Outline

Generalized Exponential Family

• Y is said to have a distribution from within the *generalized exponential family* if its p.d.f/p.m.f., $f(y)$ say, can be expressed as

$$f(y;\theta,\phi) = \exp\left\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\right\}$$

- θ: *natural* or *canonical*; ϕ: *scale* or *dispersion*
- Some examples: Normal, Binomial, Poisson, Exponential distribution, Gamma distribution

• Suppose Y is a r.v. from the generalized exponential family. Then

$$\mathrm{E}[Y] = b'(\theta)$$
$$\mathrm{Var}(Y) = a(\phi)b''(\theta).$$

Generalized Linear Models

• Three Conditions:

(I) The $\{Y_i\}$ are independent r.v. sharing the same form of distribution from the Gen. Exp. family.

(II) The *linear predictor* for the i-th observation is

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \quad i = 1, \ldots, n.$$

(III)

$$g(\mu_i) = \eta_i,$$

where $\mu_i = E[Y_i]$. The function $g(\cdot)$ is known as the *link function* and is required to be invertible.

• The *Canonical link* is defined to be that function which maps the mean μ to the canonical parameter, θ. That is, the function g is

$$g(\mu) = \theta = \eta.$$

- Examples of link function.
  - $Bin(n, \pi)$

    Logit: $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$;

    Probit: $g(\pi) = \Phi^{-1}(\pi)$;

    Complementary log-log: $g(\pi) = \log[-\log(1-\pi)]$.
  - $Possion(\lambda)$

    Log-linear: $g(\lambda) = \log(\lambda)$.
  - $N(\mu, \sigma^2)$

    Identity: $g(\mu) = \mu$.

Parameter Estimation ($\widehat{\beta}$)

• Estimate the unknown parameters $\beta_1, \beta_2, \ldots, \beta_p$ by *maximum likelihood*.

• In general score functions are non-linear in the $\{\beta_i\}$, and require numerical iterative techniques for their solution.

- Newton-Raphson Procedure.
- The Method of Scoring.

Goodness-of-fit

- Test H_0: Current model *versus* H_1: Saturated model
- Generalized Likelihood Ratio Test.
  - $$\Lambda(\boldsymbol{y}) = \frac{\sup_{\beta \in \mathbb{R}^p} L(\boldsymbol{\beta}; \phi, \boldsymbol{y})}{\sup_{\beta \in \mathbb{R}^n} L(\boldsymbol{\beta}; \phi, \boldsymbol{y})} = \frac{\hat{L}_c}{\hat{L}_s}$$
  - under H_0,
    $$-2 \log \Lambda(\boldsymbol{Y}) = -2(l_c - l_s) \sim \chi^2_{n-p} \quad \text{asymptotically}$$
  - If ϕ is unknown, we cannot carry out a goodness-of-fit test.

Model comparison

It is required to test

$$H_0 : \beta_{p+1} = \beta_{p+2} = \ldots = \beta_q = 0 \quad \text{vs} \quad H_1 : H_0 \text{ false}.$$

When ϕ is **known**, we reject H_0 at the $100\alpha\%$ level of significance if

$$W = \frac{D_r - D_f}{\phi} > \chi^2_{q-p}(\alpha).$$

When ϕ is **unknown**, we reject H_0 at the $100\alpha\%$ level of significance if

$$W_1 = \frac{(D_r - D_f)/(q-p)}{D_f/(n-q)} > F_{q-p,n-q}(\alpha).$$

Note that, in general, the assumed distributional results under H_0 are approximate.

• Two-way Contingency Table: consider two factors, A and B, where the former occurs at J levels, and the latter at K levels. Let Y_{jk} be the frequency for the (j,k)-th cell of the table.

	B_1	B_2	$\dots$	B_K	**Total**
A_1	Y_{11}	Y_{12}	$\dots$	Y_{1K}	$Y_{1.}$
A_2	Y_{21}	Y_{22}	$\dots$	Y_{2K}	$Y_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_J	Y_{J1}	Y_{J2}	$\dots$	Y_{JK}	$Y_{J.}$
Total	$Y_{.1}$	$Y_{.2}$	$\dots$	$Y_{.K}$	$Y_{..} = n$

- Case (a): Nothing fixed. Poisson distribution.
- Case (b): Total sample size fixed. Multinomial distribution: Poisson distribution conditional on n.
- Case (c): One margin fixed. Product multinomial distribution.

• The main hypothesis of interest with two factors (random variables) is whether they are *independent*, that is, we want to test

$$H_0 : \theta_{jk} = \theta_{j.}\theta_{.k} \quad versus \quad H : \theta_{jk} \geq 0$$

for all $j = 1, \cdots, J$ and $k = 1, \cdots, K$.

• Note that $E(Y_{jk}) = n\theta_{jk}$, so under H_0 the logarithms of the expected cell counts are given by

$$\log E(Y_{jk}) = \log n + \log \theta_{j.} + \log \theta_{.k}$$

which depends only on quantities indexed by j and k but not on the combination jk.

• An alternative-log-linear models

To test the independence of A and B is equivalent to test

$$HP_0 : \log E(Y_{jk}) = \mu + \alpha_j + \beta_k$$

for all j, k, against

$$HP : \log E(Y_{jk}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}.$$

The data are classified according to 3 factors, A, B, and C, each having J,K, and L levels, respectively. So the table has $J \times K \times L$ cells.

- Let Y_{jkl} be the count for cell (j,k,l) and let y_{jkl} be its realization. Y_{jkl}'s have a multinomial distribution when total sample size is fixed.
- Let θ_{jkl} be the probability that an observation falls in cell (j,k,l), that is,

$$\theta_{jkl} = P(A = j, B = k, C = l).$$

1. Complete Independence

$$H_0 : \theta_{jkl} = \theta_{j..}\theta_{.k.}\theta_{..l},$$

The corresponding poisson model is

$$\log \lambda_{jkl} = \mu + \alpha_j + \beta_k + \gamma_l$$

2. Joint Independence
One variable is independent of the other two, i.e,

$$H_1 : \theta_{jkl} = \theta_{..l}\theta_{jk.}$$

The corresponding Poisson model is

$$\log E(Y_{jkl}) = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk}.$$

3. Conditional Independence

   Suppose that given A, B and C are independent. We say that B and C are *conditionally independent* of A (There are partial associations between A and B, and A and C), i.e,

$$H_2 : \theta_{jkl} = \theta_{j..}\widetilde{\theta}_{jk.}\widetilde{\theta}_{j.l}$$

   The corresponding log-linear model is

$$\log E(Y_{jkl}) = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl}$$

4. Partial Association

   This model is also called the *homogeneous association model*. Here, each pair of factors is unaffected by the level of the third. The testing hypothesis is

$$H_3 : \theta_{jkl} = \widetilde{\theta}_{jk.}\widetilde{\theta}_{j.l}\widetilde{\theta}_{.kl}$$

   The corresponding log-linear model is

$$\log E(Y_{jkl}) = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl}$$

- Equivalence between log-linear models and logistic regression
    - Residual deviances will be the same for the corresponding models, and that the corresponding parameter estimates will be the same as well.
    - The main effects of explanatory variables in the logistic regression are identified with interactions between the response variable and explanatory variables in the loglinear model.

- Diagnostics and Model Checking
  - Raw Residuals
$$r_i = y_i - \widehat{\mu}_i$$
  - Pearson Residuals
$$r_i^P = \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}}$$
  - Deviance Residuals
$$r_i^D = \mathrm{sgn}(y_i - \widehat{\mu}_i)\sqrt{d_i}$$
  - Standardized Residuals
    -
$$r_i^{PS} = \frac{r_i^P}{\sqrt{1 - h_i}}$$
    -
$$r_i^{DS} = \frac{r_i^D}{\sqrt{1 - h_i}}$$

Multivariate Normal Distribution

• Suppose $\boldsymbol{X}$ is a $(p \times 1)$ random vector with population mean $\boldsymbol{\mu}$ and population covariance matrix $\boldsymbol{\Sigma}$. It is defined to have the *multivariate normal distribution* if and only if every linear compound $\boldsymbol{a}^T\boldsymbol{X}$ has a univariate normal distribution. That is if and only if

$$Y = \boldsymbol{a}^T\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{a}^T\boldsymbol{\mu}, \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}), \forall \boldsymbol{a}.$$

ML estimates of μ and Σ

$$\hat{\mu} = \bar{\boldsymbol{x}}; \quad \hat{\Sigma} = \boldsymbol{S}.$$

Here $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$ which is a biased estimator.

$\boldsymbol{S}_u = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$ is an unbiased estimator.

Property

If $\overline{\boldsymbol{X}}$ is based on a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from MVN $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\overline{\boldsymbol{X}} \sim$ MVN$(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$.

Principal Components

- Suppose that $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$ is the general observation vector in a random sample of size n with sample covariance matrix $\boldsymbol{S}$. Then the principal components, denoted by $y_1, \ldots, y_p$, satisfy the following conditions:

(I) $y_j = a_{1j}x_1 + a_{2j}x_2 + \ldots + a_{pj}x_p = \boldsymbol{a}_j'\boldsymbol{x}$, for $j = 1, \ldots, p$, where $\boldsymbol{a}_j = (a_{1j}, a_{2j}, \ldots, a_{pj})'$ is a vector of constants satisfying

$$\|\boldsymbol{a}_j\|^2 = \boldsymbol{a}_j'\boldsymbol{a}_j = \sum_{k=1}^{p} a_{kj}^2 = 1;$$

(II) All the principle components have the order

$$Var(y_1) > Var(y_2) > \cdots > Var(y_p);$$

(III) All the principle components are uncorrelated, that is,

$$Cov(y_i, y_j) = \boldsymbol{a}_i'\boldsymbol{S}\boldsymbol{a}_j = 0 \quad for \quad i \neq j.$$

Find all the principle components

- Result: Assume that $\boldsymbol{S}$ has distinct eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_p \geq 0$ (with corresponding eigenvectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p$), then all the possible principle components are

$$y_1 = \boldsymbol{a}_1'\boldsymbol{x}, \quad y_2 = \boldsymbol{a}_2'\boldsymbol{x}, \quad \cdots, \quad y_p = \boldsymbol{a}_p'\boldsymbol{x}$$

and

$$Var(y_j) = \boldsymbol{a_j'Sa_j} = \lambda_j, \quad j = 1, \cdots, p.$$

Identity for the sum of variances

- Let $s_{x_j}^2 = Var(x_j)$ and $s_{y_j}^2 = Var(y_j)$ for $j = 1, \cdots, p$, then

$$\text{trace}(\boldsymbol{S}) = \sum_{j=1}^{p} s_{x_j}^2 = \sum_{j=1}^{p} s_{y_j}^2 = \sum_{j=1}^{p} \lambda_j = \text{trace}(\boldsymbol{S}_y).$$

Principal Components Uses

- Interpretation

If all the covariances are positive and variances are comparable, we would expect that all the coefficients in the first principle component are positive(or negative) and some of the coefficients in the second principle component are positive and some negative. Therefore, the first component can be interpreted as an overall and the second as a contrast.

- Reduction in dimensionality
    - If an eigenvalue is zero, then there is an exact linear relationship between the variables. One variable is therefore redundant.
    - Methods for choosing components to be retained

Choice of $\boldsymbol{R}$ versus $\boldsymbol{S}$

- The use of $\boldsymbol{R}$ may be desirable particularly

(1) when the components of $\boldsymbol{x}$ have variances of quite different orders of magnitude.

(2) when the components of $\boldsymbol{x}$ have different scales.

Hotelling T^2 distribution

- If $\boldsymbol{Y} \sim N_p(\boldsymbol{0}, \boldsymbol{I})$ and *independently* $\boldsymbol{W} \sim W_p(f, \boldsymbol{I})$, then

$$T^2 = f \boldsymbol{Y}^T \boldsymbol{W}^{-1} \boldsymbol{Y}$$

is said to have a HOTELLING T^2-distribution. We write $T^2 \sim T_p^2(f)$.

- If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are iid $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$(n-1)(\overline{\boldsymbol{X}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \mathrm{T}_p^2(n-1)$$

- If $m > (p-1)$, then

$$\mathrm{T}_p^2(m) = \frac{mp}{m-p+1} \mathrm{F}(p, m-p+1)$$

One Sample Test

- Hypotheses of test: $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$
- Test statistic: Under H_0

$$T^2 = (n-1)(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0) \sim T_p^2(n-1)$$

with the rejection (critical) region

$$\left\{ T^2 : T^2 > \frac{(n-1)p}{n-p} \mathrm{F}_\alpha(p, n-p) \right\}$$

Two Sample Test

- Hypotheses of test: $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.
- Test statistic: Under H_0

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)^T \boldsymbol{S}_U^{-1} (\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2) \sim \mathrm{T}_p^2(n_1 + n_2 - 2)$$

where $\boldsymbol{S}_U = \frac{(n_1-1)\boldsymbol{S}_{1U} + (n_2-1)\boldsymbol{S}_{2U}}{n_1+n_2-2}$ is the *unbiased* estimate of $\boldsymbol{\Sigma}$. The rejection region is

$$\left\{ T^2 : T^2 > \frac{(n_1 + n_2 - 2)\, p}{n_1 + n_2 - p - 1} \mathrm{F}_\alpha(p, n_1 + n_2 - p - 1) \right\}.$$

- Assuming multivariate normality, so that the $\boldsymbol{x}_{kj}$ are independently drawn from $\mathrm{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, we can test

$$H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_K \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_k \neq \boldsymbol{\mu}_\ell \text{ for some } k \neq \ell.$$

This is the one-way multivariate analysis of variance (MANOVA).

MANOVA Table

Source	SSP Matrix	df	?
Between Groups	$\boldsymbol{B}$	$K-1$	
Within Groups	$\boldsymbol{W}$	$n-K$	
Total	$\boldsymbol{T}$	$n-1$	

- Union intersection test

Let $y = \boldsymbol{a}^T\boldsymbol{x}$, where $\boldsymbol{x}$ is the general $(p \times 1)$ observation vector. Hence $y_k = \boldsymbol{a}^T\boldsymbol{x}_k \sim N(\boldsymbol{a}^T\boldsymbol{\mu}_k, \boldsymbol{a}^T\Sigma\boldsymbol{a})$ for group k.

Let $H_{0\boldsymbol{a}} : \boldsymbol{a}^T \boldsymbol{\mu}_1 = \ldots = \boldsymbol{a}^T \boldsymbol{\mu}_K$. Then

$$H_0 = \bigcap_{\boldsymbol{a}} H_{0\boldsymbol{a}}.$$

To test $H_{0\boldsymbol{a}}$, we use the usual *univariate* one-way ANOVA F statistic, $F_{\boldsymbol{a}}$, that is,

$$\begin{aligned} F_{\boldsymbol{a}} &= \frac{\text{between-groups MS for } \boldsymbol{a}^T \boldsymbol{x}}{\text{within-groups MS for } \boldsymbol{a}^T \boldsymbol{x}} \\ &= \frac{\text{between-groups sample variance of } \boldsymbol{a}^T \boldsymbol{x}}{\text{within-groups sample variance of } \boldsymbol{a}^T \boldsymbol{x}} \\ &= \frac{\boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a}/(K-1)}{\boldsymbol{a}^T \boldsymbol{W} \boldsymbol{a}/(n-K)} \end{aligned}$$

Let F be the statistic for testing H_0, then the acceptance region of F has

$$\begin{aligned} \{F \leqslant c^2\} &= \bigcap_{\boldsymbol{a}} \{F_{\boldsymbol{a}} \leqslant c^2\} \\ &= \{\max_{\boldsymbol{a}} F_{\boldsymbol{a}} \leqslant c^2\} \end{aligned}$$

Hence, the statistic used to test H_0 is

$$\max_{\boldsymbol{a}} F_{\boldsymbol{a}} = \max_{\boldsymbol{a}} \frac{\boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a}/(K-1)}{\boldsymbol{a}^T \boldsymbol{W} \boldsymbol{a}/(n-K)}$$

- Finding the UIT statistic

Maximizing the ratio

$$R(\ell) = \frac{\ell^T \boldsymbol{B} \ell/(K-1)}{\ell^T \boldsymbol{W} \ell/(n-K)}$$

is equivalent to maximizing

$$\max_{\boldsymbol{u}} \frac{n-K}{K-1} \boldsymbol{u}^T \boldsymbol{W}^{-\frac{1}{2}} \boldsymbol{B} \boldsymbol{W}^{-\frac{1}{2}} \boldsymbol{u}$$

$$\text{subject to} \quad \boldsymbol{u}^T \boldsymbol{u} = 1,$$

where $\boldsymbol{u} = \frac{1}{\sqrt{n-K}} \boldsymbol{W}^{\frac{1}{2}} \ell$ or $\ell = \sqrt{n-K} \boldsymbol{W}^{-\frac{1}{2}} \boldsymbol{u}$.

Suppose the symmetric p.s.d. matrix $\boldsymbol{W}^{-\frac{1}{2}} \boldsymbol{B} \boldsymbol{W}^{-\frac{1}{2}}$ has p non-negative eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ and corresponding eigenvectors $\gamma_1, \ldots, \gamma_p$. Then the eigenvalues of $\frac{n-K}{K-1} \boldsymbol{W}^{-\frac{1}{2}} \boldsymbol{B} \boldsymbol{W}^{-\frac{1}{2}}$ are $\frac{n-K}{K-1}\lambda_1 \geq \frac{n-K}{K-1}\lambda_2 \geq \ldots \geq \frac{n-K}{K-1}\lambda_p \geq 0$ with the same eigenvectors $\gamma_1, \ldots \gamma_p$. Therefore the maximized ratio of between to within-groups sample variance is given by

$$\frac{n-K}{K-1}\lambda_1,$$

- Canonical Variates

Consider

$$\ell_r = \sqrt{n-K}\boldsymbol{W}^{-\frac{1}{2}}\gamma_r, r = 1, \ldots, p.$$

These are the eigenvectors of $\boldsymbol{W}^{-1}\boldsymbol{B}$ corresponding to the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$. Then the r'th canonical variate is

$$y_r = \ell_r^T \boldsymbol{x}.$$

- The r'th canonical variate has the maximum between-groups variance $\frac{n-K}{K-1}\lambda_r$ such that the within-groups variance is 1 and such that the within-groups correlations with $y_1, \ldots, y_{r-1}$ are zero.
- $(\lambda_1 + \ldots + \lambda_r)/(\lambda_1 + \ldots + \lambda_p)$ represents the proportion of between-group variation accounted for by the first r canonical variates.

Bayes Rule

• The classification rule that minimizes total expected misclassification cost is the rule that sets $c(\boldsymbol{x}) = i$ where i minimizes

$$\sum_{j=1}^{2} C(i|j)\, p(j|\boldsymbol{x}) \quad \text{with respect to } i = 1, 2.$$

That is, the rule sets $c(\boldsymbol{x}) = 1$ if $C(1|2)p(2|\boldsymbol{x}) \leq C(2|1)p(1|\boldsymbol{x})$, and $c(\boldsymbol{x}) = 2$ otherwise.

• This is known as *Bayes rule*, and the corresponding total expected misclassification cost is known as the *Bayes risk*.

• Further, allocate the individual with observation vector $\boldsymbol{x}$ to population 1 if

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{\pi_2 C(1|2)}{\pi_1 C(2|1)}.$$

Otherwise allocate to population 2.

Fisher's Linear Discriminant Function

• Suppose that observation vectors from the i-th population have the $N_p(\boldsymbol{\mu}_i, \Sigma)$ distribution, $i = 1, 2$. (So the two populations share the same covariance matrix). Then the allocation rule becomes:

$$\begin{cases} \boldsymbol{L}^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{L}^T(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq k & \text{allocate to population 1} \\ \text{otherwise} & \text{allocate to population 2.} \end{cases}$$

where $\boldsymbol{L}^T\boldsymbol{x} = \left(\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right)^T \boldsymbol{x}$ (*Fisher's Linear Discriminant function*) and $k = \ln\left(\frac{\pi_2 C(1|2)}{\pi_1 C(2|1)}\right)$.

• Invoking the estimates of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{S}$ leads to an allocation rule of the form:

$$\begin{cases} \widehat{\boldsymbol{L}}^T\boldsymbol{x} - \frac{1}{2}\widehat{\boldsymbol{L}}^T(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2) \geq k & \text{allocate to population 1} \\ \text{otherwise} & \text{allocate to population 2} \end{cases}$$

where $\widehat{\boldsymbol{L}} = \boldsymbol{S}_U^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)$.

Misclassification Probabilities

• Let $p(i|j)$ be the probability that an individual from population j is misallocated to population i. We can have the following table

		True Class	
		1	2
Predicted Class	1	$p(1\|1)$	$p(1\|2)$
	2	$p(2\|1)$	$p(2\|2)$
Associated prior probabilities		π_1	π_2

Therefore, the total probability of misclassification is

$$\pi_1 p(2|1) + \pi_2 p(1|2).$$

• For multivariate normal distribution, when $k = 0$, the total probability of misclassification can be found as $\Phi(-\frac{1}{2}D)$ where

$$D^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T \boldsymbol{S}_U^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2).$$

Related to Exam

- Past exam papers can be found on http://www.bbk.ac.uk/lib/elib/exam
- Read the questions thoroughly. For example, be careful with words such as "state without proof", "Please show ...",etc.
- Use your exam time wisely. Do not spend too much time on one question.
- Your answers should be clear and sufficient for the questions. Do not write anything irrelevant to the questions.