}(①, ④) = \frac{1}{4}(x_{1i} - x_{1k})^2 + (x_{2i} - x_{2k})^2, \quad i, k \in \{1, 2, 3, 4\}$

$$\tilde{D} = \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix}$$

Now depends on how we initialize the starting points.

we can get:

cluster 1 : ① ② or ① ③

cluster 2 : ③ ④ or ② ④

(d) define $D(①, ⓚ) = (x_{1i} - x_{1k})^2 + (x_{2i} - x_{2k})^2, \quad i, k = 1, 2, 3, 4.$

$G(①, ⓚ) = (x_{1i} - x_{1k})^2 + (x_{2i} - x_{2k})^2, \quad i, k \in \{1, 2, 3, 4\}$

$D(①, ⓚ) = \exp\{-G(①, ⓚ)\}$

When $G(\cdot,\cdot)$ is the largest, $D(①, ⓚ)$ is the smallest.

Thus the clustering result is like : cluster 1 : ① ④

cluster 2 ② ③

Question 4 [6 marks]

(a). **[3 marks]** There are five observations (denoted as A or B in Figure 2) for two features (x_1, x_2). These five observations belong to two clusters, i.e. cluster A or cluster B.

Please propose a clustering method with these five observations as training data. Please describe the procedure of your proposed method.

(b). **[3 marks]** Consider a classification problem as illustrated in Figure 2. There are two covariates x_1 and x_2. The response variable takes value A or B. In total, we have five observations (as training data) in Figure 2.

Please propose a method for this classification problem. Please describe the procedure of your proposed method.

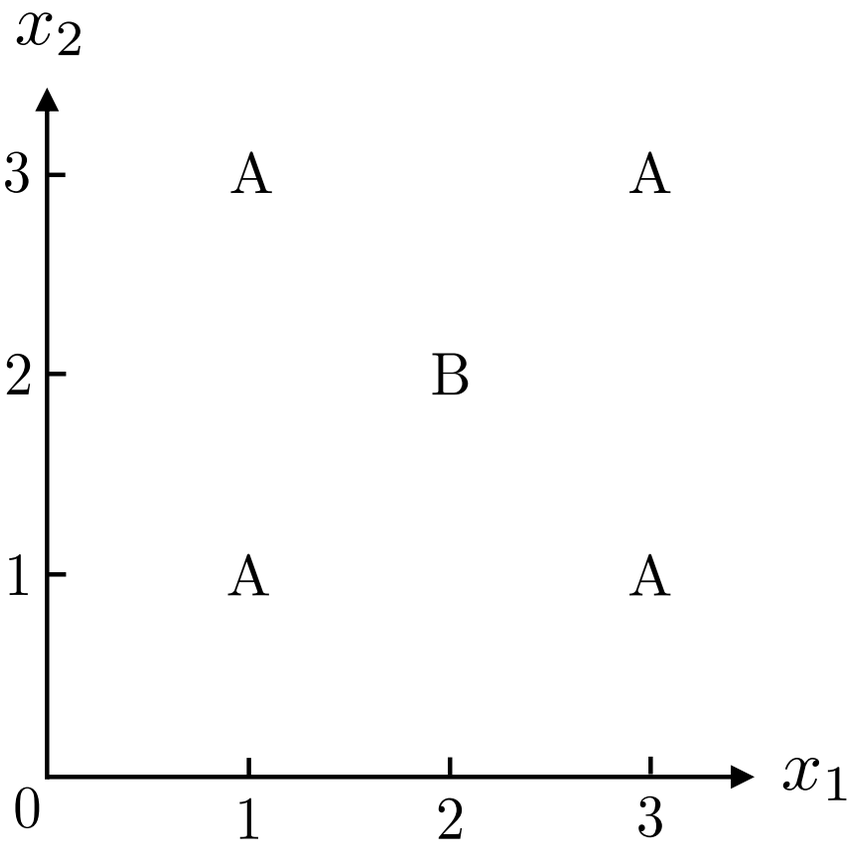


Figure 2: Question 4

(a) Non linear boundary : Spectral clustering

① Graph fully connected pairwise edges for all points.

$$W_{ii'} = S_{ii'}$$

$$W = \{ W_{ii'} \}$$

② Compute $g_i = \sum_{i'} W_{ii'}$

Write $G :=$ a diagonal matrix with digonal elements g_i.

③ Write graph Laplacian Matrix:

$$L = I - G^{-1} W$$

④ Finds m eigen vectors $Z_{N \times m}$ regarding m smallest non-zero eigenvalue of L

⑤ Do normal clustering on row of Z to yield a result

(b) We can do a SVM

① Define a Kernel mapping $\phi(x)$

② map the feature into the RKHS

③ Do a SVC in the RKHS

Question 5 [6 marks]

Consider a simple linear regression model

$$y_i = \beta x_i + \varepsilon_i, \quad i = 1, 2, 3, 4,$$

where $\beta = 1$. The four observations A, B, C, D are illustrated in Figure 3. One researcher intends to apply Gradient Boosting, but he/she finds that, the simple model (individual tree with depth 1) in the first step has attained zero training error.

(a). **[3 marks]** Please help this researcher to analyse the reason why Gradient Boosting fails in this example.

(b). **[3 marks]** Please propose an alternative estimation approach for this linear regression model. Please describe the procedure of your estimation approach.

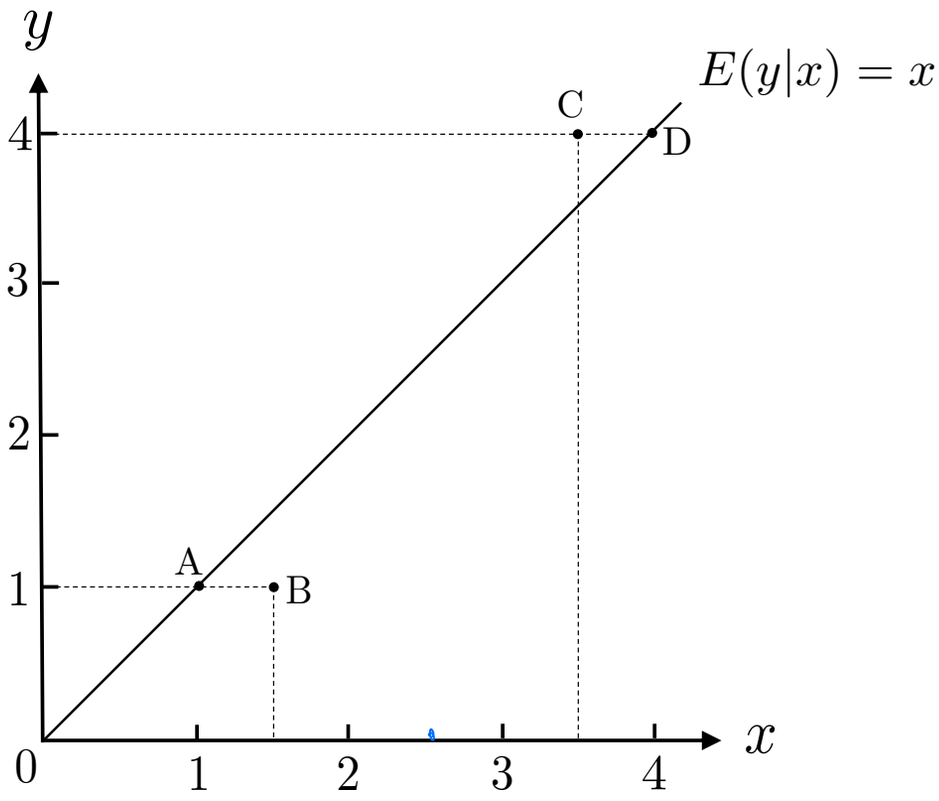


Figure 3: Question 5

(a)

$y = 1$ or $y = 4$

If we apply gradient boosting,

$$T^{(0)} = I\{x < 2.5\} \cdot 1 + I\{x > 2.5\} \cdot 4$$

Then our residual:

$$\{r_1, r_2, r_3, r_4\} = \{0, 0, 0, 0\}$$

We don't have residual for next round.

$\Rightarrow$ Gradient boosting should not have zero training error

Result for Gradient boosting to fail,

① too less covariates

② too less observations.

③ the y_i has only two classes makes it a classification instead of regression problem.

(b) We can apply a generalized lasso.

Objective function:

$$\sum_{i=1}^{n} (y_i - \beta x_i)^2 + \lambda \|\beta - 1\|_1$$

We can minimize the objective function and get β. as we know β is sufficiently closed to 1, the lasso shrinkage will shrink it to 1.

Question 6 [8 marks]

Figure 4 shows a 2-hidden layer neural network estimator for the regression function $g(x)$ ($x = (x_1, x_2, x_3, x_4)$) in the regression model

$$\mathbb{E}(Y|x) = g(x).$$

(a). **[4 marks]** Please write a formula for the constructed neural network in Figure 4.

(b). **[4 marks]** One researcher completes the estimation for this neural network. He finds that

(1). This neural network performs well in terms of small test MSE.

(2). The weights parameters have the sparsity property: for weights parameters between the input layer and the first hidden layer, only weights connected with x_2 and x_4 are nonzero.

Based on this information, this researcher decides to simplify the regression model into an additive model

$$g(x) = g_1(x_2) + g_1(x_4),$$

and then uses a backfitting algorithm to estimate the unknown function $g(x)$. The researcher claims this alternative method will also produce accurate prediction in the sense of small test MSE.

Do you agree with this researcher? Please justify your conclusion.

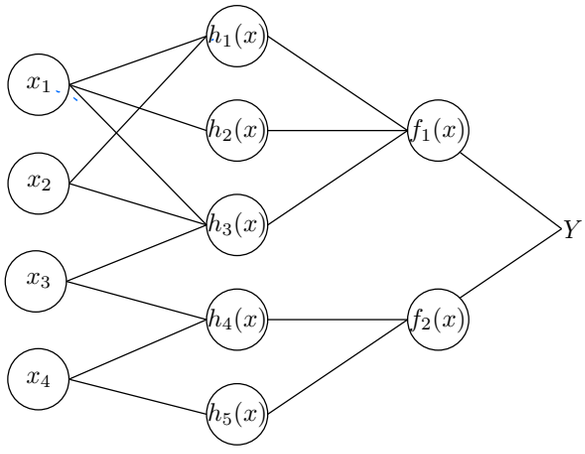


Figure 4: Question 6

(1) Define activation function as $S(\cdot)$.

first layers:

$$h_1(x) = S[w_{10} + (w_{11}x_1 + w_{12}x_2)]$$

$$h_2(x) = S(w_{20} + w_{22}x_2)$$

$$h_3(x) = S[w_{30} + (w_{31}x_1 + w_{32}x_2 + w_{33}x_3)]$$

$$h_4(x) = S[w_{40} + (w_{43}x_3 + w_{44}x_4)]$$

$$h_5(x) = S(w_{50} + w_{55}x_5)$$

2nd layers:

$$f_1(x) = S([w'_{10} + w'_{11}h_1(x) + w'_{12}h_2(x) + w'_{13}h_3(x)])$$

$$f_2(x) = S(w'_{20} + w'_{24}h_4(x) + w'_{25}h_5(x))$$

Output:

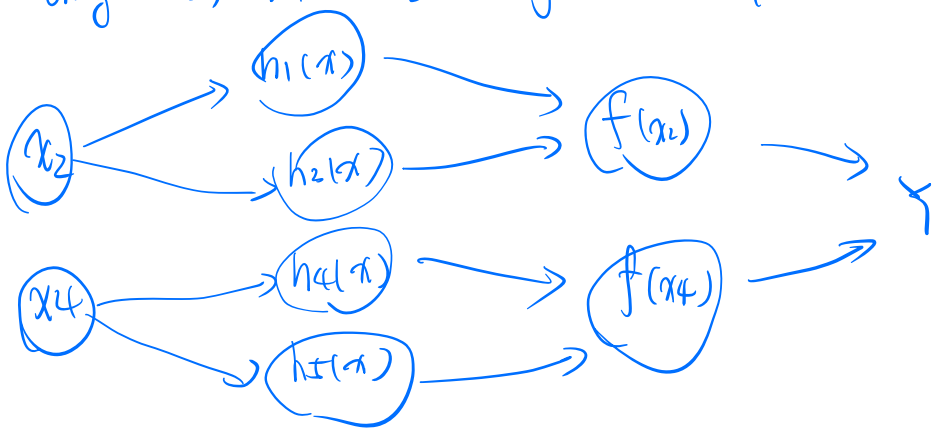
$$Y = \beta_0 + \beta_1 f_1(x) + \beta_2 f_2(x)$$

One formula:

$$y(x) = \beta_0 + \sum_{k=1}^{2} \beta_k S\left(w'_{k0} + \sum_{j=1}^{p} w'_{kj} S\left(w_{k0} + \sum_{i=1}^{m} w_{ki}x_i\right)\right)$$

(2) No, not neccessarily.

If only x_2, x_4 has weight not equal to 0.



We see that:

$f(x_2)$ is only about x_2

$f(x_4)$ is only about x_4

However the last layer:

$$Y = \beta_0 + \beta_1 f(x_2) + \beta_2 f(x_4)$$

the addictive model can only capture

$$\beta_1 f(x_2) + \beta_2 f(x_4)$$

but not β_0.

So the addictive model may not perform the same.

Question 7 [6 marks]

Consider a linear regression model

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i, \quad i = 1, 2, \ldots, 60,$$

where $\beta_1 = 1.5$, $\beta_2 = 1.5$, $\beta_3 = 10$ and $\beta_4 = \beta_5 = 0$.

(a). **[3 marks]** Suppose the data matrix comprised of all covariates $\mathbf{X} = (x_{ji})_{5\times 60}$ is an orthonormal matrix, i.e. $\mathbf{X}\mathbf{X}^\top$ is an identity matrix. Under this case, one researcher applies the naive elastic net estimation with tuning parameters $\lambda_1 = 3$ and $\lambda_2 = 3$. However, the estimation results for β_1, β_2 and β_3 are quite bad.

Please help the researcher analyse the reason and modify the estimation approach to improve the estimation results.

(b). **[3 marks]** Suppose the two covariates x_{1i} and x_{2i} are highly correlated. Under this case, one researchers utilizes the lasso estimation and also derive incorrect estimation results for coefficients.

Please help the researcher analyse the reason and propose an alternative estimation approach to improve the estimation results.

(a) Under orthonormal design:

$$\tilde{\beta}_i \text{ (naive elastic net)} = \frac{\left(|\hat{\beta}_i \text{ (OLS)}| - \frac{\lambda_1}{2}\right)_+}{1+\lambda_2} \operatorname{sign}\left(\hat{\beta}_i \text{ (OLS)}\right).$$

This is equal to a lasso and the ridge shrinkage.

So $\beta_4 = \beta_5 = 0$, but $\beta_1, \beta_2, \beta_3$ are bad.

If we use elastic net:

$$\hat{\beta}_i \text{ (elastic net)} = (1+\lambda_2)\, \tilde{\beta}_i \text{ (naive elastic net)}$$

then we can solve this problem.

(b) Grouping effect. If two covariates are highly correlated, then lasso randomly picks one, but elastic net will yield similar estimation to group of highly correlated covariates

=> Use elastic net

Question 8 [8 marks]

The data $(x_i, y_i), i = 1, 2, \ldots, n$ are generated from the following regression model

$$y_i = \beta_1 x_i^2 + \beta_2 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where $\beta_1 = 2$; $\beta_2 = 3$; x_i and ε_i both follow standard normal distribution; and x_i is independent of the error component ε_i.

One researcher applies least squares estimation to regress y_i on the covariate x_i^2, and obtains the estimator

$$\widehat{\beta}_1 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^4\right)^{-1} \times \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 y_i\right).$$

This researcher has the following two claims:

- The estimator $\widehat{\beta}_1$, as an estimator for the parameter β_1, is unbiased but of large variance;
- bagging can be applied to decrease the variance of this estimator.

Do you think these claims are correct? Please provide your reasoning.

$Cov(X, X^2) = E(X \cdot X^2) - E(X) \cdot E(X^2)$

$= 0$

We get that X and X^2 are not correlated.

$\Rightarrow$ no multi-collinearity

$\Rightarrow$ unbiased

If biased is 0. $MSE(M) = Var(M) + Bias(M)$ (model M)

$\Rightarrow MSE(M) = Var(M)$

Yes, large variance

- Bagging can be applied this case as we don't have bias just variance and $Var(bagging) = \frac{\sigma^2}{B}$, B is numbe of bootstrap sample.

Question 9 [8 marks]

(a). **[4 marks]** Consider four random variables x_1, x_2, x_3, x_4 that satisfy the following relations

$$x_1 = z_1 + y_1, \quad x_2 = z_1 + y_2, \quad x_3 = y_2 + z_2, \quad x_4 = z_2 + y_1,$$

where z_1, z_2, y_1, y_2 are independent and follow standard normal distributions.

Plot the undirected markov graph for these four random variables x_1, x_2, x_3, x_4, which can reflect the pairwise conditional independence or conditional dependence. Please provide the reasoning as well.

(b). **[4 marks]** Please construct four random variables x_1, x_2, x_3, x_4, which satisfy the undirected markov graph given in Figure 5. Please justify your result.

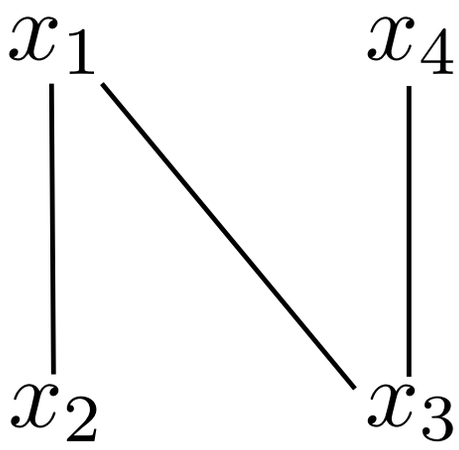


Figure 5: Question 8

(a) Assume independent:

$$P(x_1, x_2 \mid x_3, x_4) = P(x_1 \mid x_3, x_4) \cdot P(x_2 \mid x_3, x_4)$$

$$= \frac{P(x_1, x_3, x_4) \cdot P(x_2, x_3, x_4)}{P(x_3, x_4)^2}$$

$$P(x_1, x_2, x_3, x_4) \triangleq P(x_1, x_2 \mid x_3, x_4) \cdot P(x_3, x_4)$$

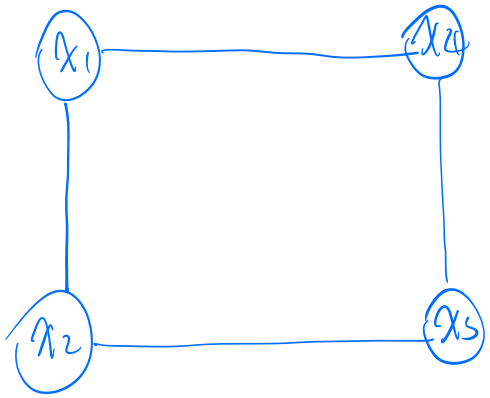
$$P(x_1, x_2 \mid x_3, x_4) = \frac{P(x_1, x_2, x_3, x_4)}{P(x_3, x_4)}$$

x_1, x_2 will only be independent given z_1

$\Rightarrow$ no z_1 in $y_2 + z_2$ and $z_2 + y_1$ (x_3, x_4)

$\Rightarrow$ x_1 and x_2 are conditionally dependent

Follow the same reason, we get:



(b) define $a \sim N(0,1)$

$b \sim N(0,1)$

$c \sim N(0,1)$

Then if $X_1 = a + b$

$x_2 = b$

$x_3 = c + a$

$x_4 = c$

Then:

$\Rightarrow$

Question 10 [8 marks] (Only for students from STAT4050/7050)

Consider the regression model

$$y_i = \sin(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

The sample size is $n = 20$. The predictor x_i follows a uniform distribution on $[-5, 5]$ while the error component ε_i is from a normal distribution with zero mean and variance 0.3.

Now we use basis expansion (fitting a natural spline) to estimate the regression function $\sin(x)$. Fitting a natural spline with d degrees of freedom amounts to fitting a least-squares regression of the response onto a set of d basis functions. The basis functions are denoted as $\{B_\ell(x_i),\ \ell = 1, 2, \ldots, d\}$.

The minimum-norm estimator for the linear coefficient in basis expansion is $\left(\mathbf{B}\mathbf{B}^\top\right)^- \mathbf{B}\mathbf{Y}$, where $\mathbf{B} = (B_\ell(x_i))_{d\times n}$, $\mathbf{Y} = (y_1, y_2, \ldots, y_n)^\top$, and $\left(\mathbf{B}\mathbf{B}^\top\right)^-$ is the generalized inverse of the matrix $\mathbf{B}\mathbf{B}^\top$.

Figure 6 shows the training error (orange line) and test error (blue line), respectively. The horizontal line is degree of freedom d.

(a). **[4 marks]** If the sample size $n = 100$, please plot training error and test error lines, as the degrees of freedom d increases from 2 to 50. (It is sufficient to plot the general tendency for these two lines.) Please provide your reasoning.

(b). **[4 marks]** If the sample size $n = 10$, please plot training error and test error lines, as the degrees of freedom d increases from 2 to 50. (It is sufficient to plot the general tendency for these two lines.) Please provide your reasoning.

(a) $\sin(x_i)$ is infinite differentiable function on the real line

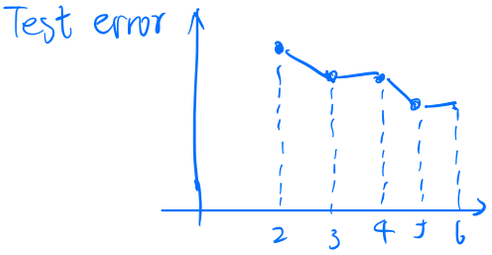
Taylor expansion: $\sin(x_i) = 0 + 1 x_i + 0 x_i^2 + \frac{-1}{3!} x^3 + 0 x^4 + \cdots$

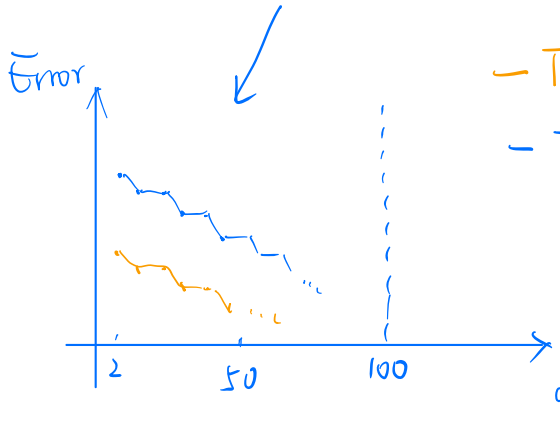
$= x_i - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} \cdots$

natural spline with d degree of freedom:

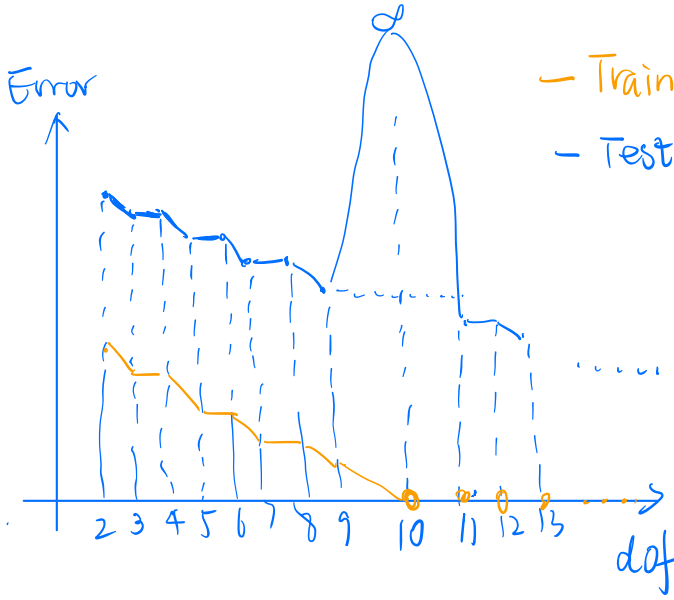
$b_1(x) = x_i^1 \cdots b_d(x) = x^d$, x^2, x^4 ... have little effect on predicting the true function

Then from 2 on, we have:





(b)



Test MSE keep decreasing until to

Traing MSE attains 0

At dof = 10

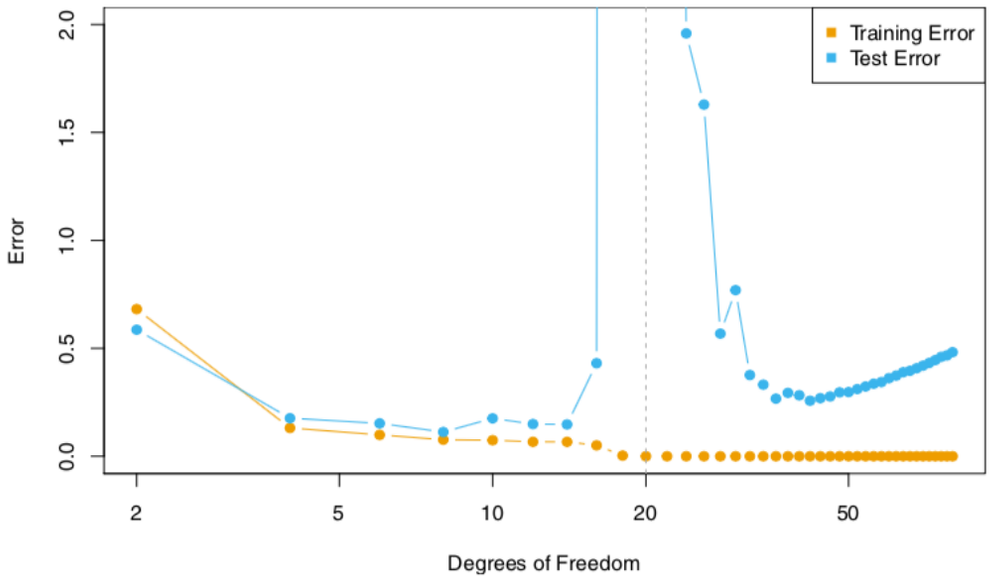


Figure 6: Question 10

END OF EXAMINATION