Multivariate Statistical Analysis: Examination: Theoretical Exercises

Deadline: Monday 2024-02-05: 13:30

Please send me the solutions in p.d.f. or html format to `noble@mimuw.edu.pl`. Scanned handwritten solutions are fine, provided they are in.pdf format, but my email cannot handle .jpg or .jpeg attachments, because they are too large. Please convert to .pdf before sending them along.

1. Assume that the ML estimators in the question are asymptotically normal.

   (a) Suppose that $Y_1, \ldots, Y_n$ are independent observations, with $\mathbb{E}[Y_i] = \mu_i$, $g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j$, where $x_{1.}, \ldots, x_{n.}$ are given covariate vectors, g is a known 'link' function, $\beta = (\beta_1, \ldots, \beta_p)^t$ is an unknown parameter vector. Suppose, furthermore, that the log-likelihood function for the data $l(\beta) := \log L(\beta; Y_1, \ldots, Y_n)$ may be written as:

   $$l(\beta) = \frac{\sum_{i=1}^{p} \beta_i T_i(Y) - \psi(\beta)}{\phi} + \text{constant}$$

   where ψ is a function that depends only on β, $T_1, \ldots, T_p$ are functions of the data and ϕ is a known positive constant.

      i. What are the sufficient statistics for β?

      ii. Show that $\mathbb{E}[T_i(Y)] = \frac{\partial \psi}{\partial \beta_i}(\beta)$.

   (b) Using the same notation as part (b), find an expression for the covariance matrix of $(T_1(Y), \ldots, T_p(Y))^t$ and hence show that $l(\beta)$ is a concave function. Why is this fact useful for evalating $\widehat{\beta}_{ML}$ (maximum likelihod estimator of β)?

   (c) Illustrate your solution to this problem by the example $Y_i \sim \text{Be}(\mu_i)$ Bernoulli trials with success probability μ_i) where $0 < \mu_i < 1$ for each $i = 1, \ldots, n$ and

   $$\log \frac{\mu_i}{1 - \mu_i} = \beta x_i.$$

   $x_1, \ldots, x_n$ are known covariates. Your solution should include a statement of the large sample distribution of $\widehat{\beta}_{ML}$ and conditions on $(x_i)_{i \geq 1}$ under which the hypotheses of asymptotic normality are satisfied.

2. **Variance Inflation** Consider the regression problem

   $$Y = X\beta + \epsilon$$

   where

   $$Y_j = \sum_{i=1}^{p} X_{ji}\beta_i + \epsilon_j$$

and assume that the $n \times p$ matrix X has been centred and scaled so that X^tX is a correlation matrix (there is no β_0 parameter in this model).

Now regress the kth column of X (i.e. $X_{.k}$) on the other $p-1$ columns of X using OLS regression. Let RSS_k denote the residual sum of squares.

Near collinearity exhibits itself when at least one of the $RSS_1, \ldots, RSS_p$ is small.

(a) Show that

$$RSS_k = \frac{1}{(X^tX)^{-1}_{kk}}$$

that is, the square root of the kth diagonal element of $(X^tX)^{-1}$.

(b) The *variance inflation factor* for factor k is defined as:

$$VIF_k = \frac{1}{1 - R_k^2}$$

where R_k^2 is the *coefficient of determination* for the column $X_{.k}$ when regressed against the other columns of X. Express VIF_k in terms of RSS_k.

3. **Support Vector Machines** Let $z \in \mathbb{R}$ and define the $(2m+1)$-dimensional Φ mapping

$$\Phi(z) = \left(\frac{1}{\sqrt{2}}, \cos(z), \ldots, \cos(mz), \sin(z), \ldots, \sin(mz) \right)'.$$

Using this mapping, show that the kernel $K(x,y) = \langle \Phi(x), \Phi(y) \rangle$, $x, y \in \mathbb{R}$ reduces to the *Dirichlet kernel*, given by:

$$K(x,y) = \frac{\sin((m+\frac{1}{2})\delta)}{2\sin(\frac{\delta}{2})} \qquad \delta = x - y$$