Summer School in Statistics for
Astronomers & Physicists
June 5–10, 2005

Center for Astrostatistics

Pennsylvania State University

Statistical Inference for Astronomers: Multivariate Analysis

Thriyambakam Krishnan
Systat Software Asia–Pacific Limited
Bangalore, India

Multivariate Statistical Analysis:

• Statistical theory, methods, algorithms, etc. for simultaneous study of more than one variable

• Descriptive statistics and graphical representation

• Inference problems similar to univariate
  ■ based on the multivariate normal

• Study of relationships between variables and finding structure

• Problems of combining variables and dimensionality reduction

Multivariate Normal Distribution

Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$ univariate normal
$\boldsymbol{X}$: p-column vector
$\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ multivariate normal

Reasons for studying Multivariate Normal:

1. p-variate generalization of univariate normal;

2. same reasons as for univariate normal in univariate analysis;

3. multivariate central limit theorem;

4. robustness of some procedures;

5. theory and methods analogous to univariate based on $\mathcal{N}$, like t and Hotelling's T^2, ANOVA and MANOVA;

6. not many other multivariate models;

7. mathematically tractable and elegant;

8. similar parameters–mean vector $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$.

Bivariate Normal

Let

$$\boldsymbol{X}^T = (X_1, X_2);$$

$$\boldsymbol{X} \sim \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma});$$

$$\boldsymbol{\mu}^T = (\mu_1, \mu_2); \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

with $\sigma_{12} = \sigma_{21}$. $\boldsymbol{\Sigma}$ is non-negative definite.
Let

$$\sigma_1^2 = \sigma_{11}; \qquad \sigma_2^2 = \sigma_{22}$$

Correlation coefficient

$$\rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$$

$\mathcal{N}_2$ Density: $f(x_1, x_2) =$ (if $\boldsymbol{\Sigma}$ is p.d.)

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} e^{-\frac{1}{2(1-\rho^2)}[(\frac{x_1-\mu_1}{\sigma_1})^2 - 2\rho(\frac{x_1-\mu_1}{\sigma_1})(\frac{x_2-\mu_2}{\sigma_2}) + (\frac{x_2-\mu_2}{\sigma_2})^2]}$$

$$(x_1, x_2) \in \Re^2$$

If $(X_1, X_2) \sim \mathcal{N}_2$ then
X_1, X_2 independent $\iff \rho = 0$
In general $\rho = 0$ does not imply independence

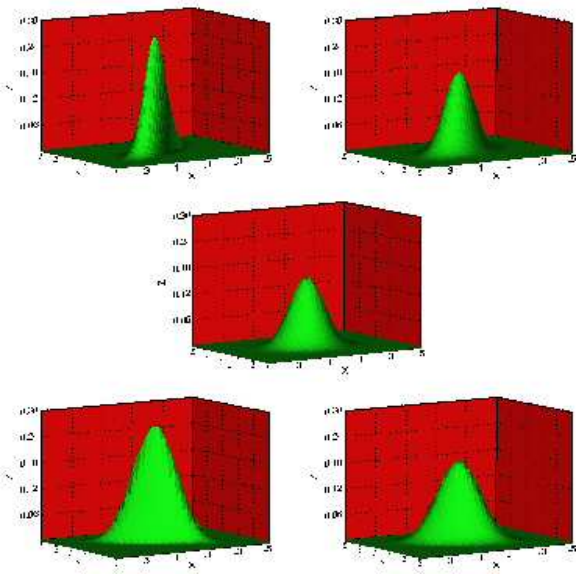
Bivariate Normal Densities

$\mu_1 = \mu_2 = 0; \sigma_{11} = \sigma_{22} = 1$

$\rho = 0.8;$ $\rho = 0.5;$

$\rho = 0;$

$\rho = -0.8;$ $\rho = -0.5;$



In $\mathcal{N}_2$ density, term inside the exponential is

$$-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

and constant is $\frac{1}{\sqrt{2\pi}^p|\boldsymbol{\Sigma}|^{\frac{1}{2}}}$, where $p=2$.
This is the form of $\mathcal{N}_p$ density:

$$\frac{1}{\sqrt{2\pi}^p|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$$

if $\boldsymbol{\Sigma}$ is strictly p.d. (it has anyway to be n.n.d., being a covariance matrix). Indeed, $\boldsymbol{x},\boldsymbol{\mu}$ are p-vectors and $\boldsymbol{\Sigma}$ is a nonsingular (symmetric) $p\times p$ matrix. The term

$$Q=(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

is a positive-definite quadratic form.

- Q is covariance-matrix adjusted distance of $\boldsymbol{x}$ from $\boldsymbol{\mu}$
- Larger this distance, smaller is probability density
- density decreases exponentially with square of distance

You can define $\mathcal{N}_p$ by this density and investigate its properties.

An **alternative** and elegant way is to use the following **definition**:

A random p-vector $\boldsymbol{X}$ is said to be multivariate normally distributed if $\forall$ p-vectors $\boldsymbol{\ell}$, $\boldsymbol{\ell}^T\boldsymbol{X}$ has a univariate normal distribution (or is a constant). This definition makes sense even if $\boldsymbol{\Sigma}$ is singular.

Properties of $\mathcal{N}_p$:

1. $\boldsymbol{\mu}$ is the vector of means of $X_1, X_2, \ldots, X_p$.

2. $\boldsymbol{\Sigma}$ is the (symmetric) matrix of variances and covariances of $X_1, X_2, \ldots, X_p$.

3. If variance-covariance matrix (also called simply covariance matrix or dispersion matrix) is singular, above density does not hold, but the alternative definition still holds. For instance, if $X \sim \mathcal{N}(0, 4)$, then $(X, 3X + 2)$ has covariance matrix $\begin{bmatrix} 4 & 12 \\ 12 & 36 \end{bmatrix}$, singular, but all linear combinations are of the form $a + bX$ for constants a, b and hence are univariate normal. $(X, 3X + 2)$ is bivariate normal by alternative definition.

4. The covariance matrix is singular (multivariate normal or not) with linear dependence of columns given by $\boldsymbol{\Sigma\ell} = \underline{0}$, iff $\boldsymbol{X}^T\boldsymbol{\ell}$ is a constant (degenerate random variable) (deterministic linear dependence of variables). In such cases, by removing deterministically dependent components, $\boldsymbol{\Sigma}$ of remaining components can be made nonsingular.
   Near-singularity of covariance matrix is a computational and conceptual problem. Some exploratory methods detect this problem. Some methods (e.g., ridge regression) overcome this problem.

5. Let us deal only with nonsingular $\boldsymbol{\Sigma}$.

6. Let $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{A}$, $k \times p$ matrix, $\boldsymbol{c} \in \Re^k$. Then

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{c} \sim \mathcal{N}_k(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{c}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$$

[If $k > p$, then $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$ is singular.]

7. $\boldsymbol{\Sigma}$ diagonal means $X_1, X_2, \ldots, X_p$ are independent random variables.

8. $\boldsymbol{X} \sim \mathcal{N}_p(\underline{0}, \boldsymbol{I}_p)$ means $X_1, X_2, \ldots X_p$ are independent standard normal variables.

9. Let $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}, \boldsymbol{I}_p)$$

$$\boldsymbol{Y} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \mathcal{N}_p(\underline{0}, \boldsymbol{I}_p)$$

10. **Marginal Distributions**: All marginal (1-dimensional and $q < p$-dimensional) are (multivariate) normal. That is, if you partition

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

analogously as q and $p - q$ dimensional vectors and matrix (note that $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^T$), then

$$\boldsymbol{X}_1 \sim \mathcal{N}_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}), \qquad \boldsymbol{X}_2 \sim \mathcal{N}_q(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

11. Under the above (multivariate normal) set-up, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent iff $\boldsymbol{\Sigma}_{12} = \underline{0}$, that is all covariances are zero.

Conditional Distributions and Regression

12. $(\boldsymbol{X}_2|\boldsymbol{X}_1 = \boldsymbol{x}_1) \sim \mathcal{N}(\boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{22.1})$, where $\boldsymbol{\mu}_{22.1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1)$, $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$
[Notation: $A|B$ stands for event A conditional on event B; also used as $X|Y = y$ for variable X given variable $Y = y$.]

    (a) if $\boldsymbol{\Sigma}$ is nonsingular, so is $\boldsymbol{\Sigma}_{11}$.

    (b) this conditional expectation is linear in $\boldsymbol{x}_1$.

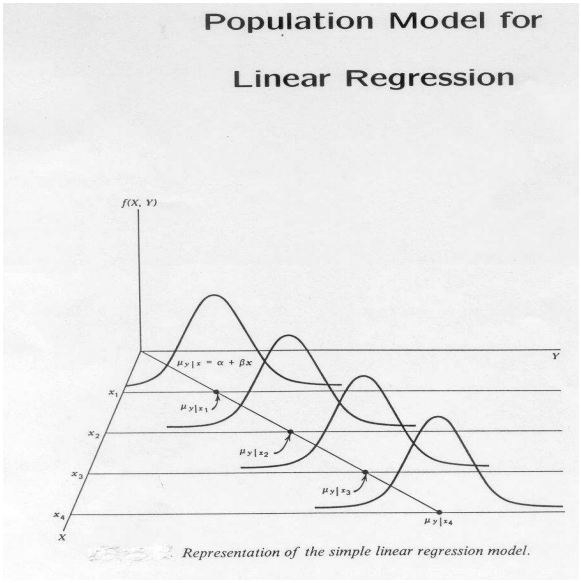
    (c) regression being defined as conditional expectation, this shows that under multivariate normality, (multiple) regression of any subset of variables on the others is linear.

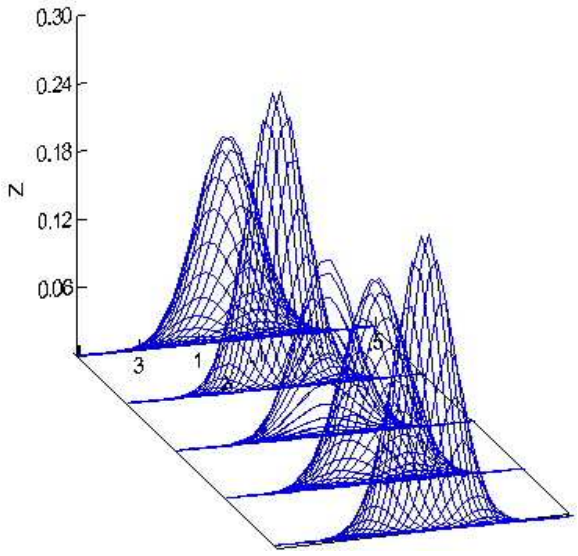
    (d) this linear regression formula is exactly the same as what you obtain by the least-squares criterion.

    (e) $p = 2$, $(X_2|X_1 = x_1) \sim \mathcal{N}(\beta_0 + \beta_1 x_1, \sigma_2^2(1 - \rho^2))$, where $\beta_1 = \frac{\sigma_{21}}{\sigma_{11}}$ and $\beta_0 = \mu_2 - \beta_1\mu_1$, the well-known formulas for (least-squares) simple linear regression.

    (f) conditional covariance matrix does not depend on $\boldsymbol{x}_1$.

    (g) These results justify linearity and homoscedasticity (common variance) assumptions in the multiple linear regression model.





- conditional means on a straight line
- conditional variances same

More Properties

1. [We know: if $X \sim N(0,1)$, then $X^2 \sim \chi^2(1)$]
   $\Delta^2(\boldsymbol{X}) = (\boldsymbol{X}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) = \boldsymbol{Y}^T\boldsymbol{Y} \sim \chi^2(p)$, being sum of squares of p independent N(0,1)'s by (9) above

2. Sample (of size n) mean vector $\bar{\boldsymbol{X}}$ and sample sum of squares and products matrix $\boldsymbol{S}$ are independently distributed.

3. $\bar{\boldsymbol{X}} \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$

4. $\boldsymbol{S} \sim \mathcal{W}_p(n-1, \boldsymbol{\Sigma})$ $\mathcal{W}$ is called the **Wishart** distribution, the multivariate analog of the χ^2 distribution—we shall not discuss it here.

5. For $\mathcal{N}(\mu, \sigma^2)$, Student's t statistic based on a sample of n with mean $\bar{X}$ and sample (mean-corrected) sum of squares S^2 is $t = \frac{\bar{X}-\mu}{S/\sqrt{n(n-1)}}$ extended to Hotelling's
   $$T^2 = (\bar{\boldsymbol{X}}-\boldsymbol{\mu})^T\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu})$$

6. Analysis of Variance (ANOVA) which decomposes observed variation into its components is analogously extended to Multivariate Analysis of Variance (MANOVA)

Estimation of $\mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$ parameters

Random sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ from $\mathcal{N}_p$
Observed values: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$
Data matrix: $n \times p$ matrix $\mathcal{U}$ with row and column names as indicated:

$$\begin{array}{c} \\ \boldsymbol{X}_1 \rightarrow \\ \boldsymbol{X}_2 \rightarrow \\ \ldots \\ \boldsymbol{X}_n \rightarrow \end{array} \begin{array}{c} \begin{array}{cccc} \boldsymbol{Y}_1 \downarrow & \boldsymbol{Y}_2 \downarrow & \ldots & \boldsymbol{Y}_p \downarrow \end{array} \\ \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix} \end{array}$$

$\bar{\boldsymbol{X}} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p)$: Sample mean vector;
$\boldsymbol{S} = ((S_{ij}))$: Sample (mean-corrected) Sum of Squares and Products Matrix

$$S_{ij} = \sum_{\ell=1}^{n} (X_{i\ell} - \bar{X}_i)(X_{j\ell} - \bar{X}_j) = \boldsymbol{Y}_i^T \boldsymbol{Y}_j - n\bar{X}_i \bar{X}_j,$$

$$i, j = 1, 2, \ldots, p$$

$$\boldsymbol{S} = \sum_{\ell=1}^{n} (\boldsymbol{X}_\ell - \bar{\boldsymbol{X}})(\boldsymbol{X}_\ell - \bar{\boldsymbol{X}})^T = \sum_{\ell=1}^{n} \boldsymbol{X}_\ell \boldsymbol{X}_\ell^T - n\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T$$

$$= \mathcal{U}^T \mathcal{U} - n\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T$$

Analogs of Univariate Normal:

- $\bar{\boldsymbol{X}}$: unbiased estimate of $\boldsymbol{\mu}$
- $\frac{1}{n-1}\boldsymbol{S}$: unbiased estimate of $\boldsymbol{\Sigma}$
- $\bar{\boldsymbol{X}}, \boldsymbol{S}$: sufficient statistics for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ [In a sense, these statistics contain all the information in the sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ in respect of $\boldsymbol{\mu}, \boldsymbol{\Sigma}$.]

Maximum Likelihood Estimation of μ, Σ: Rao (1973)

Density:

$$(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp[-\frac{1}{2}tr\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\}]$$

Joint density of observations (but for constant not involving parameters), which is the likelihood function in terms of parameters:

$$L = \boldsymbol{\Sigma}^{-n/2}\exp[-\frac{1}{2}\text{tr}\{\boldsymbol{\Sigma}^{-1}\sum_{\ell=1}^{n}(\boldsymbol{x}_\ell-\boldsymbol{\mu})(\boldsymbol{x}_\ell-\boldsymbol{\mu})^T\}]$$

$$\sum_{\ell=1}^{n}(\boldsymbol{x}_\ell-\boldsymbol{\mu})(\boldsymbol{x}_\ell-\boldsymbol{\mu})^T$$

$$=\sum_{\ell=1}^{n}(\boldsymbol{x}_\ell-\bar{\boldsymbol{x}})(\boldsymbol{x}_\ell-\bar{\boldsymbol{x}})^T+n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T$$

$$=\boldsymbol{S}+n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T$$

$$\text{tr}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\}$$

$$=\text{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\}+n\text{tr}\{\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\}$$

$$=\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}S_{ij}+n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})$$

where $\boldsymbol{\Sigma}^{-1}=((\sigma^{ij}))$.

$$\log L = \frac{n}{2}\log|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}S_{ij} - \frac{n}{2}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}) \quad (A)$$

$$= \frac{n}{2}\log|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}[S_{ij} + n(\bar{\boldsymbol{x}}_i-\boldsymbol{\mu}_i)(\bar{\boldsymbol{x}}_j-\boldsymbol{\mu}_j)] \quad (B)$$

Differentiating (A) w.r.t. $\boldsymbol{\mu}$ leads to

$$\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}) = \mathbf{0} \Rightarrow \bar{\boldsymbol{x}} = \boldsymbol{\mu} \Rightarrow \bar{\boldsymbol{x}} = \hat{\boldsymbol{\mu}} \quad (C)$$

Differentiating (B) w.r.t. σ^{ij} leads to

$$\frac{n}{|\boldsymbol{\Sigma}^{-1}|}\frac{\partial|\boldsymbol{\Sigma}^{-1}|}{\partial\sigma^{ij}} = [S_{ij} + n(\bar{x}_i-\mu_i)(\bar{x}_j-\mu_j)] \quad (D)$$

$$\frac{\partial|\boldsymbol{\Sigma}^{-1}|}{\partial\sigma^{ij}} = \text{cofactor of } \sigma^{ij} \text{ in } \boldsymbol{\Sigma}^{-1}$$

$$\frac{n}{|\boldsymbol{\Sigma}^{-1}|}\frac{\partial|\boldsymbol{\Sigma}^{-1}|}{\partial\sigma^{ij}} = n\sigma_{ij} \quad (E)$$

Equations (C), (D) and (E) lead to

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{S}$$

a slightly biased estimate as in the univariate case.

To show the solution is actually a maximum: $\log L$ at the estimated value is:

$$\frac{n}{2}\log n^p + \frac{n}{2}\log|\boldsymbol{S}^{-1}| - \frac{n}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} S^{ij}S_{ij}$$

$$= \frac{n}{2}\log n^p - \frac{n}{2}\log|\boldsymbol{S}| - \frac{np}{2}$$

The difference between this and $\log L$ at an arbitrary value is shown below to be ≥ 0, showing that $\log L$ is maximum at this solution. This difference is:

$$-\frac{n}{2}\log\frac{|n^{-1}\boldsymbol{S}|}{|\boldsymbol{\Sigma}|} - \frac{np}{2} + \frac{1}{2}\text{tr}\{\boldsymbol{\Sigma}^{-1}[\boldsymbol{S} + n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^T]\}$$

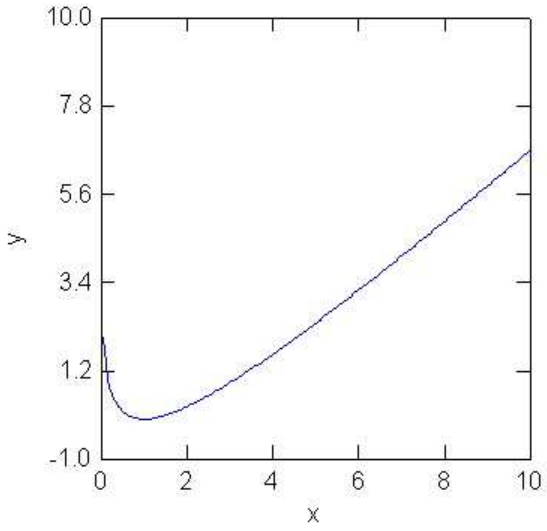
$$\geq -\frac{n}{2}\log\frac{|n^{-1}\boldsymbol{S}|}{|\boldsymbol{\Sigma}|} - \frac{np}{2} + \frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S})$$

$$= \frac{n}{2}[-\log(\lambda_1\lambda_2\ldots\lambda_p) - p + (\lambda_1 + \lambda_2 + \ldots + \lambda_p)] \quad (F)$$

where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are eigenvalues of $\frac{1}{n}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}$. [The λ_i's are positive with probability 1, since they are eigenvalues of positive definite matrix.] Since for any nonnegative $x, x \leq e^{x-1}$, or $-\log x - 1 + x \leq 0$ the quantity inside [ ] in (F) above is $\sum_{i=1}^{p}(-\log\lambda_i - 1 + \lambda_i) \geq 0$, (see figure on next page) implying left-side of (F) is ≥ 0.

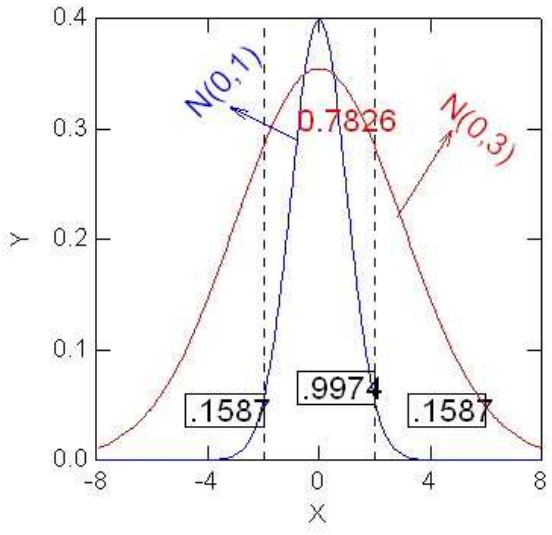
Hence $\bar{\boldsymbol{X}}, \frac{1}{n}\boldsymbol{S}$ are indeed MLEs of $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ respectively.

Graph of $-\log x - 1 + x$ for $x \geq 0$



Statistical or Standardized Distance:

- $X \sim \mathcal{N}(0, 1)$
  - Point 3 is at distance of 3 s.d. from 0;
  - Prob$(0 < X < 3) = 0.4987$
- $Y \sim \mathcal{N}(0, 9)$,
  - Point 3 is at distance of 1 s.d. from 0;
  - Prob$(0 < Y < 3) = 0.3413$
- Statistically 3 is closer in the second case than in first
- since it is closer in s.d. units; and
- equivalently, it is probabilistically closer.
- If variance is larger, points at same Euclidean distance are statistically nearer.



- Correct for varying s.d.
- standardize variables; $\frac{Y}{3} \sim \mathcal{N}(0,1)$
- point 3 in new scale $\frac{3}{3} = 1$
- statistical distance in second case:
  - Prob$(0 < X < 1) = 0.3413$
- Point at Euclidean distance of y from 0 in case of s.d. σ is at statistical distance $\frac{y}{\sigma}$.
- Squared statistical or standardized distance between y_1, y_2 under s.d. σ

$$(y_1 - y_2)(\sigma^2)^{-1}(y_1 - y_2)$$

Statistical Distance between two vectors:

• p-vectors $\boldsymbol{x}_1, \boldsymbol{x}_2$, under a covariance matrix $\boldsymbol{\Sigma}$

• if $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, then uncorrelated
(under multinormality, also independent)

• Euclidean distance is a reasonable distance

• If $\boldsymbol{\Sigma}$ is diagonal with elements σ_i^2, then

■ Euclidean distance based on standardized components is a statistical distance

• Then squared distance is

$$D^2(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_1 - \boldsymbol{x}_2)$$

• If the components are not independent, correlations $\neq 0$

■ how to adjust for correlations?

• Transform $\boldsymbol{X}$ with nonsingular $\boldsymbol{\Sigma}$ to uncorrelated

$$\boldsymbol{Y} = \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{X} \sim (\boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\mu}, \boldsymbol{I}_p)$$

• Using Euclidean distance on $\boldsymbol{Y}$ gives $D^2(\boldsymbol{y}_1, \boldsymbol{y}_2)$

$$= (\boldsymbol{y}_1 - \boldsymbol{y}_2)^T(\boldsymbol{y}_1 - \boldsymbol{y}_2)$$

$$= (\boldsymbol{x}_1 - \boldsymbol{x}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_1 - \boldsymbol{x}_2) = D^2(\boldsymbol{x}_1, \boldsymbol{x}_2)$$

• This is Mahalanobis D^2

• Adjustment depends only on the covariance matrix

• Used whether the distribution is multinormal or not

• In $\mathcal{N}_p$ density, the term (with a negative sign) in the exponential, which is a quadratic form

$$D^2(\boldsymbol{x}, \boldsymbol{\mu}) = (\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}),$$

the Mahalanobis Distance of $\boldsymbol{x}$ from the mean vector (centroid)

• Larger this distance, smaller is probability density;

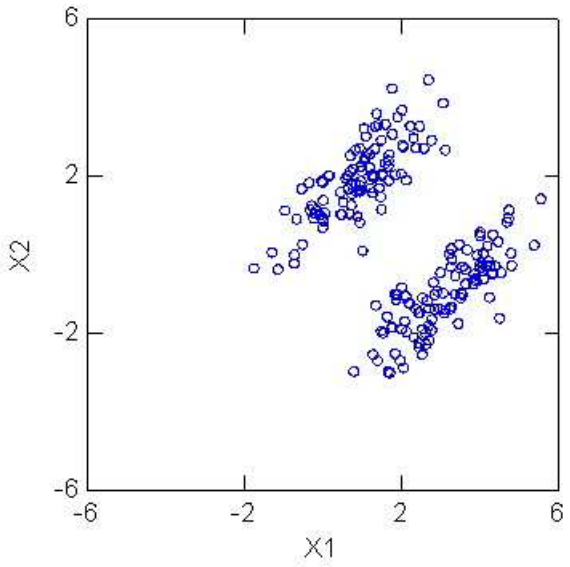
• density decreases exponentially with square of distance

• Mahalanobis distance is used

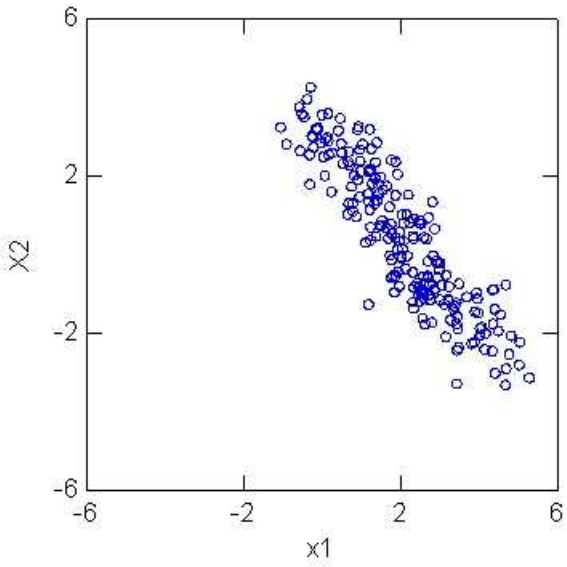
- in classification problems
- in assessing multivariate normality

Scatterplots of bivariate normal, centroids (1, 2), (3,−1)

(a) correlation +0.8



(b) correlation -0.8



Mahalanobis Squared Distance between (1,2) & (3,1) and (1,2) & (3,-1)

ρ	$D^2((1,2),(3,1)) = \frac{5+4\rho}{1-\rho^2}$	$D^2((1,2),(3,-1)) = \frac{13+12\rho}{1-\rho^2}$
-0.8	5.000	9.444
-0.5	4.000	9.333
0.0	5.000	13.000
0.5	9.333	25.333
0.8	22.778	62.778

- p.d. $\boldsymbol{\Sigma}^{-1}$ induces inner product
- Mahalanobis D^2 induced by this inner product
- Mahalanobis angle between $\boldsymbol{x}_1, \boldsymbol{x}_2$ could be defined
- In classification problems Mahalanobis D^2 between two populations
- with common covariance matrix $\boldsymbol{\Sigma}$
- Mahalanobis distance between centroids (mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$)

$$D^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

THE CLASSIFICATION PROBLEM:

Cluster Analysis (Unsupervised Pattern Recognition):

- Inherent or natural grouping of cases in the data set
- Hierarchical (tree) or one-level grouping
- No prior knowledge of grouping or even number of groups

Examples:

1. Taxonomy of flora and fauna (Linnaeus, 1707–78)
   - kingdom, (phyla), classes, orders, families, genera, species
2. Classification of books in a library (Dewey-decimal)
3. Medical science: Classification human diseases
4. Classification of celestial objects
   - stars, galaxies, nebulae

Discriminant Analysis (Supervised Pattern Recognition):

- Grouping already known
- Data: On each case, measurements $\boldsymbol{x}$ and group index z
- Objective:
  - Develop a formula for identifying a new case with
  - observation $\boldsymbol{x}$ into one of k groups
- Criterion: Overall loss (chance of misclassification) minimum

Examples:

1. Is this orange tree mandarin or tangerine?
2. Does this book by Babu–Fiegelson go into the Astronomy or Statistics shelves?
3. Is this case a schizophrenic, neurotic, hysteric or paranoid?
4. Is this celestial object a star, a galaxy or a nebula?

Object: To develop a method of identifying a new case $\boldsymbol{x}$ into one of the k groups, using training data containing prototypes of cases

Some astronomical applications:

- Variable vs nonvariable stars
- Galaxies vs stars
- "Classification of 39 new variable stars with spectral type between A2 and F8 discovered by Hipparcos with the aim of finding new γDoradus stars" (Aerts et al., 1998) using the following variables:
  - Geneva color B2-V1
  - Geneva Y index
  - Geneva Z index
- Training Data: Geneva data on above for
  - known *bona fide* γDor stars (6)
  - δScuti stars (107)
  - variable CP stars (23)
- Develop a formula (discriminant functions)
- Use Geneva data on each of 39 new variable stars
- Apply discriminant functions to put them into one of the 3 classes
- Aerts et al. identified 14 of the 39 as γDor
- Handler (1999) found 70 new γDoradus by discriminant analysis

Basics of Discriminant Analysis

• Measurements used $\boldsymbol{x}$

• First consider the case where we know k and the distributions $F_i(\boldsymbol{x})$ of $\boldsymbol{x}$ in groups $i = 1, 2, \ldots, k$

• Also needed: p_i, proportion of elements in group i

• Discuss later how to estimate F_i, p_i

A simple example:

- $p = 2; X_1, X_2$ both binary
- 2 Groups A and B
- Suppose we know the distribution of $\boldsymbol{X} = (X_1, X_2)$ under each group as follows:

[Recall notation $A|B$ for " given B"]

x_1	x_2	$P(\boldsymbol{x}\|A)$	$P(\boldsymbol{x}\|B)$	Identify as
0	0	0.25	0.4	A
0	1	0.25	0.05	A
1	0	0.45	0.05	A
1	1	0.05	0.5	B
Prior		0.8	0.2	

- Where will you classify a future case with (1,0)? A or B?
- What is a suitable criterion?
- Question: Given $\boldsymbol{X} = \boldsymbol{x}$, what is the probability of A? of B?
- Conditional probabilities

$P(A|\boldsymbol{X} = \boldsymbol{x}), P(B|\boldsymbol{X} = \boldsymbol{x})$ needed

- We are given $P(\boldsymbol{x}|A), P(\boldsymbol{x}|B)$
- Apply Bayes Theorem

Bayes Rule or Bayes Theorem:

$$P(m|\boldsymbol{X} = \boldsymbol{x}) = \frac{P(m \cap \boldsymbol{X} = \boldsymbol{x})}{P(\boldsymbol{X} = \boldsymbol{x})}$$

But given $P(\boldsymbol{X} = \boldsymbol{x}|m)$,

$$P(m \cap \boldsymbol{X} = \boldsymbol{x}) = P(\boldsymbol{X} = \boldsymbol{x}|m) \times P(m)$$

$$P(\boldsymbol{X} = \boldsymbol{x}) = \sum_m P(\boldsymbol{X} = \boldsymbol{x}|m) \times P(m)$$

- Needed: $P(m)$ **Prior Probabilities**
- $P(m|\boldsymbol{X} = \boldsymbol{x})$ called **Posterior Probabilities** or **Inverse Probabilities**
- Bayes Theorem of Probability:

$$P(m \cap \boldsymbol{X} = \boldsymbol{x}) = \frac{P(\boldsymbol{X} = \boldsymbol{x}|m) \times P(m)}{\sum_m P(\boldsymbol{X} = \boldsymbol{x}|m) \times P(m)}$$

- If $\boldsymbol{X}$ is continuous the above formula holds with $P(\boldsymbol{X} = \boldsymbol{x}|.)$ replaced by density $f(x|.)$.

Discrimination with a continuous variable:

$$X|A \sim \mathcal{N}(\mu_A, \sigma^2), X|B \sim \mathcal{N}(\mu_B, \sigma^2)$$

• priors equal.

• Then classify x according to the larger of $f(x|A)$ and $f(x|B)$.

• Comparing $\log f(x|A)$ and $\log f(x|B)$ we see that:

• classify corresponding to shorter of (Euclidean) distance of x from μ_A and μ_B.

• If $(X|A) \sim \mathcal{N}(\mu_A, \sigma_1^2), (X|B) \sim \mathcal{N}(\mu_B, \sigma_2^2)$.

• priors equal

• then the comparison is on the basis of a quadratic in x.

• See figures and next page.

• Two normal groups with means μ_1, μ_2, variances σ_1^2, σ_2^2

• Equal priors

• Log likelihood of x in group i is

$$\log L_i = -\frac{1}{2}[\log \sigma_i + \frac{x^2}{\sigma_i^2} + \frac{\mu_i^2}{\sigma_i^2} - 2\frac{\mu x}{\sigma_i^2}]$$

• If $\sigma_1 = \sigma_2$, x^2 cancels in $\log L_1 - \log L_2$, leading to

• Linear rule: Classify x in group 1 if $x - \mu_1 < x - \mu_2$

• If $\sigma_1^2 \neq \sigma_2^2$, x^2 does not cancel in
$\log L_1 - \log L_2$ which is

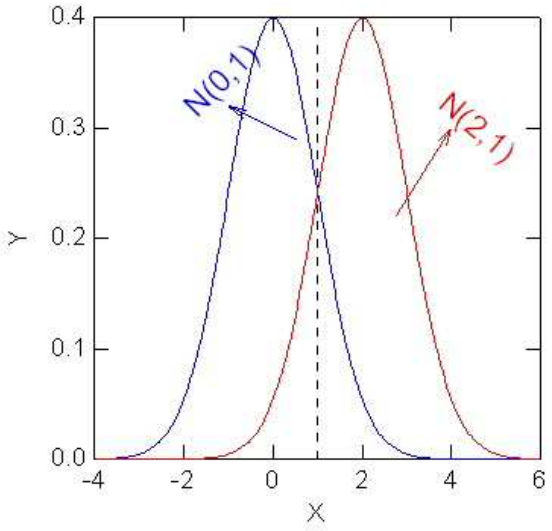
• a quadratic in x with roots a and b, leading to

• a quadratic rule: Classify x in group 2 if $x < a$ or $x > b$,
and in group 2 otherwise.

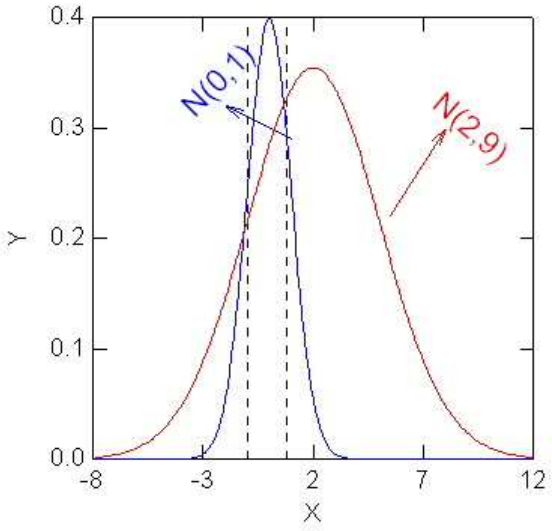
• $\mu_1 = 0, \mu_2 = 2, \sigma_1 = 1, \sigma_2 = 3$ leads to $8x^2 + 4x + 4 - 18 \log 3$
See figures.

Normal distributions with different means

(a) same variance



(b) different variances



Discrimination with several continuous variables:

Let us assume that $(\boldsymbol{X}|A) \sim \mathcal{N}_p(\boldsymbol{\mu}_A, \boldsymbol{\Sigma})$ and $(\boldsymbol{X}|A) \sim \mathcal{N}_p(\boldsymbol{\mu}_A, \boldsymbol{\Sigma})$ in the groups A and B respectively. Priors: $p_A, p_B, p_A + p_B = 1$.

Bayes Rule: Compare $p_A f_A(\boldsymbol{x})$ and $p_B f_B(\boldsymbol{x})$ or $\log p_A + \log f_A(\boldsymbol{x})$ and $\log p_B + \log f_B(\boldsymbol{x})$ To compare take difference between

$$\log p_A - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) \qquad (G1)$$

$$\log p_B - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2) \qquad (G2)$$

leading to

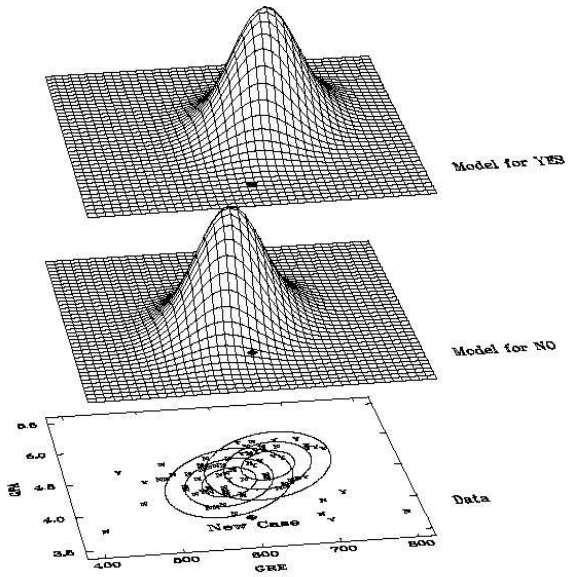
$$d(\boldsymbol{x}) = \log \frac{p_A}{p_B} - \frac{1}{2}[\boldsymbol{\mu}_1 \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2] + \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1}[\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2]$$

which is of the form $\ell_0 + \boldsymbol{\ell}_1 \boldsymbol{x}$ a linear function of $\boldsymbol{x}$; more precisely

$$constant + \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

This is the well-known **LINEAR DISCRIMINANT FUNCTION**(LDF). Fisher derived it as the solution to a linear function of $\boldsymbol{x}$, say $y = \boldsymbol{\ell}^T \boldsymbol{x}$ which maximizes the ratio:

$$\frac{\text{between group sum of squares in } y}{\text{within group sum of squares in } y}$$


Model for YES
Model for NO
Data
New Case
CRE
400
500
600
700
800
3.5
4.0
4.5
5.0
5.5

Unknown Parameters:

- Supervised training sample
- Estimate parameters; plug in
- Prior prob can be estimated if
  - random sample from the mixture
  - else, external estimate
  - software choices:
    equal, data based, user-specified
- Loss unequal, L_{ij}:
  loss if actual i, classified j
  - minimize expected loss
  - still linear for equal covariance

Training Samples: Data on p measurements $\boldsymbol{X}$ on n cases, together with supervisor classification of each case into one of k groups

$$\begin{array}{cc} \text{Group} & \\ & \begin{array}{cccc} \boldsymbol{Y}_1 \downarrow & \boldsymbol{Y}_2 \downarrow & \dots & \boldsymbol{Y}_p \downarrow \end{array} \\ \begin{array}{cc} z_1 & \boldsymbol{X}_1 \rightarrow \\ z_2 & \boldsymbol{X}_2 \rightarrow \\ \dots & \dots \\ z_n & \boldsymbol{X}_n \rightarrow \end{array} & \left[\begin{array}{cccc} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \dots & \dots & \dots & \dots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{array}\right] \end{array}$$

Use estimates (say, MLE) of $\mu_i, \boldsymbol{\Sigma}_i$ or common $\boldsymbol{\Sigma}$;
use estimates of p_i if estimable from data or from external sources

Mahalanobis Distance and LDF:

- second term in (G1) is Mahalanobis $D^2(\boldsymbol{x}, \boldsymbol{\mu})$
- Similarly (G2)
- for equal prior probability $\log p_A, \log p_B$ can be ignored
- comparison on the basis of D^2
- classify an object with $\boldsymbol{x}$ into group whose centroid (mean vector) is closer
- similar to univariate normal
- same principle and method holds for k groups;
  - comparison of posterior prob
  - common covariance: linear functions
- unequal covariance: quadratic discriminant function

Error Rates:

- How good is the classification?
- Conditional error rates; Apparent error rate:
  - apply developed formula to data set
  - compute proportion of misclassification
  - group-wise (conditional);
  - overall (apparent)
  - favorably biased to the method
  - classification table in computer output
- Split data into two halves; develop with one, test with another
- Jackknifed error rate: leave-one-out method
- Bootstrap error rate
- Theoretical error rate for two normal distributions with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$,with same covariance matrix $\boldsymbol{\Sigma}$ with equal priors is $\Phi(-\frac{1}{2}\Delta)$ (Φ: standard normal CDF; Δ: Mahalanobis distance between the two mean vectors).

Similarity to Regression:

- Same problem, but for qualitative dependent
- Robust extensions

IDENTIFICATION OF HYADES WITHIN HIPPARCOS STARS:
Application of Discriminant Analysis

Object: To develop a formula to identify Hyades in Hipparcos
Variables used: RA, DE, PMRA, PMDE
Training Set: HIP star dat, with an extra column of 0 NONHYADES, 1: HYADES
Supervisor: Eric Feigelson (who identified 114 of them as Hyades)

Here is what Professor Feigelson says:

" I have compared this to the published list of Hyades cluster (here I mean the "Hyades star cluster") members from Perryman et al. (1998); see

`http://astrostatistics.psu.edu/datasets/HIP_star.html`

for access to this paper. Using the Vizier online archive, I found the HIP number (first column of HIP star.dat of Perryman's cluster members adding the constraint that $20 < PLX < 25$ which defined our 2719 Hipparcos dataset.

The file gives Perryman's 114 HIP numbers, and an indicator I added including the 92 which were in David Hunter's heuristic sample. The agreement is EXCELLENT!! We missed some because of our RA and DE cuts, and we didn't have any members Perryman's omitted. I'm very happy with this agreement."

Before analysis, we explore the data with a few **multivariate** plots (Wilkinson, 1999).

- displays of multivariate data—all dimensions
- each observation a line or function across display
- useful for detecting clusters, outliers, test significance
- > 2 dimension need special techniques

Andrews's Fourier Plot: For each multivariate observation $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ consider a function $f_{\boldsymbol{x}}(t) =$

$$x_1\sqrt{2} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin 2(t) + x_5 \cos(2t) + \ldots,$$

t ranges from $-\pi(-3.142)$ to $\pi(3.142)$ ($\frac{1}{4}$ radians on either side of 0)

- preserves distances; linear
- each observation consists of one line across display
- a set of waveforms with sine and cosine components for each $\boldsymbol{x}_i$
- a waveform for each case
- similar cases have similar waveforms
- different cases have contrasting waveforms
- Fourier plot shows distinctness of the two groups

Parallel Coordinate Display:

• One scaled horizontal axis for each variable

• Each observation is represented as unbroken series of line segments which intersect horizontal axes

• Points are then connected using line segments

• Result is a "signature" across p dimensions for each observation

• Similar observations share similar signatures

• Clusters and group differences can be discerned

• Associations and correlations among variables can be visualized

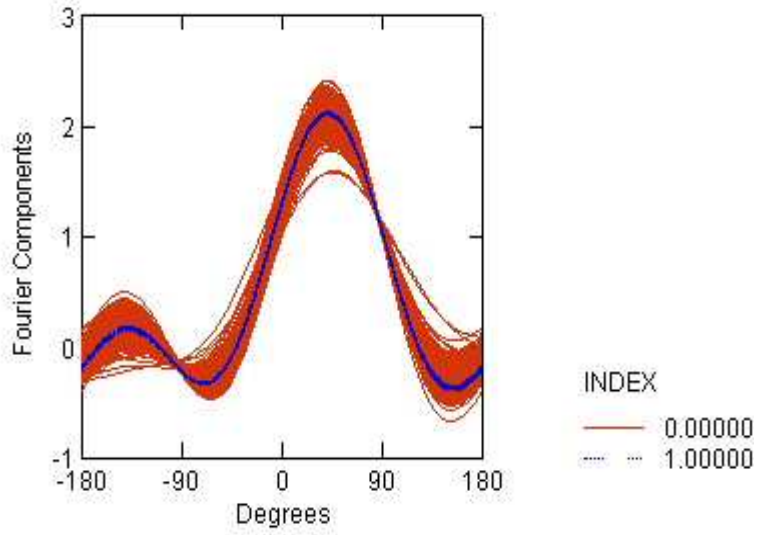
• Two related variables will be connected by parallel line segments

• Number of crossings of line segments is directly related to correlation coefficient

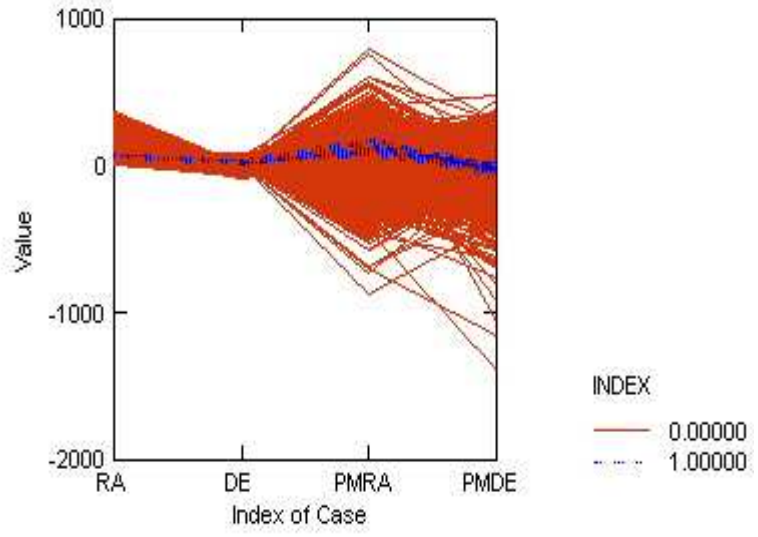
• Display shows distinctness of the two groups

In both the plots, the Hyades group exhibit patterns in the middle of the plots. We examine the implications of this later.

Andrews's Fourier Plot



Parallel Coordinate Display



RESULTS OF LINEAR DISCRIMINANT ANALYSIS

USING RA, DE, PMRA, PMDE

(with some annotation)

• below is a typical output from a software (with our annotations)
• the sample mean vectors for the groups followed by a test of significance of differences between them, using Hotelling's T^2 and its F distribution
• the two groups are highly significantly different
For **prior probabilities** we used the sample proportions, viz., NONHYADES: $\frac{2605}{2719}$ HYADES : $\frac{114}{2719}$

```
Group frequencies
        NONHYADES        HYADES
          2605           114
Group means
                    NONHYADES      HYADES
 RA                178.15341      66.04449
 DE                  2.83470      16.56296
 PMRA                0.73344     111.46579
 PMDE              -65.52675     -27.72886
```

[Considerable difference in the means, confirmed by the following Hypothesis Test]

Between groups F(4,2717): 59.14502 p-value=0.0000

Classification functions

	NONHYADES	HYADES
CONSTANT	-1.59435	-3.92643
RA	0.01630	0.00599
DE	-0.00483	0.03230
PMRA	0.00122	0.00496
PMDE	-0.00324	-0.00091

[Coefficients for the two linear functions $\ell_0 + \ell_1 RA + \ell_2 DE + \ell_3 PMRA + \ell_4 PMDE$]

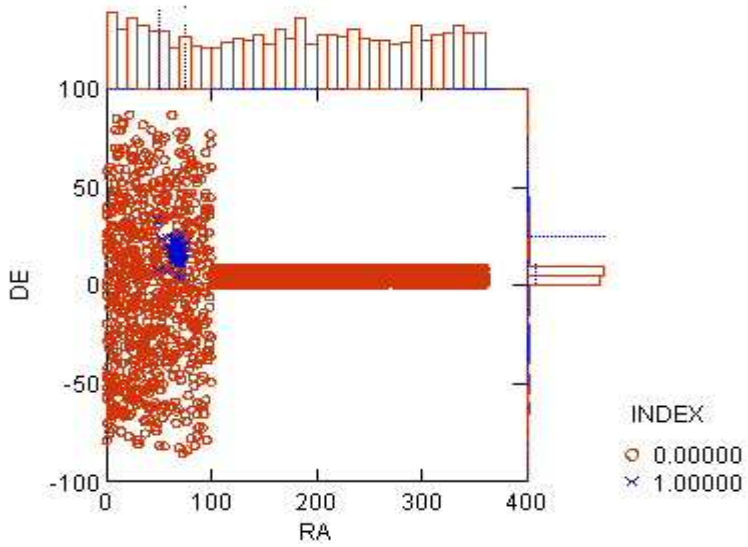
Classification matrix
(cases in row categories classified into columns)

	NONHYADES	HYADES	%correct
NONHYADES	2570	35	99
HYADES	114	0	0
Total	2684	35	95

Jackknifed classification matrix

	NONHYADES	HYADES	%correct
NONHYADES	2570	35	99
HYADES	114	0	0
Total	2684	35	95

RA, DE are the main discriminators.
Scatterplots of RA, DE in different colors for the two groups



The Hyades stars are in the center of the scatterplot.

• All Hyades stars have been (mis)classified as NonHaydes
• Reason: Prior probability for NonHyades is very large
• Plots indicate quadratic nature of discriminant function
• Try Quadratic Discriminant Analysis
• Same prior as before

Results of Quadratic Discriminant Analysis:

Group NONHYADES Quadratic discriminant function coefficients

	RA	DE	PMRA	PMDE
RA	-0.00004			
DE	0.00002	-0.00104		
PMRA	-0.00000	-0.00000	-0.00002	
PMDE	-0.00000	-0.00002	0.00000	-0.00002

[The diagonal are coefficients of $(RA)^2$, etc.
The off-diagonals are coefficients of (RA)(DE), etc.]

Linear	0.01562	-0.00463	0.00117	-0.00310

[These are coefficients of the linear terms of RA, DE, PMRA, PMDE]

-19.34770

[This is the constant term]

Group HYADES Quadratic discriminant function coefficients

	RA	DE	PMRA	PMDE
RA	-0.12213			
DE	-0.06303	-0.22389		
PMRA	-0.02544	-0.01143	-0.00678	
PMDE	-0.01927	-0.06477	-0.00350	-0.02158
Linear	22.82360	14.69899	5.05737	4.27297
Constant	-1107.89539			

Classification matrix (cases in row categories classified into columns)

	NONHYADES	HYADES	%correct
NONHYADES	2603	2	100
HYADES	6	108	95
Total	2609	110	100

Jackknifed classification matrix

	NONHYADES	HYADES	%correct
NONHYADES	2603	2	100
HYADES	6	108	95
Total	2609	110	100

- Much improved classification

HOW TO IDENTIFY A NEW STAR:

You have observations on RA, DE, PMRA, PMDE of your candidate star

Suppose they are:

RA = 170; DE=3.1; PMRA=5.4; PMDE=-2

Compute quadratic classification function for group non-Hyades and group Hyades

Classify into that group for which this quantity is larger

Illustration:

NONHYADES: $-0.00004 \times (RA)^2 + 2 \times 0.00002 \times RA \times DE - 2 \times 0 \times RA \times PMRA - 2 \times 0 \times RA \times PMDE$ $-0.00104 \times (DE)^2 + 2 \times 0 \times DE \times PMRA - 2 \times 0.00002 \times DE \times PMDE - 0.00002 \times (PMRA)^2 + 2 \times 0 \times (PMRA) \times (PMDE) - 0.00002 \times (PMDE)^2$ $+0.01562 \times RA - 0.00463 \times DE + 0.00117 \times PMRA - 0.00310 \times PMDE - 19.3477 = -17.82946$

HYADES: $-0.12213 \times (RA)^2 - 2 \times 0.06303 \times RA \times DE - 2 \times 0.02544 \times RA \times PMRA - 2 \times 0.01927 \times RA \times PMDE$ $-0.22389 \times (DE)^2 - 2 \times 0.01143 \times DE \times PMRA - 2 \times 0.06477 \times DE \times PMDE - 0.00678 \times (PMRA)^2 - 2 \times 0.00350 \times (PMRA) \times (PMDE) - 0.02158 \times (PMDE)^2$ $+22.82360 \times RA + 14.69899 \times DE + 5.05737 \times PMRA + 4.27297 \times PMDE - 1107.89539 = -795.08706$

Since NonHaydes value is larger we classify this star as NonHaydes.

Halo of Hyades Stars:

In the NonHyades stars, there are 35 which are nearer the Hyades centre (by Mahalanobis distance) than the NonHyades centre and hence they were (mis)classified as Hyades. This Mahalanobis distance was computed on the basis of a common covariance matrix.

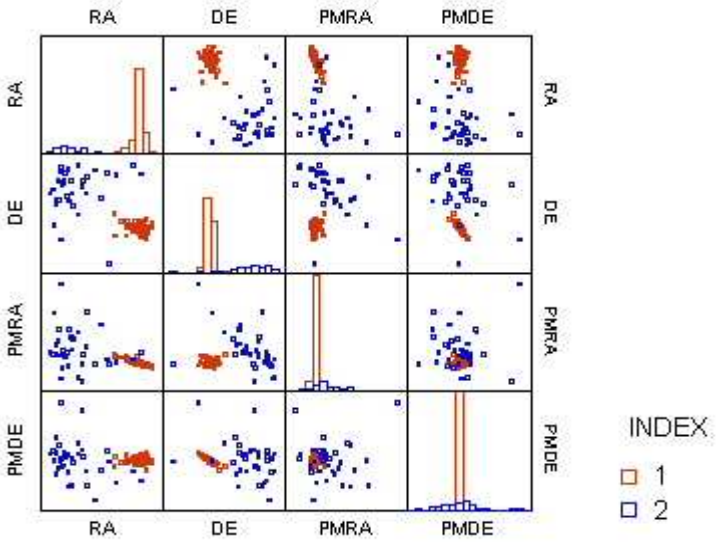
The following is a list of those stars—their hip no in the data file, their Mahalanobis distance from NonHyades centre, probability it belongs to NonHyades, Mahalanobis distance from Hyades centre, probability it belongs to Hyades. Notice that the distance from Hyades is smaller and hence the probability of belonging to Hyades is larger.

These stars may be considered to be the halo of Hyades stars.

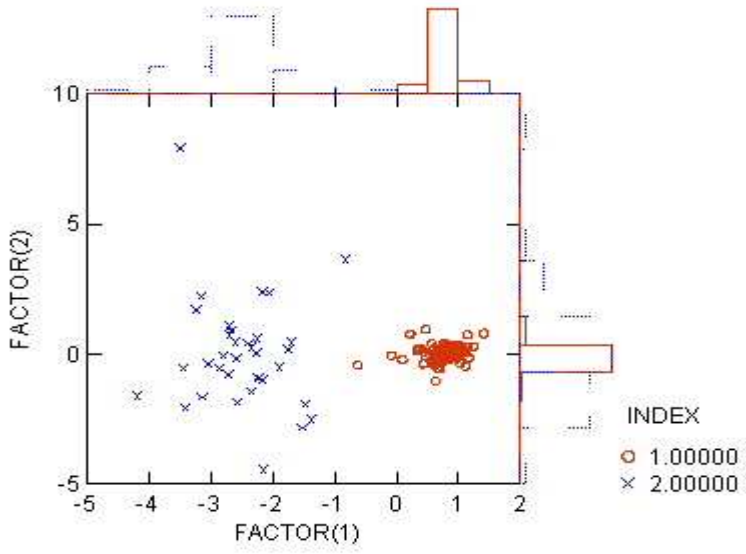
HIP		Distance from NonH	Prob	Distance from Hyades	Prob
641	-->	11.6	0.38	4.4	0.62
1148	-->	12.8	0.35	5.3	0.65
1389	-->	13.9	0.44	7.2	0.56
1871	-->	9.9	0.46	3.4	0.54
1978	-->	9.8	0.49	3.4	0.51
1987	-->	10.9	0.39	3.8	0.61
2453	-->	16.5	0.25	8.0	0.75
2712	-->	14.5	0.27	6.3	0.73
2797	-->	17.8	0.25	9.3	0.75
3006	-->	18.9	0.35	11.5	0.65
3086	-->	30.9	0.21	22.1	0.79
3267	-->	12.3	0.47	5.8	0.53
3509	-->	14.2	0.33	6.4	0.67
3641	-->	14.8	0.31	7.0	0.69
4041	-->	17.5	0.26	9.2	0.74
4393	-->	10.4	0.44	3.7	0.56
4617	-->	17.6	0.34	10.0	0.66
4900	-->	13.4	0.42	6.5	0.58
5273	-->	17.2	0.28	9.0	0.72
5542	-->	11.1	0.42	4.2	0.58
5684	-->	9.8	0.47	3.3	0.53
6890	-->	10.4	0.46	3.8	0.54
6913	-->	19.7	0.34	12.1	0.66
7670	-->	12.3	0.45	5.6	0.55
7744	-->	18.7	0.44	12.0	0.56
7949	-->	14.2	0.34	6.7	0.66
8582	-->	11.1	0.45	4.4	0.55
8825	-->	25.4	0.15	15.6	0.85
10778	-->	17.3	0.47	10.8	0.53
11923	-->	14.1	0.37	6.7	0.63
13456	-->	16.8	0.32	9.0	0.68
13785	-->	13.4	0.43	6.6	0.57
15415	-->	14.3	0.49	7.9	0.51
19454	-->	18.0	0.36	10.6	0.64
21165	-->	15.3	0.43	8.5	0.57

Various plots show the difference between the Hyades stars and their halo. Hyades stars are plotted in blue and the halo in red. Principal components are linear combinations of the four variables which preserve as much of the original total variation as possible.

Hyades and Halo Scatter Plots

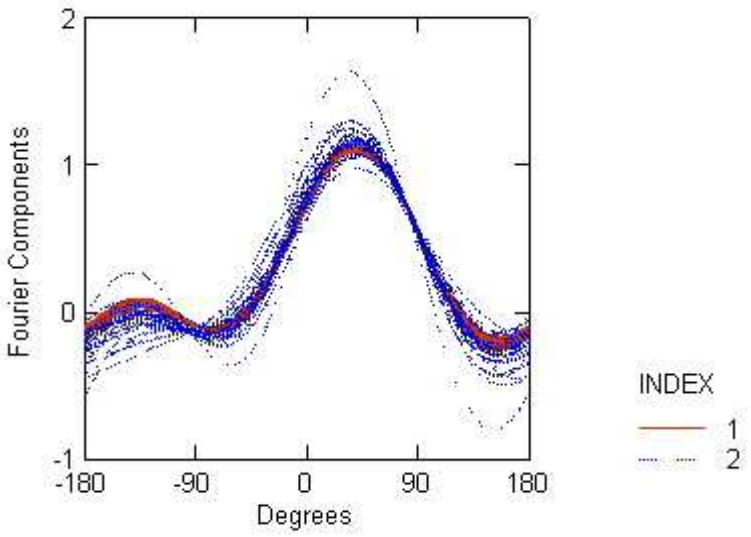


Hyades and Halo Principal Components Plots

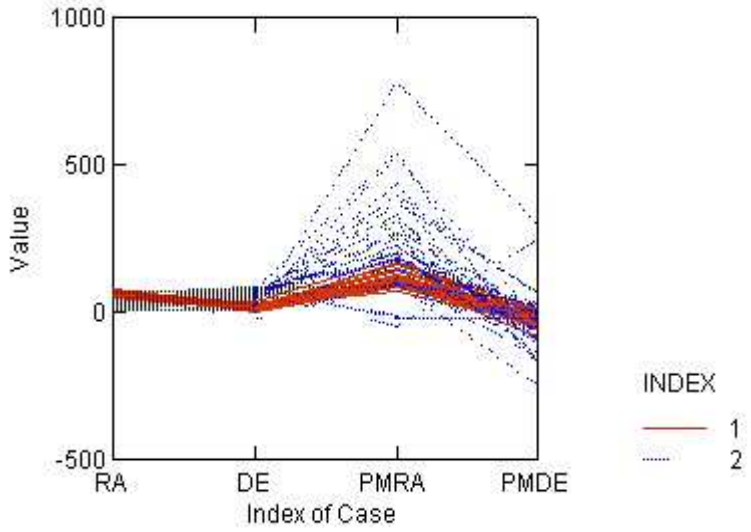


Hyades and Halo Andrews's

Fourier Plot



Hyades and Halo Parallel Coordinate Display



Other Approaches to Discriminant Analysis:

1. Step-wise discriminant analysis:
   Variables are added (or removed) one by one depending on their usefulness.

2. Logistic discriminant (regression) analysis:
   A model linear in the (predictor) variables $\boldsymbol{x}$ is assumed for $\log \frac{p}{1-p}$, (log of odds ratio) where p and $1-p$ are the proportions in the two groups.

3. k-Nearest Neighbor classification:
   Classify a new observation $\boldsymbol{x}$ into that group to which a majority of k nearest neighbors (according to a suitable distance function) belong.

4. Classification (and Regression) tress (CART)
   A highly nonparametric method where a tree is constructed with branches determined by values of the x_i variables and where leaves suggest which group a $\boldsymbol{x}$ value will be classified into.

5. There are many other statistical and non-statistical approaches

   (a) Neural Networks;

   (b) Support Vector Machines (SVM): A hyperplane separating the two classes is computed such that the distance between the closest vector to the hyperplane is maximized.

Concluding Remarks:

1. Most quantitative problems involve more than one measurement; need multivariate analysis (MVA).

2. Variables may be of different types—nominal, ordinal, interval, ratio, discrete, continuous.

3. MVA

   (a) deals with description of multivariate data (graphical representation; various types of association and correlation);

   (b) explores relationships between variables, reduces dimensions, and finds patterns (component analysis, factor analysis, structural equation modelling);

   (c) explores relationships between cases (observational units), finds patterns (cluster analysis);

   (d) solves classical inferential and decision-theoretic problems relating to several variables (estimation, hypothesis testing, discriminant analysis); normality-based theory and asymptotics generalize comfortably;

   (e) deals with problems of simultaneous prediction of several variables (multivariate regression).

4. Useful models and nice methods are available mainly for continuous (based on multivariate normal) and all nominal (log-linear).

5. Classical MVA is based on the multivariate normal model.

6. Bayesian approaches in the multivariate case are not so well developed despite improved computing facilities. Even for the $\mathcal{N}_p$ case, say, for the 2-group discriminant analysis problem, even for a simplistic prior—for $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}, p$—analysis of posterior is very complicated, even with MCMC!

7. Computations are heavy; great deal of matrix computations (with symmetric matrices).

8. Methods for non-continuous variables, for mixed type of variables are not available or messy.

References:

B.Flury (1997): *A First Course in Multivariate Statistics.* New York: Springer-Verlag.

R.A.Johnson & D.W.Wichern (2002): *Applied Multivariate Statistical Analysis.* Fifth Edition. Englewood Cliffs, NJ: Prentice-Hall.

C.R.Rao (1973): *Linear Statistical Inference and its Applications.* New York: John Wiley.

G.A.F.Seber (2004): *Multivariate Observations.* New York: John Wiley.

M.S.Srivastava (2002): *Methods of Multivariate Statistics.* New York: John Wiley.

N.H.Timm (2002): *Applied Multivariate Analysis.* New York: Springer-Verlag.

L.Wilkinson (1999): *The Grammar of Graphics*. New York: Springer-Verlag.

Some Astronomy References:

C.Aerts, L.Eyer & E.Kestens (1998): The discovery of new γDoradus stars from the HIPPARCOS mission. *Astronomy and Astrophysics*, **337**, 790–796.

G.Handler (1999): The domain of γDoradus variables in the Hertzsprung–Russell diagram. *Monthly Notices of the Royal Astronomical Society*, **309**, L19–L23.

F.Murtagh & A.Heck (1987): *Multivariate Data Analysis*. Dordrecht: D.Reidel.

D.Qin, Z.Hu & Y.Zhao (2002): A new automated classification technique of galaxy spectra. In J.-L.Starck & F.D.Murtagh (Eds.): *Astronomical Data Analysis*. Bellingham, WA: SPIE. pp. 362–370.

Y.Zhang, C.cui & Y.Zhao (2002): Classification of AGNs from stars and normal galaxies by support vector machines. In J.-L.Starck & F.D.Murtagh (Eds.): *Astronomical Data Analysis*. Bellingham, WA: SPIE. pp. 371–378.