COLUMBIA UNIVERSITY
Statistical Inference and Modeling
Graded Homework 2

Prof. Marco Avella
TA: Ian Kinsella, Shun Xu
Spring 2019

You are free to present your solutions to the exercises below in *groups of at most three*. This homework is due on March 9. Buena suerte!

Exercise 1

The dataset **cancer.txt** contains head and neck cancer data from the Northeren California Oncology Group (NCOG). The data matrix gives detailed time information, with `t` (months till death/censoring) being used in our analysis. The column **d** is the death(1)/censoring(0) indicator, and ”`arm`” the treatment arm (A Chemotherapy, B Chemotherapy + Radiation). Use this datset to provide a minimum amount of `R` output to justify your answers.

1. Do you think that it is reasonable to assume that the right censoring in this data set is random? why?
2. Plot the Kaplan-Meier estimators of the survival curves of the two treatment groups. Does the plot seem to indicate a difference between these groups? which treatment seems to be more efficient?
3. Fit an exponential model to using the function `survreg`. Do the signs of the fitted parameters agree with your intuition from last point?
4. Using point 3, do we observe a significant difference between the two groups based on the likelihood ratio statistic? Does this conclusion depend on parametric model assumptions?
5. Do a visual model check using the fitted exponential models and the Kaplan-Meier fits. Comment briefly what you observe.
6. Try fitting a Weibull model now. Does this model fit the data better? Does this new model provide evidence against the relevance on an exponential model?

Exercise 2

The purpose of this exercise is to explore the numerical performance of the following approaches to missing data:

(a) Complete case analysis.

(b) Available case analysis.

(c) Mean imputation

(d) Mean imputation with the bootstrap.

(e) The EM-algorithm

We will start by looking at the Student Score Data[1] which has both a small sample size and missing observations. Let λ_1 denote the largest eigenvalue of the population covariance matrix Σ and $\hat{\lambda}_1$ its estimated value using the sample covariance[2]. In the absence of missing data, it can be shown that

$$\sqrt{n}(\hat{\lambda}_1 - \lambda_1) \xrightarrow[n\to\infty]{\mathcal{D}} N(0, 2\lambda^2).$$

Solutions to 1–3 below should be presented in *no more than two pages.* Provide a minimum amount of `R` output to justify your answers.

1. Estimate the covariance matrix of the 5 variables in the Student Score Data using methods (a)–(e) and comment your results.

2. How would you contruct an aymptotic confidence interval leveraging the asymptotic normality of $\hat{\lambda}_1$? Use this result to construct confidence intervals for λ_1 using the estimated covariances from point 1. Comment your findings.

3. Note that the student score data is just a subset of the mathmarks data[3] with artificially generated missing entries. Use the full data to compute the sample covariance and give a confidence interval for λ_1. Compare this with your findings in the previous two questions.

4. Can you derive the form of the EM-algorithm for and i.i.d. normal sample with missing data?

   Remember that with partially observed vectors $X_i = (X_{io}^T, X_{im}^T)^T$ the EM-algorithm simplifies to the iterations:

$$\mu^{(k+1)} : \sum_{i=1}^{n} (\hat{X}_i - \mu) = 0$$

$$\Sigma^{(k+1)} : \sum_{i=1}^{n} \left(\Sigma - (\hat{X}_i - \mu)(\hat{X}_i - \mu)^T - \mathbf{C}_i^{(k)} \right) = 0$$

[1] Table 1 in Efron (1994), *Journal of the American Statistical Association.*

[2] These quantities are interesting for multivariate analysis and unsupervised learning. For example, in a principal component analysis, the largest empirical eigenvalue is the variance explained by the first principal component

[3] Available in the `R` package ”SMPracticals”.

with $\hat{X}_{io} = X_{io}$ and

$$\hat{X}_{im} = \mu_{im}^{(k)} + \Sigma_{imo}^{(k)}(\Sigma_{ioo}^{(k)})^{-1}(X_{io} - \mu_{io}^{(k)})$$

and $C_{ijk}^{(k)} = 0$ if X_{ij} or X_{ik} are observed and

$$\mathbf{C}_{imm}^{(k)} = \Sigma_{imm}^{(k)} - \Sigma_{imo}^{(k)}(\Sigma_{ioo}^{(k)})^{-1}\Sigma_{iom}^{(k)}$$

Exercise 3

The data set **CentralPark.csv** consists of precipitation data from weather station at Central Park, New York. The data was collected from the National Oceanic and Atmospheric Administration (NOAA). The variable PRCP shows the observed amount of rain at time t in mm. Consider a first order Markov Chain model with a two dimensional state space corresponding to the states $\{0, 1\} = \{$"snow day", "no snow"$\}$, where we define a snow day as one with a snow depth of at least 50mm. Consider the following estimated transition probability matrices

$$\mathbf{T}_{IL} = \begin{bmatrix} 0.966 & 0.034 \\ 0.045 & 0.955 \end{bmatrix}, \quad \mathbf{T}_{WI} = \begin{bmatrix} 0.861 & 0.139 \\ 0.011 & 0.989 \end{bmatrix}, \quad \mathbf{T}_{NY} = \begin{bmatrix} 0.964 & 0.036 \\ 0.224 & 0.776 \end{bmatrix}.$$

obtained using snow data collected at the Weather stations of Charleston (IL), Medford (WI) and New York (NY), for the period between December 17 and December 31. The sample sizes are 1373, 1067 and 1346 respectively.
Use these estimates to answer the first 4 questions below.

1. Interpret the estimated values of the transition probabilities.

2. If a significant snow depth is observe on Christmas Eve, what is the probability of observing a White Christmas in Central Park?

3. Give an estimate of the long-term probability of observing a snow day in Medford.

4. Are the probability laws of "$X_{t+1}|X_t = 1$" and "$1-X_{t+1}|X_t = 0$" significantly different in Charleston?

5. Can you reproduce the above results for the Central Park data[4]?

6. Does a higher order chain improve the fit of the data?

[4] Note that the SNWD has only been recorded since 1912 and records are unavailable from 1999–2002

Exercise 4

Consider the following model for a DNA sequencing experiment. Assume that the probability of obtaining one of the four bases A,C,G,T are $p_A = 1-\theta$, $p_C = \theta-\theta^2$, $p_G = \theta^2 - \theta^3$ and $p_T = \theta^3$, where $0 \leq \theta \leq 1$. Further assume that we have n independent realizations where n_A, n_C, n_G, n_T are the observed occurrences for each base and $n_A + n_C + n_G + n_T = n$.

1. Give the joint distribution of (N_A, N_C, N_G, N_T)

2. Show that the MLE of θ is

$$\hat{\theta} = \frac{N_C + 2N_G + 3N_T}{N_A + 2N_C + 3N_G + 3N_T}$$

3. Find the asymptotic distribution of $\hat{\theta}$

4. Find constants a_A, a_C, a_G, a_T such that $T = a_A N_A + a_C N_C + a_G N_G + a_T N_T$ is unbiased for θ.

5. Find the variance of T and five the asymptotic relative efficiency between T and $\hat{\theta}$

6. Compare the two estimators discussed above with the MLE that does not assume that p_A, p_C p_G and p_T depend on a common unknown parameter θ.

7. Using the last point, propose a test statistics for the null hypothesis H_0 : $\mathbf{p} = \mathbf{p}(\theta)$, where $\mathbf{p} = (p_A, p_C, p_G, p_T)^T$ and $\mathbf{p}(\theta) = (1-\theta, \theta-\theta^2, \theta^2-\theta^3, \theta^3)^T$ for some fixed value $\theta \in (0, 1)$.

Exercise 5 (Optional bonus question)

milk.txt reports monthly milk production (pounds per cow) from Jan 62 to Dec 75.

1. Fit a linear model to the data and comment the estimated coefficients. Do you see some structure in the residuals?

2. Plot the correlogram and partial correlogram of the residuals obtained in the previous point and comment them.

3. Try fitting an AR(1) to the data. Does it provides a better model? what about an AR(2) model?

4. Try different ARMA models and present two models that best fit the data in your analysis.