Homework 3

STA 521: Predictive Modeling

Due on Thursday September 17 at 11:59 pm

Your homework must be submitted in R Markdown format. Your responses must be supported by both textual explanations and the code you generate to produce your result. (Just examining your various objects in the "Environment" section of RStudio is insufficient–you must use scripted commands.) Use the naming convention from Lab 1 or points will be deducted. Finally, please show all steps in your write up for full credit. (You must upload your .Rmd and .pdf file to Sakai).

Hints: In problem 3, you may find the following commands useful: `colMeans, var, round, dvmnorm, contour`.

1. The total variance of $\boldsymbol{X}$ is the MSE of the mean:

$$E||\boldsymbol{X} - E(\boldsymbol{X})||^2 = \text{tr}(\text{Var}(\boldsymbol{X})).$$

   Prove this.

2. Suppose $\boldsymbol{Y} \sim N_n(0, I)$ and let A be symmetric. Show that

$$\boldsymbol{Y}^T A \boldsymbol{Y} \overset{dist}{=} \sum_i \lambda_i W_i,$$

   where $W_i \overset{iid}{\sim} \chi^2_1$ and $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A.

3. A very simple model of poverty rates would be as follows: for state i in year t, the poverty rate Y_{it} is

$$Y_{it} = \mu_t + U_i + \epsilon_{it}$$

   where μ_t is a nation-wide poverty rate that varies over time, U_i is a state-specific poverty rate reflecting enduring characteristics of state i, and ϵ_{it} is some combination of fluctuations and measurement error. Note that U_i and ϵ_{it} are independent. (We will see later how to decompose U_i and ϵ_{it} to reflect state-level covariates.)[1]

   For simplicity, assume that $U_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_u)$ and $\epsilon_{it} \overset{ind}{\sim} \mathcal{N}(0, \sigma^2_{e_t})$.

[1] You can download the data from the course webpage. It might also to be helpful to know that the fips codes for the states may be useful as well as this command: colnames(saipe98)[1] $<-$ "fips"

(a) Show that this model implies that there is a positive covariance for within-state poverty rates over time, $\mathrm{Cov}\left[Y_{i1}, Y_{i2}\right] = \mathrm{Var}\left[U_i\right] = \sigma_u^2$.

(b) Let that $\mathbf{Y}_i = (Y_{i1}, Y_{i2})^T$. Show that $\mathbf{Y}_i$ has an $\mathcal{MVN}$ distribution, and express its μ and Σ in terms of μ_1, μ_2, σ_u^2, $\sigma_{e_1}^2$ and $\sigma_{e_2}^2$.

(c) Load the Small Area Income and Poverty Estimate (SAIPE) data for 1998 and 2005, and extract the CPS poverty rates for children aged 5–17; let these be Y_{i1} and Y_{i2}, respectively. Construct a data frame which just contains these values.

(d) Find the mean poverty rates, across states, for 1998 and 2005.

(e) Find the 2×2 covariance matrix for $\mathbf{Y}$.

(f) Report estimates of μ_1, μ_2, σ_u^2, $\sigma_{e_1}^2$ and $\sigma_{e_2}^2$ from the sample mean vector and the sample covariance matrix.

(g) Make a contour plot of the MVN distribution with the mean and covariance you estimated from the data. (The x axis should be 1998 child poverty rates and the y axis should be 2005 rates.) Add points representing the data. Comment on whether the multivariate normal looks like a good fit to this data.