Matrix Factorization Applications in Data Mining

Uzair Ahmad

Contents

- Linear Algebra Primer
- Singular Value Decomposition
- Principal Component Analysis
- Exploratory Factor Analysis
- Independent Component Analysis

System of linear Equations

- $Ax = b$
- Where $A \in R^{m \times n}$ is a known matrix,
- $b \in R^{m}$ is a known vector, and
- $x \in R^{n}$ is a vector of unknown variables we would like to solve for.
- Each element x_i of x is one of these unknown variables.
- Each row of A and each element of b are constraints.
  - $A_{1,1} x_1 + A_{1,2} x_2 + \cdots + A_{1,n} x_n = b_1$

Definitions

- Scalars, Vectors, Matrices, Tensors
- Matrix Rank
  - # of Non Zero and Linearly Independent Rows/Columns
- Matrix Multiplication
  - C = AB ➔ $C_{i,j} = \sum_k A_{i,k} B_{k,j}$
  - Vector Multiplication → Dot Product
    - Similar Vs Orthogonal
- Matrix Multiplication Attributes
  - Distributive ➔ A(B + C) = AB + AC
  - Associative ➔ A(BC) = (AB)C
  - not commutative ➔ AB != BA
    - Vector multiplication is $x^T y = y^T x$
  - $A^T B^T = (AB)^T$
- Transpose: $(A^T)_{i,j} = A_{j,i}$

Definitions

- Broadcasting C = A + b (Addition of Matrix and a vector)
- Identity Matrix I
  - IA = A ➔ 1s at the diagonal and 0s elsewhere
- Square Matrix
  - Rows == Columns
- Inverse Matrix
  - For every square matrix A
  - There is an inverse matrix B
  - Such that AB = I or BA = I and BAB = B
- Symmetric Matrix
  - A^T= A

Definitions

- Unit Vectors
  - A unit vector is a vector with unit norm; $\|x\|_2 = 1$
- Orthogonal Vectors
  - $x^T y = 0$
- Orthonormal
  - Orthogonal and Unit norm
- Orthogonal Matrix
  - Rows/colus are mutually orthonormal
  - $AA^T = A^T A = I \Rightarrow A^{-1} = A^T$
- Linear Independence

Solving a System of linear Equations

- $Ax = b$
  - $A^{-1} A = I_n$
  - $A^{-1} Ax = A^{-1}b$
  - $I_n x = A^{-1}b$
- $x = A^{-1}b$

If A^{-1} exists, finding it in closed form is possible algorithmically.

Theoretically, the same inverse matrix can then be used to solve the equation for different values of b

However, A−1 is primarily useful as a theoretical tool, and should not actually be used in practice for most software applications. Because A^{-1} can be represented with only limited precision on a digital computer, algorithms that make use of the value of b can usually obtain more accurate estimates of x.

Matrix Decomposition

- Central to many fundamental Linear Algebra problems
- Eigen Value Decomposition
- Singular Value Decomposition

Eigenvectors and Eigenvalues

$$\begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & -3 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 0 \end{bmatrix} = 3 \times \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

$$\boxed{A\,v = \lambda\,v}$$

Eigenvectors and Eigenvalues

$$\begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & -3 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 0 \end{bmatrix} = 3 \times \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

Eigenvalue

Eigenvector

$$A\,v = \lambda\,v$$

An eigenvector of a square matrix A is a non-zero vector v such that multiplication by A alters only the scale of v:

Eigenvectors and Eigenvalues

$$\begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & -3 \end{bmatrix} \text{X} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 0 \end{bmatrix} = 3 \text{ X} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

Eigenvalue

Eigenvector

$$A\,v = \lambda\,v$$

If V is an eigenvector of A, then so is any rescaled vector sV (s != 0). sV still has the same eigenvalue.
For this reason, we usually only look for unit eigenvectors.

Eigenvectors and Eigenvalues

- Singular Matrix
  - The matrix is singular if and only if any of the eigenvalues are zero.
- Positive Definite
  - eigenvalues are all +ve
- Positive Negative
  - eigenvalues are all –ve
- Positive Semi-Definite
  - +ves and some zeros

Singular Value Decomposition

- singular vectors and singular values
- some of the same kind of information as the eigen-decomposition
  - But more generally applicable
  - Eigen-decomposition requires Square matrix
- Every real matrix has a singular value decomposition, but the same is not true of the eigenvalue decomposition.
- For example, if a matrix is not square, the eigen-decomposition is not defined, and we must use a singular value decomposition instead.

SVD

Non –ve, Diagonal, Orthogonal, Decreasing, not necessarily square

~100 Years

$$X = USV^T$$

Orthogonal

U is an m × m matrix: Left Singular vectors
S is an m × n matrix: singular values
V is an n × n matrix: Right Singular vectors

SVD

Non –ve, Diagonal, Orthogonal, Decreasing, not necessarily square

~100 Years

$$X = USV^T$$

Orthogonal

U ➔ eigenvectors of XX^T

V ➔ eigenvectors of X^TX

S ➔ Square roots of eigenvalues of X^TX or XX^T

SVD

Set 2 | Set 1 SVs | Set 2 SVs

Set 1

$$\underset{m \times n}{\overset{X}{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & \\ \vdots & \vdots & \ddots & \\ x_{m1} & & & x_{mn} \end{pmatrix}}} = \underset{\boxed{m} \times r}{\overset{U}{\begin{pmatrix} u_{11} & \cdots & u_{1r} \\ \vdots & \ddots & \\ u_{m1} & & u_{mr} \end{pmatrix}}} \underset{r \times r}{\overset{S}{\begin{pmatrix} s_{11} & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & s_{rr} \end{pmatrix}}} \underset{r \times \boxed{n}}{\overset{V^{\mathsf{T}}}{\begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \\ v_{r1} & & v_{rn} \end{pmatrix}}}$$

Rank of X

A hidden feature space where the **Set 1** and **Set 2** have feature vectors that are closely aligned.
So, when we compute X=U×s×V,
U ➔ The feature vectors corresponding to the **Set 1** in the hidden feature space
V ➔ The feature vectors corresponding to the **Set 2** in the hidden feature space.

Question: Is i^{th} member of Set 1 is related to j^{th} member of Set 2 ? Set1SVs$i dotproduct Set2SVs$j

Note: Set 1 and Set 2 should be known

SVD Applications

- Collaborative Filtering
- Data Compression
- Document Clustering
- Topic Modelling
- Recommender Systems: B.M. Sarwar, G. Karypis, J.A. Konstan, and J.Reidl. Application of dimensionality reduction in recommender system - a case study. In ACM WebKDD 2000 Web Mining for E-Commerce Workshop, 2000
- Linear Algebra

SVD Applications

	M1	M2	M3	M…	Mn
U1	1	1	0	0	0
U2	0	1	1	0	1
U3	0	0	1	1	0
…	1	1	0	0	0
Um	0	1	0	1	1

1. 500k X 17k Sparse Matrix (84 in 85 cells are empty: Netflix)
2. SVD De-sparses the large matrix → **K***(17K+500K)
3. Precomputing SVD: Incremental solutions

SVD Applications

Social Graph: 0s are **missing values** to be predicted

	U1	U2	U3	U…	Un
U1	1	1	0	0	0
U2	0	1	1	0	1
U3	0	0	1	1	0
…	1	1	0	0	0
Um	0	1	0	1	1

1. If Ui is friends with Uj
2. SVD De-sparses the large matrix
3. Precomputing SVD: Incremental solutions

SVD Applications

- If meaningful generalities can help you represent your data with fewer numbers, finding a way to represent your data in fewer numbers can often help you find meaningful generalities. (Sifter)
- Compression == Understanding

Why Dimensionality Reduction

- Not handful observations but sheer data sizes 100s x 100ks ?
- Groups in data
  - How Variables / Observations hang together
  - Multiple variable explain similar things

Dimensions Reduction Goals

- Reduce dimensions of Population
  - Cluster Analysis
  - K-Means, Hierarchical, Density

- Reduce dimensions of the Problem/Model/Construct
  - Singular Value Decomposition
  - Principal Component Analysis (PCA)
  - Exploratory Factor Analysis (EFA)

Vocabulary

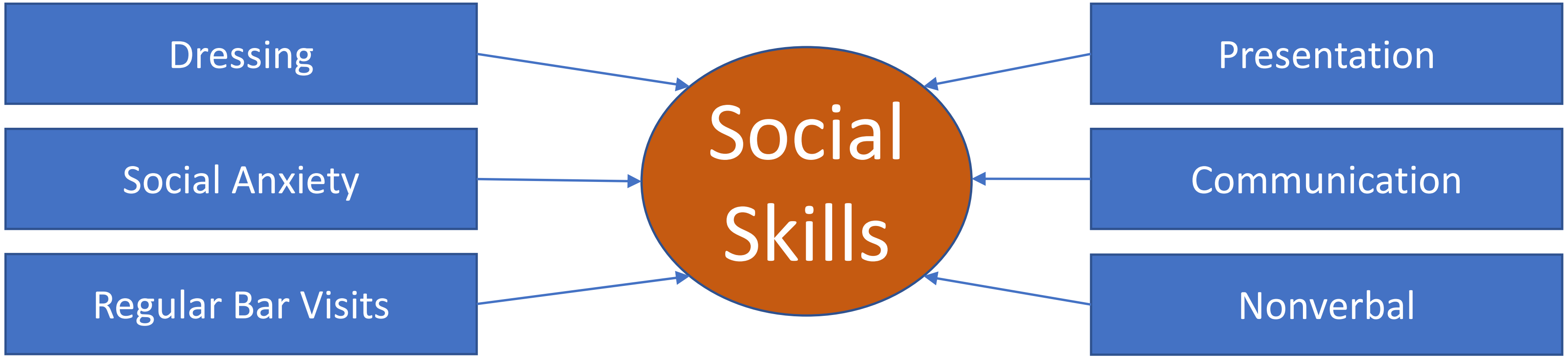
- Latent Variables:
  - Hidden / Unobservable variables
- Loadings
  - Weight of Relationship between each variable and component
- Communality
  - The amount of variance of each variable explained by the factor structure

Important in factor analysis to measure error

- Uniqueness
  - The amount of variance of each variable not explained by the component/factor (1 – Communality)

Dimensions Reduction Goals

- Reduce dimensions of the Problem/Model/Construct
  - Principal Component Analysis (PCA)
  - Exploratory Factor Analysis (EFA)

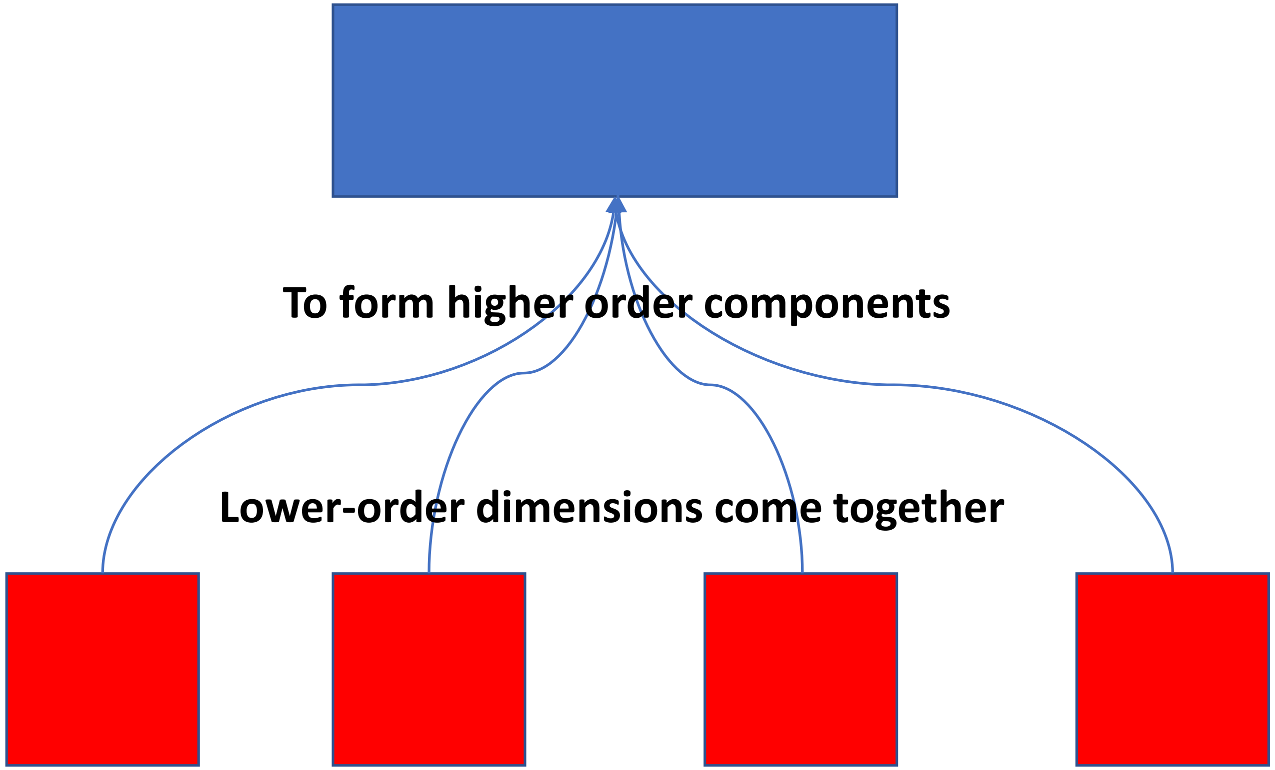


Can 6 different variables represented by just 1; the Social Skills

PCA

Fewer higher-order components EXPLAIN/CAPTRUE the maximum amount of variance in lower order variables

selected



Teacher Quality

- Present
- Explain
- Communicate
- Teach
- Workload
- Difficulty

	Present	Explain	Communi	Teach	Workload	Difficulty
Present	1.000	0.855	0.603	0.800	0.151	0.043
Explain	0.855	1.000	0.756	0.891	0.056	-0.026
Communi	0.603	0.756	1.000	0.819	0.128	0.060
Teach	0.800	0.891	0.819	1.000	0.138	0.081
Workload	0.151	0.056	0.128	0.138	1.000	0.719
Difficulty	0.043	-0.026	0.060	0.081	0.719	1.000

Rotation:

	PC1	PC2	PC3	PC4	PC5	PC6
Present	0.4812739	0.04614410	-0.63443650	-0.03777691	0.55990582	-0.22094091
Explain	0.5127898	0.12651133	-0.16342075	-0.06968290	-0.34260959	0.75637171
Communi	0.4670223	0.04023162	0.73502722	0.15324099	0.46160796	0.05866729
Teach	0.5188259	0.04304370	0.10233296	-0.10077104	-0.58073509	-0.60916901
Workload	0.1167782	-0.69214922	-0.11178861	0.69377824	-0.11331686	0.02503335
Difficulty	0.0670437	-0.70662842	0.08689213	-0.69191860	0.07712509	0.06270205

Applying PCA

- Correlation Matrix: Pairwise Correlation of all variables
- Calculate Eigenvectors and Eigenvalues
- Eigenvectors: are the Principal Components (Transformed Variables/New coordinate system)
- Eigenvalues: are the amount of variance each PC explains

Applying PCA

- How many components ?
- Eigenvalue > 1 (Kaiser-Guttman Rule) [Norman Cliff's critic]
- Percentage of variance: Combined variance is more than T (50% ?)
- Scree plot: discard scree

Applying PCA

- cor()
  - Data Standardization

- Orthogonality and correlation

- Orthogonality and components

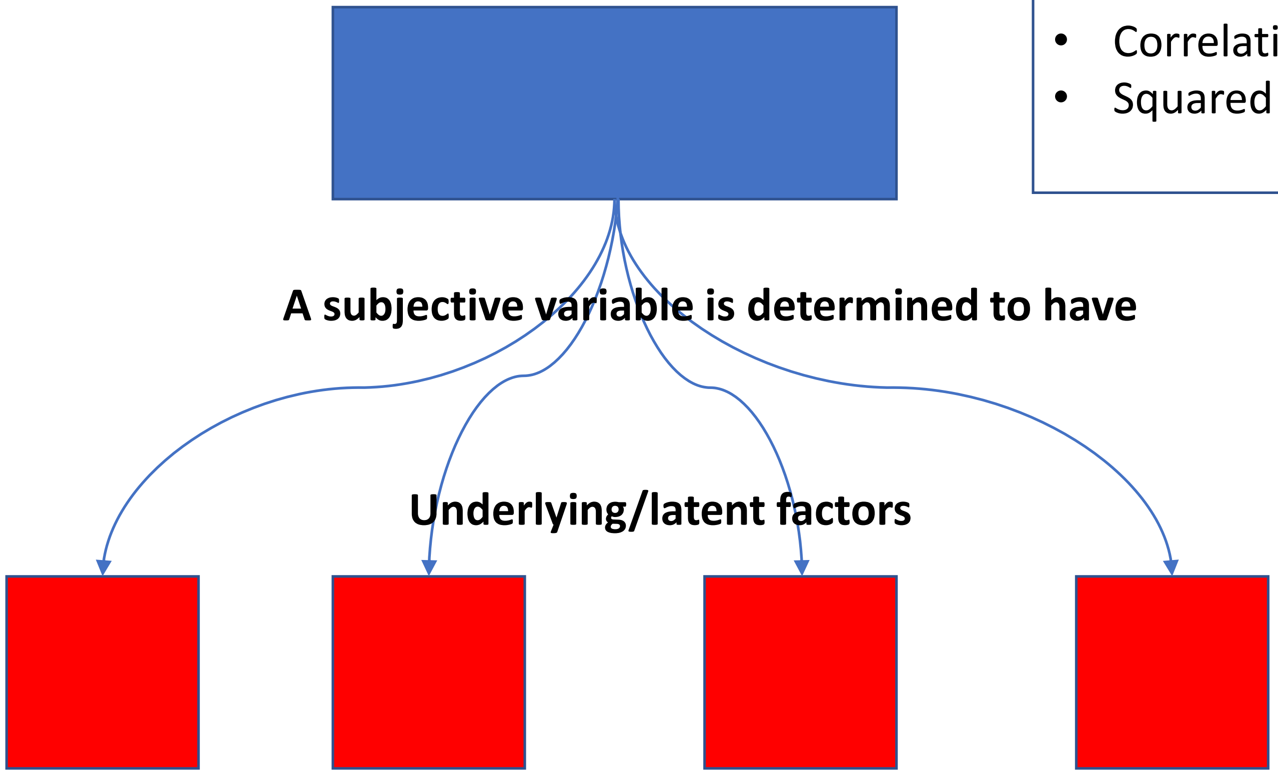
EFA

Thurstone (1931)

- Dimension Reduction
- Structure (of variable relationships) discovery

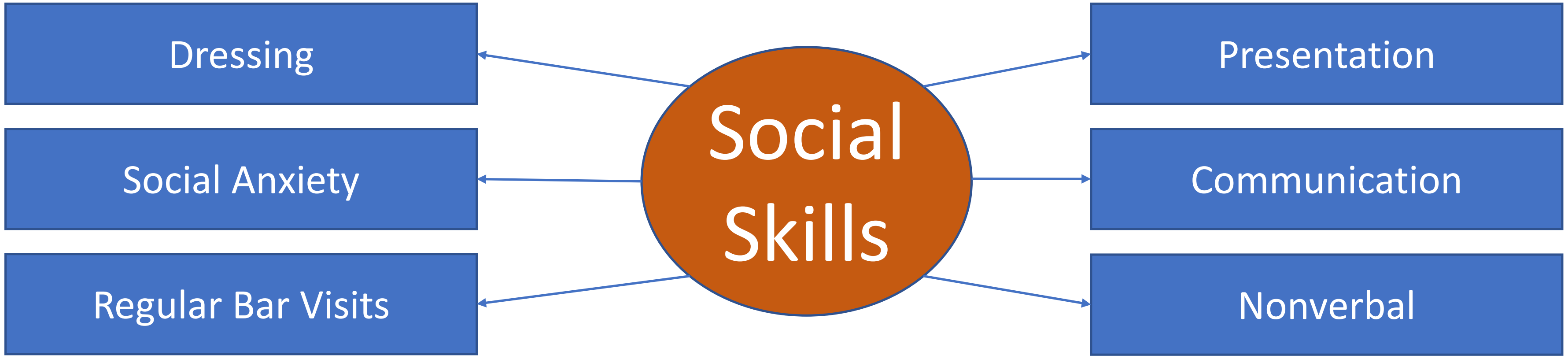
The concept of shared variance

- Standard Scores (Z)
- Correlation ← → Z
- Squared Correlation (R^2)



Dimensions Reduction Goals

- Reduce dimensions of the Problem/Model/Construct
  - Principal Component Analysis (PCA)
  - Exploratory Factor Analysis (EFA)



Can 6 different variables represented by just 1; the Social Skills

EFA

- PCA Vs EFA
  - Objective Vs Subjective
  - No Measurement Error Vs Error is allowed
  - Total Vsd Common/Shared Variability/Variance
  - PCA Robust, EFA Sensitive to Non-normality

- EFA Subsumes all of the dimensions within a single factor

	PA1	PA2
SS loadings	3.19	1.45
Proportion Var	0.53	0.24
Cumulative Var	0.53	0.77
Proportion Explained	0.69	0.31
Cumulative Proportion	0.69	1.00

EFA Example

```
Factor Analysis using method =  pa
Call: fa(r = corMat, nfactors = 2, rotate = "varimax", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
```

	PA1	PA2
Present	0.83	0.06
Explain	0.97	-0.04
Communi	0.79	0.06
Teach	0.96	0.07
Workload	0.10	0.85
Difficulty	0.01	0.85

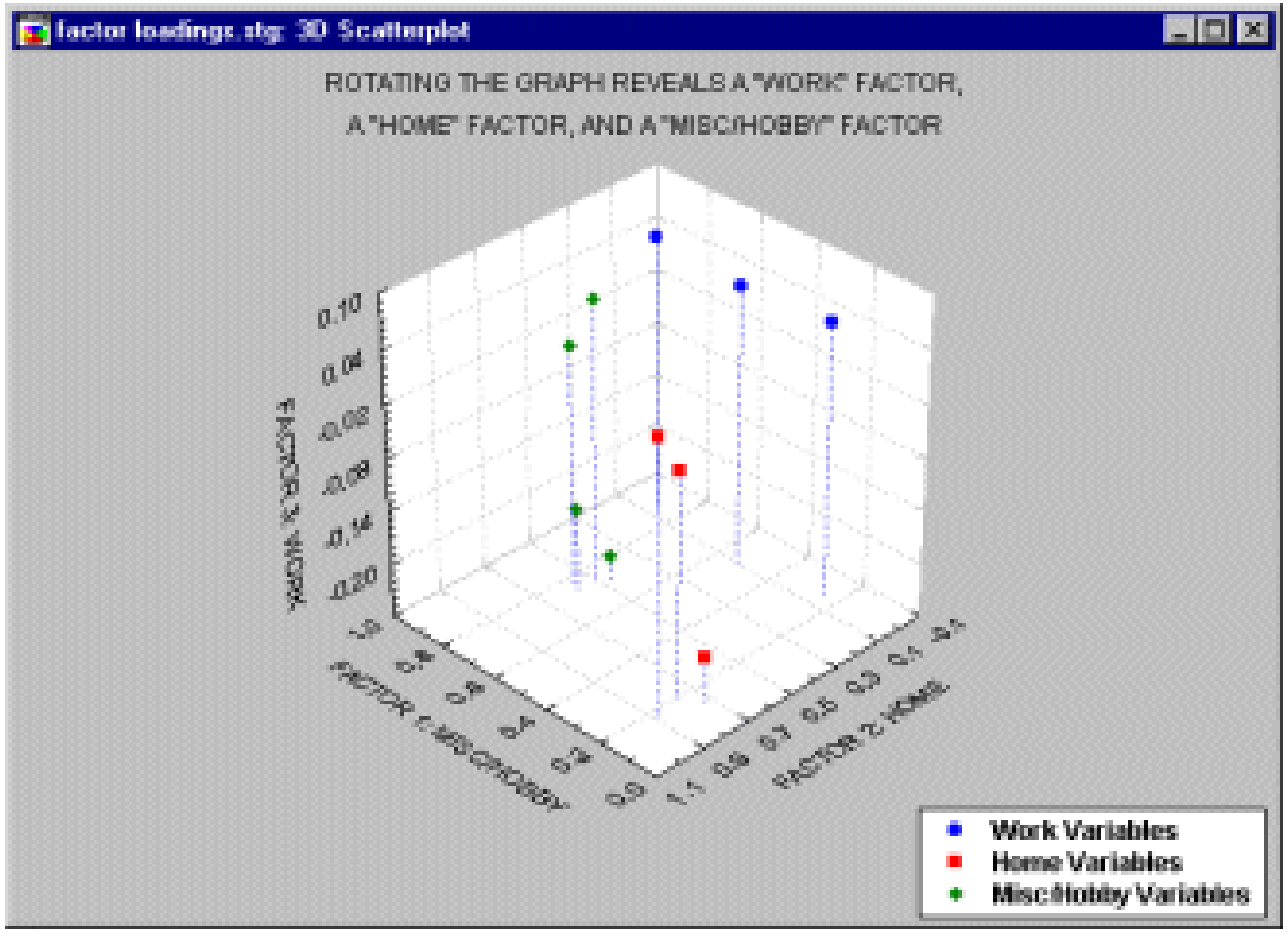
	PA1	PA2
SS loadings	3.19	1.45
Proportion Var	0.53	0.24
Cumulative Var	0.53	0.77
Proportion Explained	0.69	0.31
Cumulative Proportion	0.69	1.00

- Only eyes are not enough in the ….
- Hang upside down to see things correct

Concept of Rotation

- Geometric spinning of axes in multidimensional space … but …
- Initial result of FA is not the most interpretable one
  - You have to rotate
- VariMax (Variance Maximization) Rotation
  - Variance is maximal on a Regression line (2d) or plane (3d) or … example
  - Leftover/Residual Variance around the line can be captured be another line that is orthogonal to regression line
  - Because each consecutive factor is defined to maximize the variability that is not captured by the preceding factor, consecutive factors are independent of each other.
    - consecutive factors are uncorrelated or *orthogonal* to each other
- Oblique
  - Allows somewhat correlated
  - represent "clusters" of variables without orthogonality condition
  - Promax

Concept of Rotation



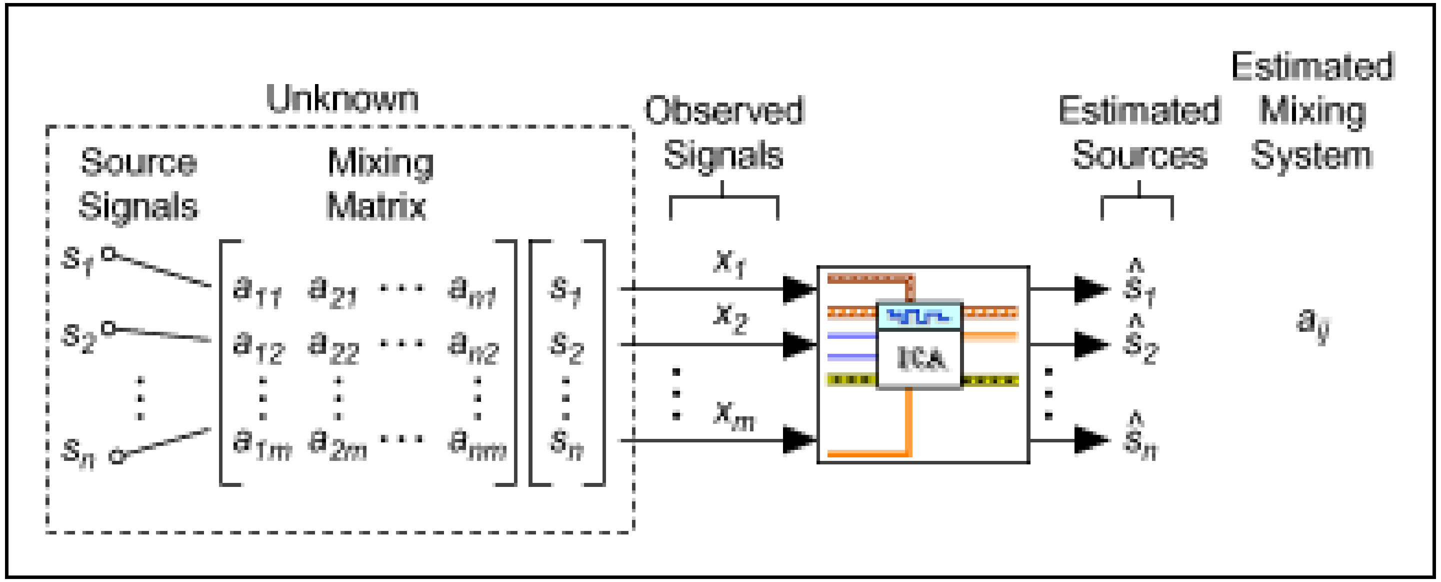
Source: dell.com

T/F?

- Which of the following techniques can be used to reduce the dimensions of the population?
- Cluster Analysis partitions the columns of the data, whereas principal component and exploratory factor analyses partition the rows of the data. True or false?
- PCA explains the total variance
- EFA explains the common variance
- EFA identifies measures that are sufficiently similar to each other to justify combination
- PCA captures latent constructs that are assumed to cause variance

ICA

- Independent Component Analysis
    - Blind Source Seperation (BSS)
- Unsupervised Learning Problem
    - Probabilistic Model → Latent Structure → Data
    - ICA discovers that structure



ICA Assumptions

- uncorrelated and independent variables

Known

Unknown
Mixture Matrix

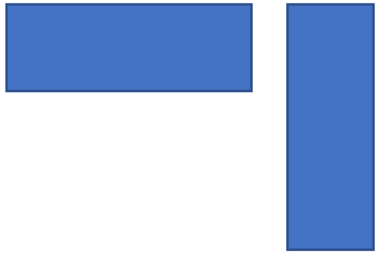
Unknown
Signal Sources

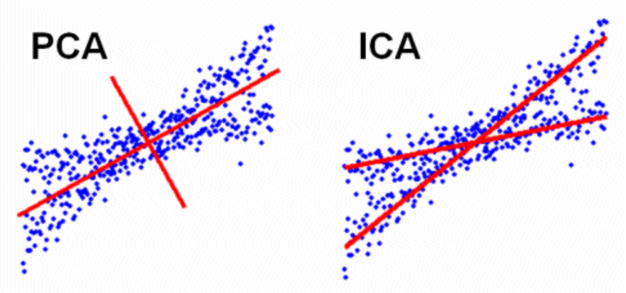
$$x = A\,s$$

$$s = W\,x$$

ICA Vs PCA

- Local/Fundamental Vs. Average features
  - Faces
  - Natural scenes
  - Text corpus
- Direction Sensitive Vs. Direction Neutral
- BSS Vs. Common Source Extraction





- http://www.ats.ucla.edu/stat/r/pages/svd_demos.htm