Homework Problems 5

May 3, 2020

1. Principal components analysis. Analyze Boston housing dataset ("HW5_BostonHousing.RDAT"). Show all your work in using R.

   (a) Remove variable "chas" from the data.

   (b) Find the sample covariance matrix S.

   (c) Find the eigenvalues and eigenvectors of S.

   (d) Numerically verify that $S = V\Lambda V^T$ (see the slides).

   (e) Find the principal components for the data.

   (f) Find the variance for each principal components.

   (g) Plot the ordered eigenvalues.

   (h) Plot the first eigenvector vs the second eigenvector and explain what you can observe from the plot.

   (i) Give a *biplot* for PCA and explain what you can see there.

2. Factor analysis. Analyze data from a survey on views about Wikipedia usage among university instructors ("HW5_wiki.xlsx"). Show all your work in using R.

   (a) Remove rows with missing data.

   (b) Conduct a factor analysis on the data, using $k = 6$.

   (c) Display loading matrix obtained in step (b) and explain their meanings.

   (d) Display "uniquenesses" obtained in step (b) and explain their meanings.

   (e) Group survey questions according to dominant latent variables.

   (f) What can say about these groups?

3. Cluster analysis. Analyze Swiss banknotes data ("HW5_banknotes.csv"). The dataset consist of 150 data points and it is known that some of these banknotes are real and some are counterfeit. Among these data points, we only know the 51th point is from a genuine banknote. The goal is to identify as many counterfeit notes as possible in the dataset. Show all your work in using R.

   (a) Perform a k-mean clustering with $k = 2$.

   (b) Divid the 150 data points into $k = 2$ subsets according to their cluster assignments obtained from step (a).

(c) Find out which cluster the 11th points is assigned to.

(d) List the indices of all the data points corresponding to counterfeit banknotes.

The following problems are for graduate students.

1. Suppose the columns of matrix $\mathcal{Y}$ are the sample principal components of data matrix $\mathcal{X}$. Let $S_{\mathcal{X}}$ and $S_{\mathcal{Y}}$ be the sample covariance matrices of $\mathcal{X}$ and $\mathcal{Y}$ respectively. Show

   (a) $S_{\mathcal{Y}} = \text{diag}(\lambda_1, ..., \lambda_p)$, where λ_j are the sample eigenvalues of $S_{\mathcal{X}}$.

   (b) $\text{trace}(S_{\mathcal{X}}) = \text{trace}(S_{\mathcal{Y}})$.

2. Suppose X is a p-dimensional random vector with mean vector μ_X and covariance matrix Σ_X which has the spectral decomposition $\Sigma_X = UEU^T$, where U is an orthonormal matrix and $E = \text{diag}(e_1, ..., e_p)$. The principal components of X are defined as the components in vector $Y = (Y_1, ..., Y_p)^T = U^T(X - \mu_X)$. Through finding the covariance matrix Σ_Y to show that

   (a) $Var(Y_j) = e_j, \forall j$.

   (b) $Cov(Y_j, Y_{j'}) = 0, \forall j \neq j'$.

3. Consider the factor analysis model $X = QF + U$, where F and U are random vectors and Q is a constant matrix. Assume the orthogonal factor model conditions from the slides. Show that
$$\Sigma_X = QQ^T + \Psi.$$

4. Suppose $\mathcal{X}$ is from a multivariate normal distribution $N(\mu, \Sigma)$. Show the log-likelihood function is
$$\ell(\mathcal{X}, \bar{x}, \Sigma) = -\frac{1}{2}\left\{n \log\left(\det\left(2\pi\Sigma\right)\right) + (n-1)\text{trace}\left(\Sigma^{-1}S\right)\right\}.$$
   Hint: you need to use
$$(x_i - \bar{x})^T\Sigma^{-1}(x_i - \bar{x}) = \text{trace}\left((x_i - \bar{x})^T\Sigma^{-1}(x_i - \bar{x})\right)$$
   and
$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T.$$