Harvard-MIT Division of Health Sciences and Technology
HST.951J: Medical Decision Support, Fall 2005
Instructors: Professor Lucila Ohno-Machado and Professor Staal Vinterbo

6.873/HST.951 Medical Decision Support

Spring 2005

Variable Compression: Principal Components Analysis Linear Discriminant Analysis

Lucila Ohno-Machado

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

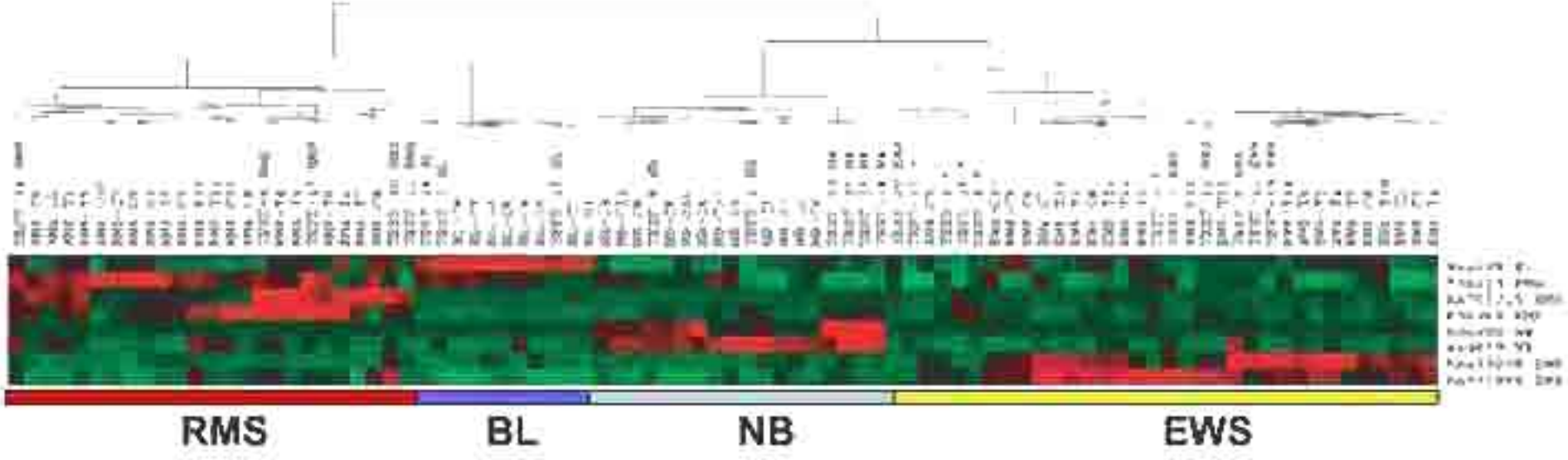
Screenshots removed due to copyright reasons.

Figures removed due to copyright reasons.
Please see:
Khan, J., et al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." *Nat Med* 7, no. 6 (Jun 2001): 673-9.

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

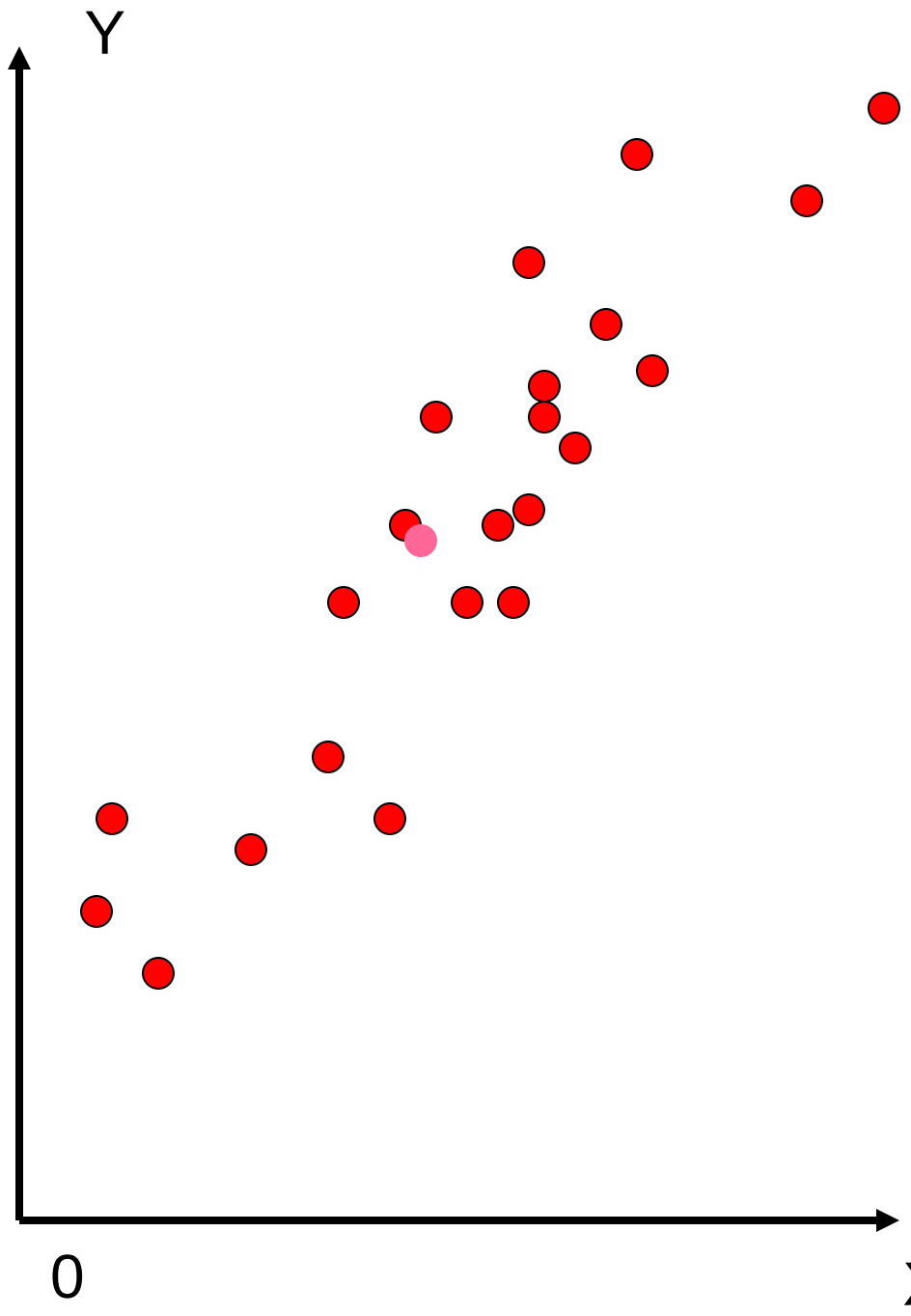
$$\sigma_{XX} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})}{n-1}$$

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

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

$$corr = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

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

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

Slope from linear regression is asymmetric, covariance and ρ are symmetric

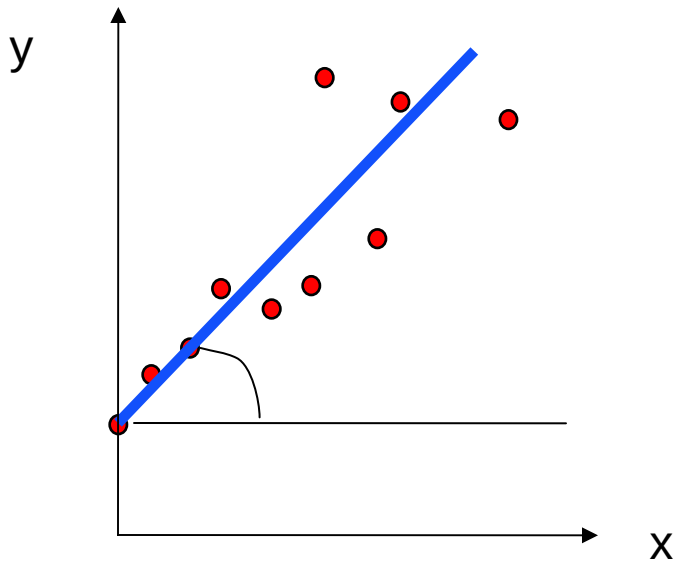
$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

$$y = \beta_0 + \beta_1 x$$

$$y = 2 + 4x$$

$$x = y/4 - 2$$

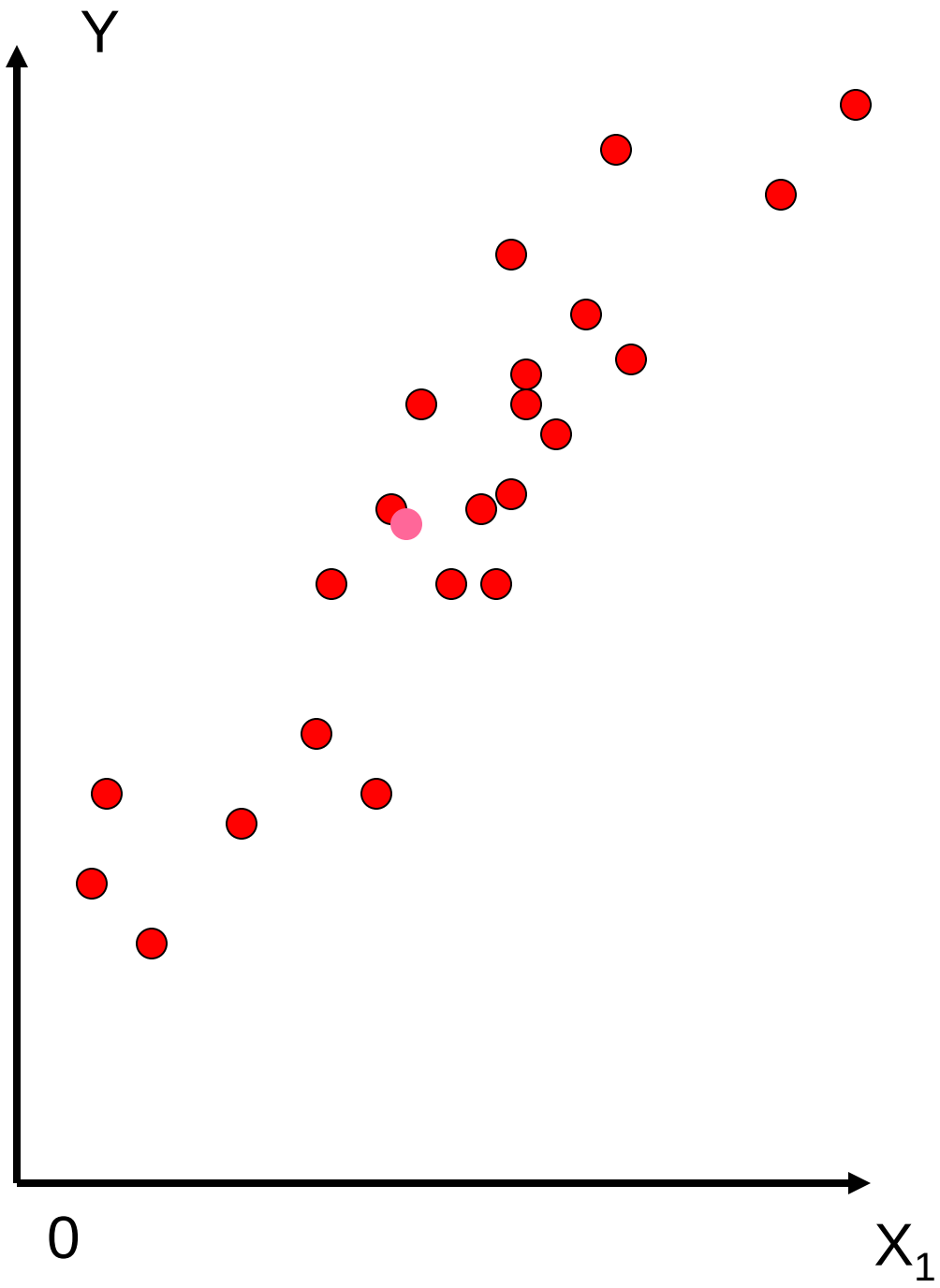


$$\text{cov} = \begin{bmatrix} 0.86 & 0.35 \\ 0.35 & 15.69 \end{bmatrix} = \Sigma$$

$$corr = \begin{bmatrix} 1 & 0.96 \\ 0.96 & 1 \end{bmatrix}$$

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)

$$Y = (X * 4) + e$$

$$\text{cov} = \begin{bmatrix} \sigma_{XX} & \sigma_{XY} \neq 0 \\ \sigma_{XY} \neq 0 & \sigma_{YY} \end{bmatrix}$$

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

$$\rho = 0.96$$

Eigenvalues

I is the identity matrix.

A is a square matrix (such as the covariance matrix).

$|\mathbf{A} - \lambda\mathbf{I}| = 0$

λ is called the *eigenvalue* (or *characteristic root*) of **A.**

$$\left| \begin{bmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{XY} & \sigma_{YY} \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

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

$[a\ b]^T$ 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.

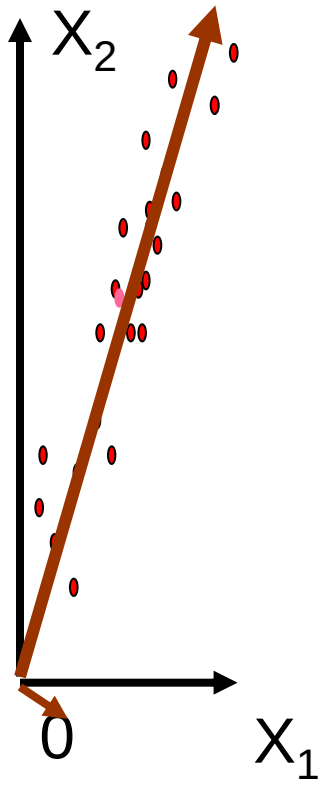
If q > m, then $[a\ b]^T$ is PC1

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



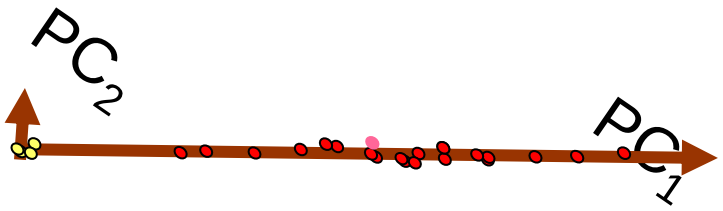
Transformed data

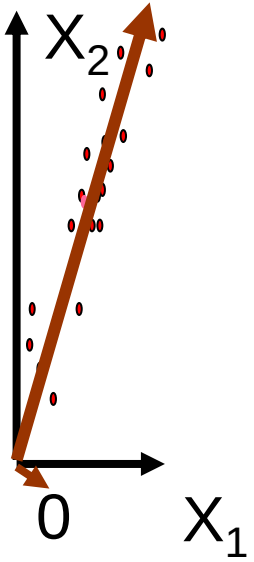
O1 O2 O3

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

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

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





Total variance is 22.92

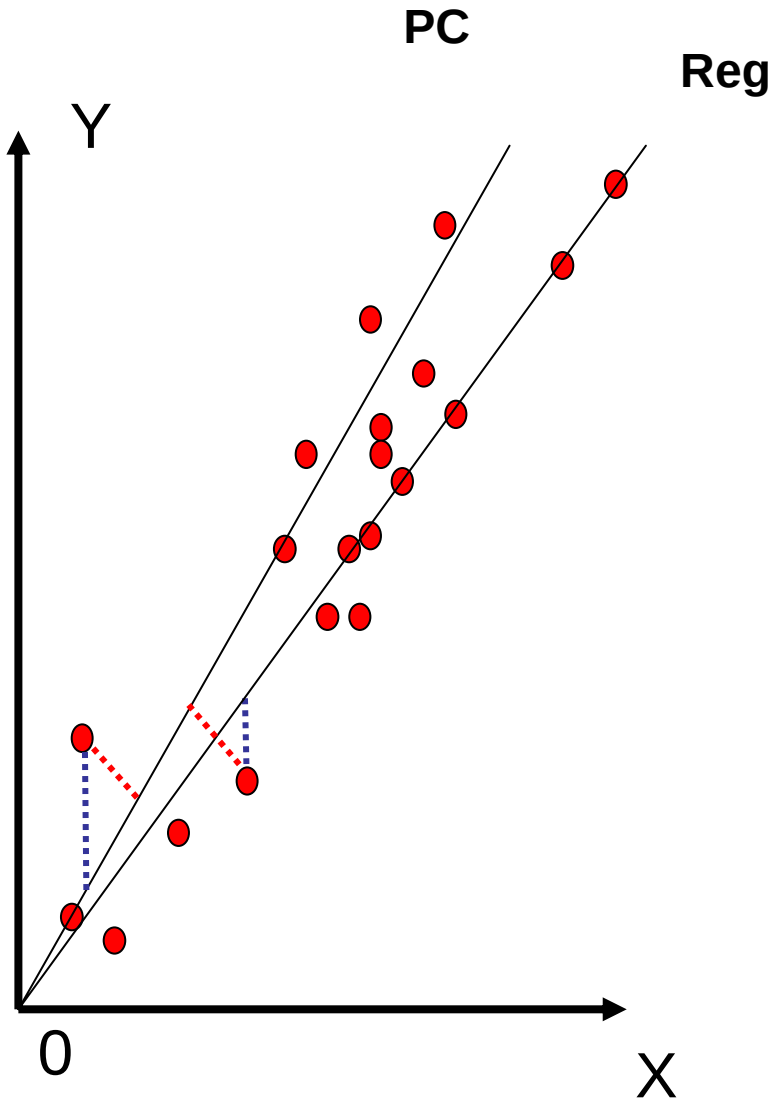
Variance of PC1 is 22.87, so it captures 99% of the variance.

PC2 can be discarded with little loss of information.

PC1 is not at the regression line

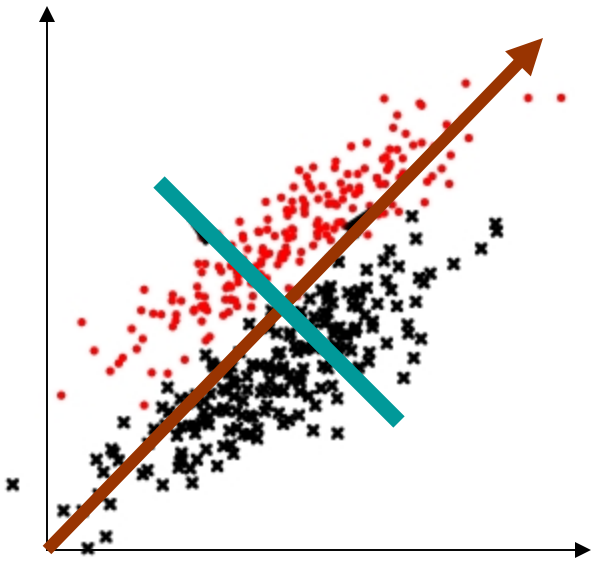
- y=4x
- $[a\ b]^T = [0.23\ 0.97]$
- Transformation is 0.23x+0.97y
- PC1 goes thru

  (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

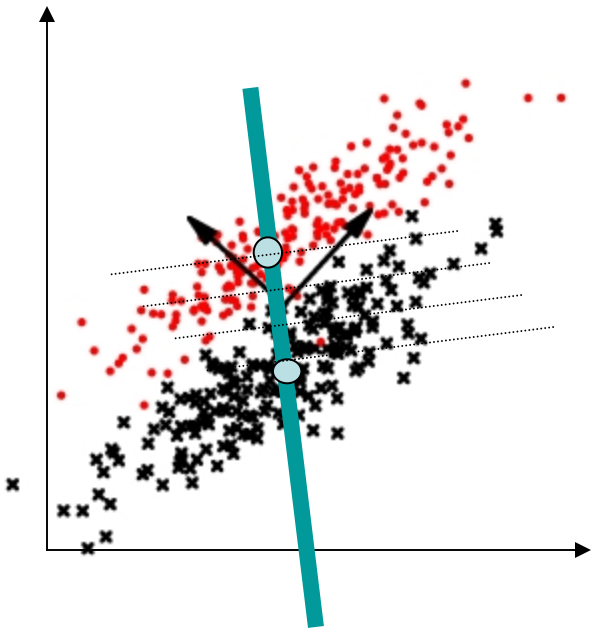


(not so good) idea: Class mean separation

- 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

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

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

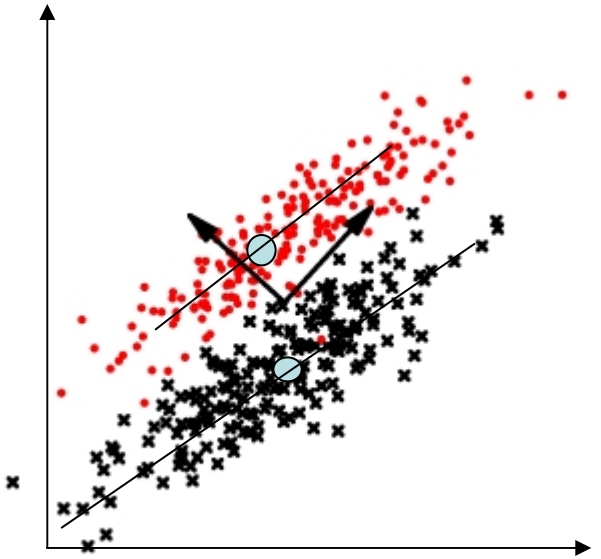
scatter

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

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

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

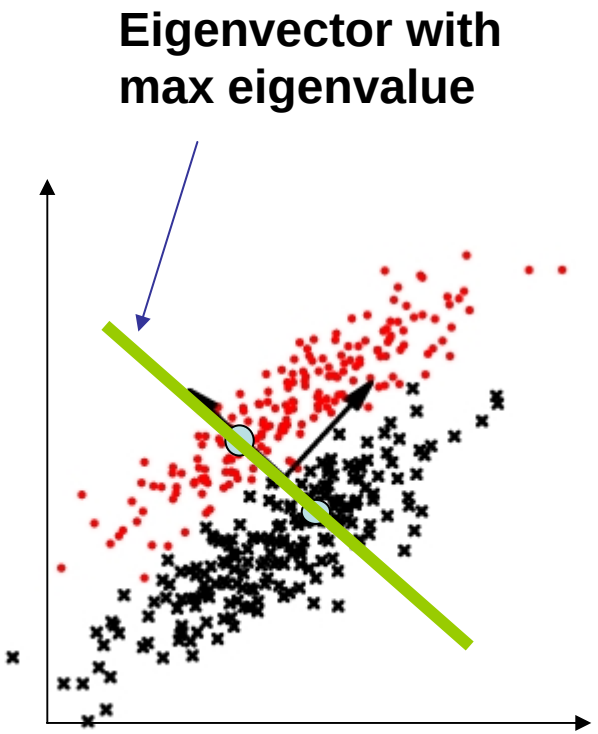
$$S_w = S_1 + S_2$$



S_b is scatter *between* the classes

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

- Maximize Sb/Sw (a square d x 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 $\mathrm{Var}(X\alpha)$,

PLS maximizes $\mathrm{Corr}^2(y,X\alpha)\mathrm{Var}(X\alpha)$

- Logistic Regression

PCA, PLS, Selection

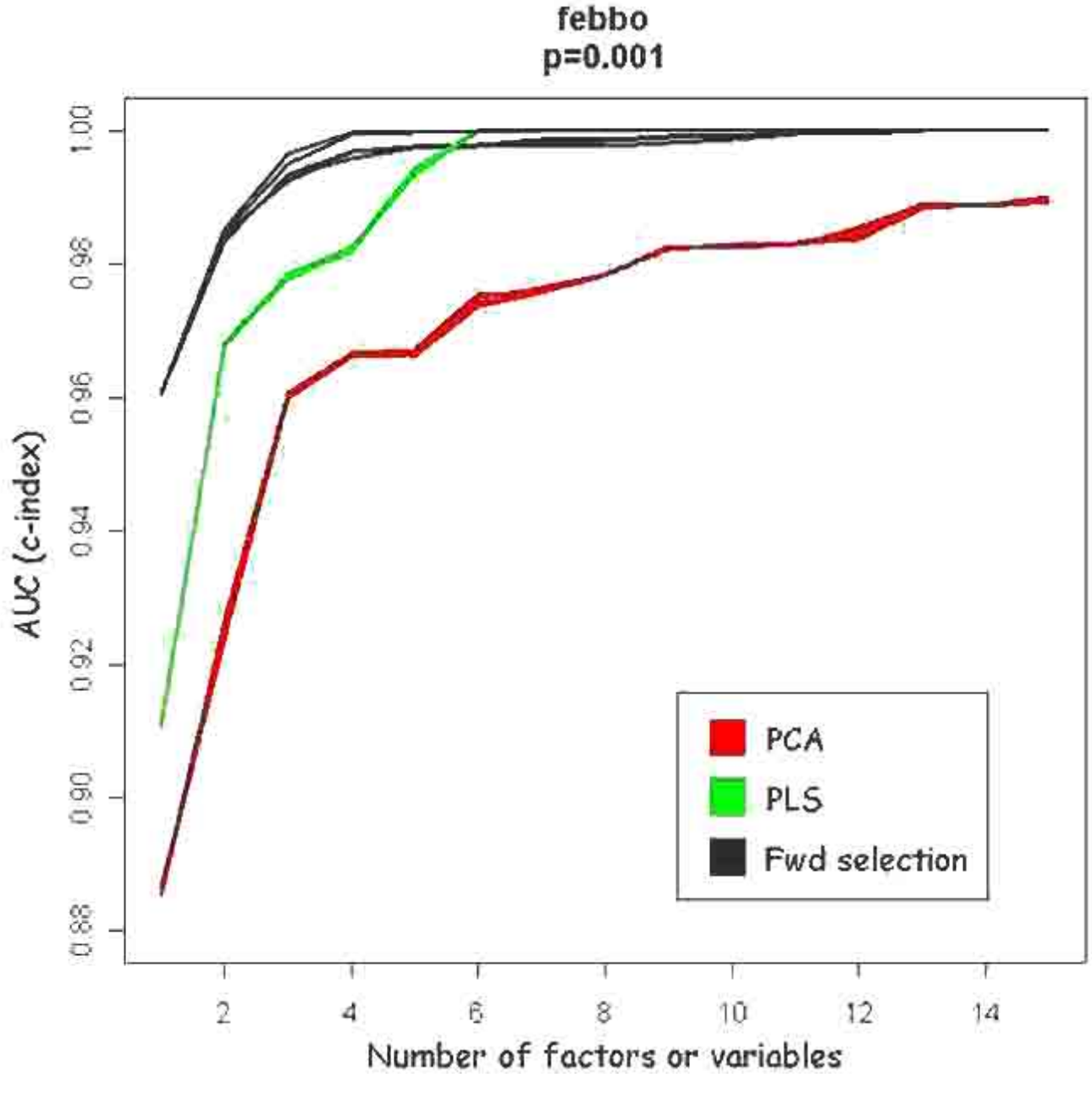
- 3 data sets
  - Singh et al. (*Cancer Cell*, 2002: 52 cases of benign and malignant prostate tumors)
  - Bhattachajee et al. (PNAS, 2001: 186 cases of different types of lung cancer)
  - Golub et al. (Science, 1999: 72 cases of acute myeloblastic and lymphoblastic leukemia)

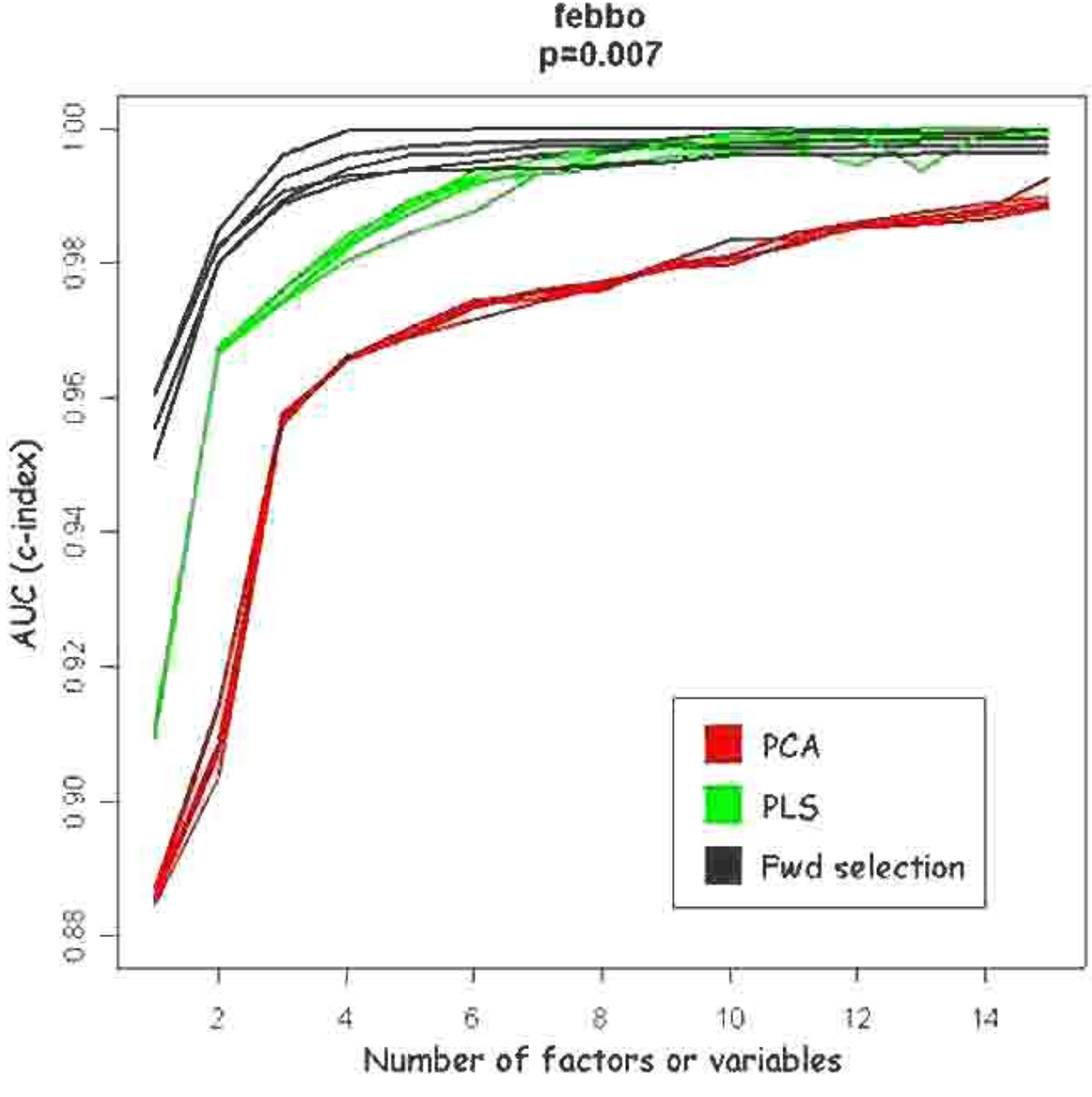
- PCA logistic regression
- PLS
- Forward selection logistic regression

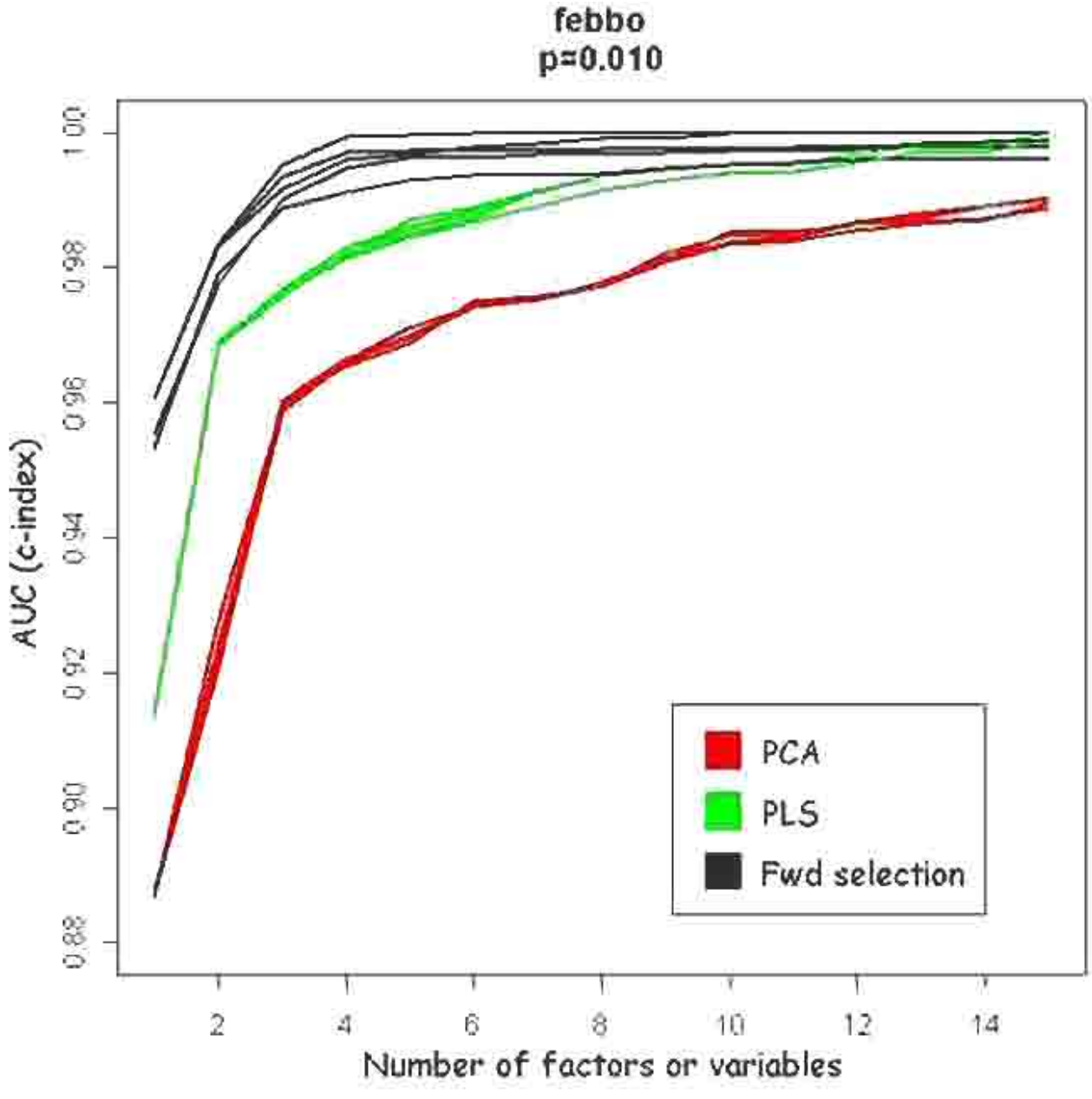
- 5-fold cross-validation

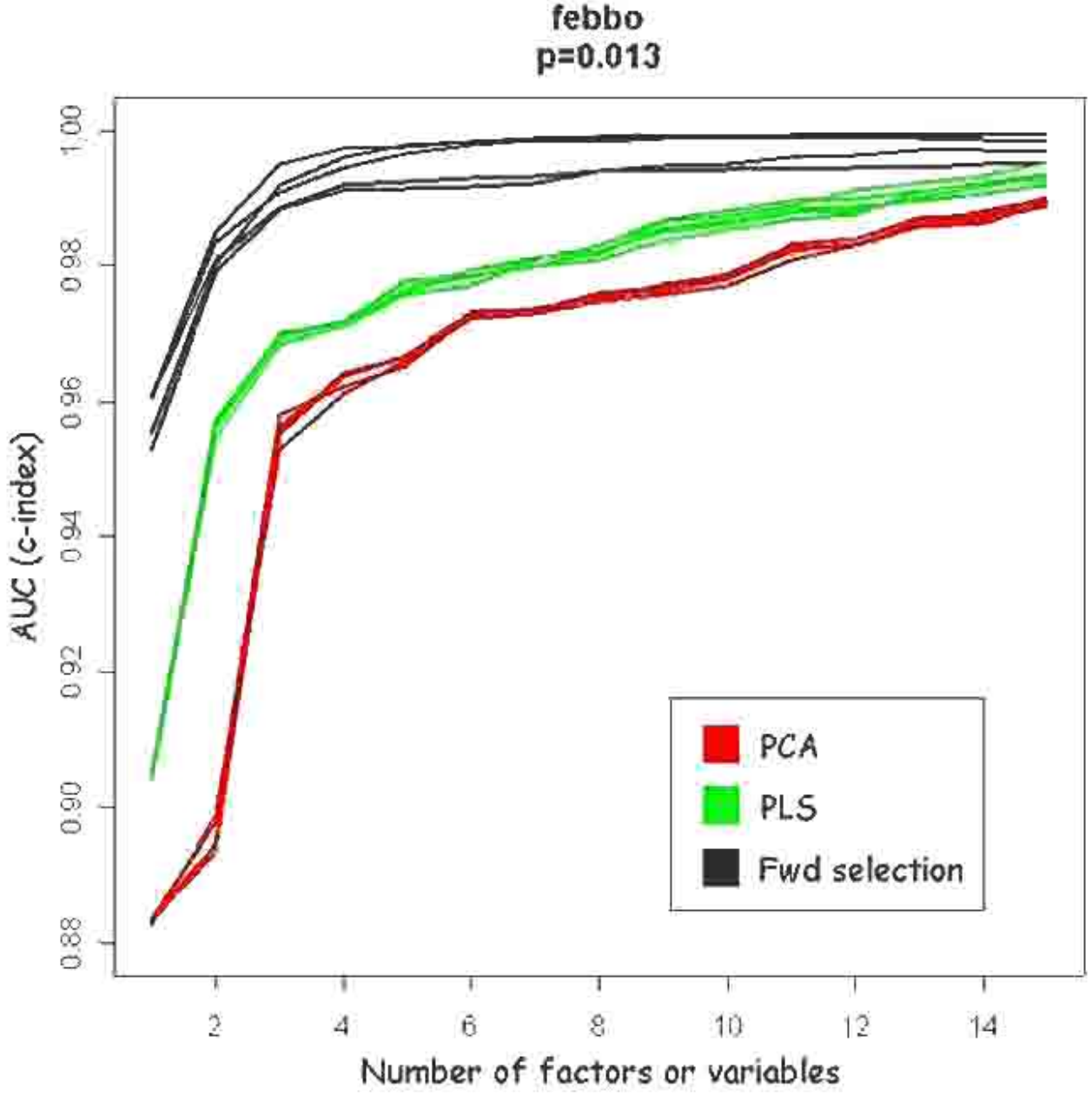
Missing Values: 0.001% - 0.02%

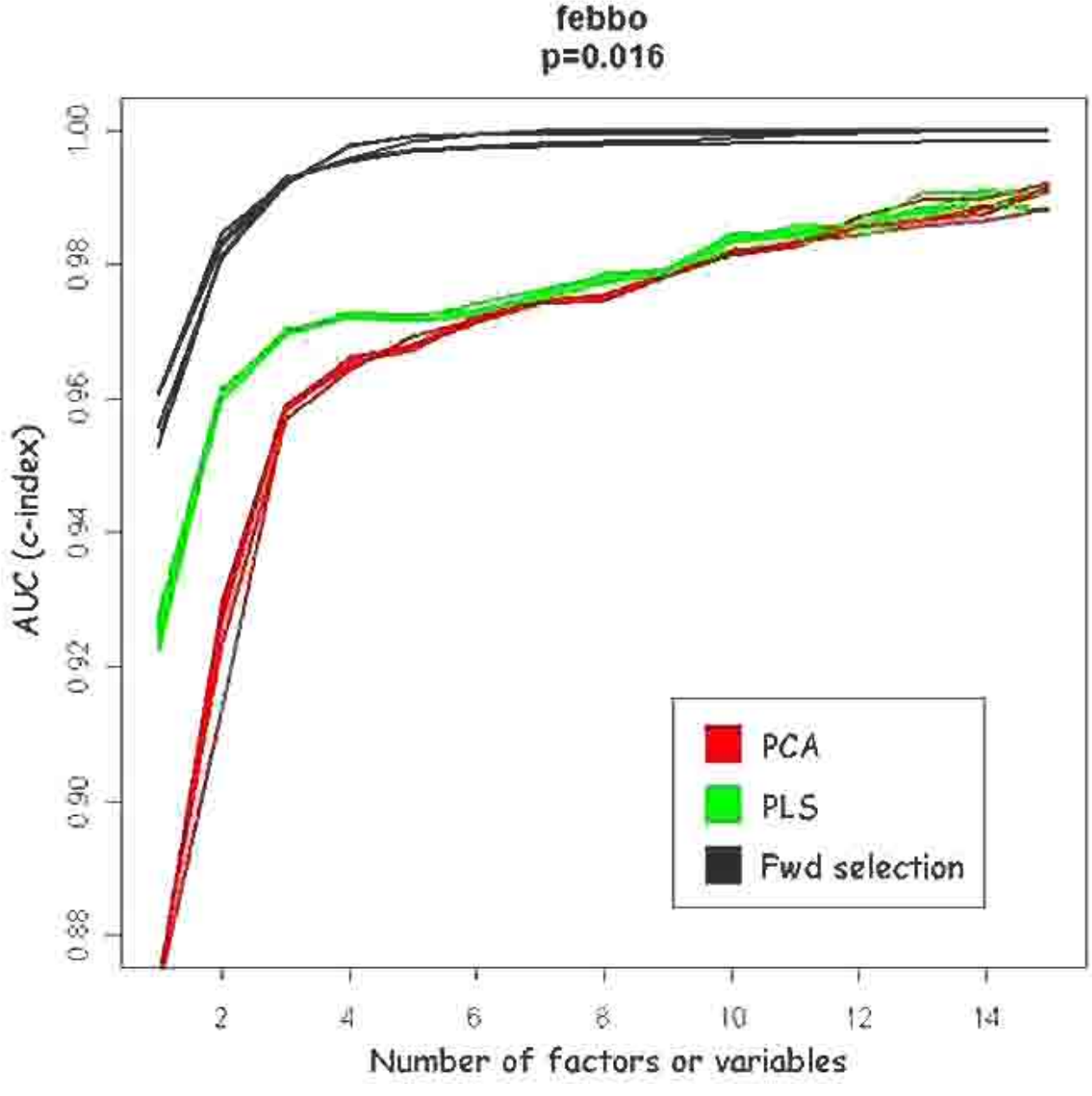
Screenshots removed due to copyright reasons.


febbo
p=0.001
AUC (c-index)
0.88
0.90
0.92
0.94
0.96
0.98
1.00
2
4
6
8
10
12
14
Number of factors or variables
PCA
PLS
Fwd selection


febbo
p=0.007
AUC (c-index)
0.88
0.90
0.92
0.94
0.96
0.98
1.00
2
4
6
8
10
12
14
Number of factors or variables
PCA
PLS
Fwd selection


febbo
p=0.010
AUC (c-index)
0.88
0.90
0.92
0.94
0.96
0.98
1.00
2
4
6
8
10
12
14
Number of factors or variables
PCA
PLS
Fwd selection


febbo
p=0.013
AUC (c-index)
0.88
0.90
0.92
0.94
0.96
0.98
1.00
2
4
6
8
10
12
14
Number of factors or variables
PCA
PLS
Fwd selection

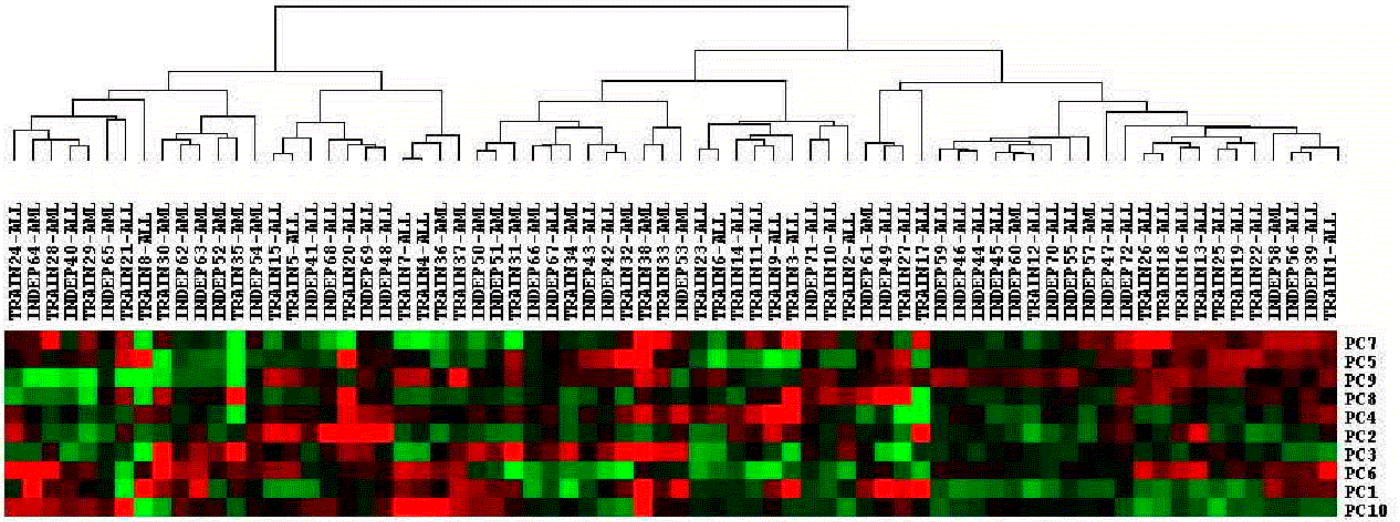

febbo
p=0.016
AUC (c-index)
0.88
0.90
0.92
0.94
0.96
0.98
1.00
2
4
6
8
10
12
14
Number of factors or variables
PCA
PLS
Fwd selection

PC6
PC1
PC10
PC7
PC2
PC3
PC8
PC5
PC4
PC9

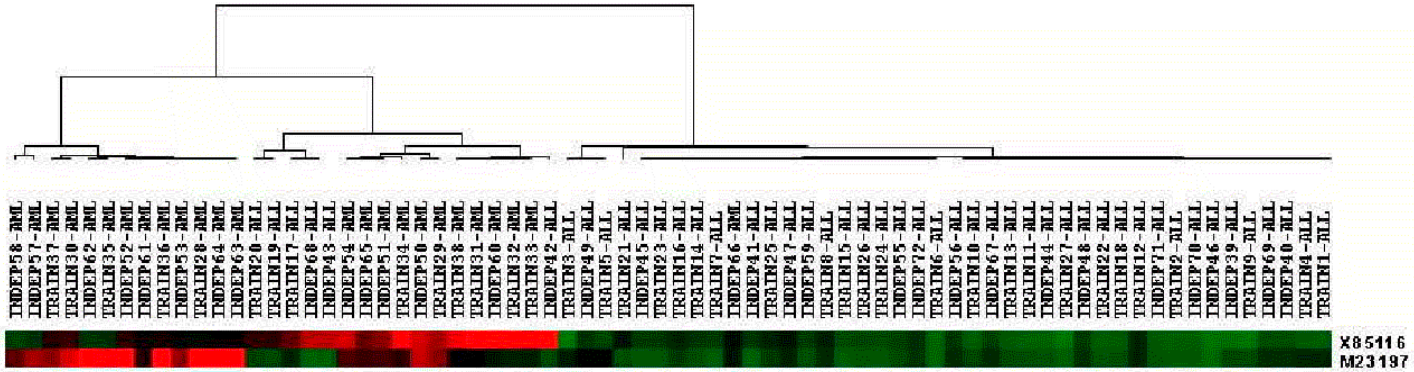
BL-NB7018
RMS-T54213
RMS-AA705225
RMS-R36980
NB-H06192
NB-W69818
EWS-AA458288
EWS-AA917895

Classification of Leukemia with Gene Expression

PCA



Variable Selection



Variable selection from ~2,300 genes