Stats 601 Homework 1

Due by 11:59pm on Canvas on Jan 19, 2022

Please scan your solution as a pdf file and submit it to Canvas. Please arrange the pages in order and put your name and uniqname on the top of the first page.

1. Consider a p-dim Gaussian random variable $X \sim N_p(\mu, \Sigma)$. Partition X, μ, Σ, and the corresponding precision matrix Σ^{-1} as

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{pmatrix}, \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} & \Lambda_{13} \\ \Lambda_{21} & \Lambda_{22} & \Lambda_{23} \\ \Lambda_{31} & \Lambda_{32} & \Lambda_{33} \end{pmatrix},$$

   (a) Show that X_1 and X_2 are independent if and only if $\Sigma_{12} = \Sigma_{21} = 0$.

   (b) Show that X_1 and X_2 are conditionally independent given X_3 if and only if $\Lambda_{12} = \Lambda_{21} = 0$.

2. Multivariate linear regression is widely used to model the relationships between multiple related responses and a set of predictors. Suppose we have N observations of m-dimensional responses $Y_i = (y_{i,1}, \ldots, y_{i,m})^\top$ and p-dimensional predictors $X_i = (x_{i,1}, \ldots, x_{i,p})^\top$, for $i = 1, \ldots, N$. Let $\mathbf{Y} = (Y_1, \ldots, Y_N)^\top$ be the $N \times m$ response matrix and $\mathbf{X} = (X_1, \ldots, X_N)^\top$ be the $N \times p$ covariates matrix. The multivariate linear regression model assumes

$$\mathbf{Y} = \mathbf{X}B + E,$$

   where B is a $p \times m$ matrix of unknown regression parameters and $E = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N)^\intercal$ is an $N \times m$ matrix of regression errors, with $\boldsymbol{\epsilon}_i$'s independently sampled from an m-dimensional Gaussian distribution $N(\mathbf{0}, \Sigma)$.

   (a) For the multivariate linear regression model, derive the Maximum Likelihood Estimator (MLE) of B and Σ.

   (b) Similarly to the unidimensional linear regression model, the Ordinary Least Squares (OLS) estimator of B is defined as the minimizer of the Residual Sum of Squares (RSS) defined as

$$RSS(B) = \sum_{i=1}^{N} (Y_i^\top - X_i^\top B)(Y_i^\top - X_i^\top B)^\top.$$

   Derive the OLS estimator of B. Is it the same as the MLE?

3. Turn in your figures and codes for the following.

   (a) Load the height/weight data from `heightWeightData.txt` on Canvas website. The first column is the gender label (1 for male and 2 for female), the second coumn is height, and the third is weight. Extract the height/weight data corresponding to the females. Fit a 2-dim Gaussian to the female data, using the empirical mean and covariance. Plot your Gaussian distribution as a 95% confidence ellipse, superimposing on your scatter plot of data points.

(b) *Standardizing* the data means ensuring the empirical variance along each dimension is 1. This can be done by computing $\frac{x_{ij}-\overline{x}_j}{\sigma_j}$, where σ_j is the empirical std of dimension j, $\overline{x}_j$ the empirical mean. Standardize the data and replot.

(c) *Whitening* or *sphereing* the data means ensuring its empirical covariance matrix is proportional to identity matrix, so the data is uncorrelated and of equal variance along each dimension. This can be done by computing $\mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{x}$ for each data vector $\mathbf{x}$, where $\mathbf{U}$ are the eigenvectors and $\mathbf{\Lambda}$ the eigenvalues of the covariance matrix $\mathbf{X}^T\mathbf{X}$. Whiten the data and replot. Note that whitening rotates the data, so people (data points) move to counter-intuitive locations in the new coordinate systems.

(d) Assume the males' data are from bivariate Gaussian $N(\mu_1, \Sigma)$ and females' data are from bivariate Gaussian $N(\mu_2, \Sigma)$. Test hypothesis $\mu_1 = \mu_2$.

4. Suppose that we observe a group of p-dimensional independent and identically distributed samples $\{X_i\}_{i=1}^N$, where $X_i \sim N_p(\mu, \Sigma)$ with μ and Σ unknown. We are interested in testing $H_0 : \mu = \mathbf{0}$. Suppose we use Hotelling's T^2-test

$$T^2 = N\bar{X}_N^\top S_N^{-1}\bar{X}_N$$

where $\bar{X}_N$ and S_N are the sample mean and sample covariance matrix.

(a) Show that the likelihood ratio test takes the form

$$LRT = \left\{\frac{1}{1+T^2/(N-1)}\right\}^{N/2}$$

and therefore is equivalent to Hotelling's T^2-test.

(b) When p is fixed, show that the test statistic T^2 has a limiting χ^2-distribution under the null hypothesis as $N \to \infty$.

(c) Under the setting with $N = 100$ and $p = 3$, perform a simulation study to check if the limiting χ^2-distribution in (b) controls the type I error well (Note that the simulation study would need at least a few hundreds replications to estimate the type I error well).

(d) Under the setting of (c) but increasing p to 10, 40, and 80, does the limiting χ^2-distribution still control the type I error well?

(e) When p is assumed to increase with N such that $p/N \to \gamma \in (0,1)$ as $N \to \infty$, show that the test statistic T^2 has a limiting normal distribution (in the sense that there exist a_N and b_N such that $a_N(T^2 - b_N) \to N(0,1)$ as $N, p \to \infty$).

(f) If $p > N$, can we still use Hotelling's T^2-test? Explain why. If not, please propose an alternative testing procedure and conduct a simple simulation study to verify (This is an open-ended question and theoretical results are not needed for your new procedure).