SOLUTIONS - MIDTERM EXAM - STA437H1S/2005H1S

1. (a) $(X_1, X_2, X_3)^T$ is multivariate normal with mean $(1, 2, 1)^T$ and covariance matrix

$$\begin{pmatrix} 55 & 7 & -5 \\ 7 & 59 & -13 \\ -5 & -13 & 19 \end{pmatrix}$$

(b) $(X_4, X_5)^T$ is multivariate normal with mean $(0, 0)^T$ and covariance matrix

$$\begin{pmatrix} 55 & -6 \\ -6 & 60 \end{pmatrix}$$

Thus $X_4 + X_5$ is normal with mean 0 and variance

$$(1 \quad 1) \begin{pmatrix} 55 & -6 \\ -6 & 60 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 103.$$

(c) There are no links (edges) between variables 1 and 2, and between variables 2 and 4.

2. (a), (b), and (c) are true but (d) is false. Two possible corrections to (d) are:

- If $\boldsymbol{X} \sim \mathcal{N}_p(\mathbf{0}, C)$ then $\boldsymbol{X}^T C^{-1} \boldsymbol{X}$ has a χ^2 distribution with p degrees of freedom.
- If $\boldsymbol{X} \sim \mathcal{N}_p(\mathbf{0}, I)$ then $\boldsymbol{X}^T \boldsymbol{X}$ has a χ^2 distribution with p degrees of freedom.

3. (a) x_2 and x_4 are most highly correlated (with correlation 0.78).

(b) We know that the sum of squares of the standard deviations is 4 (the number of variables) so $1.6336942^2 + \mathtt{A}^2 + 0.59706982^2 + 0.41051087^2 = 4$. This gives $\mathtt{A} = 0.90$. Alternatively, we could compute $\mathtt{A}$ by $\mathtt{A}^2 = 4 \times 0.2015079$, which again gives $\mathtt{A} = 0.90$.
To compute $\mathtt{B}$, we can either use the fact that the loadings are orthogonal or that the loadings are eigenvectors of $\widehat{R}$. The former approach (using the loadings for the first and third PCs) gives

$$(-0.360) \times 0.239 + \mathtt{B} \times (-0.726) + (-0.529) \times 0.642 = 0,$$

which gives

$$\mathtt{B} = \frac{0.360 \times 0.239 + 0.529 \times 0.642}{-0.726} = -0.586.$$

(c) If $\boldsymbol{\ell}_1, \cdots, \boldsymbol{\ell}_4$ are the loadings then the PC scores are $\boldsymbol{\ell}_j^T \boldsymbol{y}_i$ for $j = 1, \cdots, 4$ and $i = 1, \cdots, 100$.

(d) By definition, the PC scores are uncorrelated and so the correlation matrix is simply the identity matrix.

4. (a) Two basic approaches:

- Assess the normality of many one-dimensional projections: informally, using quantile-quantile plots or more formally, using tests of normality such as the Shapiro-Wilk test.
- Compare quantiles of a χ^2 distribution with p degrees of freedom to order values of $(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T S^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})$ $(i = 1, \cdots, n)$. If the data are multivariate normal the points should fall close to a straight line.

(b) The biplot is a plot of the first versus second PC scores with vectors indicating how the individual variables are correlated with the first two PCs.