Variable Compression:
Principal Components Analysis
Linear Discriminant Analysis
Variable Selection

• Use few variables

• Interpretation is easier
Ranking Variables Univariately

- Remove one variable from the model at a time
- Compare performance of \([n-1]\) model with full \([n]\) model
- Rank variables according to performance difference
Figures removed due to copyright reasons.

Please see:
Variable Selection

• Ideal: consider all variable combinations
  – Not feasible in most data sets with large number of n variables: $2^n$

• Greedy Forward:
  – Select most important variable as the “first component”, Select other variables conditioned on the previous ones
  – Stepwise: consider backtracking

• Greedy Backward:
  – Start with all variables and remove one at a time.
  – Stepwise: consider backtracking

• Other search methods: genetic algorithms that optimize classification performance and # variables
Variable Selection on Small Round Blood Cell Tumors
Variable compression

- Direction of maximum variability
  - PCA
    - PCA regression
  - LDA
  - (Partial Least Squares)
correlation_coefficient

\[ r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \rho \]

**VARIANCE**

\[ \sigma_{XX} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})}{n-1} \]

**COVARIANCE**

\[ \sigma_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \]

**ST.DEVIATION**

\[ \sigma_X = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})}{n-1}} \]
Covariance and Correlation Matrices

\[
\text{cov} = \begin{bmatrix}
\sigma_{XX} & \sigma_{XY} \\
\sigma_{YX} & \sigma_{YY}
\end{bmatrix}
\]

\[
\text{corr} = \begin{bmatrix}
1 & \rho \\
\rho & 1
\end{bmatrix}
\]

\[
\sigma_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}
\]

\[
\sigma_{XX} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})}{n-1}
\]
Slope from linear regression is asymmetric, covariance and ρ are symmetric

\[ \beta_0 = \bar{y} - \beta_1 \bar{x} \]
\[ \beta_1 = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} \]

\[ y = \beta_0 + \beta_1 x \]
\[ y = 2 + 4x \]
\[ x = \frac{y}{4 - 2} \]

\[
\begin{bmatrix}
0.86 & 0.35 \\
0.35 & 15.69
\end{bmatrix} = \Sigma
\]
\[
\begin{bmatrix}
1 & 0.96 \\
0.96 & 1
\end{bmatrix} = \text{corr}
\]
Principal Components Analysis

- Our motivation: Reduce the number of variables so that we can run interesting algorithms
- The goal is to build linear combinations of the variables (transformation vectors)
- First component should represent the direction with largest variance
- Second component is orthogonal to (independent of) the first, and is the next one with largest variance
- and so on…
X and Y are not independent
(covariance is not 0)

\[ Y = (X \times 4) + e \]

\[
\text{cov} = \begin{bmatrix}
\sigma_{XX} & \sigma_{XY} \\
\sigma_{XY} & \sigma_{YY}
\end{bmatrix}
\]

\[
\text{cov} = \begin{bmatrix}
0.86 & 0.35 \\
0.35 & 15.69
\end{bmatrix}
\]

\[ \rho = 0.96 \]
Eigenvalues

\( \mathbf{I} \) is the identity matrix.
\( \mathbf{A} \) is a square matrix (such as the covariance matrix).

\[
|\mathbf{A} - \lambda \mathbf{I}| = 0
\]

\( \lambda \) is called the eigenvalue (or characteristic root) of \( \mathbf{A} \).

\[
\begin{bmatrix}
\sigma_{XX} & \sigma_{XY} \\
\sigma_{XY} & \sigma_{YY}
\end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0
\]
Eigenvectors

\[
\begin{bmatrix}
\sigma_{XX} & \sigma_{XY} \\
\sigma_{XY} & \sigma_{YY}
\end{bmatrix}
\begin{bmatrix}
a \\
b
\end{bmatrix} = q
\begin{bmatrix}
a \\
b
\end{bmatrix}
\]

\[
\begin{bmatrix}
\sigma_{XX} & \sigma_{XY} \\
\sigma_{XY} & \sigma_{YY}
\end{bmatrix}
\begin{bmatrix}
c \\
d
\end{bmatrix} = m
\begin{bmatrix}
c \\
d
\end{bmatrix}
\]

q and m are eigenvalues

\([a \ b]^T \text{ and } [c \ d]^T\) are eigenvectors, they are orthogonal (independent of each other, do not contain redundant information)

The eigenvector associated with the largest eigenvalue will point in the direction of largest variance in the data.

If q > m, then \([a \ b]^T\) is PC1
Principal Components

\[
\begin{bmatrix}
1.27 & 5.12 \\
5.12 & 21.65 \\
\end{bmatrix}
\begin{bmatrix}
0.23 \\
0.97 \\
\end{bmatrix}
= 22.87
\begin{bmatrix}
0.23 \\
0.97 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
1.27 & 5.12 \\
5.12 & 21.65 \\
\end{bmatrix}
\begin{bmatrix}
0.97 \\
-0.23 \\
\end{bmatrix}
= 0.05
\begin{bmatrix}
0.97 \\
-0.23 \\
\end{bmatrix}
\]

Total variance is \(21.65 + 1.27 = 22.92\)
Transformed data

\[ x = \begin{bmatrix} x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{32} \end{bmatrix}, \quad PC = \begin{bmatrix} a & c \\ b & d \end{bmatrix} \]

\[
\begin{bmatrix} a & b \\ c & d \end{bmatrix} x = \begin{bmatrix} ax_{11} + bx_{12} & ax_{21} + bx_{22} & ax_{31} + bx_{32} \\ cx_{11} + dx_{12} & cx_{21} + dx_{22} & cx_{31} + dx_{32} \end{bmatrix}
\]

\[
\begin{bmatrix} 0.23 & 0.97 \\ 0.97 & -0.23 \end{bmatrix} x = \begin{bmatrix} 0.23x_{11} + 0.97x_{12} & 0.23x_{21} + 0.97x_{22} & 0.23x_{31} + 0.97x_{32} \\ 0.97x_{11} - 0.23x_{12} & 0.97x_{21} - 0.23x_{22} & 0.97x_{31} - 0.23x_{32} \end{bmatrix}
\]

Total variance is 22.92

Variance of PC1 is 22.87, so it captures 99% of the variance.

PC2 can be discarded with little loss of information.
PC1 is not at the regression line

- $y=4x$
- $[a \ b]^T = [0.23 \ 0.97]$
- Transformation is $0.23x + 0.97y$
- PC1 goes thru $(0,0)$ and $(0.23,0.97)$
- Its slope is $-0.97/0.23 = 4.217$
PCA regression

- Reduce original dimensionality $n$ (number of variables) finding $n$ PCs, such that $n<d$
- Perform regression on PCs
- Problem: Direction of greater overall variance is not necessarily best for classification
- Solution: Consider also direction of greater separation between two classes
(not so good) idea: Class mean separation

- Find means of each category
- Draw the line that passes through the 2 means
- Project points on the line (a.k.a. orthogonalize points with respect to the line)
- Find point that best separates data
Fisher’s Linear Discriminant

• Use classes to define discrimination line, but criterion to maximize is:
  – ratio of (between classes variation) and (within classes variation)

• Project all objects into the line
• Find point in the line that best separates classes
$S_w$ is the sum of scatters within the classes

\[ \text{cov}_1 = \begin{bmatrix} 1 & 4 \\ 4 & 16 \end{bmatrix} = \frac{(x_i - \mu_1)(x_i - \mu_1)^T}{n-1} \]

\[ \text{cov}_2 = \begin{bmatrix} 1.1 & 3.9 \\ 3.9 & 14 \end{bmatrix} = \frac{(x_i - \mu_2)(x_i - \mu_2)^T}{n-1} \]

\[ S_w = \sum_{j=1}^{k} p_j \text{cov}_j(n - 1) \]

\[ S_1 = .5 \begin{bmatrix} 1 & 4 \\ 4 & 16 \end{bmatrix} (99) \]

\[ S_2 = .5 \begin{bmatrix} 1.1 & 3.9 \\ 3.9 & 14 \end{bmatrix} (99) \]

\[ S_w = S_1 + S_2 \]
$S_b$ is scatter *between* the classes

$$S_b = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

- Maximize $S_b/S_w$ (a square $d \times d$ matrix, where $d$ is the number of dimensions)
- Find maximum eigenvalue and respective eigenvector
- This is the direction that maximizes Fisher’s criterion
- Points are projected over this line
- Calculate distance from every projected point to projected class means and decide class corresponding to smaller distance
- Assumption: class distributions are normal
Classification Models

- Quadratic Discriminant Analysis
- Partial Least Squares
  - PCA uses $X$ to calculate directions of greater variation
  - PLS uses $X$ and $Y$ to calculate these directions
    - It is a variation of multiple linear regression

PCA maximizes $\text{Var}(X_\alpha)$,
PLS maximizes $\text{Corr}^2(y,X_\alpha)\text{Var}(X_\alpha)$

- Logistic Regression
PCA, PLS, Selection

• 3 data sets
  – Singh et al. (Cancer Cell, 2002: 52 cases of benign and malignant prostate tumors)
  – Bhattachajee et al. (PNAS, 2001: 186 cases of different types of lung cancer)
  – Golub et al. (Science, 1999: 72 cases of acute myeloblastic and lymphoblastic leukemia)

• PCA logistic regression
• PLS
• Forward selection logistic regression

• 5-fold cross-validation
Missing Values: 0.001% - 0.02%

Screenshots removed due to copyright reasons.
The graph shows the AUC (c-index) for different methods as a function of the number of factors or variables. The methods compared are PCA, PLS, and Fwd selection. The significance of the difference is indicated by p=0.001.
febbo
p=0.007
febbo

$p = 0.013$

**Diagram Description:**

- **Y-axis:** AUC (c-index)
- **X-axis:** Number of factors or variables
- **Legend:**
  - Red: PCA
  - Green: PLS
  - Black: Fwd selection

The graph shows the performance metrics (AUC) achieved by different methods (PCA, PLS, Fwd selection) as the number of factors or variables increases. The AUC values are indicated by lines on the graph, with different colors representing each method. The p-value of 0.013 suggests statistical significance between the methods.
Classification of Leukemia with Gene Expression

Variable selection from ~2,300 genes