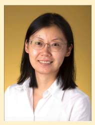
BIG DATA ANALYTICS

Dimension Reduction and Variable Selection



Sherry Ni, Ph.D.
Professor in Statistics and Interim Chair
Department of Statistics and Analytical Sciences
Kennesaw State University

1

1

Unsupervised Dimension Reduction

- 1.1 Introduction
- 1.2 Principal component analysis
- 1.3 Variable Clustering

2

2

Unsupervised Dimension Reduction

- 1.1 Introduction
- 1.2 Principal component analysis
- 1.3 Variable Clustering

3

3

Introduction Objectives

- Discuss reasons for variable reduction.
- Describe unsupervised versus supervised methods.
- Describe variable selection versus dimension reduction methods.

4

4

Huge Amounts of Data …

- A curse or a blessing?

5

5

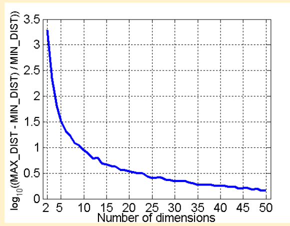
Problems with Many Variables

- Correlation
- Overfitting
- Sparseness

6

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful



- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

7

Feature Subset Selection

- One way to reduce dimensionality of data
- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA

8

8

Basic Variable Reduction Techniques

- Regression: forward, backward, stepwise selection
- Decision tree
- Variable Selection node

9

9

An Architecture for Feature Subset Selection

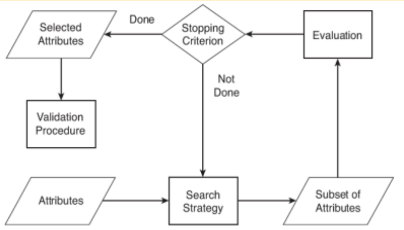


Figure 2.11. Flowchart of a feature subset selection process.

10

10

Variable Reduction: Target Used?

- Some variable reduction methods use the target variable ⇨ Supervised
- Some variable reduction methods ignore the target variable ⇨ Unsupervised

11

11

Variable Reduction: Output Variables?

- Some variable reduction methods use the original variables as inputs into subsequent models ⇨ **Variable Selection**
- Some variable reduction methods use combinations of the original variables as inputs into subsequent models ⇨ **Dimension Reduction**

12

12

Comparison of Methods

Method	Target?	Outputs?	Theoretical Basis
PCA	Not Used	Constructed	Interval Inputs
Variable Clustering	Not Used	Original or Constructed	Interval Inputs
PLS	Used	Original	Interval Inputs
LAR/LASSO	Used	Original	Interval Inputs
Logits	Used	Original or Constructed	Nominal Inputs
SWOE	Used	Constructed	Nominal Inputs

13

Our Class

Method	Target?	Outputs?	Theoretical Basis
PCA	Not Used	Constructed	Interval Inputs
Variable Clustering	Not Used	Original or Constructed	Interval Inputs
PLS	Used	Original	Interval Inputs
LAR/LASSO	Used	Original	Interval Inputs
Logits	Used	Original or Constructed	Nominal Inputs
SWOE	Used	Constructed	Nominal Inputs

14

Unsupervised Dimension Reduction

1.1 Introduction

1.2 Principal Component Analysis

1.3 Variable Clustering

15

15

Objectives

- Describe principal components analysis.
- Explain how to use principal components analysis in SAS.
- Discuss advantages and disadvantages with this variable selection method.

16

16

Principal Component Analysis

- Target used or not?
  - Not used
- Original or constructed variables as output?
  - Constructed variables

17

17

Principal Component Analysis: Main Features

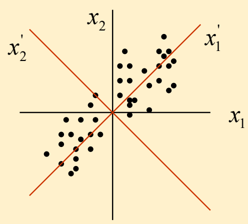
- Principal components are constructed as mathematical transformations of the input variables.
- The first principal component is constructed in such a way that it captures as much of the variation in the input variables (the X-space) set as possible.
- The second principal component is orthogonal to the first principal component.
- The second principal component captures as much as possible of the variation in the input data not captured by the first principal component.
- And so on ...

18

18

Dimension Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data



19

19

Input and Output Variables

– Input variables:
x_1, x_2, x_3

– Principal component 1:
$pc_1 = a_1x_1 + b_1x_2 + c_1x_3$

– Principal component 2:
$pc_2 = a_2x_1 + b_2x_2 + c_2x_3$

– Principal component 3:
$pc_3 = a_3x_1 + b_3x_2 + c_3x_3$

20

20

Mathematical Background

- Let $\mathbf{x}=(x_1,\ldots,x_p)^T$ be a random vector with mean m and covariance matrix V.
- First principal component:

$a_1 = \arg\max \mathrm{Var}(a'x) = \arg\max[a'Va]$ subject to $||a|| = 1$

Introduce the Lagrange multiplier λ:

$$\frac{\partial}{\partial a_i}\{a^T V a + \lambda(1 - a^T a)\} = 0 \text{ for } i = 1, \ldots, p$$
$$\Longrightarrow Va = \lambda a$$

This implies that

- λ is an eigenvalue of V
- **a** is the corresponding eigenvector

21

21

Mathematical Background

- Second principal component:

$$a_2 = \arg\max \mathrm{Var}(a'x) = \arg\max[a'Va]$$

subject to $a_1'a_2 = 0$ and $||a_2|| = 1$.

Introduce the Lagrange multiplier λ_1 and λ_2:

$$\frac{\partial}{\partial a_i}\{a^T V a + \lambda_1(1 - a^T a) + \lambda_2 a_1' a_2\} = 0 \text{ for } i = 1, ..., p$$

$$\Longrightarrow V a_2 = \lambda a_2, \quad a_1' a_2 = 0 \quad \text{and } \lambda_2 = 0$$

- Third principal component, …

22

22

Review: Linear Algebra

- Since V is symmetric, its eigenvalues (solutions of the polynomial equation det(V- λ I)=0) are real and can be ordered as λ_1, … , λ_p.
  They are all nonnegative since V is nonnegative definite.
  Moreover,
  $$tr(V) = \lambda_1 + ... + \lambda_p, \quad det(V) = \lambda_1 ... \lambda_p.$$
  Let a_j be the eigenvector corresponding to λ_j, then the eigenvectors are orthogonal to each other.
- Definition:
  - $a_i^T x$ is called the i-th principal component of x.

23

23

Review: Linear Algebra

- Basic Facts:

(a) $V = \lambda_1 a_1 a_1^T + ... + \lambda_p a_p a_p^T, \quad I = a_1 a_1^T + ... + a_p a_p^T$

(b) $\sum_{i=1}^{p} Var(x_i) = tr(V) = \lambda_1 + ... + \lambda_p.$

(c) $Var(a_i^T x) = \lambda_i$

(d) We hope that only a few principal components account for most of the overall variance.
$(\sum_{i=1}^{k} \lambda_i)/tr(V)$ is near 1 for some small k.

(e) Factor loadings are columns giving the elements of the column vectors a_i for the principal components $a_i^T x$.

24

24

Review: Linear Algebra

$(a)\ V = \lambda_1 \mathrm{a}_1 \mathrm{a}_1^{\mathrm{T}} + ... + \lambda_p \mathrm{a_p} \mathrm{a_p^T}, \quad I = \mathrm{a}_1 \mathrm{a}_1^{\mathrm{T}} + ... + \mathrm{a_p a_p^T}$

Proof:

$$V \cdot (\mathrm{a}_1 ... \mathrm{a}_p) = (\mathrm{a}_1 ... \mathrm{a}_p) \begin{pmatrix} \lambda_1 & & \\ & ... & \\ & & \lambda_p \end{pmatrix}$$

$$\Longrightarrow\ V = (\mathrm{a}_1 ... \mathrm{a}_p) \begin{pmatrix} \lambda_1 & & \\ & ... & \\ & & \lambda_p \end{pmatrix} \begin{pmatrix} \mathrm{a}_1^T \\ ... \\ \mathrm{a}_p^T \end{pmatrix}$$

$$\Longrightarrow\ V = (\mathrm{a}_1 \lambda_1 ... \mathrm{a}_p \lambda_p) \begin{pmatrix} \mathrm{a}_1^T \\ ... \\ \mathrm{a}_p^T \end{pmatrix} = \sum_{i=1}^{p} \mathrm{a}_i \lambda_i \mathrm{a}_i^T$$

25

25

Implementation

- Suppose $x_1,\dots,x_n$ are a sample of n independent observations from a multivariate population with mean m and covariance matrix V.

$$\hat{\mu} = \bar{x} = \sum_{i=1}^{n} x_i / n, \qquad \hat{V} = X^T X / (n-1)$$

$$X = \begin{pmatrix} x_{11} - \bar{x}_1 & ... & x_{1p} - \bar{x}_p \\ ... & ... & ... \\ x_{n1} - \bar{x}_1 & ... & x_{np} - \bar{x}_p \end{pmatrix} = (\mathbf{X}_1, ..., \mathbf{X}_p)$$

- The jth principal component of $\mathbf{X}_1,\dots,\mathbf{X}_p$ is the linear combination

$$\mathbf{Y}_j = \hat{a}_{1j} \mathbf{X}_1 + ... + \hat{a}_{pj} \mathbf{X}_p$$

where $\hat{\mathrm{a}}_j = (\hat{a}_{1j}, ..., \hat{a}_{pj})^T$ is the eigenvector corresponding to the jth largest eigenvalue $\hat{\lambda}_j$ of the sample covariance matrix $\hat{V}$.

26

26

Summary

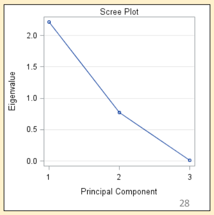
- With *p* input variables, you can compute *p* principal components.
- Each of the principal components is an uncorrelated, linear combination of all original input variables.
- The coefficients of such a linear combination are the eigenvectors of the correlation or covariance matrix.
- The principal components are sorted by descending order of the eigenvalues.
- The eigenvalues represent the variances of the principal components.

27

27

Selection of the Number of Principal Components

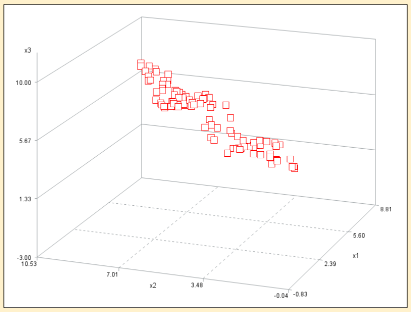
- The number of principal components used as input variables for the successor modeling nodes can be selected using one of the following:
  - Proportion of variance explained
  - Scree plot
  - Eigenvalue > 1



28

28

Example 1



29

29

Example 1

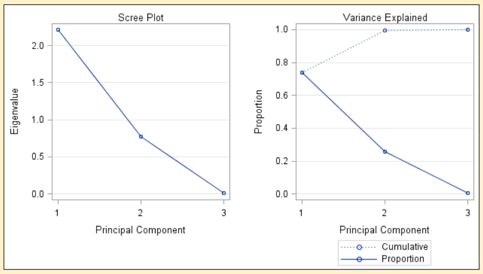
Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.21358154	1.44403082	0.7379	0.7379
2	0.76955072	0.75268297	0.2565	0.9944
3	0.01686775		0.0056	1.0000

Eigenvectors			
	Prin1	Prin2	Prin3
x1	0.650940	0.263685	0.711862
x2	0.645235	0.301851	-.701825
x3	-.399937	0.916164	0.026348

30

30

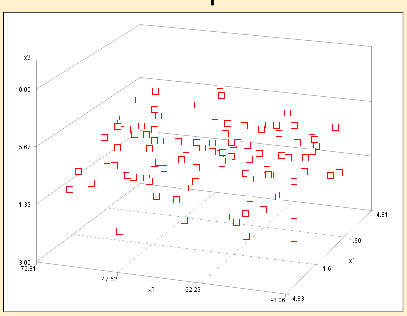
Example 1



31

31

Example 2



32

32

Example 2

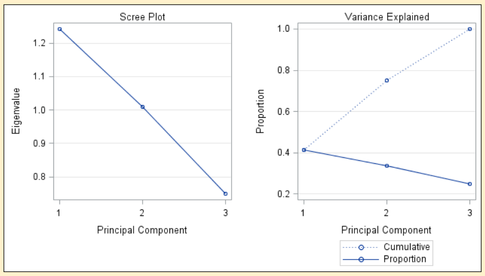
	Eigenvalues of the Correlation Matrix			
	Eigenvalue	Difference	Proportion	Cumulative
1	1.24205697	0.23274599	0.4140	0.4140
2	1.00931098	0.26067894	0.3364	0.7505
3	0.74863205		0.2495	1.0000

	Eigenvectors		
	Prin1	Prin2	Prin3
x1	0.704619	-.156546	0.692102
x2	0.708942	0.113759	-.696032
x3	0.030228	0.981097	0.191139

33

33

Example 2



34

34

Example 3 – Step 1 Get Some Data

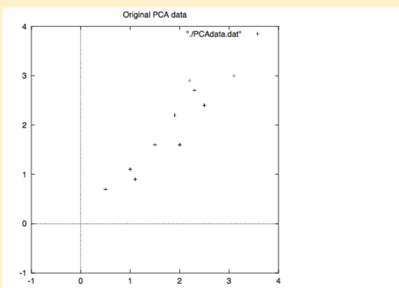


Figure 3.1: PCA example data, original data on the left, data with the means subtracted on the right, and a plot of the data

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

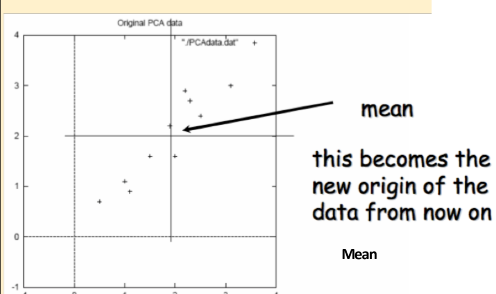
35

35

Example 3 – Step 2 Subtract the Mean

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01



36

36

Example 3 – Step 3 Calculate the Covariance Matrix V

$$\text{cov} = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

37

37

Example 3 – Step 4 Calculate the Eigenvalues and Eigenvectors of the Covariance Matrix V

$$\text{eigenvalues} = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$\text{eigenvectors} = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -735178656 \end{pmatrix}$$

Question: Proportion of variance explained by the first PC?

38

38

Example 3 – Step 4 Calculate the Eigenvalues and Eigenvectors of the Covariance Matrix V

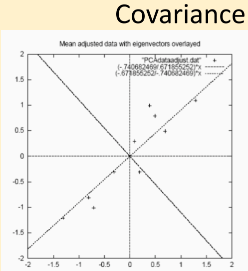


Figure 3.2: A plot of the normalised data (mean subtracted) with the eigenvectors of the covariance matrix overlayed on top.

- eigenvectors are plotted as diagonal dotted lines on the plot.
- Note they are perpendicular to each other.
- Note one of the eigenvectors goes through the middle of the points, like drawing a line of best fit.
- The second eigenvector gives us the other, less important, pattern in the data, that all the points follow the main line, but are off to the side of the main line by some amount.

39

39

Example 3 – Step 5 Choose Components

FeatureVector = (eig_1 eig_2 eig_3 … eig_n)
We can either form a feature vector with both of the eigenvectors:

$$\begin{pmatrix} -.677873399 & -.735178656 \\ -.735178656 & .677873399 \end{pmatrix}$$

or, we can choose to leave out the smaller, less significant component and only have a single column:

$$\begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix}$$

40

40

Example 3 – Step 6 Derive the New Data Coordinates

If choose to keep both components

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

$$\begin{pmatrix} -.677873399 & -.735178656 \\ -.735178656 & .677873399 \end{pmatrix}$$

PC1 = -0.677873399*x + (-0.7351786)*y

PC2 = -0.735178656*x + 0.677873399*y

41

41

Example 3 – Step 6 Derive the New Data Coordinates

If choose to keep one component only – dimension reduction

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

$$\begin{pmatrix} -.677873399 & \cancel{-.735178656} \\ -.735178656 & \cancel{.677873399} \end{pmatrix}$$

PC1 = -0.677873399*x + (-0.7351786)*y

PC2 = -0.735178656*x + 0.677873399*y

42

42

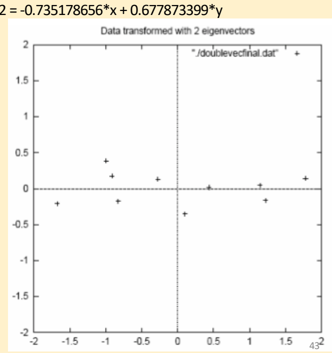
DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

PC1 = -0.677873399*x + (-0.7351786)*y

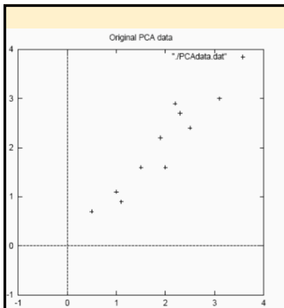
PC2 = -0.735178656*x + 0.677873399*y

PC1	PC2
-.827970186	-.175115307
1.77758033	.142857227
-.992197494	.384374989
-.274210416	.130417207
-1.67580142	-.209498461
-.912949103	.175282444
.0991094375	-.349824698
1.14457216	.0464172582
.438046137	.0177646297
1.22382056	-.162675287

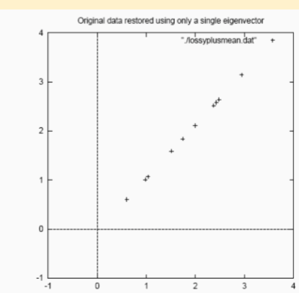


43

43



2D point cloud



Approximation using one eigenvector basis

44

44

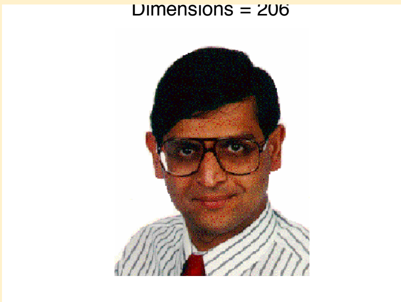
Discussion

- Do you have any experience with principal component analysis?
- Have you ever used PCA in an analysis?
- What do you think are the benefits? …the limitations?

45

45

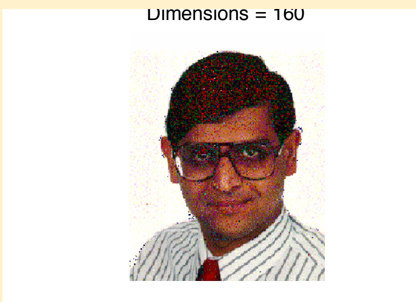
Dimension Reduction: PCA



46

46

Dimension Reduction: PCA



47

47

Dimension Reduction: PCA – Iris Data

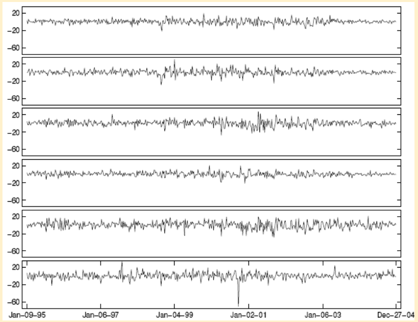
48

48

PCA: A Portfolio of Stocks

Weekly log returns:

JP Morgan Chase,
Lehman brothers,
Cisco system Inc.,
Microsoft Corp.,
Sun Microsystems Inc., and
Apple Computer Inc..



49

49

PCA: A Portfolio of Stocks

- Denote

$\mathbf{r}' = (\text{JPM, LEH, CSCO, MSFT, SUNW, AAPL}).$

- The sample mean and sample covariance matrix

$$\boldsymbol{\mu} = \begin{pmatrix} 0.2327 & 0.4786 & 0.4458 & 0.3996 & 0.3116 & 0.2140 \end{pmatrix},$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 26.6124 & 17.9980 & 13.5795 & 8.1036 & 14.1662 & 6.2569 \\ 17.9980 & 34.8167 & 18.0034 & 10.5591 & 20.1398 & 9.5653 \\ 13.5795 & 18.0034 & 43.5496 & 16.6841 & 34.5675 & 16.0576 \\ 8.1036 & 10.5591 & 16.6841 & 23.7764 & 18.6738 & 10.8830 \\ 14.1662 & 20.1398 & 34.5675 & 18.6738 & 63.4471 & 19.2954 \\ 6.2569 & 9.5653 & 16.0576 & 10.8830 & 19.2954 & 56.9298 \end{pmatrix},$$

50

50

PCA: A Portfolio of Stocks

- The sample correlation matrix

$$\boldsymbol{\rho} = \begin{pmatrix} 1.0000 & 0.5913 & 0.3989 & 0.3222 & 0.3448 & 0.1607 \\ 0.5913 & 1.0000 & 0.4623 & 0.3670 & 0.4285 & 0.2149 \\ 0.3989 & 0.4623 & 1.0000 & 0.5185 & 0.6576 & 0.3225 \\ 0.3222 & 0.3670 & 0.5185 & 1.0000 & 0.4808 & 0.2958 \\ 0.3448 & 0.4285 & 0.6576 & 0.4808 & 1.0000 & 0.3211 \\ 0.1607 & 0.2149 & 0.3225 & 0.2958 & 0.3211 & 1.0000 \end{pmatrix}$$

51

51

PCA: A Portfolio of Stocks

(a) Sample covariance matrix

Eigenvalue	128.8825	45.3876	30.8143	17.7629	14.1168	12.1679
Proportion	51.7326	18.2183	12.3687	7.1299	5.6664	4.8841
Eigenvector	0.2530	-0.2090	0.5420	-0.0909	-0.0044	0.7683
	0.3401	-0.2347	0.6151	-0.2163	0.0269	-0.6352
	0.4828	-0.1571	-0.1414	0.6570	0.5387	-0.0211
	0.2771	-0.0366	-0.0101	0.4697	-0.8362	-0.0432
	0.6054	-0.2099	-0.5385	-0.5368	-0.0882	0.0594
	0.3792	0.9117	0.1339	-0.0680	0.0449	-0.0208

52

52

PCA: A Portfolio of Stocks

$\mathbf{r}' =$(JPM, LEH, CSCO, MSFT, SUNW, AAPL).

78.50% of total variation

General movement of the market

Difference of industrial sectors

(b) Sample correlation matrix

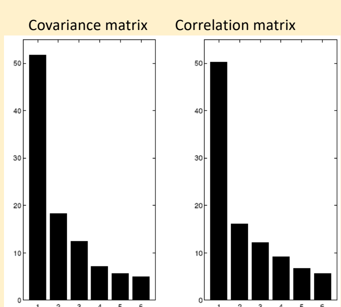
Eigenvalue	3.0130	0.9664	0.7303	0.5524	0.4009	0.3370
Proportion	50.2168	16.1060	12.1722	9.2060	6.6818	5.6171
Eigenvector	0.3831	-0.5729	0.2983	0.1051	0.6429	-0.1085
	0.4220	-0.4524	0.2330	-0.0475	-0.7475	0.0435
	0.4715	0.1357	-0.2985	-0.3034	0.1537	0.7446
	0.4094	0.2106	-0.3237	0.8207	-0.0541	-0.0819
	0.4525	0.1958	-0.3317	-0.4702	0.0255	-0.6520
	0.2835	0.6051	0.7435	-0.0015	0.0259	0.0023

53

53

PCA: A Portfolio of Stocks

- Fraction of total variation of each principal component



54

54

Principal Component Analysis: Pros

- Constructed output variables are definitely uncorrelated.
- The selection order of the principal components is automatically determined.
- The principal components are constructed in such a way that the first principal component represents more of the variation in the data cloud than the second one, and so on.
- Often, a very small number of principal components must be kept in order to explain a lot of the variation in the data cloud.

55

55

Principal Component Analysis: Cons

- It is difficult or impossible to interpret the constructed principal components.
- It is difficult to know how many principal components should be selected as new input variables.
- All original input variables are still used because they build the principal components.
- Misinterpretation of the coefficients of the linear combinations is common.

56

56

Auto Insurance Claim Demonstration

Analysis goal:

An automobile insurance company would like to predict if customers will have a future claim as well as the amount of the claim, if one is made. Historical claim information about customers is used to predict future claims.

57

...

57

Auto Insurance Claim Demonstration

For this demonstration:

- Dimension reduction and variable selection techniques will be used on input variables.
- The final predictive model will not be built in this demonstration.
- Some modeling best practices are ignored for instructional purposes.

58

58

Principal Components Analysis

This demonstration illustrates how a principal component analysis can be used for dimension reduction of the input space. Property settings are discussed, and the results of the node are interpreted.

59

59

Other Dimension Reduction Techniques

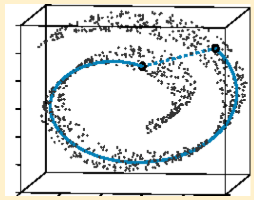
- Techniques
  - Principle Component Analysis (PCA)
  - Singular Value Decomposition (SVD)
  - Others: supervised and non-linear techniques
    - Factor analysis
    - Locally linear embedding (LLE)
    - Multidimensional scaling, FastMap, ISOMAP
- Details:
  - Appendix B in Tan et. Al.
  - STAT8320 Multivariate Data Analysis

60

60

Dimension Reduction: Swiss Roll

By: Tenenbaum, de Silva, Langford (2000)

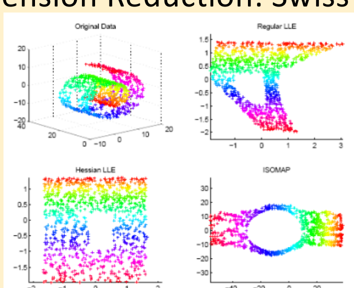


- ISOMAP:
  - Construct a neighbourhood graph
  - For each pair of points in the graph, compute the shortest path distances – **geodesic** distances

61

61

Dimension Reduction: Swiss Roll

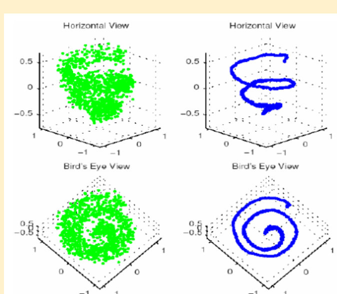


Donoho and Grimes (2003)

62

62

Dimension Reduction: LLP



Huo and Chen (2002)

63

63

Additional References

- *Advanced Predictive Modeling Using SAS® Enterprise Minter™ Course Notes.*
- Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (2005). *Introduction to Data Mining*.
- Robert T. Collins (2010) http://www.cse.psu.edu/~rtc12/CSE586Spring2010/lectures/pcaLectureShort.pdf
- Xiaoming Huo (2016) *Quantitative Financial Data Analysis Class Notes.*

64

64