Assignment 7 (3 pages)

Statistics 32950-24620 (Spring 2024)

Due 9 am Tuesday, May 14. (Reminder: In-class final Thursday May 16th.)

References (ref. books are listed in Syllabus on Canvas)
Sections 14.5 (on PCA and generalizations), 14.7 (on ICA) in Hastie, Tibshirani and Friedman.
Sections 10.1-10.4 (on ICA), 13.4 (on Sparse PCA) in Koch.
Chapter 9 (on Mixture models and EM) in *Pattern Recognition and Machine Learning* by Bishop.

Problem assignments:

1. (*Least squares vs Ridge regression*) Consider the linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Generate the data using the following R commands.

```
x1=rnorm(30)
x2=x1+rnorm(30,sd=0.01)
Y=rnorm(30,mean=3+x1+x2)
```

(a) Write the fitted model with estimated parameters by the Lease Squares method (LS).
The R command for fitting the LS model is `lm(Y∼x1+x2)`.

(b) What is the true model with the true β_i's? Are the parameter estimates of the LS model in (a) good? Why so?

(c) Compute the residual sum of squares (RSS) of the fitted LS model and the RSS of the true model.

$$RSS = \sum_{j=1}^{n} \left[y_j - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} \right) \right]^2$$

Are the two RSS comparable (or close in numerical values)? Give a reason (or an excuse) of performance of the LS parameter estimates in (a) (i.e. on whether bad parameter estimates could yield not so bad prediction values).

(d) Use the R function `lm.ridge` to fit a Ridge regression model with $\lambda = 1$:

```
library(MASS)
lm.ridge(Y~x1+x2, lambda=1)
```

Write out the fitted Ridge model. Are the parameter estimates good?

(e) (Comparison and comments) What is the criterion of LS method? That is, which function of the model parameters does LS method try to optimize? What is the function of model parameters that Ridge regression method tries to optimize? Compare the two methods using the results in (a) and (d). What is the effect on parameter estimates by the Ridge regression method?

2. (*LASSO regression exercise*) The dataset `Boston` of 506 observations and 14 variables is on housing values in the suburbs of Boston. The data and variable names can be obtained in R by the commands below.

```
library(MASS)
data(Boston)
colnames(Boston)
 [1] "crim"    "zn"     "indus"   "chas"    "nox"    "rm"     "age"    "dis"    "rad"    "tax"
[11] "ptratio"  "black"   "lstat"   "medv"
```

The following describe the variables (variable names in capital letters).

```
1. CRIM: per capita crime rate by town
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX: nitric oxides concentration (parts per 10 million)
6. RM: average number of rooms per dwelling
7. AGE: proportion of owner-occupied units built prior to 1940
```

```
8. DIS: weighted distances to five Boston employment centres
9. RAD: index of accessibility to radial highways
10. TAX: full-value property-tax rate per $10,000
11. PTRATIO: pupil-teacher ratio by town
12. BLACK: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT: % lower status of the population
14. MEDV: Median value of owner-occupied homes in $1000's
```

More reference information about the data can be found at
https://www.rdocumentation.org/packages/mlbench/versions/2.1-1/topics/BostonHousing.

(a) Take the variable `medv` (median value of owner-occupied homes) as the response variable to fit LASSO regression models, using the first 300 observations as the training set and the rest (206 observations) for validation (or calibration; this part of the data is not to be used in cross validation). Interpret your results.

(b) Compare your fitted LASSO model with the linear model fitted by the ordinary least squares method. Comment.

The following commands are for your reference.

```
Tdata = Boston[1:300,]
Cdata = Boston[301:506,]
X=as.matrix(Tdata[,1:13])
Y=Tdata[,14]
```

3. (*PCA vs Sparse PCA*)

The data set hearlossData.csv (may automatic download when clicked, also available next to the link of this p-set in Canvas) can be input into R by the following commands.

```
data = read.csv("hearlossData.csv")   # or data = read.csv("hearlossData.csv", header=FALSE)
colnames(data)=c("Left5c","Left1k","Left2k","Left4k","Right5c","Right1k","Right2k","Right4k")
```

The data consists of 100 observations from males, aged 39. The measurements are decibel loss (in comparison to a reference standard) at frequencies 500Hz, 1000Hz, 2000Hz and 4000Hz for the left and the right ear, respectively. More detailed information can be found in Chapter 5 in the library e-book A user's guide to principal components by Jackson.

(a) Conduct a principal component analysis.

(b) Conduct a sparse principal component analysis to highlight important frequency relationship in hearing and hearing loss.

4. (*PCA vs ICA*)

The data tableICA (may automatic download when clicked, also available next to the link of this p-set in Canvas) can be read in R by the command below.

```
X = read.table("tableICA")
```

(a) Conduct a Principal Component Analysis. Plot the observations in the space of the first two principal components. Provide the screeplot. Comment on the PCA results.

(b) Conduct an Independent Component Analysis. Interpret your results. Plot the three independent components recovered.

(c) Compare (a) and (b) and comment.

5. (*Entropy as a measure of non-Gaussian-ness*)

In Independent Component Analysis, we may use Excess Kurtosis to seek components that are as far away from normal distributions as possible. Entropy can also be used as a measure of non-Gaussian-ness.

The Differential Entropy (or continuous entropy) for a continuous random variable X with probability density function $f(x)$ is defined as

$$H(X) = -\int_{\mathbb{R}} f(x) \log f(x) dx$$

where $0 \log 0$ is defined as 0, log stands for $\log_a$ for some constant a such as $a = 2$. Here let's use $a = e$.

(a) Derive that univariate normal random variable with mean μ and variance σ^2 has entropy $\log\left(\sigma\sqrt{2\pi e}\right)$.

(b) Let X be a continuous random variable with mean zero, variance σ^2, and density function $f(x)$. Let $\phi(x)$ be the density function of a normal random variable with mean zero and variance σ^2. Show that

$$-\int_{\mathbb{R}} f(x)\log\phi(x)dx = \log\left(\sigma\sqrt{2\pi e}\right)$$

(c) (*Maximum property of normal random variables*)
Let X be any continuous random variable on the real line with mean zero and variance σ^2. Show that the differential entropy of X is smaller than the differential entropy of a normal random variable with mean zero and variance σ^2; equality holds if and only if X is of normal distribution.

(Hint: Use Jensen's Inequality $h(\mathbb{E}(X)) \leq \mathbb{E}(h(X))$ for any convex function h and integrable continuous random variable X.)

(d) (*Maximize entropy for sums of random variables*)
For $i = 1, 2$, let X_i be any continuous random variable on the real line with mean zero and variance σ_i^2. Let $Y = X_1 + X_2$. Find $\max_{X_1,X_2} H(Y)$. Describe your choice of X_1, X_2 which yield the maximum entropy of Y.

6. (*Mixture Bernoulli likelihood exercise, preliminary steps in EM*)

In this exercise, you will derive preparation setup in maximum likelihood estimation (by the EM algorithm) for a mixture of multivariate Bernoulli distributions, widely used in high-dimensional binary data network models.

Suppose $X = [X_1 \ \cdots \ X_p]'$, where each component X_i is an independent Bernoulli random variable with parameter ν_i, so $X_i = \begin{cases} 1, & \text{with probability } \nu_i, \\ 0, & \text{with probability } 1-\nu_i, \end{cases}$ for $i = 1, \cdots, p$. Then $X = [X_1 \ \cdots \ X_p]'$ is a p-variate Bernoulli vector.

(a) Write out $\mathbb{E}(X)$. Indicate the dimensions of the vector.

(b) Derive $Cov(X)$. Indicate the dimensions of the matrix.

(c) Now consider Y to be of mixture distribution: with probability π_c, Y is from a p-variate Bernoulli distribution with mean $\boldsymbol{\mu}_c = [\mu_{c1}, \cdots, \mu_{cp}]$ and covariance Σ_c, $\sum_{c=1}^K \pi_c = 1$. Denote $\boldsymbol{\pi} = [\pi_1 \ \cdots \ \pi_K]$.

i. Write an expression for the mean vector $\boldsymbol{\mu} = \mathbb{E}(Y)$.

ii. Write an expression for the conditional probability $P(\boldsymbol{y}|\boldsymbol{\mu}_c) = \mathbb{P}(Y = \boldsymbol{y}|\boldsymbol{y}$ is from cluster c with mean $\boldsymbol{\mu}_c)$, where $\boldsymbol{y} = [y_1 \ \cdots \ y_p]'$ is a realization of Y.

iii. Write an expression for
$P(\boldsymbol{y}|\boldsymbol{\pi}, \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K) = \mathbb{P}(Y = \boldsymbol{y}|\boldsymbol{y}$ is from a mixture model with parameters $\boldsymbol{\pi}, \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K)$.

iv. Let C be the class membership variable, $\mathbb{P}(C = c) = \pi_c, c = 1, \cdots, K$. Use vector version of the formula for random variable $Var(X) = \mathbb{E}(Var(X|C)) + Var(\mathbb{E}(X|C))$ to find an expression for $Cov(Y)$.

(d) Suppose $\boldsymbol{y}^{(i)}$, $i = 1, \cdots, n$ are n independent observations of Y in (c) (which is of the mixture distribution). Each $\boldsymbol{y}^{(i)} = [y_{i1}, \cdots, y_{ip}]'$ is from a p-variate Bernoulli distribution with mean $\boldsymbol{\mu}_c$ and covariance Σ_c with probability π_c, $\sum_{c=1}^K \pi_c = 1$. Denote the data as $\boldsymbol{Y} = \{\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(n)}\}$.

i. Write an expression (in terms of y_{ij}, π_c, μ_{ci}) for the likelihood function $L(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\pi}|\boldsymbol{Y}) = P(\boldsymbol{Y}|\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\pi})$.

ii. Create (i.e. make up, invent) a (non-degenerate) numerical data set $[y_{ij}]_{n\times p}$ with $p = 3, n = 5$.

iii. Write out the likelihood function in (i) in terms of π_c, μ_{ci} and the numerical values of your y_{ij}.