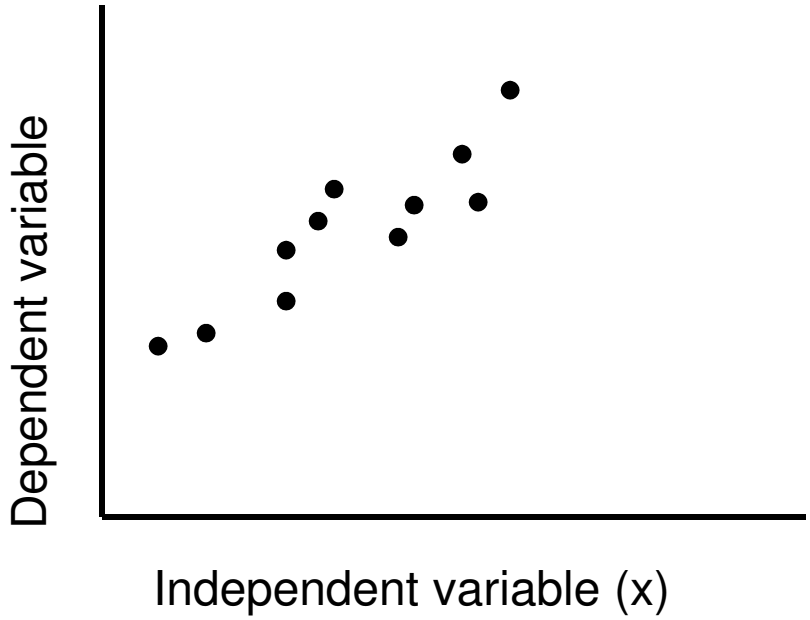
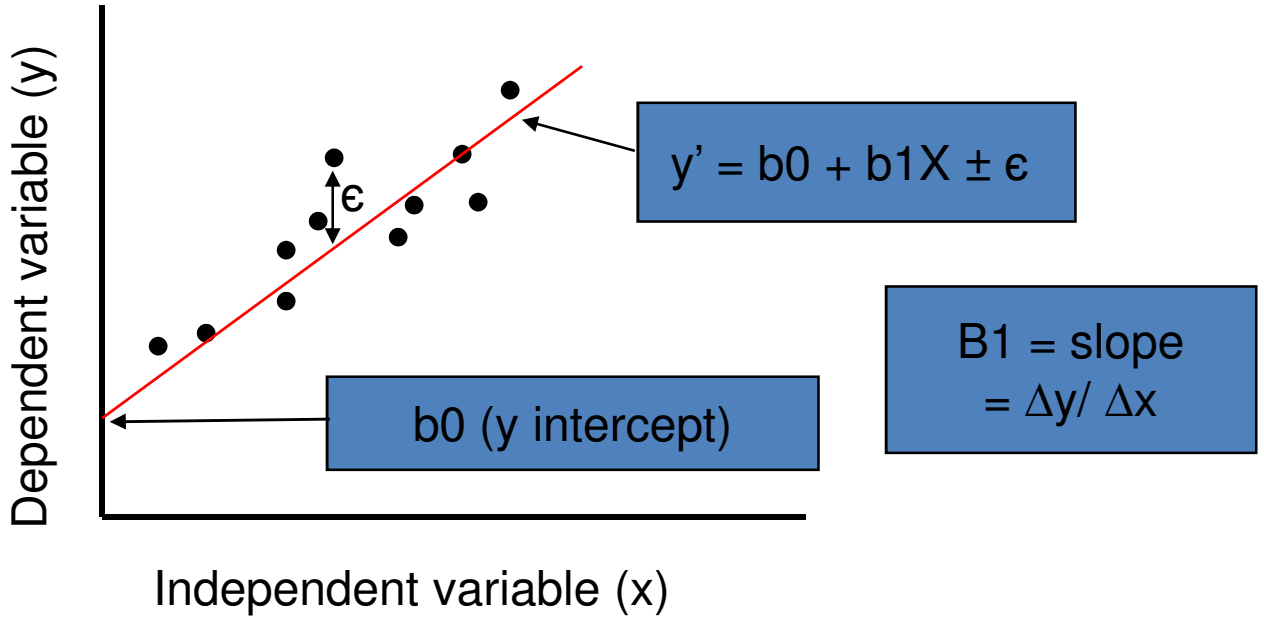
Regression Analysis

Dr. Saurabh Pal
Associate Professor & Head
Department of Computer Applications
VBS Purvanchal University, Jaunpur



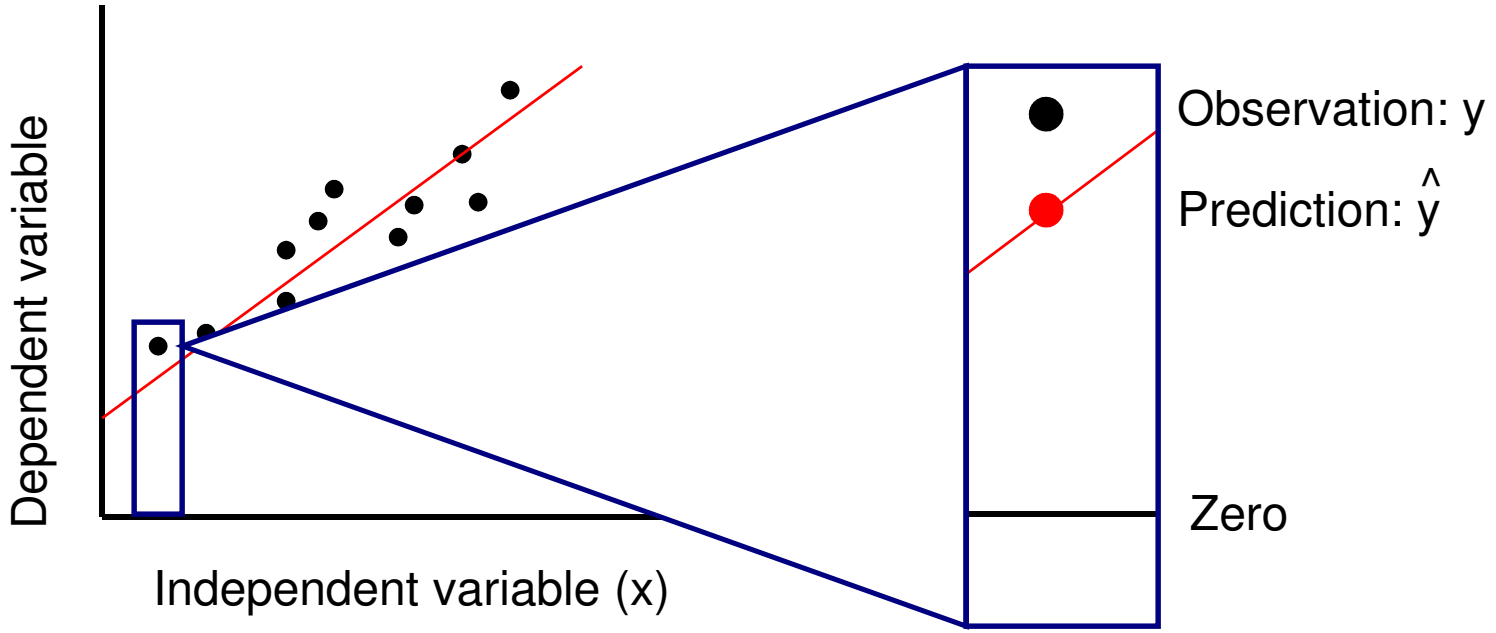
Regression is the attempt to explain the variation in a dependent variable using the variation in independent variables.

If the independent variable(s) sufficiently explain the variation in the dependent variable, the model can be used for prediction.



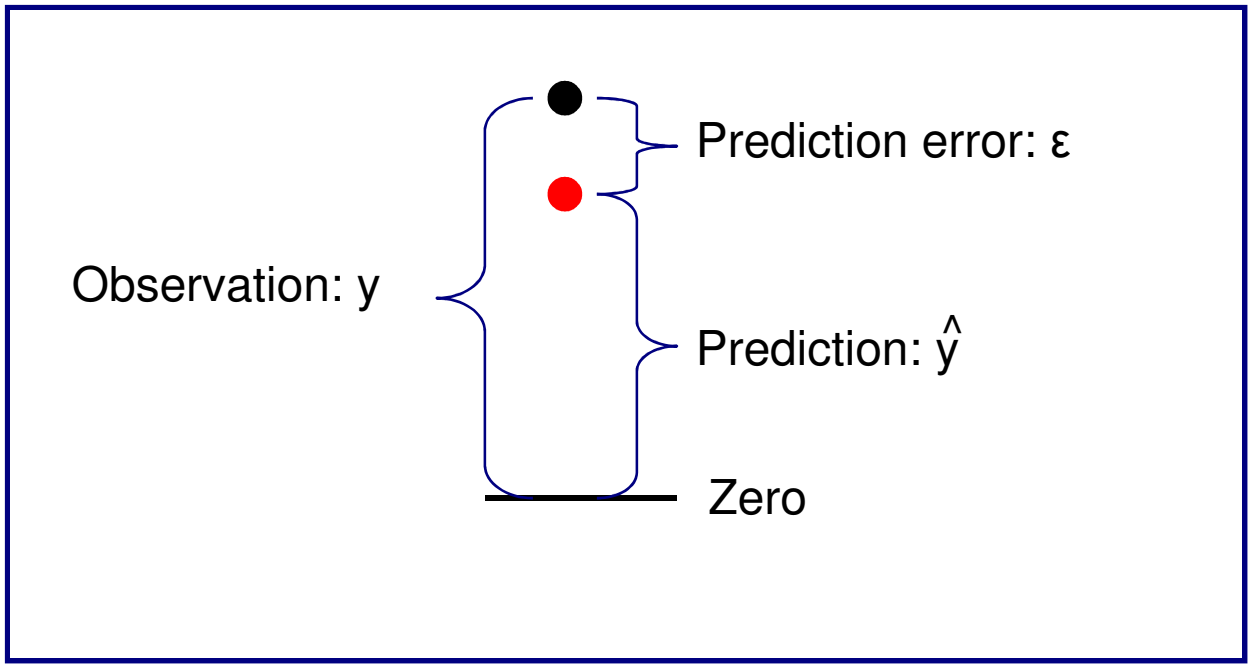
The output of a regression is a function that predicts the dependent variable based upon values of the independent variables.

Simple regression fits a straight line to the data.



The function will make a prediction for each observed data point.

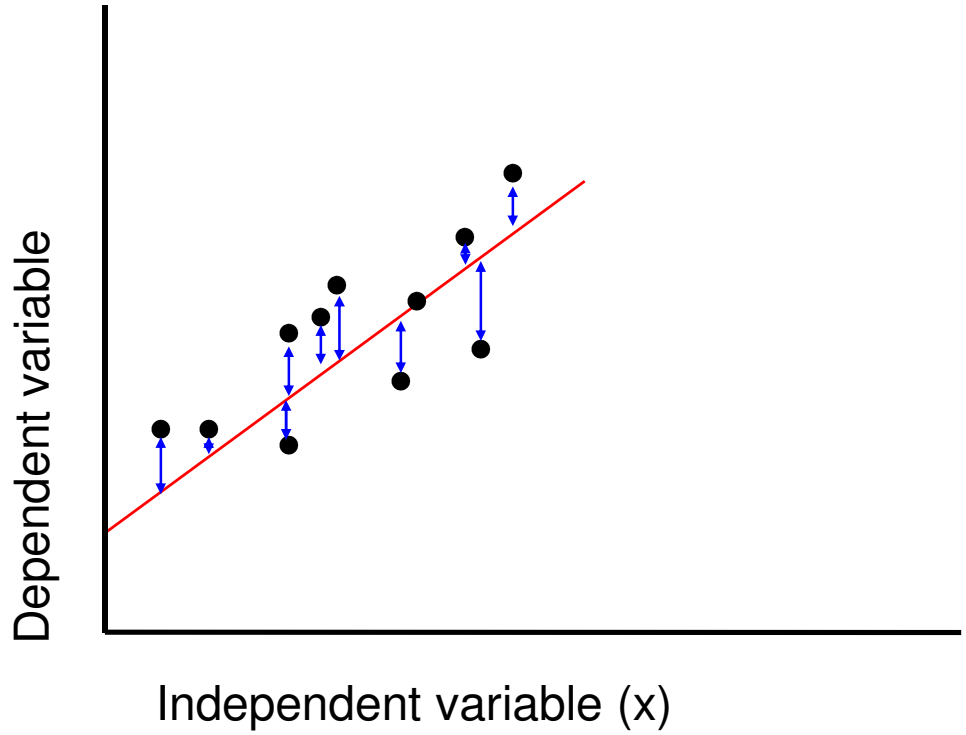
The observation is denoted by y and the prediction is denoted by $\hat{y}$.



For each observation, the variation can be described as:

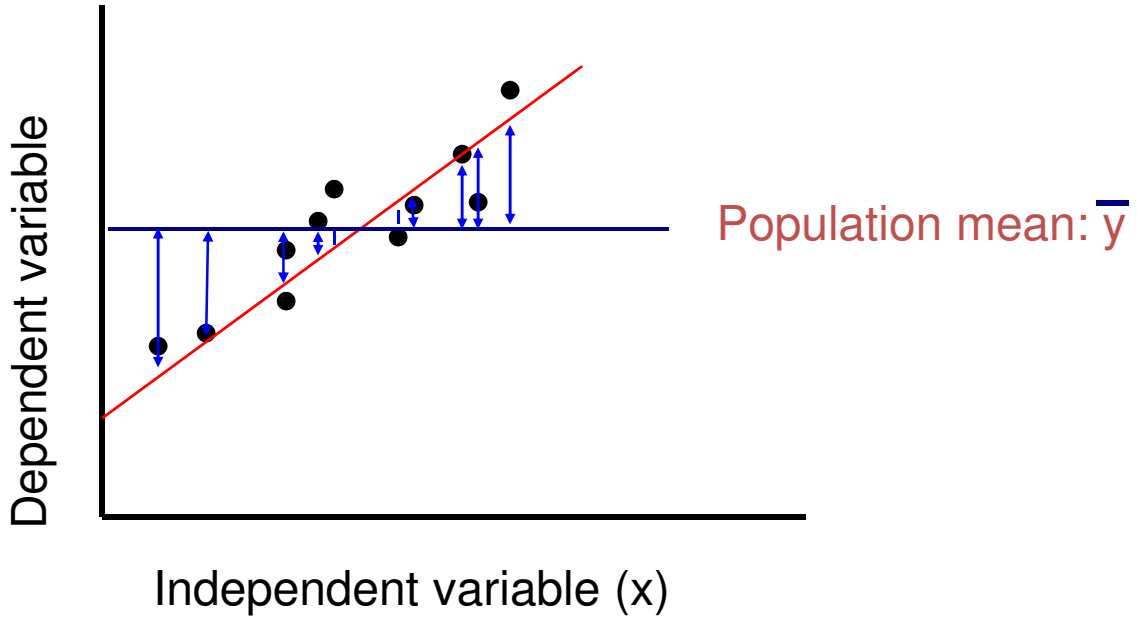
$$y = \hat{y} + \varepsilon$$

Actual = Explained + Error



A least squares regression selects the line with the lowest total sum of squared prediction errors.

This value is called the Sum of Squares of Error, or SSE.



The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean.

The Total Sum of Squares (SST) is equal to SSR + SSE.

Mathematically,

$$SSR = \sum (\hat{y} - \bar{y})^2$$ **(measure of explained variation)**

$$SSE = \sum (y - \hat{y})$$ **(measure of unexplained variation)**

$$SST = SSR + SSE = \sum (y - \bar{y})^2$$ **(measure of total variation in y)**

The proportion of total variation (SST) that is explained by the regression (SSR) is known as the Coefficient of Determination, and is often referred to as R^2.

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE}$$

The value of R^2 can range between 0 and 1, and the higher its value the more accurate the regression model is. It is often referred to as a percentage.

The Standard Error of a regression is a measure of its variability. It can be used in a similar manner to standard deviation, allowing for prediction intervals.

y ± 2 standard errors will provide approximately 95% accuracy, and 3 standard errors will provide a 99% confidence interval.

Standard Error is calculated by taking the square root of the average prediction error.

$$\text{Standard Error} = \sqrt{\frac{SSE}{n-k}}$$

Where n is the number of observations in the sample and k is the total number of variables in the model

Parameter estimation

- Assume a particular form for the density (e.g. Gaussian), so only the parameters (e.g., mean and variance) need to be estimated

- Maximum Likelihood

- Bayesian Estimation

- Introduction
  - Bayesian framework
    - We could design an optimal classifier if we knew:
      - $P(\omega_i)$: priors
      - $P(x \mid \omega_i)$: class-conditional densities

      Unfortunately, we rarely have this complete information!

  - Design a classifier based on a set of labeled training samples (supervised learning)
    - Assume priors are known
    - Need sufficient no. of training samples for estimating class-conditional densities, especially when the dimensionality of the feature space is large

– Assumption about the problem: parametric model of $P(x \mid \omega_i)$ is available

– Assume $P(x \mid \omega_i)$ is multivariate Gaussian

$$P(x \mid \omega_i) \sim N(\mu_i, \Sigma_i)$$

- Characterized by 2 parameters

– Parameter estimation techniques
- Maximum-Likelihood (ML) and Bayesian estimation
- Results of the two procedures are nearly identical, but there is a subtle difference

- In ML estimation parameters are assumed to be fixed but unknown! Bayesian parameter estimation procedure, by its nature, utilizes whatever prior information is available about the unknown parameter
- MLE: Best parameters are obtained by maximizing the probability of obtaining the samples observed
- Bayesian methods view the parameters as random variables having some known prior distribution; How do we know the priors?
- In either approach, we use $P(\omega_i \mid x)$ for our classification rule!

- Maximum-Likelihood Estimation
  - Has good convergence properties as the sample size increases; estimated parameter value approaches the true value as n increases
  - Simpler than any other alternative technique

– General principle

- Assume we have c classes and

  $P(x \mid \omega_j) \sim N(\mu_j, \Sigma_j)$

  $P(x \mid \omega_j) \equiv P(x \mid \omega_j, \theta_j)$, where

$$\theta = (\mu_j, \Sigma_j) = (\mu_j^1, \mu_j^2, ..., \sigma_j^{11}, \sigma_j^{22}, \mathbf{cov}(x_j^m, x_j^n)...)$$

Use class ω_j samples to estimate class ω_j parameters

- Use the information in training samples to estimate

  $\theta = (\theta_1, \theta_2, \ldots, \theta_c)$; θ_i (i = 1, 2, …, c) is associated with the ith category

- Suppose sample set D contains n iid samples, $x_1, x_2, \ldots, x_n$

$$\mathbf{P(D \mid \theta) = \prod_{k=1}^{k=n} P(x_k \mid \theta) = F(\theta)}$$

P(D | θ) is called the likelihood of θ w.r.t. the set of samples)

- ML estimate of θ is, by definition, the value that maximizes P(D | θ)

  "It is the value of θ that best agrees with the actually observed training samples"

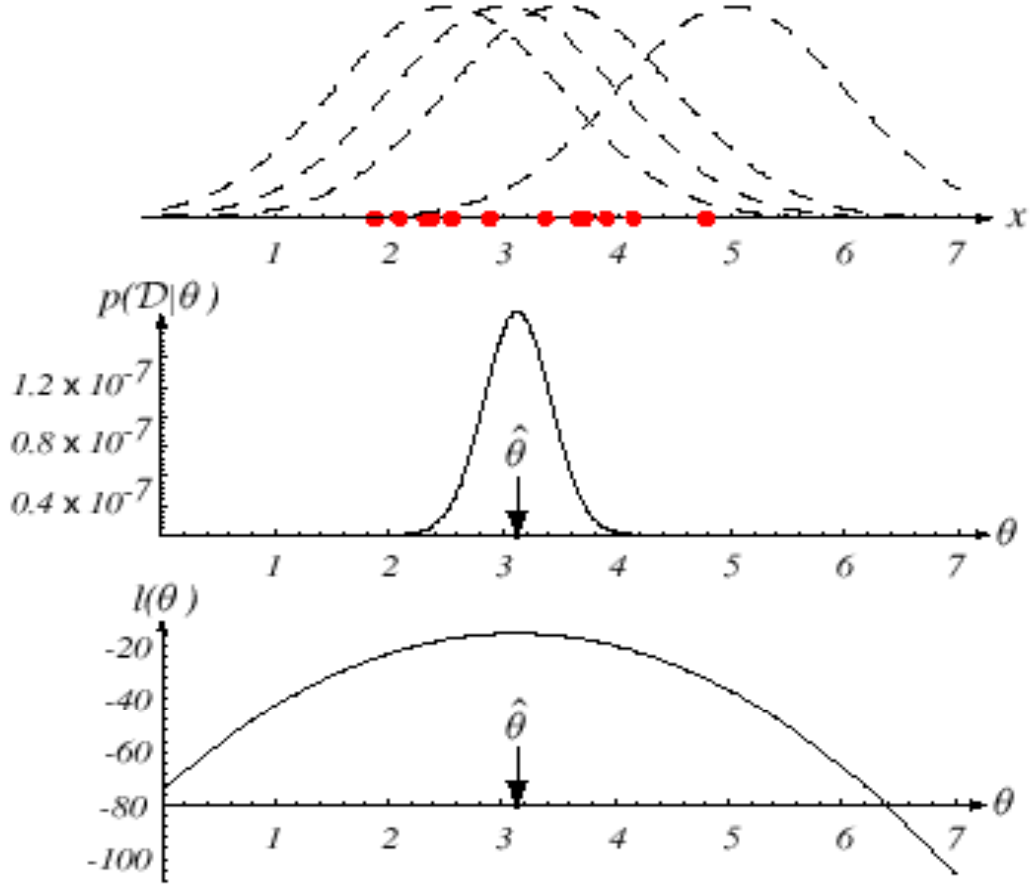


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.

- Optimal estimation
  - Let $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$ and ∇_θ be the gradient operator

$$\nabla_\theta = \left[\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, \ldots, \frac{\partial}{\partial \theta_p} \right]^t$$

  - We define $l(\theta)$ as the log-likelihood function

$$l(\theta) = \ln P(D \mid \theta)$$

  - New problem statement:

    determine θ that maximizes the log-likelihood

$$\hat{\theta} = \arg\max_{\theta} l(\theta)$$

Set of necessary conditions for an optimum is:

$$(\nabla_\theta \mathbf{l} = \sum_{k=1}^{k=n} \nabla_\theta \mathbf{ln P(x_k \mid \theta)})$$

$$\nabla_\theta l = 0$$

- Example of a specific case: unknown μ
  - $P(x \mid \mu) \sim N(\mu, \Sigma)$
    (Samples are drawn from a multivariate normal population)

$$\ln P(\mathbf{x_k} \mid \mu) = -\frac{1}{2}\ln\left[(2\pi)^d |\Sigma|\right] - \frac{1}{2}(\mathbf{x_k} - \mu)^t \Sigma^{-1}(\mathbf{x_k} - \mu)$$

$$\text{and} \quad \nabla_{\theta\mu} \ln P(\mathbf{x_k} \mid \mu) = \Sigma^{-1}(\mathbf{x_k} - \mu)$$

$\theta = \mu$, therefore the ML estimate for μ must satisfy:

$$\sum_{k=1}^{k=n} \Sigma^{-1}(\mathbf{x_k} - \hat{\mu}) = 0$$

- Multiplying by Σ and rearranging, we obtain:

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^{k=n} \mathbf{x}_k$$

which is the arithmetic average or the mean of the samples of the training samples!

Conclusion:

Given $P(x_k \mid \omega_j)$, j = 1, 2, …, c to be Gaussian in a *d*-dimensional feature space, estimate the vector $\theta = (\theta_1, \theta_2, \ldots, \theta_c)^t$ and perform a classification using the Bayes decision rule of chapter 2!

- ML Estimation:
  - Univariate Gaussian Case: *unknown* μ & σ

    $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$

$$l = \ln P(x_k \mid \theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_\theta l = \begin{pmatrix} \frac{\sigma}{\sigma\theta_1}(\ln P(x_k \mid \theta)) \\ \frac{\sigma}{\sigma\theta_2}(\ln P(x_k \mid \theta)) \end{pmatrix} = 0$$

$$\begin{cases} \frac{1}{\theta_2}(x_k - \theta_1) = 0 \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} = 0 \end{cases}$$

Summation:

$$\begin{cases} \sum_{k=1}^{k=n} \frac{1}{\hat{\theta}_2}(x_k - \theta_1) = 0 & (1) \\ -\sum_{k=1}^{k=n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{k=n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 & (2) \end{cases}$$

Combining (1) and (2), one obtains:

$$\mu = \sum_{k=1}^{k=n} \frac{x_k}{n} \quad ; \quad \sigma^2 = \frac{\sum_{k=1}^{k=n} (x_k - \mu)^2}{n}$$

- Bias
  - ML estimate for σ^2 is biased

$$E\left[\frac{1}{n}\Sigma(x_i - \bar{x})^2\right] = \frac{n-1}{n}.\sigma^2 \neq \sigma^2$$

  - An unbiased estimator for Σ is:

$$\underbrace{C = \frac{1}{n-1}\sum_{k=1}^{k=n}(x_k - \mu)(x_k - \hat{\mu})^t}_{\text{Sample covariance matrix}}$$

- Bayesian Estimation (Bayesian learning approach for pattern classification problems)
  - In MLE θ was supposed to have a fixed value
  - In BE θ is a random variable
  - The computation of posterior probabilities $P(\omega_i \mid x)$ lies at the heart of Bayesian classification
  - Goal: compute $P(\omega_i \mid x, D)$

    Given the training sample set D, Bayes formula can be written

$$P(\omega_i \mid x, D) = \frac{P(x \mid \omega_i, D).P(\omega_i \mid D)}{\sum_{j=1}^{c} P(x \mid \omega_j, D).P(\omega_j \mid D)}$$

- To demonstrate the preceding equation, use:

$$\mathbf{P}(\mathbf{x}, D \mid \omega_i) = \mathbf{P}(\mathbf{x} \mid D\,\omega_i).\mathbf{P}(D \mid \omega_i)$$

$$\mathbf{P}(\mathbf{x} \mid D) = \sum_{j} \mathbf{P}(\mathbf{x}, \omega_j \mid D)$$

$$\mathbf{P}(\omega_i) = \mathbf{P}(\omega_i \mid D) \ \textbf{(Training sample provides this!)}$$

Thus :

$$\mathbf{P}(\omega_i \mid \mathbf{x}, D) = \frac{\mathbf{P}(\mathbf{x} \mid \omega_i, D_i).\mathbf{P}(\omega_i)}{\sum_{j=1}^{c} \mathbf{P}(\mathbf{x} \mid \omega_j, D).\mathbf{P}(\omega_j)}$$

- Bayesian Parameter Estimation: Gaussian Case

Goal: Estimate θ using the a-posteriori density $P(\theta \mid D)$

– The univariate Gaussian case: $P(\mu \mid D)$

μ is the only unknown parameter

$$P(x \mid \mu) \sim N(\mu, \sigma^2)$$

$$P(\mu) \sim N(\mu_0, \sigma_0^2)$$

μ_0 and σ_0 are known!

$$P(\mu \mid D) = \frac{P(D \mid \mu).P(\mu)}{\int P(D \mid \mu).P(\mu)d\mu} \qquad (1)$$

$$= \alpha \prod_{k=1}^{k=n} P(x_k \mid \mu).P(\mu)$$

– Reproducing density

$$P(\mu \mid D) \sim N(\mu_n, \sigma_n^2) \qquad (2)$$

The updated parameters of the prior:

$$\mu_n = \left(\frac{n\sigma_0^2}{n_0\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} .\mu_0$$

$$\text{and } \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

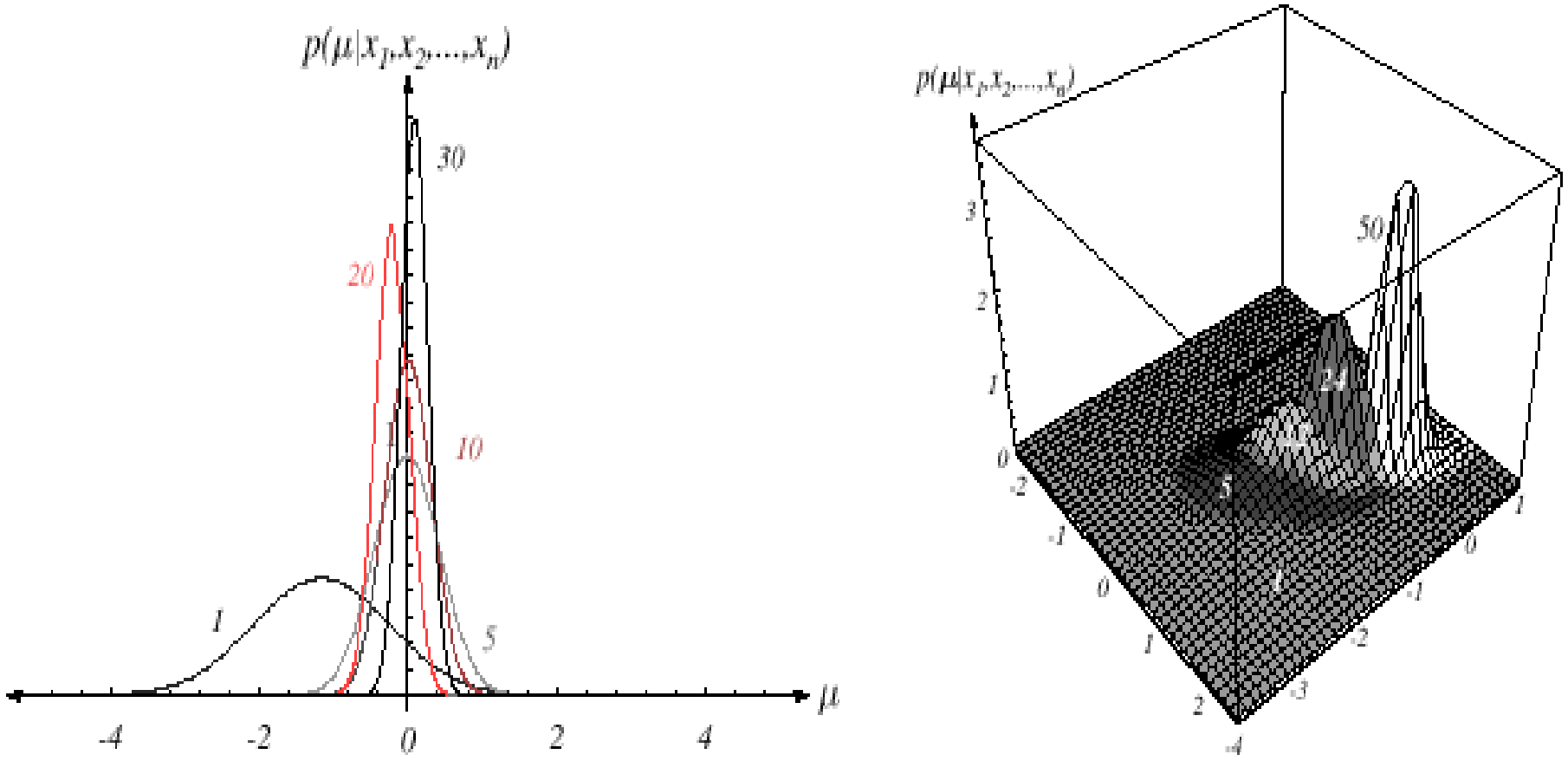


FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

– The univariate case P(x | D)

- P(μ | D) has been computed
- P(x | D) remains to be computed!

$$\mathbf{P(x \mid D) = \int P(x \mid \mu).P(\mu \mid D)d\mu \ \ is\ Gaussian}$$

It provides:

$$\mathbf{P(x \mid D) \sim N(\mu_n, \sigma^2 + \sigma_n^2)}$$

Desired class-conditional density $P(x \mid D_j, \omega_j)$

$P(x \mid D_j, \omega_j)$ together with $P(\omega_j)$ and using Bayes formula, we obtain the Bayesian classification rule:

$$\mathbf{\underset{\omega_j}{Max}\ [P(\omega_j \mid x, D] \equiv \underset{\omega_j}{Max}\ [P(x \mid \omega_j, D_j).P(\omega_j)]}$$

- Bayesian Parameter Estimation: General Theory

  - $P(x \mid D)$ computation can be applied to any situation in which the unknown density can be parametrized: the basic assumptions are:

    - The form of $P(x \mid \theta)$ is assumed known, but the value of θ is not known exactly
    - Our knowledge about θ is assumed to be contained in a known prior density $P(\theta)$
    - The rest of our knowledge about θ is contained in a set D of n random variables $x_1, x_2, \ldots, x_n$ that follows $P(x)$

The basic problem is:

"Compute the posterior density $P(\theta \mid D)$"

then "Derive $P(x \mid D)$"

Using Bayes formula, we have:

$$\mathbf{P}(\theta \mid D) = \frac{\mathbf{P}(D \mid \theta).\mathbf{P}(\theta)}{\int \mathbf{P}(D \mid \theta).\mathbf{P}(\theta)\mathbf{d}\theta},$$

And by ind

$$\mathbf{P}(D \mid \theta) = \prod_{k=1}^{k=n} \mathbf{P}(\mathbf{x}_k \mid \theta)$$

Multivariate Analysis

- Many statistical techniques focus on just one or two variables
- Multivariate analysis (MVA) techniques allow more than two variables to be analysed at once
  - Multiple regression is not typically included under this heading, but can be thought of as a multivariate analysis

Outline of Lectures

- We will cover
  - Why MVA is useful and important
    - Simpson's Paradox
  - Some commonly used techniques
    - Principal components
    - Cluster analysis
    - Correspondence analysis
    - Others if time permits
  - Market segmentation methods
  - An overview of MVA methods and their niches

Simpson's Paradox

- Example: 44% of male applicants are admitted by a university, but only 33% of female applicants
- Does this mean there is unfair discrimination?
- University investigates and breaks down figures for Engineering and English programmes

	Male	Female
Accept	35	20
Refuse entry	45	40
Total	80	60

Simpson's Paradox

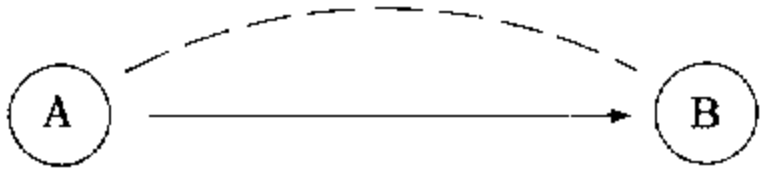
- No relationship between sex and acceptance for either programme
  - So no evidence of discrimination
- Why?
  - More females apply for the English programme, but it it hard to get into
  - More males applied to Engineering, which has a higher acceptance rate than English
- Must look deeper than single cross-tab to find this out

Engineer-ing	Male	Female
Accept	30	10
Refuse entry	30	10
Total	60	20

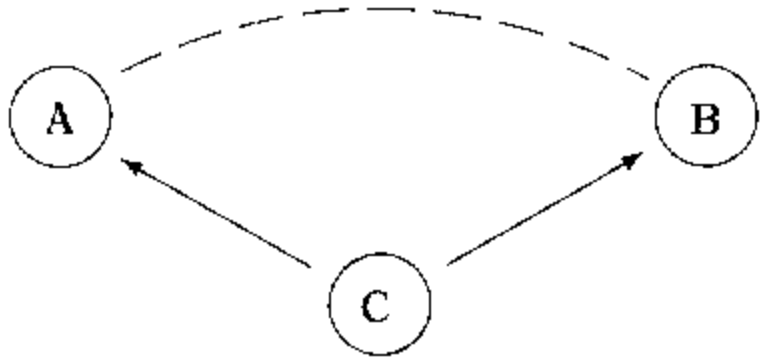
English	Male	Female
Accept	5	10
Refuse entry	15	30
Total	20	40

Another Example

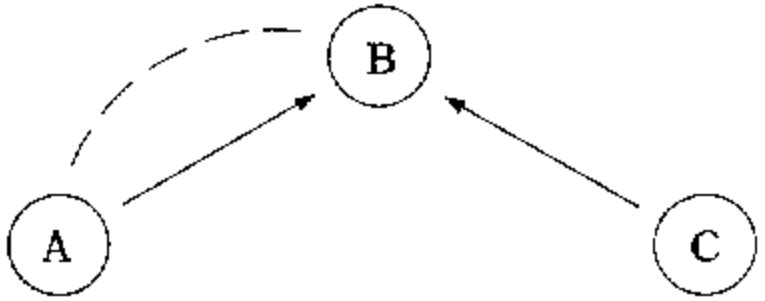
- A study of graduates' salaries showed negative association between economists' starting salary and the level of the degree
  - i.e. PhDs earned less than Masters degree holders, who in turn earned less than those with just a Bachelor's degree
  - Why?
- The data was split into three employment sectors
  - Teaching, government and private industry
  - Each sector showed a positive relationship
  - Employer type was confounded with degree level



CAUSATION—Changes in A cause changes in B.



COMMON RESPONSE—Changes in both A and B are caused by changes in a third variable, C.



CONFOUNDING—Changes in B are caused both by changes in A and by changes in third variable C.

Simpson's Paradox

- In each of these examples, the bivariate analysis (cross-tabulation or correlation) gave misleading results
- Introducing another variable gave a better understanding of the data
  - It even reversed the initial conclusions

Many Variables

- Commonly have many relevant variables in market research surveys
  - E.g. one not atypical survey had ~2000 variables
  - Typically researchers pore over many crosstabs
  - However it can be difficult to make sense of these, and the crosstabs may be misleading
- MVA can help summarise the data
  - E.g. factor analysis and segmentation based on agreement ratings on 20 attitude statements
- MVA can also reduce the chance of obtaining spurious results

Multivariate Analysis Methods

- Two general types of MVA technique
  - Analysis of dependence
    - Where one (or more) variables are dependent variables, to be explained or predicted by others
      - E.g. Multiple regression, PLS, MDA
  - Analysis of interdependence
    - No variables thought of as “dependent”
    - Look at the relationships among variables, objects or cases
      - E.g. cluster analysis, factor analysis

Principal Components

- Identify underlying dimensions or principal components of a distribution
- Helps understand the joint or common variation among a set of variables
- Probably the most commonly used method of deriving “factors” in factor analysis (before rotation)

Principal Components

- The first principal component is identified as the vector (or equivalently the linear combination of variables) on which the most data variation can be projected
- The 2nd principal component is a vector perpendicular to the first, chosen so that it contains as much of the remaining variation as possible
- And so on for the 3rd principal component, the 4th, the 5th etc.

Principal Components - Examples

- Ellipse, ellipsoid, sphere
- Rugby ball
- Pen
- Frying pan
- Banana
- CD
- Book

Multivariate Normal Distribution

- Generalisation of the univariate normal
- Determined by the mean (vector) and covariance matrix

$$X \sim N(\mu, \Sigma)$$

- E.g. Standard bivariate normal

$$X \sim N((0,0), I_2), \quad p(x) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$$

Example – Crime Rates by State

Crime Rates per 100,000 Population by State

Obs	State	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto_Theft
1	Alabama	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
2	Alaska	10.8	51.6	96.8	284.0	1331.7	3369.8	753.3
3	Arizona	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
4	Arkansas	8.8	27.6	83.2	203.4	972.6	1862.1	183.4
5	California	11.5	49.4	287.0	358.0	2139.4	3499.8	663.5
…	…	…	…	…	…	…	…	…

The PRINCOMP Procedure

Observations	50
Variables	7

Simple Statistics							
	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto_Theft
Mean	7.444000000	25.73400000	124.0920000	211.3000000	1291.904000	2671.288000	377.5260000
StD	3.866768941	10.75962995	88.3485672	100.2530492	432.455711	725.908707	193.3944175

Correlation Matrix							
	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto_Theft
Murder	1.0000	0.6012	0.4837	0.6486	0.3858	0.1019	0.0688
Rape	0.6012	1.0000	0.5919	0.7403	0.7121	0.6140	0.3489
Robbery	0.4837	0.5919	1.0000	0.5571	0.6372	0.4467	0.5907
Assault	0.6486	0.7403	0.5571	1.0000	0.6229	0.4044	0.2758
Burglary	0.3858	0.7121	0.6372	0.6229	1.0000	0.7921	0.5580
Larceny	0.1019	0.6140	0.4467	0.4044	0.7921	1.0000	0.4442
Auto_Theft	0.0688	0.3489	0.5907	0.2758	0.5580	0.4442	1.0000

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.11495951	2.87623768	0.5879	0.5879
2	1.23872183	0.51290521	0.1770	0.7648
3	0.72581663	0.40938458	0.1037	0.8685
4	0.31643205	0.05845759	0.0452	0.9137
5	0.25797446	0.03593499	0.0369	0.9506
6	0.22203947	0.09798342	0.0317	0.9823
7	0.12405606		0.0177	1.0000

Eigenvectors							
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
Murder	0.300279	-.629174	0.178245	-.232114	0.538123	0.259117	0.267593
Rape	0.431759	-.169435	-.244198	0.062216	0.188471	-.773271	-.296485
Robbery	0.396875	0.042247	0.495861	-.557989	-.519977	-.114385	-.003903
Assault	0.396652	-.343528	-.069510	0.629804	-.506651	0.172363	0.191745
Burglary	0.440157	0.203341	-.209895	-.057555	0.101033	0.535987	-.648117
Larceny	0.357360	0.402319	-.539231	-.234890	0.030099	0.039406	0.601690
Auto_Theft	0.295177	0.502421	0.568384	0.419238	0.369753	-.057298	0.147046

- 2-3 components explain 76%-87% of the variance
- First principal component has uniform variable weights, so is a general crime level indicator
- Second principal component appears to contrast violent versus property crimes
- Third component is harder to interpret

Cluster Analysis

- Techniques for identifying separate groups of similar cases
  - Similarity of cases is either specified directly in a distance matrix, or defined in terms of some distance function
- Also used to summarise data by defining segments of similar cases in the data
  - This use of cluster analysis is known as “dissection”

Clustering Techniques

- Two main types of cluster analysis methods
  - Hierarchical cluster analysis
    - Each cluster (starting with the whole dataset) is divided into two, then divided again, and so on
  - Iterative methods
    - k-means clustering (PROC FASTCLUS)
    - Analogous non-parametric density estimation method
  - Also other methods
    - Overlapping clusters
    - Fuzzy clusters

Applications

- Market segmentation is usually conducted using some form of cluster analysis to divide people into segments
  - Other methods such as latent class models or archetypal analysis are sometimes used instead
- It is also possible to cluster other items such as products/SKUs, image attributes, brands

Tandem Segmentation

- One general method is to conduct a factor analysis, followed by a cluster analysis
- This approach has been criticised for losing information and not yielding as much discrimination as cluster analysis alone
- However it can make it easier to design the distance function, and to interpret the results

Tandem *k*-means Example

```
proc factor data=datafile n=6 rotate=varimax round reorder flag=.54 scree out=scores;
  var reasons1-reasons15 usage1-usage10;
run;

proc fastclus data=scores maxc=4 seed=109162319 maxiter=50;
  var factor1-factor6;
run;
```

- Have used the default unweighted Euclidean distance function, which is not sensible in every context
- Also note that *k*-means results depend on the initial cluster centroids (determined here by the seed)
- Typically *k*-means is very prone to local maxima
  - Run at least 20 times to ensure reasonable maximum

Selected Outputs

19th run of 5 segments

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Distance Between Cluster Centroids
1	433	0.9010	4.5524	4	2.0325
2	471	0.8487	4.5902	4	1.8959
3	505	0.9080	5.3159	4	2.0486
4	870	0.6982	4.2724	2	1.8959
5	433	0.9300	4.9425	4	2.0308

Selected Outputs

19th run of 5 segments

FASTCLUS Procedure: Replace=RANDOM Radius=0 Maxclusters=5 Maxiter=100 Converge=0.02

Statistics for Variables

Variable	Total STD	Within STD	R-Squared	RSQ/(1-RSQ)
FACTOR1	1.000000	0.788183	0.379684	0.612082
FACTOR2	1.000000	0.893187	0.203395	0.255327
FACTOR3	1.000000	0.809710	0.345337	0.527503
FACTOR4	1.000000	0.733956	0.462104	0.859095
FACTOR5	1.000000	0.948424	0.101820	0.113363
FACTOR6	1.000000	0.838418	0.298092	0.424689
OVER-ALL	1.000000	0.838231	0.298405	0.425324

Pseudo F Statistic = 287.84
Approximate Expected Over-All R-Squared = 0.37027
Cubic Clustering Criterion = -26.135
WARNING: The two above values are invalid for correlated variables.

Selected Outputs

19th run of 5 segments

Cluster Means

Cluster	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6
1	-0.17151	0.86945	-0.06349	0.08168	0.14407	1.17640
2	-0.96441	-0.62497	-0.02967	0.67086	-0.44314	0.05906
3	-0.41435	0.09450	0.15077	-1.34799	-0.23659	-0.35995
4	0.39794	-0.00661	0.56672	0.37168	0.39152	-0.40369
5	0.90424	-0.28657	-1.21874	0.01393	-0.17278	-0.00972

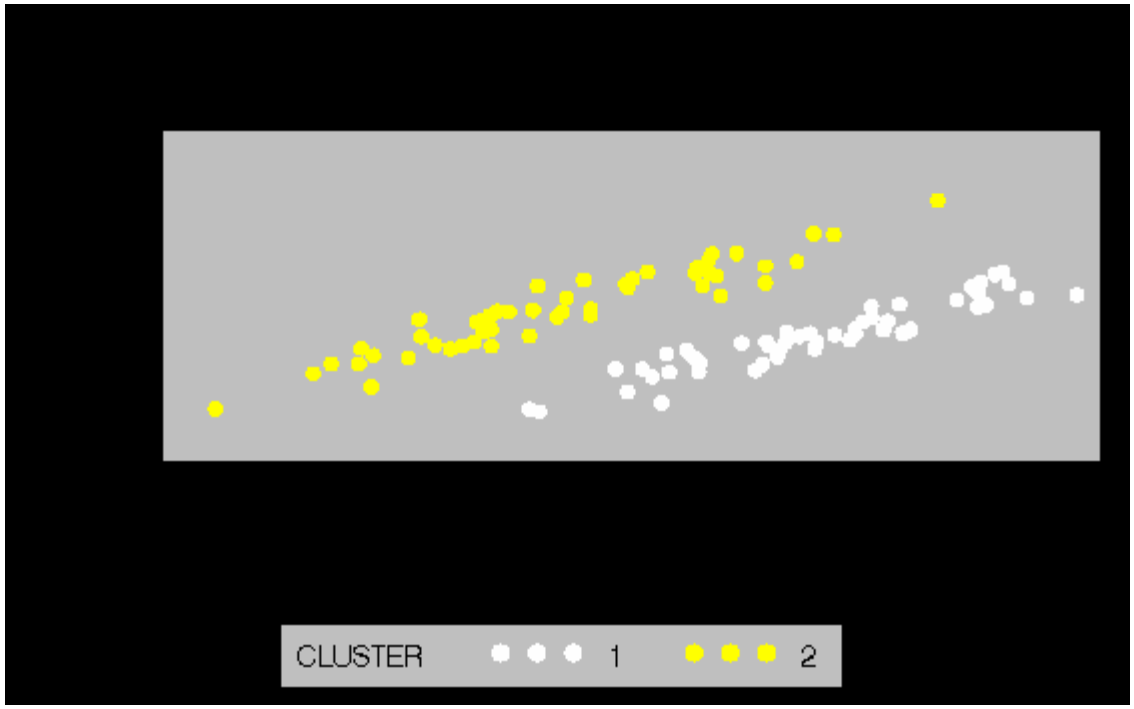
Cluster Standard Deviations

Cluster	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6
1	0.95604	0.79061	0.95515	0.81100	1.08437	0.76555
2	0.79216	0.97414	0.88440	0.71032	0.88449	0.82223
3	0.89084	0.98873	0.90514	0.74950	0.92269	0.97107
4	0.59849	0.74758	0.56576	0.58258	0.89372	0.74160
5	0.80602	1.03771	0.86331	0.91149	1.00476	0.93635

Cluster Analysis Options

- There are several choices of how to form clusters in hierarchical cluster analysis
  - Single linkage
  - Average linkage
  - Density linkage
  - Ward's method
  - Many others
- Ward's method (like k-means) tends to form equal sized, roundish clusters
- Average linkage generally forms roundish clusters with equal variance
- Density linkage can identify clusters of different shapes

Density Linkage



Cluster Analysis Issues

- Distance definition
  - Weighted Euclidean distance often works well, if weights are chosen intelligently
- Cluster shape
  - Shape of clusters found is determined by method, so choose method appropriately
- Hierarchical methods usually take more computation time than *k*-means
- However multiple runs are more important for *k*-means, since it can be badly affected by local minima
- Adjusting for response styles can also be worthwhile
  - Some people give more positive responses overall than others
  - Clusters may simply reflect these response styles unless this is adjusted for, e.g. by standardising responses across attributes for each respondent

MVA - FASTCLUS

- PROC FASTCLUS in SAS tries to minimise the root mean square difference between the data points and their corresponding cluster means
  - Iterates until convergence is reached on this criterion
  - However it often reaches a local minimum
  - Can be useful to run many times with different seeds and choose the best set of clusters based on this RMS criterion
- See http://www.clustan.com/k-means_critique.html for more k-means issues

Iteration History from FASTCLUS

Relative Change in Cluster Seeds

Iteration	Criterion	1	2	3	4	5
1	0.9645	1.0436	0.7366	0.6440	0.6343	0.5666
2	0.8596	0.3549	0.1727	0.1227	0.1246	0.0731
3	0.8499	0.2091	0.1047	0.1047	0.0656	0.0584
4	0.8454	0.1534	0.0701	0.0785	0.0276	0.0439
5	0.8430	0.1153	0.0640	0.0727	0.0331	0.0276
6	0.8414	0.0878	0.0613	0.0488	0.0253	0.0327
7	0.8402	0.0840	0.0547	0.0522	0.0249	0.0340
8	0.8392	0.0657	0.0396	0.0440	0.0188	0.0286
9	0.8386	0.0429	0.0267	0.0324	0.0149	0.0223
10	0.8383	0.0197	0.0139	0.0170	0.0119	0.0173

Convergence criterion is satisfied.

Criterion Based on Final Seeds = 0.83824

Results from Different Initial Seeds

19th run of 5 segments

Cluster Means

Cluster	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6
1	-0.17151	0.86945	-0.06349	0.08168	0.14407	1.17640
2	-0.96441	-0.62497	-0.02967	0.67086	-0.44314	0.05906
3	-0.41435	0.09450	0.15077	-1.34799	-0.23659	-0.35995
4	0.39794	-0.00661	0.56672	0.37168	0.39152	-0.40369
5	0.90424	-0.28657	-1.21874	0.01393	-0.17278	-0.00972

20th run of 5 segments

Cluster Means

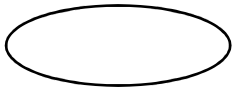
Cluster	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6
1	0.08281	-0.76563	0.48252	-0.51242	-0.55281	0.64635
2	0.39409	0.00337	0.54491	0.38299	0.64039	-0.26904
3	-0.12413	0.30691	-0.36373	-0.85776	-0.31476	-0.94927
4	0.63249	0.42335	-1.27301	0.18563	0.15973	0.77637
5	-1.20912	0.21018	-0.07423	0.75704	-0.26377	0.13729

Howard-Harris Approach

- Provides automatic approach to choosing seeds for *k*-means clustering
- Chooses initial seeds by fixed procedure
  - Takes variable with highest variance, splits the data at the mean, and calculates centroids of the resulting two groups
  - Applies *k*-means with these centroids as initial seeds
  - This yields a 2 cluster solution
  - Choose the cluster with the higher within-cluster variance
  - Choose the variable with the highest variance within that cluster, split the cluster as above, and repeat to give a 3 cluster solution
  - Repeat until have reached a set number of clusters
- I believe this approach is used by the ESPRI software package (after variables are standardised by their range)

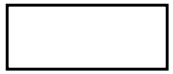
Another “Clustering” Method

- One alternative approach to identifying clusters is to fit a finite mixture model
  - Assume the overall distribution is a mixture of several normal distributions
  - Typically this model is fit using some variant of the EM algorithm
    - E.g. weka.clusterers.EM method in WEKA data mining package
    - See WEKA tutorial for an example using Fisher’s iris data
- Advantages of this method include:
  - Probability model allows for statistical tests
  - Handles missing data within model fitting process
  - Can extend this approach to define clusters based on model parameters, e.g. regression coefficients
- Also known as latent class modeling



=max.

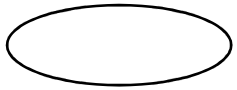
Cluster Means

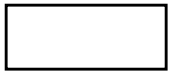


=min.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Reason 1	4.55	2.65	4.21	4.50
Reason 2	4.32	4.32	4.12	4.02
Reason 3	4.43	3.28	3.90	4.06
Reason 4	3.85	3.89	2.15	3.35
Reason 5	4.10	3.77	2.19	3.80
Reason 6	4.50	4.57	4.09	4.28
Reason 7	3.93	4.10	1.94	3.66
Reason 8	4.09	3.17	2.30	3.77
Reason 9	4.17	4.27	3.51	3.82
Reason 10	4.12	3.75	2.66	3.47
Reason 11	4.58	3.79	3.84	4.37
Reason 12	3.51	2.78	1.86	2.60
Reason 13	4.14	3.95	3.06	3.45
Reason 14	3.96	3.75	2.06	3.83
Reason 15	4.19	2.42	2.93	4.04

Cluster Means

 =max.

 =min.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Usage 1	3.43	3.66	3.48	4.00
Usage 2	3.91	3.94	3.86	4.26
Usage 3	3.07	2.95	2.61	3.13
Usage 4	3.85	3.02	2.62	2.50
Usage 5	3.86	3.55	3.52	3.56
Usage 6	3.87	4.25	4.14	4.56
Usage 7	3.88	3.29	2.78	2.59
Usage 8	3.71	2.88	2.58	2.34
Usage 9	4.09	3.38	3.19	2.68
Usage 10	4.58	4.26	4.00	3.91

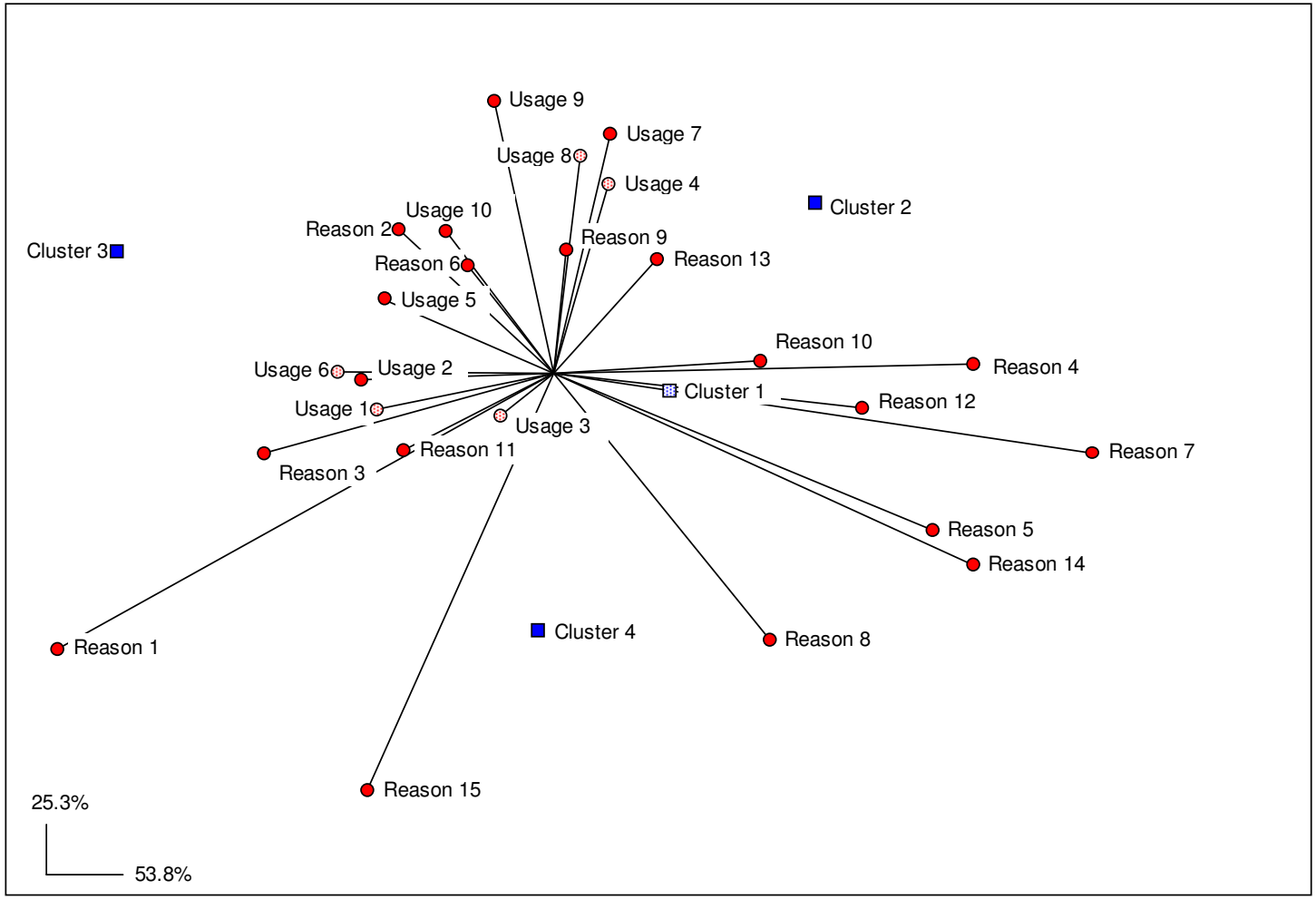
Cluster Means

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Usage 1	3.43	3.66	3.48	4.00
Usage 2	3.91	3.94	3.86	4.26
Usage 3	3.07	2.95	2.61	3.13
Usage 4	3.85	3.02	2.62	2.50
Usage 5	3.86	3.55	3.52	3.56
Usage 6	3.87	4.25	4.14	4.56
Usage 7	3.88	3.29	2.78	2.59
Usage 8	3.71	2.88	2.58	2.34
Usage 9	4.09	3.38	3.19	2.68
Usage 10	4.58	4.26	4.00	3.91

Correspondence Analysis

- Provides a graphical summary of the interactions in a table
- Also known as a perceptual map
  - But so are many other charts
- Can be very useful
  - E.g. to provide overview of cluster results
- However the correct interpretation is less than intuitive, and this leads many researchers astray

Four Clusters (imputed, normalised)



= Correlation < 0.50 2D Fit = 79.1%

Interpretation

- Correspondence analysis plots should be interpreted by looking at points relative to the origin
  - Points that are in similar directions are positively associated
  - Points that are on opposite sides of the origin are negatively associated
  - Points that are far from the origin exhibit the strongest associations
- Also the results reflect relative associations, not just which rows are highest or lowest overall

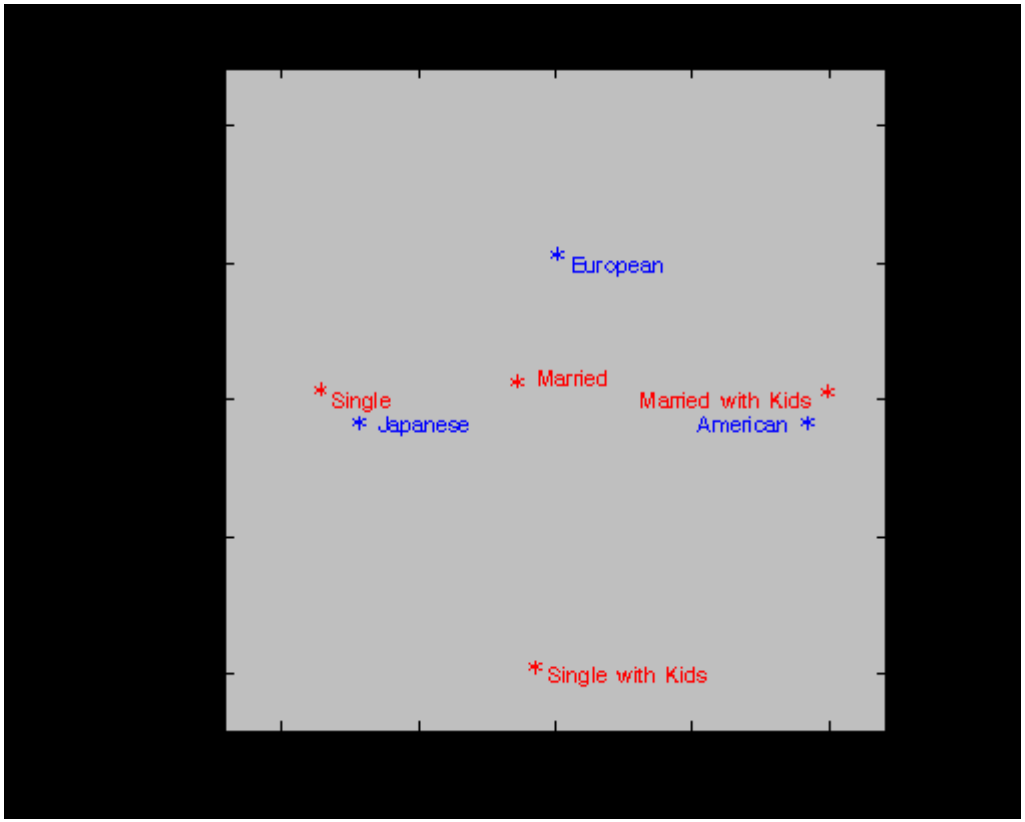
Software for Correspondence Analysis

- Earlier chart was created using a specialised package called BRANDMAP
- Can also do correspondence analysis in most major statistical packages
- For example, using PROC CORRESP in SAS:

```
*---Perform Simple Correspondence Analysis—Example 1 in SAS OnlineDoc;
proc corresp all data=Cars outc=Coor;
   tables Marital, Origin;
run;

*---Plot the Simple Correspondence Analysis Results---;
%plotit(data=Coor, datatype=corresp)
```

Cars by Marital Status

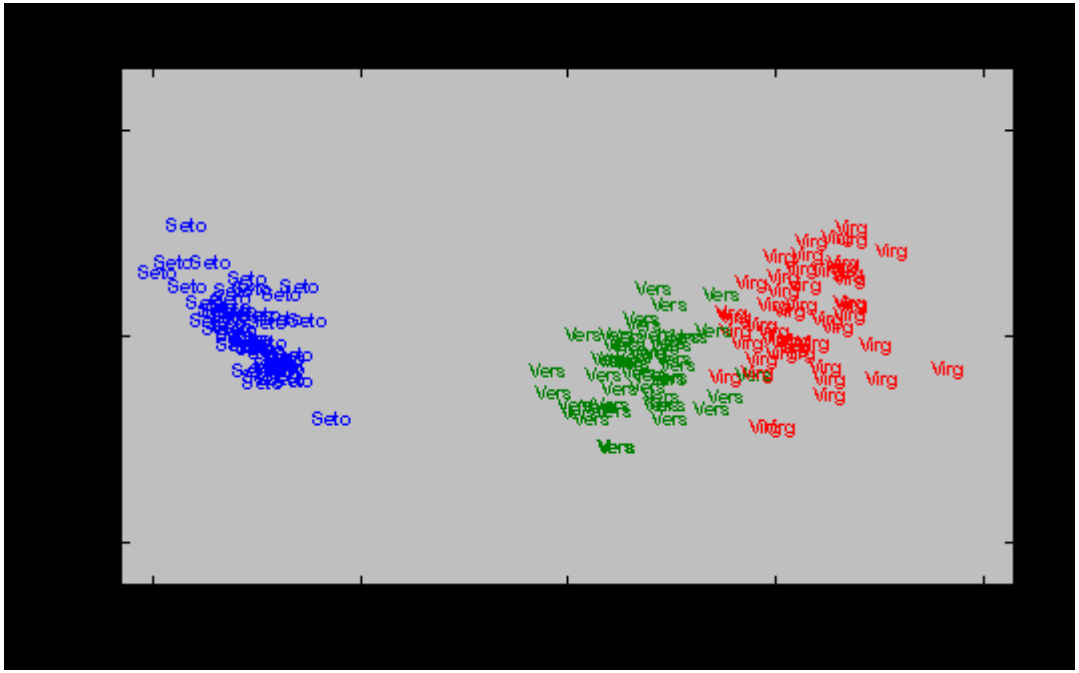


Canonical Discriminant Analysis

- Predicts a discrete response from continuous predictor variables
- Aims to determine which of *g* groups each respondent belongs to, based on the predictors
- Finds the linear combination of the predictors with the highest correlation with group membership
  - Called the first canonical variate
- Repeat to find further canonical variates that are uncorrelated with the previous ones
  - Produces maximum of g-1 canonical variates

CDA Plot

Canonical
Var 2



Canonical Var 1

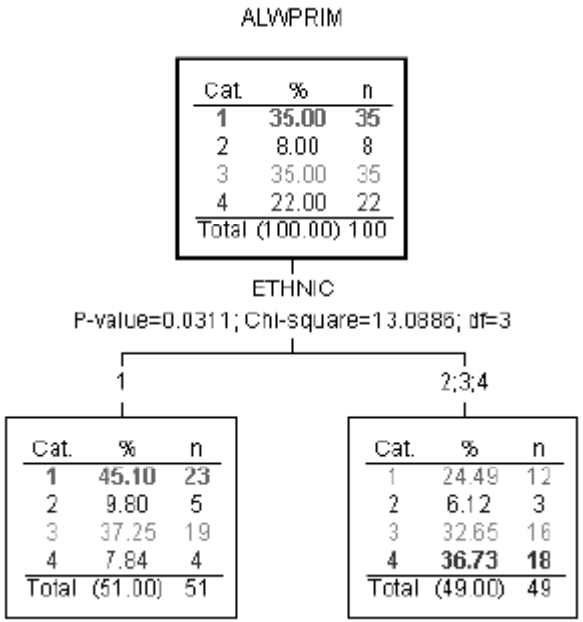
Discriminant Analysis

- Discriminant analysis also refers to a wider family of techniques
  - Still for discrete response, continuous predictors
  - Produces discriminant functions that classify observations into groups
    - These can be linear or quadratic functions
    - Can also be based on non-parametric techniques
  - Often train on one dataset, then test on another

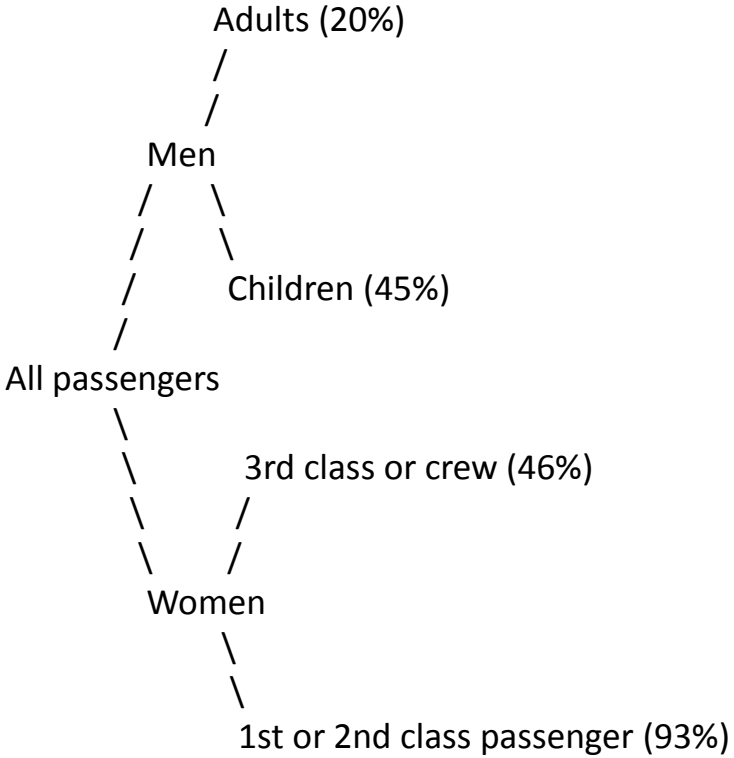
CHAID

- Chi-squared Automatic Interaction Detection
- For discrete response and many discrete predictors
  - Common situation in market research
- Produces a tree structure
  - Nodes get purer, more different from each other
- Uses a chi-squared test statistic to determine best variable to split on at each node
  - Also tries various ways of merging categories, making a Bonferroni adjustment for multiple tests
  - Stops when no more “statistically significant” splits can be found

Example of CHAID Output



Titanic Survival Example



CHAID Software

- Available in SAS Enterprise Miner (if you have enough money)
  - Was provided as a free macro until SAS decided to market it as a data mining technique
  - TREEDISC.SAS – still available on the web, although apparently not on the SAS web site
- Also implemented in at least one standalone package
- Developed in 1970s
- Other tree-based techniques available
  - Will discuss these later

TREEDISC Macro

```
%treedisc(data=survey2, depvar=bs,
        nominal=c o p q x ae af ag ai: aj al am ao ap aw bf_1 bf_2 ck cn:,
        ordinal=lifestag t u v w y ab ah ak,
        ordfloat=ac ad an aq ar as av,
        options=list noformat read,maxdepth=3,
        trace=medium, draw=gr, leaf=50,
        outtree=all);
```

- Need to specify type of each variable
  - Nominal, Ordinal, Ordinal with a floating value

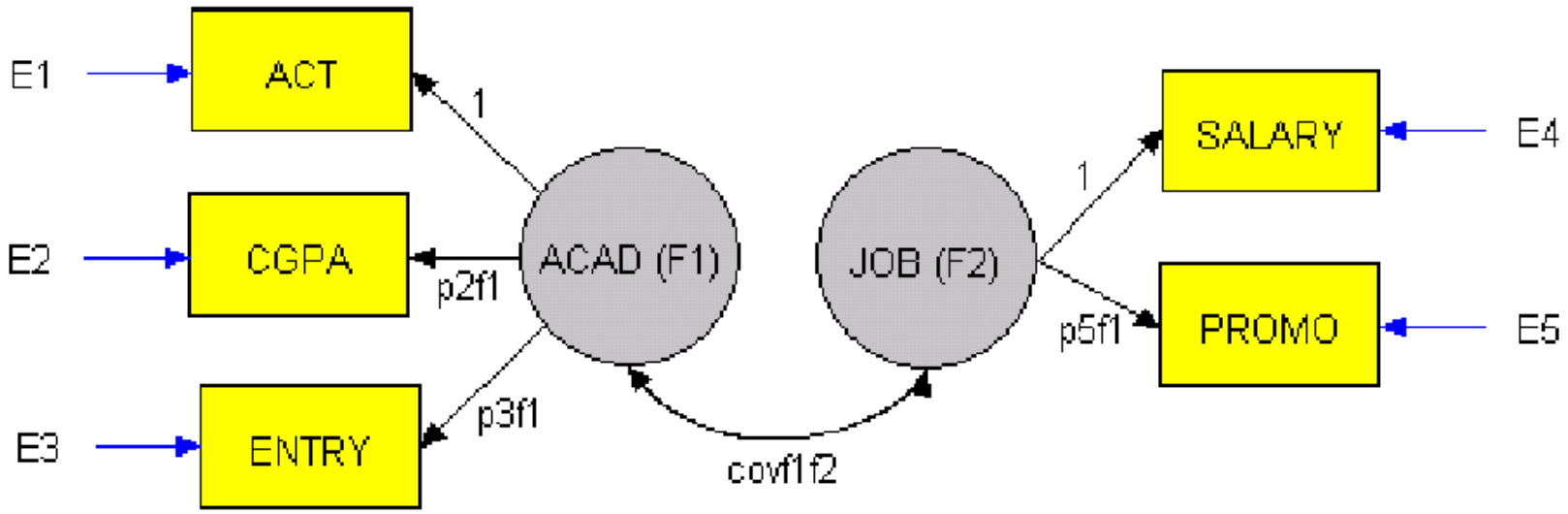
Partial Least Squares (PLS)

- Multivariate generalisation of regression
  - Have model of form $Y=XB+E$
  - Also extract factors underlying the predictors
  - These are chosen to explain both the response variation and the variation among predictors
- Results are often more powerful than principal components regression
- PLS also refers to a more general technique for fitting general path models, not discussed here

Structural Equation Modeling (SEM)

- General method for fitting and testing path analysis models, based on covariances
- Also known as LISREL
- Implemented in SAS in PROC CALIS
- Fits specified causal structures (path models) that usually involve factors or latent variables
  - Confirmatory analysis

SEM Example:
Relationship between
Academic and Job Success



SAS Code

```
data jobfl (type=cov);
 input _type_ $ _name_ $ act cgpa entry
   salary promo;
cards;
n 500 500 500 500 500
cov act 1.024
cov cgpa 0.792 1.077
cov entry 0.567 0.537 0.852
cov salary 0.445 0.424 0.518 0.670
cov promo 0.434 0.389 0.475 0.545 0.716
;
```

```
proc calis data=jobfl cov stderr;
 lineqs
   act = 1*F1 + e1,
   cgpa = p2f1*F1 + e2,
   entry = p3f1*F1 + e3,
   salary = 1*F2 + e4,
   promo = p5f1*F2 + e5;
 std
   e1 = vare1,
   e2 = vare2,
   e3 = vare3,
   e4 = vare4,
   e5 = vare5,
   F1 = varF1,
   F2 = varF2;
 cov
   f1 f2 = covf1f2;
 var act cgpa entry salary promo;
run;
```

Results

Table 3. Parameter Estimates – Business Model

Variable	*Estimate*	*Standard Error*	*z-value*
act←F1	1.000*	--	--
Cgpa←F1 (p2f1)	0.972	0.048	20.265
Entry←F1 (p3f1)	0.785	0.043	18.122
Salary←F2	1.000*	--	--
Promo←F2 (p5f1)	0.943	0.046	20.341
Variance			
e1	0.260	0.029	8.780
e2	0.355	0.033	10.820
e3	0.381	0.030	12.900
e4	0.092	0.023	4.050
e5	0.202	0.024	8.690
F1	0.764	0.067	11.350
F2	0.578	0.047	12.200
Covariance			
F1F2	0.485	0.042	11.470

*fixed value

- All parameters are statistically significant, with a high correlation being found between the latent traits of academic and job success
- However the overall chi-squared value for the model is 111.3, with 4 d.f., so the model does not fit the observed covariances perfectly

Latent Variable Models

- Have seen that both latent trait and latent class models can be useful
  - Latent traits for factor analysis and SEM
  - Latent class for probabilistic segmentation
- Mplus software can now fit combined latent trait and latent class models
  - Appears very powerful
  - Subsumes a wide range of multivariate analyses

Broader MVA Issues

- Preliminaries
  - EDA is usually very worthwhile
    - Univariate summaries, e.g. histograms
    - Scatterplot matrix
    - Multivariate profiles, spider-web plots
  - Missing data
    - Establish amount (by variable, and overall) and pattern (across individuals)
    - Think about reasons for missing data
    - Treat missing data appropriately – e.g. impute, or build into model fitting

MVA Issues

- Preliminaries (continued)
  - Check for outliers
    - Large values of Mahalonobis' D^2
- Testing results
  - Some methods provide statistical tests
  - But others do not
    - Cross-validation gives a useful check on the results
      - Leave-1-out cross-validation
      - Split-sample training and test datasets
        - » Sometimes 3 groups needed
        - » For model building, training and testing