Unsupervised Dimension Reduction

1.1 Introduction

1.2 Principal component analysis

1.3 Variable Clustering

1

1

Objectives

- Describe variable clustering.
- Explain how to use variable clustering in in SAS.
- Discuss advantages and disadvantages with this dimension reduction method.

2

2

Variable Clustering

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

3

3

Variable Clustering: Main Features

- The Variable Clustering node/procedure in SAS divides the input variables into hierarchical clusters.
- The main idea is to select one variable (or the cluster component) from each cluster as a cluster representative.
- The representative variables (or components) are used as input variables in successor nodes.
- The other input variables are rejected.

4

4

Variable Clustering

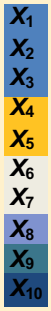
X_1
X_2
X_3
X_4
X_5
X_6
X_7
X_8
X_9
X_{10}

5

...

5

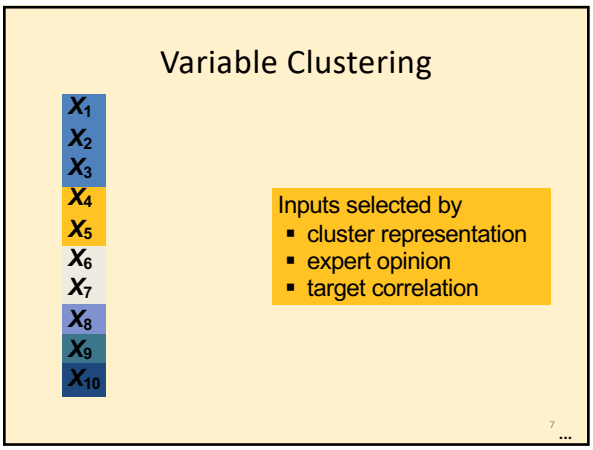
Variable Clustering



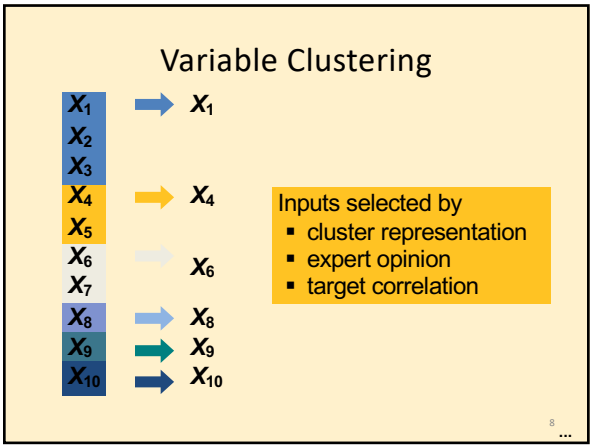
6

...

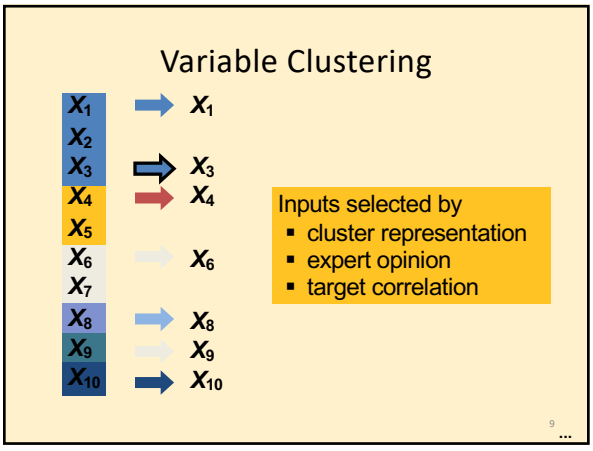
6


Variable Clustering
X_1
X_2
X_3
X_4
X_5
X_6
X_7
X_8
X_9
X_{10}
Inputs selected by
▪ cluster representation
▪ expert opinion
▪ target correlation
7

7


Variable Clustering
X_1
X_2
X_3
X_4
X_5
X_6
X_7
X_8
X_9
X_{10}
X_1
X_4
X_6
X_8
X_9
X_{10}
Inputs selected by
▪ cluster representation
▪ expert opinion
▪ target correlation
8

8


Variable Clustering
X_1
X_2
X_3
X_4
X_5
X_6
X_7
X_8
X_9
X_{10}
X_1
X_3
X_4
X_6
X_8
X_9
X_{10}
Inputs selected by
▪ cluster representation
▪ expert opinion
▪ target correlation
9

9

What Is a Cluster Component?

– Each cluster can be described as a linear combination of the variables in the cluster.

– This is **the first principal component** of the cluster.

– In this context, it is called the *cluster component*.

10

10

Variable Clustering Algorithm – a K-Mean Approach (Partitional)

- **Step 1** -- Set K (the number of groups)
- **Step 2** -- Choose randomly K variables as latent variable for each group. This variable is the first latent component of the group
- **Step 3**

DO WHILE no convergence
  FOR EACH variable
    Assign the variable to the closest latent component (r^2)
  END FOR
  Update the latent component for each group (the 1st PC)
END DO

11

11

Variable Clustering Algorithm – the SAS Approach (Hierarchical)

– The algorithm is divisive; at the start, all variables are in one single cluster.

– The following steps are repeated until convergence:

1. A cluster is chosen for splitting.
2. The chosen cluster is split into two clusters.
3. The variables are iteratively (re)assigned to the clusters.

12

12

Variable Clustering Algorithm – SAS Approach

– The following steps are repeated until convergence:

1. A cluster is chosen for splitting.
   1) Variation Proportion Property
   2) Maximum Eigen Value Property
2. The chosen cluster is split into two clusters.
   According to the first two principal components and their rotation.
3. The variables are iteratively (re)assigned to the clusters.
   1. Nearest component sorting phase
   2. Search phase (alternate assignment)

13

13

Variable Clustering Algorithm – SAS Approach

– Stop Criteria:

- The maximum number of clusters is reached, or
- Both of the following:
  – For each cluster, the % of variance explained > the pre-set value
  – All clusters have a 2nd eigegnvalue < the pre-set value (defacult is 1)

14

14

Variable Clustering Algorithm – SAS Approach

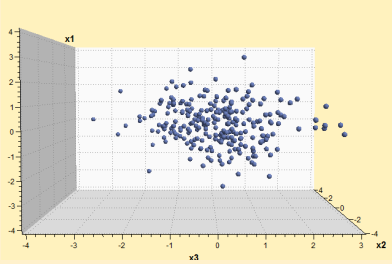
- Summary

VARCLUS (L variables)
  PCA with the L variables
  Rotation (QUARTIMAX) on the 2 first components
  IF (stopping criteria is not met) THEN
    Subdivision according to " r^2 " of the variables with the components (L1 and L2)
    VARCLUS (L1 variables)
    VARCLUS (L2 variables)
  END IF
RETURN

15

15

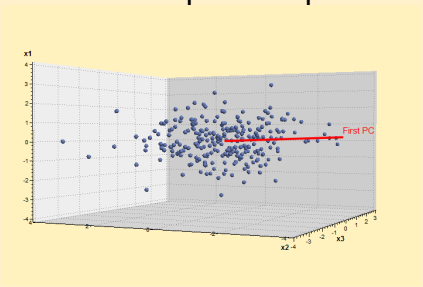
Example



16

16

First Principal Component

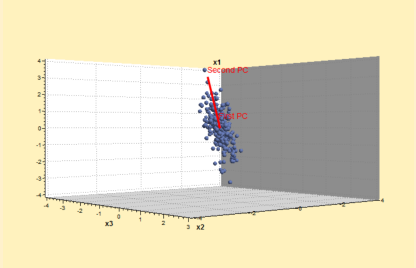


First Eigenvalue=1.94

17

17

Second Principal Component

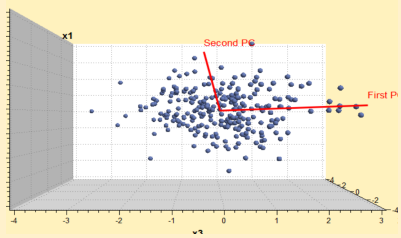


First Eigenvalue=1.94 Second Eigenvalue=1.02

18

18

Is This Cluster Split?

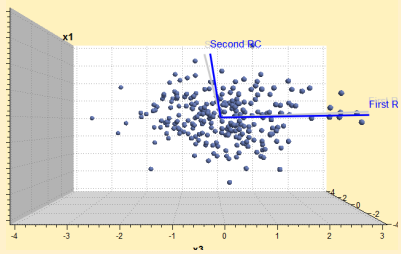


First Eigenvalue=1.94 Second Eigenvalue=1.02

19

19

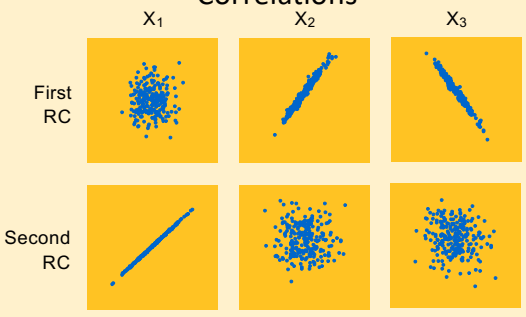
Rotated Components



20

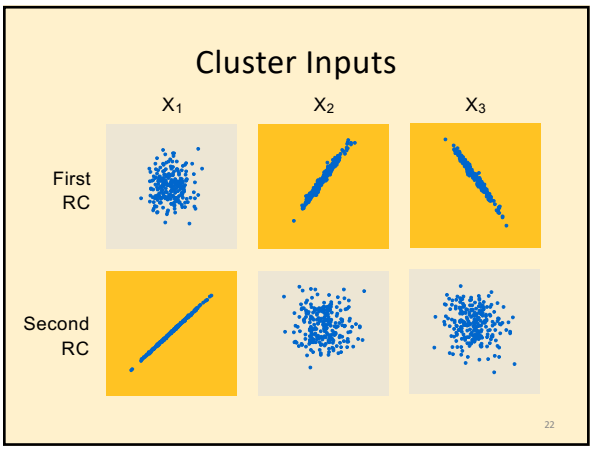
20

Input/Rotated Component Correlations

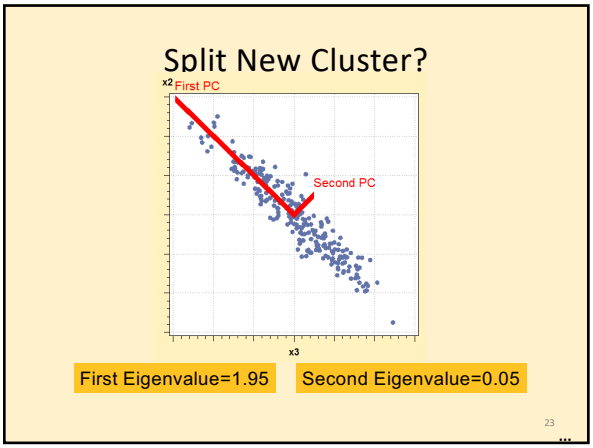


21

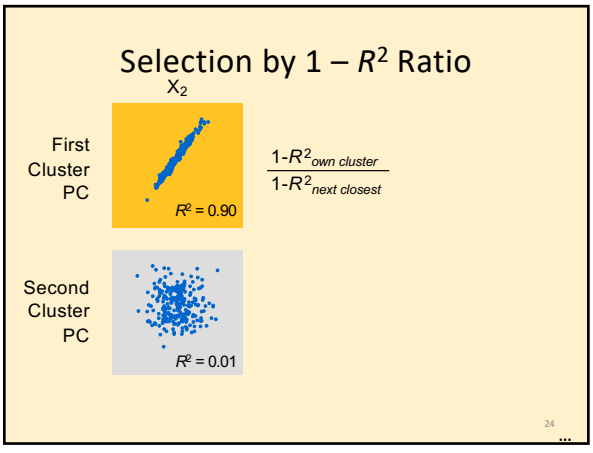
21


Cluster Inputs
X1
X2
X3
First RC
Second RC
22

22

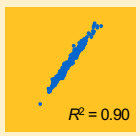

Split New Cluster?
x2
First PC
Second PC
x3
First Eigenvalue=1.95
Second Eigenvalue=0.05
23

23

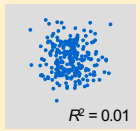

Selection by 1 – R2 Ratio
X2
First Cluster PC
R2 = 0.90
1-R2own cluster
1-R2next closest
Second Cluster PC
R2 = 0.01
24

24

Selection by 1 – R^2 Ratio



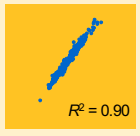
$$\frac{1-R^2_{own\ cluster}}{1-R^2_{next\ closest}} = \frac{1-0.90}{1-0.01}$$



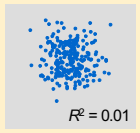
25

25

Selection by 1 – R^2 Ratio



$$\frac{1-R^2_{own\ cluster}}{1-R^2_{next\ closest}} = \frac{1-0.90}{1-0.01} = 0.101$$



26

26

Additional Comments

- Nominal Variables
  - Need dummy variables
  - Dummy for different levels of the **same** nominal variable can be put into different clusters
  - By default, EM passes all the nominal variables to successor nodes as inputs
- Clustering Source:
  - By default, the correlation matrix
  - The alternative is to use the covariance matrix

27

27

Large Data Sets

– Computationally efficient if the data set has fewer than 100 variables and fewer than 100,000 observations.

– If you have more than 100 variables:
  • Use two-stage variable clustering.

– If you have more than 100,000 observations:
  • Sample the data.

28

28

Methods for Reducing Processing Time (Only in Enterprise Miner™)

- If the data set has more than 30 variables:
  – If the number of clusters is known, specify the number of clusters.

Maximum Clusters	3

  – Set the **Keep Hierarchies** property to **Yes**.

Keeps Hierarchies	Yes

  – Set the **Two Stage Clustering** property to **Yes.**

Two Stage Clustering	Yes

29

29

Two-Stage Variable Clustering

– This four-step approach is used to speed up variable clustering with more than 100 input variables.
  1. Calculate the correlation matrix for all variables.
  2. Initialize the *Global Clusters -- # of clusters = K = (# of variables/100 + 2.*
  3. Perform variable clustering on each of the global clusters.
  4. Cluster components are calculated for each of the global clusters and are then used to reconstruct hierarchy clusters among the global clusters

– Or
  • Generate the global clusters (k-mean or hierarchical)
  • Hierarchical approach on each global cluster

30

30

Discussion

- Variable clustering is related to PCA, but the results can be very different. Which would you use in the presence of a large number of inputs? Why?

31

31

Variable Clustering: Pros

- Reduction of collinearity
- Redundancy reduction with low information loss
- Identification of underlying data structure
- Interpretation of original input variables can be kept in successor nodes.

32

32

Variable Clustering: Cons

- One-stage clustering is not computationally efficient if more than 100 input variables.
- Node cannot be used on data with more than 100,000 observations.
- Method is not so well-known. You need to explain it.
- Levels of categorical variables can be located in different clusters.

33

33

Poll

PCA and variable clustering are both forms of supervised dimension reduction?

- ❍ True
- ❍ False

34

34

Poll – Correct Answer

PCA and variable clustering are both forms of supervised dimension reduction?

- ◉ True
- ❍ False

35

35

Simple Variable Clustering

This demonstration illustrates how to use the Variable Clustering node for dimension reduction.

36

36

Questions?

37

37

References

- *Advanced Predictive Modeling Using SAS® Enterprise Minter™ Course Notes.*
- *SAS/STAT® User's Guide, The VARCLUS Procedure.*

38

38