Structured Low Rank Approximation

Lecture I: Introduction

Moody T. Chu

North Carolina State University

presented at

XXII School of Computational Mathematics
Numerical Linear Algebra and Its Applications
September 13, 2004

Syllabus

- Objectives:
  - ◇ To provide some preliminaries.
  - ◇ To treat some mathematics.
  - ◇ To point out some applications.
  - ◇ To describe some algorithms.
- Topics:
  - ♦ Lecture I: Introduction
  - ◇ Lecture II: General Approach
  - ◇ Lecture III: Distance Geometry and Protein Structure
  - ◇ Lecture IV: Singular Value Reassignment with Low Rank Matrices
  - ◇ Lecture V: Nonnegative Matrix Factorization
- Assignments:
  - ◇ Be able to relate the various subjects.
  - ◇ Answer the various open problems.
  - ◇ Get excited and enjoy the lectures.

Lecture I:

Introduction

Outline

- Problem Description
- The Meaning of Rank Reduction
  - ◇ Karhunen-Loève Expansion
  - ◇ Low Rank Modeling
  - ◇ Random sampling
  - ◇ TSVD
- Applications
  - ◇ Data Mining
  - ◇ Factor Retrieval
- The Involvement of Structure
  - ◇ Types of Structures
  - ◇ Difficulties

Structure Preserving Rank Reduction Problem

- Given
  - ◇ A target matrix $A \in \mathbb{R}^{n\times n}$,
  - ◇ An integer k, $1 \leq k < \mathrm{rank}(A)$,
  - ◇ A class of matrices Ω with a specified structure,
  - ◇ a fixed matrix norm $\|\cdot\|$;

  Find
  - ◇ A matrix $\hat{B} \in \Omega$ of rank k, such that

$$\|A - \hat{B}\| = \min_{B\in\Omega, \mathrm{rank}(B)=k} \|A - B\|. \tag{1}$$

Some Known Facts

- It seems that we know a lot about low rank approximations.
  - ◇ Truncated singular value decomposition.
  - ◇ Noise removal.
  - ◇ ⋮
- Exactly what do we know about low rank approximations?
  - ◇ Why is low rank doing what we think it is doing?

Karhunen-Loève Expansion

- Let $\mathcal{X} \in \mathbb{R}^n$ denote a random column vector.
  - ◇ $C := \mathcal{E}[(\mathcal{X} - \mathcal{E}[\mathcal{X}])(\mathcal{X} - \mathcal{E}[\mathcal{X}])^\top] \in \mathbb{R}^{n \times n}$ is defined as the *covariance matrix* of $\mathcal{X}$.
  - ◇ Being symmetric and positive semi-definite, C has a spectral decomposition

$$cov(\mathcal{X}) = \sum_{j=1}^{n} \lambda_j \boldsymbol{u}_j \boldsymbol{u}_j^\top.$$

  - ◇ $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ are deterministic and form an orthonormal basis for $\mathbb{R}^n$.
- The random column vector $\mathcal{X}$ can be expressed as

$$\mathcal{X} = \sum_{j=1}^{n} (\boldsymbol{u}_j^\top \mathcal{X}) \boldsymbol{u}_j.$$

- Each coefficient $\alpha_j := \boldsymbol{u}_j^\top \mathcal{X}$ itself is a random variable.
- Properties of $\boldsymbol{\alpha}$:

$$\begin{aligned} \mathcal{E}[\boldsymbol{\alpha}] &= U^\top \mathcal{E}[\mathcal{X}], \\ cov(\boldsymbol{\alpha}) &= \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}. \end{aligned}$$

Low Rank Modeling

- Since the basis vectors $\boldsymbol{u}_j$, $j = 1, \ldots n$, are deterministic,
  - ◇ Stochastic properties of $\mathcal{X}$ are caused only by stochastic properties of coefficients $\boldsymbol{\alpha}$.
  - ◇ The randomness of $\mathcal{X}$ is due to the randomness of $\boldsymbol{\alpha}$.
- Random variables $\alpha_j, j = 1, \ldots, n$, are mutually stochastically independent.
  - ◇ Variance measures the unpredictability or scattering of a random variable.
  - ◇ The larger the eigenvalue λ_j is, the larger the variance of α_j is and, hence, the more randomness it contributes.

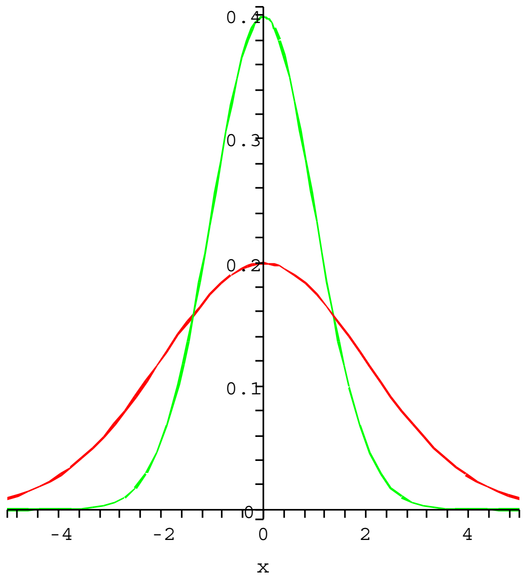


Figure 1: Normal distributions with different variances ($\sigma = 1, 2$).

Feature Selection

- The effectiveness of basis vectors:
  - Those coefficients with larger variances and the corresponding directions are the more important components in representing the stochastic nature of $\mathcal{X}$.
  - Rank the importance of corresponding eigenvectors $\boldsymbol{u}_j$ as *essential* components for the variable $\mathcal{X}$ according to the magnitude of λ_j.
- If truncation is necessary, those eigenvectors corresponding to smaller variances should be thrown away first.
  - A Maple demonstration.

Minimum-Variance Approximation

- Assume $\lambda_1 \geq \ldots \geq \lambda_r \geq \ldots \geq \lambda_n$. The low rank model

$$\hat{\mathcal{X}} := \sum_{j=1}^{r} (\boldsymbol{u}_j^\top \mathcal{X}) \boldsymbol{u}_j \tag{2}$$

  is an approximate reconstruction of $\mathcal{X}$.

- Consider the error

$$\mathbf{e} := \mathcal{X} - \hat{\mathcal{X}} = \sum_{j=r+1}^{n} \alpha_j \boldsymbol{u}_j.$$

- Assume $\mathcal{E}[\mathcal{X}] = 0$. Among *all* unbiased variables restricted to *any* r-dimensional subspaces in R^n, the random variable $\hat{\mathcal{X}}$ defined in (2) is the best linear minimum-variance estimate of $\mathcal{X}$ in the sense that

$$\mathcal{E}[\|\mathcal{X} - \hat{\mathcal{X}}\|^2] = \sum_{j=r+1}^{n} \lambda_j$$

  is minimized.

An Application of Low Rank Modeling

- Consider the transmission of a data vector $\mathcal{X}$ over a noisy communication channel.
  - ◇ $\mathcal{X}$ resides in the data space.
  - ◇ The reduced-rank approximation $\hat{\mathcal{X}}$, represented by

    $$\hat{\boldsymbol{\alpha}} = [\alpha_1, \ldots, \alpha_r]^\top,$$

    resides in the feature space.
- Assume the signal is corrupted by an additive white noise of zero mean $\boldsymbol{v}$, that is,
  - ◇ The noise vector $\boldsymbol{v}$ is uncorrelated to the data vector $\mathcal{X}$,

    $$\mathcal{E}[\mathcal{X}\boldsymbol{v}^\top] = 0.$$
  - ◇ Elements of the noise vector $\boldsymbol{v}$ are identical independently distributed random variables,

    $$\mathcal{E}[\boldsymbol{v}\boldsymbol{v}^\top] = \sigma^2 I.$$

Direct Transmission

- In the direct transmission, the received signal is given by

$$\mathcal{Y} = \mathcal{X} + \mathbf{v}.$$

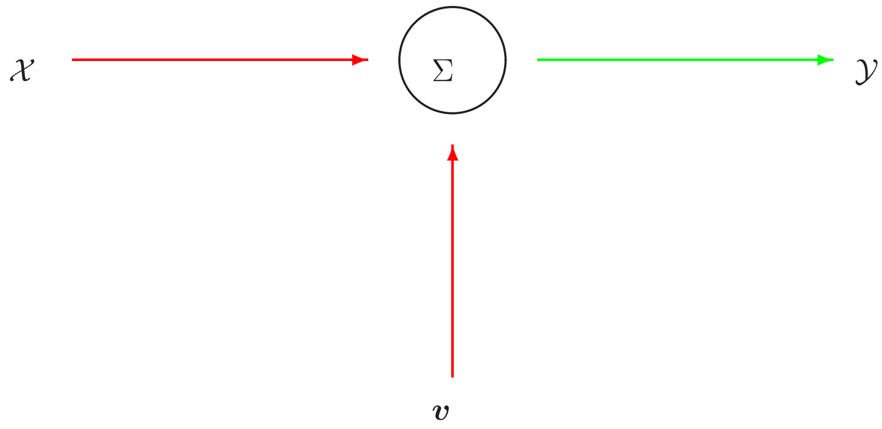


Figure 2: Direct data transmission.

- The mean-square transmission error is

$$\begin{aligned} \epsilon_{direct} &= \mathcal{E}[\|\mathcal{Y} - \mathcal{X}\|^2] = \mathcal{E}[\|\boldsymbol{v}\|^2] \\ &= \sum_{j=1}^{n} \mathcal{E}[v_j^2] \quad \text{(by independence)} \\ &= n\sigma^2. \end{aligned}$$

Indirect Transmission

- In the indirect transmission, the signal $\mathcal{X}$ is first applied to a transmit filter bank made of $U_1^\top = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r]^\top$. The resulting $\hat{\boldsymbol{\alpha}} = U_1^\top \mathcal{X}$ is sent through the noisy channel and the received signal is given by

$$\boldsymbol{\beta} = \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{v}},$$

which is then applied to the receive filter bank U_1 to obtain

$$\mathcal{Z} = U_1 \boldsymbol{\beta} = U_1 \hat{\boldsymbol{\alpha}} + U_1 \hat{\boldsymbol{v}}.$$

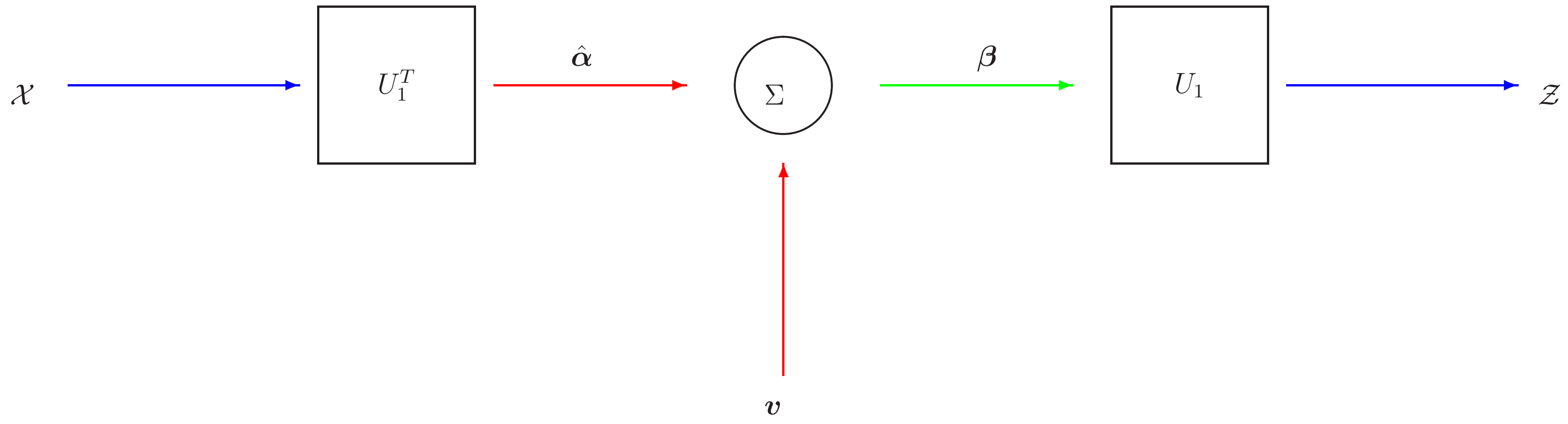


Figure 3: Indirect data transmission.

- The mean-square transmission error of the entire process is given by

$$\epsilon_{indirect} = \mathcal{E}[\|\mathcal{Z} - \mathcal{X}\|^2] = \mathcal{E}[\|(U_1 U_1^\top - I)\mathcal{X} + U_1 \hat{\boldsymbol{v}}\|^2] = \sum_{j=r+1}^{n} \lambda_j + r\sigma^2.$$

Bias-Variance Tradeoff

- Note that $\epsilon_{indicrect} < \epsilon_{direct}$ if and only if
$$\sum_{j=r+1}^{n} \lambda_j < (n-r)\sigma^2.$$
  - ⋄ If the tail-end eigenvalues of the correlation matrix C are sufficient small, then truncation is better then no-truncation.
- Truncation introduces a bias.
- The reduced yet biased low rank model is less susceptible to noise.

Random Sampling in Sample Space

- The distribution of a random variable is often simulated by a collection of ℓ random samples.
  - ◇ Samples are recorded in a $n \times \ell$ matrix X.
  - ◇ Each column of X represents one random sample of the underlying random (column vector) variable $\mathcal{X} \in \mathbb{R}^n$.
- Law of large numbers: When ℓ is large enough, many of the stochastic properties of $\mathcal{X}$ can be recouped from X.
  - ◇ Sample mean $\boldsymbol{\mu} = X\frac{\mathbf{1}}{\ell}$ converges stochastically to $\mathcal{E}(\mathcal{X})$.
  - ◇ Sample covariance $R = \frac{1}{\ell}(X - \boldsymbol{\mu}\mathbf{1}^\top)(X - \boldsymbol{\mu}\mathbf{1}^\top)^\top$ converges stochastically to $cov(\mathcal{X})$.

Low Rank Representation of Sample Data

- How to retrieve a sample data matrix from X to represent the minimum-variance approximation $\hat{\mathcal{X}}$ of $\mathcal{X}$?
  - ⋄ Spectral decomposition of sample covariance:

$$R = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top. \tag{3}$$

- Comparison:
  - ⋄ Best *low dimension* minimum-variance estimate of the *continuous* random *variable* $\mathcal{X}$.

$$\hat{\mathcal{X}} := \sum_{j=1}^{r} (\boldsymbol{u}_j^\top \mathcal{X}) \boldsymbol{u}_j.$$

  - ⋄ Best *low rank* minimum-variance estimate of the *discrete* random *sample* of X?

$$\hat{X} := \sum_{j=1}^{r} \mathbf{u}_j (\mathbf{u}_j^\top X). \tag{4}$$

TSVD

- The singular value decomposition of X:

$$X = U\Sigma V^\top = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \tag{5}$$

  - ◇ Share the same eigenvectors of R as its left singular vectors, i.e., $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$.
  - ◇ Singular values $\sigma_i = \sqrt{\ell \lambda_i}$ are ranked in the same ordering as eigenvalues λ_i, $i = 1, \ldots n$.
  - ◇ The notion of the truncated singular value decomposition of X is simply the partial sum

$$\hat{X} = \sum_{i=1}^{r} \mathbf{u}_i (\sigma_i \mathbf{v}_i^\top).$$

- The TSVD of a give data matrix X representing random samples of an unknown random variable $\mathcal{X}$ has a statistical meaning.
  - ◇ The truncated rank-r SVD $\hat{X}$ represents random samples of the best minimum-variance linear estimate $\hat{\mathcal{X}}$ to $\mathcal{X}$ among all possible r-dimensional subspaces.

Data Mining

- Data mining is about extracting interesting information from raw data.
- What constitutes "information"?
  - ◇ Patterns of appearance.
  - ◇ Association rules between sets of items.
  - ◇ Clustering of the data points.
  - ◇ Concepts or categories.
  - ◇ Principal components or factors.
  - ◇ ...
- What should be counted as "interesting"?
  - ◇ Confidence and support.
  - ◇ Information content.
  - ◇ Unexpectedness.
  - ◇ Actionability — The ability to suggest concrete and profitable decision-making.
  - ◇ ...
- For different information retrievals, different techniques should be used.
  - ◇ Factors — Rank reduction or lower dimension approximation.
  - ◇ Clusters — Centroids or k-means.
  - ◇ ...

From Complexity to Simplicity

- Data analysis:
  - An indispensable task in almost every discipline of science.
  - Search for relationships between a set of externally caused and internal variables.
  - Especially important in this era of information and digital technologies when massive amounts of data are generated at almost all levels of applications.
- Data observed from complex phenomena:
  - Often represent the integrated result of several interrelated variables acting together.
  - These variables sometimes are less precisely defined.
- What to distinguish which variable is related to which and how the variables are related.

Two Classical Approaches

- Factor analysis:
  - A class of procedures that can help identify and test what *constructs*, or *factors*, might be used to explain the interrelationships among the variables.
  - Each construct itself is a complex image, idea, or theory formed from a number of simpler elements.
- Cluster analysis:
  - A procedure used to organize information about cases so that relatively *homogenous groups*, or *clusters*, can be formed.
    - Group members should be highly internally homogenous (members are "similar" to one another in their characteristics) and highly externally heterogenous (members are not "like" members of other clusters).
    - Need a measurement of similarity or dissimilarity.
    - Need a decision on how many clusters to keep.
- Either analysis is meant to bring forth the effect of reducing the size of the data table.

Basic Model

- Making observation, data gathering and processing:
  - ◇ Assume ℓ entities and n variable.
  - ◇ Record raw scores that entity j received from all variables.
  - ◇ Normalize raw scores to have mean 0 and standard deviation 1 (standardized score).
  - ◇ Let $Y = [y_{ij}] \in \mathbb{R}^{n \times \ell}$ denote the matrix of observed data.
    - ▷ y_{ij} = *standard score* of entity j on variable i.
- Correlation matrix of all n variables:

$$R := \frac{1}{\ell} Y Y^T. \tag{6}$$

Linear Relationship (Only an Assumption)

- Assume that y_{ij} is a linearly weighted score of entity j on several factors.

$$Y = AF. \tag{7}$$

- $A = [a_{ik}] \in \mathbb{R}^{n\times m}$ is the factor loading matrix.
  - ◇ a_{ik} = the loading of variable i on factor k, or the influence of factor k on variable i.
- $F = [f_{kj}] \in \mathbb{R}^{m\times \ell}$is the factor scoring matrix.
  - ◇ f_{kj} = the score of factor k on entity j, or the response of entity j to factor k.

$$\begin{bmatrix} & & y_{1j} & & \\ & & \vdots & & \\ \dots & \dots & \boxed{y_{ij}} & \dots & \dots \\ & & \vdots & & \\ & & y_{nj} & & \end{bmatrix} = \underbrace{\begin{bmatrix} \\ a_{i1} & \dots & a_{ik} & \dots & a_{im} \\ \\ \end{bmatrix}}_{\text{influence of factors}} \left.\begin{bmatrix} & & f_{1j} & & \\ & & \vdots & & \\ \dots & \dots & f_{kj} & \dots & \dots \\ & & \vdots & & \\ & & f_{mj} & & \end{bmatrix}\right\} \text{response to factors}$$

Four Examples

- Factor retrieval
- Latent semantic indexing
- Cluster analysis
- Receptor model

Example 1: What Factors Affect Students' Academic Performance?

- Y represents the transcripts of ℓ college students (the entities) at the end of freshman year. Each column denote one student's grades on n fixed subjects (the variables), e.g., Calculus, English, Chemistry, and so on.
- A college freshman's academic performance depends on a number of factors including, for instance, family social status, finance, high school GPA, cultural background, and so on.
- Upon entering the college, each student could be asked to fill out a questionnaire inquiring these factors of his/her background. In turn, individual responses to those factors are translated into scores and placed in the corresponding column of the scoring matrix F.
- What is not clear to the educators/administrators is how to choose the factors to compose the questionnaire or how each of the chosen factors would be weighted (the loadings) to reflect the effect on each particular subject.
- In practice, we usually do not have a priori knowledge about the number and character of underlying factors in A. Sometimes we do not even know the factor scores in F.
- Only the data matrix Y is observable.
- Explaining the complex phenomena observed in Y with the help of a minimal number of factors extracted from the data matrix is the primary and most important goal of factor analysis.

Factor Analysis and Matrix Decomposition

- Two additional assumptions:
  - ◇ All sets of factors being considered are uncorrelated with each other.
  - ◇ Similar to Y, the scores in F for each factor are normalized.

$$\frac{1}{\ell}FF^T = I_m. \tag{8}$$

- The correlation matrix R can be expressed directly in terms of the loading matrix A, i.e.,

$$R = AA^T. \tag{9}$$

  - ◇ Factor extraction now becomes a problem of decomposing the correlation matrix R into the product AA^T.
  - ◇ Would like to use as few factors as possible.

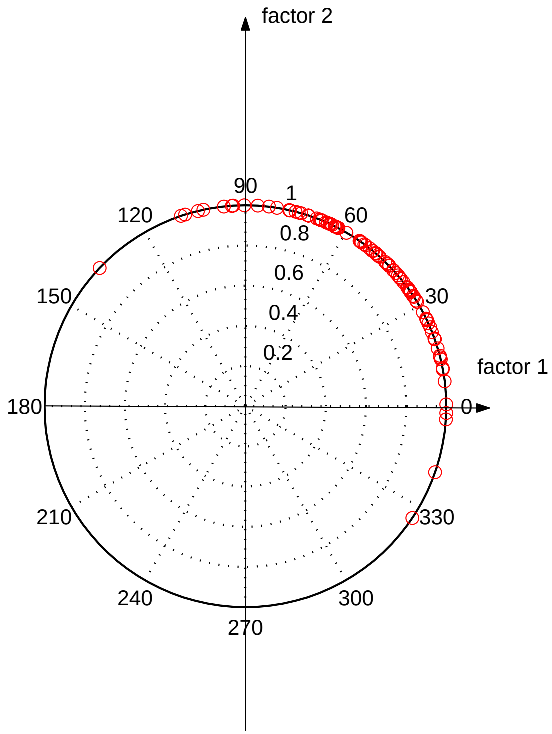
Interpretation of the Loading Matrix A

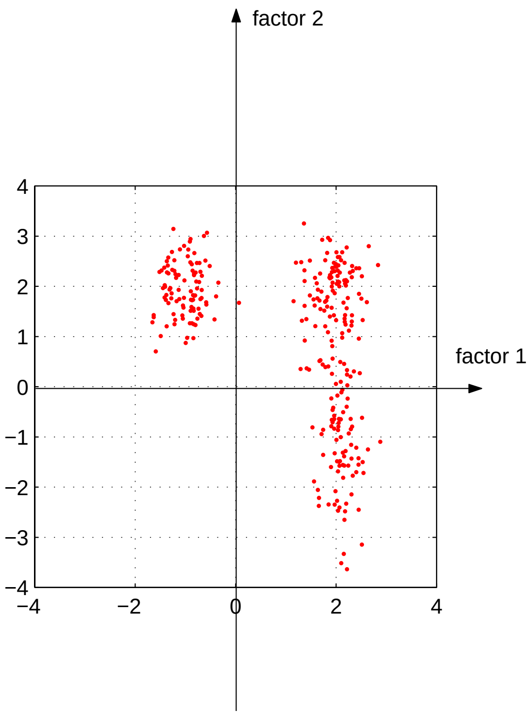
- a_{i*} = how the data variable i is weighted across/influenced by the list of current factors.
  - ◇ $\|a_{i*}\|_2^2$ = the *communality* of variable i.
    - ▷ If $\|a_{i*}\|_2$ is small, variable i is of little consequence to the current factors.
- a_{*k} = correlations of the data variables with that particular kth factor.
  - ◇ $\|a_{*k}\|$ = the *significance* of factor k.
    - ▷ Variables with high factor loadings are more "like" the factor in some sense.
    - ▷ Variables with zero or near-zero loadings are treated as being unlike the factor.

$$\begin{bmatrix} & & & y_{1j} & & & \\ & & & \vdots & & & \\ y_{i1} & \cdots & \cdots & y_{ij} & \cdots & \cdots & y_{i\ell} \\ & & & \vdots & & & \\ & & & y_{nj} & & & \end{bmatrix} = \underbrace{\begin{bmatrix} & & a_{1k} & & \\ & & \vdots & & \\ a_{i1} & \cdots & a_{ik} & \cdots & a_{im} \\ & & \vdots & & \\ & & a_{nk} & & \end{bmatrix}}_{factors} \begin{bmatrix} & & & f_{1j} & & & \\ & & & \vdots & & & \\ f_{k1} & \cdots & \cdots & f_{kj} & \cdots & \cdots & f_{k\ell} \\ & & & \vdots & & & \\ & & & f_{mj} & & & \end{bmatrix}$$

Tasks to Do in Factor Analysis

- Want to rewrite the loadings of variables over some *newly selected* factors.
  - Fewer factors.
  - Manifest more clearly the correlation between variables and factors.
- Represent the loading of each variable (each row of A) as a single point in the factor space $\mathbb{R}^m$.
  - What does it mean if these points cluster around a certain direction?
  - How to find the clustering direction?





What Is Going On?

- Suppose the newly selected factors are expressed in terms of columns of the orthogonal matrix

$$V := [\mathbf{v}_1, \ldots, \mathbf{v}_m] \in \mathbb{R}^{m\times m}. \tag{10}$$

  - ◇ Determine some appropriate new basis for V.
  - ◇ The rewriting of factor loadings with respect to V is mathematically equivalent to a change of basis, i.e., A is now written as $B := AV$.
  - ◇ Because $Y = AF = (AV)(V^T F) = BG$,
    - ▷ $B = AV$ denotes new factor loadings.
    - ▷ $G = V^T F$ denotes new factor scores.
  - ◇ The correlation matrix $R = AA^T = BB^T \in \mathbb{R}^{n\times n}$ is independent of factors selected.
    - ▷ Would like that the significance of factors concentrates on "fewer" columns of B.
    - ▷ Lower rank approximation of A.
- In the process of defining new factors it is often desirable to retrieve information ...
  - ◇ Directly from the correlation matrix R rather than from any particular loading matrix A, if A is not readily available; or
  - ◇ Approximate A, if A is too large or too expensive.

Example 2: Latent Semantic Indexing

- Indexing matrix $H = [h_{ik}] \in \mathbb{R}^{n \times m}$:
  - ◇ Each document is represented by one row in H.
  - ◇ h_{ik} = the *weight* of one particular *term* k in document i.
    - ▷ Each term could be just one single word or a string of phrases.
    - ▷ The weight h_{ik} could simply be the number of occurrence of term k in document i.
    - ▷ More elaborate weighting schemes are available and yield better performance.

$$h_{ik} = t_{ik} g_k n_i,$$

$$\text{documents} \longrightarrow \begin{bmatrix} & & \overset{\text{terms}}{\overset{\downarrow}{h_{1k}}} & & \\ & & \vdots & & \\ h_{i1} & \ldots & \boxed{h_{ik}} & \ldots & h_{im} \\ & & \vdots & & \\ & & h_{nk} & & \end{bmatrix}$$

- Watch out the constraint: Each row should be normalized to unit length.

Search Similarities

- Queries $\mathbf{q}_j = [q_{1j}, \ldots, q_{mj}]^T \in \mathbb{R}^m$:
  - ◇ q_{kj} = the weight of term k in the query j.
- Would like to find documents relevant to given queries.
  - ◇ To measure how the query $\mathbf{q}_j$ matches the documents,
    - ▷ Calculate the dots product

$$\mathbf{s}_j = H\mathbf{q}_j. \tag{11}$$

    - ▷ Rank the relevance of documents to $\mathbf{q}_j$ according to the *scores* in $\mathbf{s}_j$.

Comparison of LSI with Linear Model

- Analogies:

$$
\begin{aligned}
\text{indexing matrix } H &\longleftrightarrow \text{loading matrix } A \\
\text{document } i &\longleftrightarrow \text{variable } i \\
\text{one query } \mathbf{q}_j &\longleftrightarrow \text{one column in scoring matrix } F \\
\text{term } k &\longleftrightarrow \text{factor } k \\
\text{weight } h_{ik} &\longleftrightarrow \text{loading of factor } k \text{ on variable } i \\
\text{weights } q_{kj} \text{ of term } k \text{ in query } \mathbf{q}_j &\longleftrightarrow \text{response } f_{kj} \text{ of entity } j \text{ on factor } k \\
\text{scores } s_{ij} \text{ in of document } i \text{ in query } \mathbf{q}_j &\longleftrightarrow \text{scores } y_{ij} \text{ of variable } i \text{ in entity } j
\end{aligned}
$$

- Differences:
  - ◇ In LSI, terms/factors are predetermined.
    - ▷ How are the terms/factors predetermined?
    - ▷ What is the notion of “orthogonal words”?
    - ▷ What is the notion of “term/factor reduction”?
  - ◇ LSI is not trying to compute factors based on the scores in $\mathbf{s}_j$, $j = 1, \ldots, \ell$.
    - ▷ Though, this information may be used as a learning process for selecting terms/factors.
  - ◇ LSI emphasizes effective vector-matrix multiplication (11).
    - ▷ Want to represent the indexing matrix and the queries in a more *compact form* so as to facilitate the computation of the scores.

Example 3: Electronic Model Design

- Consider the switch manufacturing in the telecommunication industry.
  - ⋄ A cabinet consists of m slots.
  - ⋄ Each slot may be filled with a selection from r types of boards.

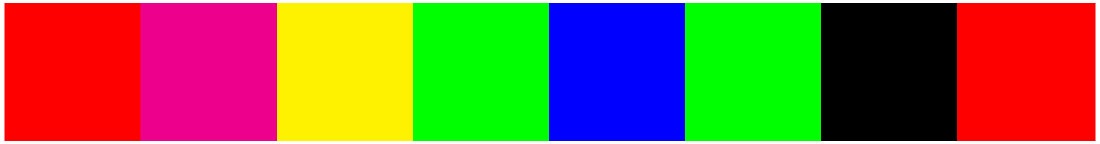


Figure 5: An electronic cabinet with 8 slots to be filled with 6 boards.

- History of past n customer orders have been recorded into a matrix $A \in \mathbb{R}^{n \times m}$.
- Would like to preassemble q semi-finished cabinet models.
- Determine the model configurations and the corresponding customer-to-model assignment of semi-finished cabinets based on A so as to minimize the total number of insertions required to manufacture the entire order.

Example 4: Receptor Model

- An observational technique within the air pollution research community.
- Make use of the ambient data and source profile data to apportion sources or source categories.
- Fundamental principle:
  - ◇ Mass conservation can be assumed.
  - ◇ Mass balance analysis can identify and apportion sources of airborne particulate matter in the atmosphere.

Ideas

- Obtain a large number of chemical constituents such as elemental concentrations in a number of samples.
- Mass balance equation describes the relationships between p sources which contribute m chemical species to n samples.

$$y_{ij} = \sum_{k=1}^{p} a_{ik} f_{kj}, \tag{12}$$

  - ◇ y_{ij} = the elemental concentration of the ith chemical measured in the jth sample.
  - ◇ a_{ik} = the gravimetric concentration of the ith chemical in the kth source.
  - ◇ f_{kj} = the airborne mass concentration that the kth source has contributed to the jth sample.
- Application:
  - ◇ Typically, only values of y_{ij} are observable.
  - ◇ Neither the sources are known nor the compositions of the local particulate emissions are measured.
  - ◇ A critical question is to estimate the number p, the compositions a_{ik}, and the contributions f_{kj} of the sources.
- The source compositions a_{ik} and the source contributions f_{kj} must all be nonnegative. The identification and apportionment, therefore, becomes a nonnegative matrix factorization problem of Y.

Types of Structures

- Structural constraints:
  - ◇ Toeplitz or circulant matrices.
  - ◇ Covariance matrices.
  - ◇ Euclidean distance matrices.
  - ◇ Data on the unit sphere.
  - ◇ Nonnegative matrix factorizations.
- Applications:
  - ◇ Signal and image processing with Toeplitz structure.
  - ◇ Model reduction problem in speech encoding and filter design with Hankel structure.
  - ◇ Protein folding problem with Euclidean distance structure.
  - ◇ Data mining with normalized information.
  - ◇ Principal component analysis with nonnegative structure.
- Each of the above topics is a major research effort.

Difficulties

- No easy way to characterize, either algebraically or analytically, a given class of structured lower rank matrices.
- Lack of explicit description of the feasible set $\Longrightarrow$ Difficult to apply classical optimization techniques.
- Little discussion on whether lower rank matrices with specified structure actually exist.