Lecture 1
Keith Knight
OH: Tuesday 10am – noon @ SS 5016 G
email: keith@utstat.utoronto.ca
evaluation : HW 40%
Midterm 25%
Final 35%

Computing: R

Multivariate observations
- measurements of $p > 1$ variables on n subjects
  - represent observations using vectors

subject

$$\underset{\sim}{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \qquad i = 1, \cdots, n$$

Examples
① Test scores : p subject tests
n students
② Imaging (eg. facial recognition)
$\left.\begin{matrix} n \text{ images} \\ p \text{ pixels} \end{matrix}\right\} \; p \gg n$

<u>Machine Learning</u> $\left\{\begin{matrix} \text{supervised learning} \\ \text{unsupervised learning} \end{matrix}\right.$

<u>Supervised learning</u>: Multivariate observations $\underset{\sim}{x}_i$

each "input" $\underset{\sim}{x}_i$ has an "output" (label) y_i

<u>Examples</u> ① Regression: y_i is a response associated with predictor $\underset{\sim}{x}_i$
② Classification: y_i is a class (e.g. spam or good) associated with $\underset{\sim}{x}_i$

<u>Unsupervised learning</u>: $\{\underset{\sim}{x}_i\}$ are unlabelled
- find structure in $\{\underset{\sim}{x}_i\}$ (which may be interesting or insightful)
- dimension-reduction

<u>Simple model for unsupervised learning</u>

<u>Model</u>: $\underset{\sim}{x}_1, \cdots, \underset{\sim}{x}_n$ are sampled from a population with multivariate dist'n F. ↗ don't know what F is

$$F(\underset{\sim}{x}) = F(x_1, \cdots, x_p) = P(X_1 \le x_1, \cdots, X_p \le x_p) = P(\underset{\sim}{X} \le \underset{\sim}{x})$$

What about F?
- F can be totally unspecified
- F can be determined by a finite number of unknown parameters

e.g. ① F is multivariate Normal
② F is a mixture dist'n.

$F(\underset{\sim}{x}) = \lambda_1 F_1(\underset{\sim}{x}; \underset{\sim}{\theta}) + \cdots + \lambda_k F_k(\underset{\sim}{x}; \underset{\sim}{\theta})$

unknown parameters

where $\lambda_1, \cdots, \lambda_k$ & $\underset{\sim}{\theta}$ are unknown

$\lambda_1 + \cdots + \lambda_k = 1$

Means, variances & covariances

$\underset{\sim}{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$ random vector $\quad \underset{\sim}{X} \overset{\text{dist'd as}}{\sim} F$

mean (expected values) : $\mu_j = E(X_j) = E_F(X_j)$

mean vector $\underset{\sim}{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$

variances : $\sigma_j^2 = Var(X_j) = E[(X_j - \mu_j)^2]$

covariances : $\sigma_{ij} = \sigma_{ji} = Cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$ ← measure of linear dependence

Variance-covariance (covariance) matrix

$$C = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

Properties:

① C is a symmetric matrix. ($C = C^T$)

② C is a non-negative definite matrix

$\underset{\sim}{a}^T C \underset{\sim}{a} \geq 0$ for all vectors $\underset{\sim}{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix}$

Proof: Look at $Var(\underbrace{\underset{\sim}{a}^T \underset{\sim}{X}}_{\sum_{j=1}^{p} a_j x_j})$

$$0 \leq Var(\underset{\sim}{a}^T \underset{\sim}{X}) = \sum_{i=1}^{p} \sum_{j=1}^{p} a_i a_j Cov(x_i, x_j) = \underset{\sim}{a}^T C \underset{\sim}{a}$$

③ Unless there are some $P(\underset{\sim}{a}^T \underset{\sim}{X} = \text{constant}) = 1$, $\underset{\sim}{a} \neq \underset{\sim}{0}$, C is also positive definite

i.e. $\underset{\sim}{a}^T C \underset{\sim}{a} > 0$ if $\underset{\sim}{a} \neq \underset{\sim}{0}$

Note: If C is positive definite then C^{-1} exists.

C^{-1} is called the concentration matrix and often contains useful information about dependence structure of $\underset{\sim}{X}$.

Estimation of $\underset{\sim}{\mu}$, C:

Method of moments: $\hat{\underset{\sim}{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{x}_i = \bar{\underset{\sim}{x}}$ ← component-wise average

$$\hat{C} = S = \overset{\nearrow \text{ in txtbook}}{\frac{1}{n-1}} \sum_{i=1}^{n} \underbrace{\underbrace{(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})}_{p\times 1 \text{ matrix}} \underbrace{(x_i - \bar{x})^T}_{1\times p \text{ matrix}}}_{p\times p \text{ matrix}}$$

can use $\frac{1}{n}$ instead

- under simple model, $\hat{\mu}$, S are unbiased-estimated

But ... the matrix properties of S (↓ data) can be very different from those of C (↓ population).

Example: If $p \geq n$ then S cannot be positive definite.

For Friday: Why?