Independent component analysis and non-negative matrix factorization

David J. Hessen

Utrecht University

Academic year 2024-2025

Applications

The analysis of

- educational or psychological test data
- EEG data
- trading prices of stocks

The data

The data matrix (the scores of N cases on k features)

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Nk} \end{bmatrix} = [\, \mathbf{x}_1 \; \mathbf{x}_2 \; \dots \; \mathbf{x}_N \,]^T$$

The vector of feature means is $\bar{\mathbf{x}} = [\, \bar{x}_1 \; \bar{x}_2 \; \dots \; \bar{x}_k \,]^T$

The data

Centered data: $\mathbf{c}_i = [\, c_{i1} \; c_{i2} \; \ldots \; c_{ik} \,]^T = \mathbf{x}_i - \bar{\mathbf{x}}$, for $i = 1, \ldots, N$

The matrix of centered data is

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \ldots & c_{1k} \\ c_{21} & c_{22} & \ldots & c_{2k} \\ \vdots & \vdots & & \vdots \\ c_{N1} & c_{N2} & \ldots & c_{Nk} \end{bmatrix} = [\, \mathbf{c}_1 \; \mathbf{c}_2 \; \ldots \; \mathbf{c}_N \,]^T = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T$$

The classical factor model

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ are the sample observations of random variables

$$
\begin{aligned}
X_1 &= \mu_1 + \lambda_{11} F_1 + \ldots + \lambda_{1q} F_q + U_1 \\
X_2 &= \mu_2 + \lambda_{21} F_1 + \ldots + \lambda_{2q} F_q + U_2 \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
X_k &= \mu_k + \lambda_{k1} F_1 + \ldots + \lambda_{kq} F_q + U_k
\end{aligned}
$$

where

- $F_1, \ldots, F_q$ are q common factors
- $U_1, \ldots, U_k$ are k unique factors
- $\mu_1, \ldots, \mu_k$ are k intercepts
- $\lambda_{11}, \ldots, \lambda_{kq}$ are factor loadings (regression slopes)

The classical factor model

Let $\boldsymbol{X} = [\, X_1 \, \ldots \, X_k \,]^T$, $\boldsymbol{F} = [\, F_1 \, \ldots \, F_q \,]^T$, and $\boldsymbol{U} = [\, U_1 \, \ldots \, U_k \,]^T$

Then, in terms of matrices

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda F} + \boldsymbol{U}$$

where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & \ldots & \lambda_{1q} \\ \vdots & & \vdots \\ \lambda_{k1} & \ldots & \lambda_{kq} \end{bmatrix}$$

The classical factor model

The elements of $\boldsymbol{F}$ and $\boldsymbol{U}$ are latent random variables

Scales must be assigned to all factors

The means of $F_1, \ldots, F_q$ are set to zero and the variances are set to one

The means of $U_1, \ldots, U_k$ are set to zero

Consequently, the means of $X_1, \ldots, X_k$ are $\mu_1, \ldots, \mu_k$ (the intercepts)

The classical factor model

It is assumed that all factors (common and unique) are uncorrelated

Consequently, the covariance matrix of $X_1, \ldots, X_k$ is

$$cov(\boldsymbol{X}) = \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Theta}$$

where $\boldsymbol{\Theta} = cov(\boldsymbol{U}) = \text{diag}\{var(U_1), \ldots, var(U_k)\}$

The diagonal elements of $\boldsymbol{\Theta}$ are called uniquenesses

The classical factor model

From

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda F} + \boldsymbol{U}$$

it follows that

$$\boldsymbol{C} = \boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{\Lambda F} + \boldsymbol{U}$$

where $\boldsymbol{C} = [\, C_1 \ \ldots \ C_k \,]^T$ is the random vector of centered features

The sample estimate of $\boldsymbol{\mu}$ is $\bar{\mathbf{x}}$

The covariance matrix of $C_1, \ldots, C_k$ is also

$$cov(\boldsymbol{C}) = \boldsymbol{\Sigma} = \boldsymbol{\Lambda\Lambda}^T + \boldsymbol{\Theta}$$

The classical factor model

Non-uniqueness due to rotational indeterminacy

Note that

$$\begin{aligned}\boldsymbol{\Sigma} &= \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Theta} \\ &= \boldsymbol{\Lambda}\mathbf{B}^T\mathbf{B}\boldsymbol{\Lambda}^T + \boldsymbol{\Theta} \quad \text{where } \mathbf{B}^T\mathbf{B} = \mathbf{I} \\ &= \boldsymbol{\Lambda}\mathbf{B}^T(\boldsymbol{\Lambda}\mathbf{B}^T)^T + \boldsymbol{\Theta} \\ &= \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*T} + \boldsymbol{\Theta}\end{aligned}$$

where $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}B^T$

Rotational indeterminacy is usually solved by constraining

$$\boldsymbol{\Lambda}^T\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}$$

to be diagonal

The normal factor model

$\boldsymbol{X}$ is assumed to have a multivariate normal distribution ($\boldsymbol{X}$ is Gaussian)

As a consequence,

- The elements of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$ can be estimated using maximum likelihood
- The fit of the model to data can be tested using a large sample chi-square goodness of fit test
- Fit indices can be calculated

Maximum likelihood estimation

Assuming $\boldsymbol{X}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, for all i, the log-likelihood function is

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\tfrac{N}{2}\big\{k\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}_N) + (\bar{\mathbf{x}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})\big\}$$

where $\mathbf{S}_N = (N-1)\mathbf{S}/N$ and $\mathbf{S}$ is the sample covariance matrix

The unconstrained log-likelihood function is maximized by $\bar{\mathbf{x}}$ and $\mathbf{S}_N$

Under the common factor model, $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Theta}$

$l(\bar{\mathbf{x}}, \boldsymbol{\Lambda}, \boldsymbol{\Theta})$ is maximized w.r.t. $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$ subject to diagonal $\boldsymbol{\Lambda}^T\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}$

Goodness of fit

Let $\hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{\Theta}}$ be the ML estimates of $\mathbf{\Lambda}$ and $\mathbf{\Theta}$, respectively

A likelihood ratio test

If $\boldsymbol{X}_i \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$, for all i, where $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Theta}$, then

$$X^2 = 2\left\{ l(\bar{\mathbf{x}}, \mathbf{S}_N) - l(\bar{\mathbf{x}}, \hat{\mathbf{\Lambda}}, \hat{\mathbf{\Theta}}) \right\} \xrightarrow{L} \chi^2_{df}$$

where $df = k(k-1)/2 + q(q-1)/2 - kq$

Problem: the power to reject a model close to the true data generating process, increases with sample size

Fit indices (TLI, CFI, RMSEA, etc.) can be used to check the badness of fit

Principal components

The random vector of k features is given by

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda F}$$

where $\boldsymbol{F} = [\, F_1 \, \ldots \, F_k \,]^T$ is a random vector of principal components

$cov(\boldsymbol{X})$ is $\boldsymbol{\Sigma} = \boldsymbol{\Lambda \Psi \Lambda}^T$, where $\boldsymbol{\Lambda}^T = \boldsymbol{\Lambda}^{-1}$ and $\boldsymbol{\Psi} = cov(\boldsymbol{F})$ is diagonal

Under multivariate normality, the ML estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ are given by the eigen-decomposition of $\mathbf{S}_N$

So, if $\mathbf{L}^T \mathbf{S}_N \mathbf{L} = \mathbf{P}$, where $\mathbf{L}^T \mathbf{L} = \mathbf{I}$ and $\mathbf{P}$ is diagonal, then $\mathbf{L}$ and $\mathbf{P}$ are the ML estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$, respectively

Total variance explained

The total variance is defined as: $\text{tr}(\mathbf{\Sigma}) = \sum_{i=1}^{k} \sigma_i^2$

Under the factor model, the proportion of total variance explained is

$$\text{tr}(\mathbf{\Lambda}\mathbf{\Lambda}^T)/\text{tr}(\mathbf{\Sigma}) = \text{tr}(\mathbf{\Lambda}^T\mathbf{\Lambda})/\text{tr}(\mathbf{\Sigma}) = \sum_{r=1}^{q} \boldsymbol{\lambda}_r^T\boldsymbol{\lambda}_r/\text{tr}(\mathbf{\Sigma})$$

Then, the proportion explained by the rth factor is $\boldsymbol{\lambda}_r^T\boldsymbol{\lambda}_r/\text{tr}(\mathbf{\Sigma})$

Factor scores

Estimators of $\boldsymbol{F}$

- Thurstone's factor scores

$$\mathbf{f}_T = \boldsymbol{\Lambda}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu})$$

- Bartlett's factor scores

$$\mathbf{f}_B = (\boldsymbol{\Lambda}^T \boldsymbol{\Theta}^{-1} \boldsymbol{\Lambda})^{-1} \boldsymbol{\Lambda}^T \boldsymbol{\Theta}^{-1} (\boldsymbol{X} - \boldsymbol{\mu})$$

- McDonald's factor scores

$$\mathbf{f}_M = \left(\boldsymbol{\Lambda}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda}\right)^{-1/2} \boldsymbol{\Lambda}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu})$$

Rotation

The rotated matrix of factor loadings is given by

$$\bar{\mathbf{\Lambda}} = \hat{\mathbf{\Lambda}}\mathbf{M}$$

where $\hat{\mathbf{\Lambda}}$ is the initial estimate of $\mathbf{\Lambda}$ and $\mathbf{M}$ is an invertible rotation matrix

In the case of orthogonal rotation, $\mathbf{M}^{-1} = \mathbf{M}^T$

- orthogonal: varimax
- oblique: promax or oblimin

Rotation

Rotation to a position as close as possible to simple structure

Simple structure $\rightarrow$ exactly $q-1$ elements of each row of $\mathbf{\Lambda}$ are zero

An example of 9 features and 3 common factors

$$\begin{bmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ \lambda_{31} & 0 & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & \lambda_{52} & 0 \\ 0 & \lambda_{62} & 0 \\ 0 & 0 & \lambda_{73} \\ 0 & 0 & \lambda_{83} \\ 0 & 0 & \lambda_{93} \end{bmatrix}$$

Independent components

$F_1, \ldots, F_q$ are assumed to be mutually independent

Independence implies that $cov(\boldsymbol{F}) = \boldsymbol{\Psi}$ is a diagonal matrix

However, a diagonal $cov(\boldsymbol{F})$ does not imply independence

$F_1, \ldots, F_q$ are independent if and only if their joint density is

$$g(f_1, \ldots, f_q) = \prod_{r=1}^{q} g_r(f_r)$$

Solution: orthogonal rotation to a position with independent factors

The classical *cocktail party problem*

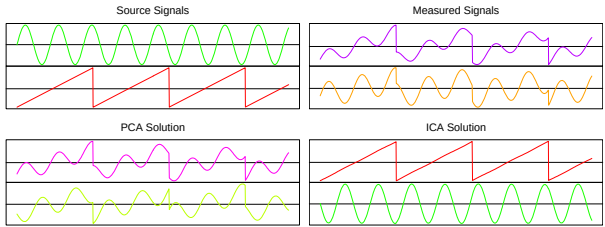

FIGURE 14.37. *Illustration of ICA vs. PCA on artificial time-series data. The upper left panel shows the two source signals, measured at* 1000 *uniformly spaced time points. The upper right panel shows the observed mixed signals. The lower two panels show the principal components and independent component solutions.*

The classical *cocktail party problem*

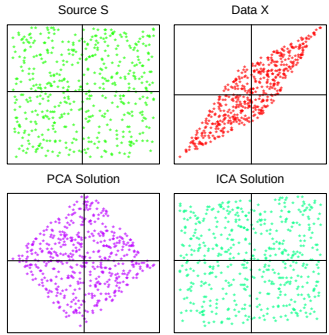

FIGURE 14.38. *Mixtures of independent uniform random variables. The upper left panel shows* 500 *realizations from the two independent uniform sources, the upper right panel their mixed versions. The lower two panels show the PCA and ICA solutions, respectively.*

Example: Places Rated

In the Places Rated Almanac, Boyer and Savageau rated 329 communities according to the following nine criteria:

1. Climate and Terrain
2. Housing
3. Health Care & the Environment
4. Crime
5. Transportation
6. Education
7. The Arts
8. Recreation
9. Economics

Note! Within the dataset, except for housing and crime, the higher the score the better. For housing and crime, the lower the score the better. Where some communities might rate better in the arts, other communities might rate better in other areas such as having a lower crime rate and good educational opportunities.

Example: Places Rated

Assuming normality, a five-factormodel is fitted to the data

Uniquenesses:

V1	V2	V3	V4	V5	V6	V7	V8	V9
0.737	0.005	0.005	0.411	0.005	0.691	0.206	0.646	0.005

Loadings:

	Factor1	Factor2	Factor3	Factor4	Factor5
V1	0.070	0.450	-0.150	-0.024	0.183
V2	0.250	0.926	0.239	0.122	-0.049
V3	0.939	0.240	-0.025	0.105	0.212
V4	0.125	0.102	0.130	0.104	0.731
V5	0.332	0.094	-0.002	0.916	0.191
V6	0.509	0.023	0.100	0.191	-0.053
V7	0.753	0.291	-0.034	0.131	0.351
V8	0.146	0.380	0.095	0.232	0.355
V9	0.054	0.030	0.982	0.010	0.166

Example: Places Rated

```
                       Factor1   Factor2   Factor3   Factor4   Factor5
SS loadings              1.925     1.368     1.081     0.985     0.931
Proportion Var           0.214     0.152     0.120     0.109     0.103
Cumulative Var           0.214     0.366     0.486     0.595     0.699

Test of the hypothesis that 5 factors are sufficient.
The chi square statistic is 5.59 on 1 degree of freedom.
The p-value is 0.018
```

Example: Places Rated

ICA with 1, 2, 3, 4, 5, and 6 components

# of components	% of explained var
1	0.753
2	0.889
3	0.939
4	0.973
5	0.987
6	0.995

Example: Places Rated

ICA with 3 components

Correlations

	component		
feature	1	2	3
1	-0.348	-0.156	-0.030
2	-0.952	-0.218	-0.197
3	-0.238	-0.810	-0.250
4	-0.040	-0.352	-0.220
5	-0.018	-0.287	-0.954
6	-0.084	-0.325	-0.256
7	-0.212	-0.958	-0.194
8	-0.341	-0.246	-0.367
9	-0.344	0.020	-0.084

The data matrix

Suppose all elements of the $N \times p$ data matrix $\mathbf{X}$ are non-negative

- counts or frequencies
- financial data or stock market data

Non-negative $\mathbf{X}$ is approximated by $\mathbf{WH}$, so

$$\mathbf{X} \approx \mathbf{WH}$$

where $\mathbf{W}$ is non-negative and $N \times r$ and $\mathbf{H}$ is non-negative and $r \times p$

r can be significantly less than both N and p

Compared to PCA, non-negative matrix factorization (NMF) extracts sparse and easily interpretable components (purely additive)

Interpretation

Basis features matrix $\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1r} \\ w_{21} & w_{22} & \dots & w_{2r} \\ \vdots & \vdots & & \vdots \\ w_{N1} & w_{N2} & \dots & w_{Nr} \end{bmatrix}$

Entry w_{ik} is the coordinate (position) of case i along the kth dimension

Coefficients (loadings) matrix $\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1p} \\ h_{21} & h_{22} & \dots & h_{2p} \\ \vdots & \vdots & & \vdots \\ h_{r1} & h_{r2} & \dots & h_{rp} \end{bmatrix}$

h_{kj} is the contribution of the kth basis feature to the jth column of $\mathbf{X}$

Interpretation

Let $\mathbf{H} = [\,\mathbf{h}_1\ \mathbf{h}_2\ \ldots\ \mathbf{h}_p\,]$

$\mathbf{X}$ is approximated by

$$\mathbf{WH} = [\,\mathbf{Wh}_1\ \mathbf{Wh}_2\ \ldots\ \mathbf{Wh}_p\,]$$

So the jth column of $\mathbf{X}$ is approximated $\mathbf{Wh}_j$, that is

$$\begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{Nj} \end{bmatrix} \approx h_{1j} \begin{bmatrix} w_{11} \\ w_{21} \\ \vdots \\ w_{N1} \end{bmatrix} + h_{2j} \begin{bmatrix} w_{12} \\ w_{22} \\ \vdots \\ w_{N2} \end{bmatrix} + \ldots + h_{rj} \begin{bmatrix} w_{1r} \\ w_{2r} \\ \vdots \\ w_{Nr} \end{bmatrix}$$

Matrix product

Let $\mathbf{W} = [\,\mathbf{w}_1\ \mathbf{w}_2\ \ldots\ \mathbf{w}_N\,]^T$ and $\mathbf{H} = [\,\mathbf{h}_1\ \mathbf{h}_2\ \ldots\ \mathbf{h}_p\,]$

$$\mathbf{WH} = \begin{bmatrix} \mathbf{w}_1^T\mathbf{h}_1 & \mathbf{w}_1^T\mathbf{h}_2 & \ldots & \mathbf{w}_1^T\mathbf{h}_p \\ \mathbf{w}_2^T\mathbf{h}_1 & \mathbf{w}_2^T\mathbf{h}_2 & \ldots & \mathbf{w}_2^T\mathbf{h}_p \\ \vdots & \vdots & & \vdots \\ \mathbf{w}_N^T\mathbf{h}_1 & \mathbf{w}_N^T\mathbf{h}_2 & \ldots & \mathbf{w}_N^T\mathbf{h}_p \end{bmatrix}$$

The entry on row i and column j is the dot product

$$\mathbf{w}_i^T\mathbf{h}_j = \sum_{k=1}^{r} w_{ik}h_{kj} = w_{i1}h_{1j} + w_{i2}h_{2j} + \ldots + h_{ir}w_{rj} = (\mathbf{WH})_{ij}$$

Maximum likelihood

If the entries of $\mathbf{X}$ are counts or frequencies, then they can be assumed to be the observations of $N \times p$ independent Poisson random variables with means $\mathbf{w}_1^T\mathbf{h}_1, \ldots, \mathbf{w}_N^T\mathbf{h}_p$

The log-likelihood function

$$l(\mathbf{W}, \mathbf{H}) = \ln L(\mathbf{W}, \mathbf{H}) = \ln \left\{ \prod_{i=1}^{N} \prod_{j=1}^{p} \frac{(\mathbf{w}_i^T\mathbf{h}_j)^{x_{ij}} \exp(-\mathbf{w}_i^T\mathbf{h}_j)}{x_{ij}!} \right\}$$

is maximized with respect to $\mathbf{W}$ and $\mathbf{H}$ subject to $\mathbf{W} \geq 0$ and $\mathbf{H} \geq 0$

Least squares

The sum of squares

$$\sum_{i=1}^{N}\sum_{j=1}^{p}\{x_{ij} - (\mathbf{WH})_{ij}\}^2$$

is minimized with respect to $\mathbf{W}$ and $\mathbf{H}$ subject to $\mathbf{W} \geq 0$ and $\mathbf{H} \geq 0$

Algorithm

Lee and Seung (2001)

1. Non-negative starting values h_{kj}^0 and w_{ik}^0 are chosen for all i, j, and $k = 1, \ldots, r$
2. The value for w_{ik}, for all i and k, is updated by

$$w_{ik}^{n+1} = w_{ik}^n \frac{\sum_{j=1}^p h_{kj}^n x_{ij} / (\mathbf{W}^n \mathbf{H}^n)_{ij}}{\sum_{j=1}^p h_{kj}^n}, \quad \text{for all } i \text{ and } k$$

3. The value for h_{kj}, for all j and k, is updated by

$$h_{kj}^{n+1} = h_{kj}^n \frac{\sum_{i=1}^N w_{ik}^n x_{ij} / (\mathbf{W}^n \mathbf{H}^n)_{ij}}{\sum_{i=1}^N w_{ik}^n}, \quad \text{for all } j \text{ and } k$$

4. Steps 2 and 3 are repeated iteratively until the updates have converged

Non-uniqueness

Even if $\mathbf{X} = \mathbf{WH}$ exactly, the decomposition may not be unique

To see this, note that

$$\mathbf{WH} = \mathbf{WBB}^{-1}\mathbf{H} = \mathbf{W}^*\mathbf{H}^*$$

where $\mathbf{W}^* = \mathbf{WB}$ and $\mathbf{H}^* = \mathbf{B}^{-1}\mathbf{H}$

If $\mathbf{W}^* \geq 0$ and $\mathbf{H}^* \geq 0$, then they form another parametrization of the factorization

The solution found by the algorithm depends on the starting values

Non-uniqueness

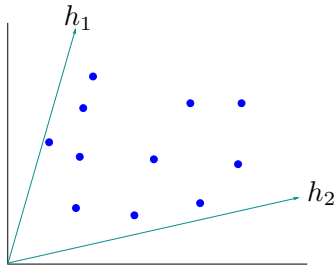

FIGURE 14.34. *Non-uniqueness of the non-negative matrix factorization. There are 11 data points in two dimensions. Any choice of the basis vectors h_1 and h_2 in the open space between the coordinate axes and data, gives an exact reconstruction of the data.*

Reconstruction accuracy

The reconstruction error is given by

$$\sum_{i=1}^{N}\sum_{j=1}^{p}\{x_{ij} - (\mathbf{WH})_{ij}\}^2 = \sum_{i=1}^{N}\sum_{j=1}^{p}x_{ij}^2 - 2\sum_{i=1}^{N}\sum_{j=1}^{p}x_{ij}(\mathbf{WH})_{ij} + \sum_{i=1}^{N}\sum_{j=1}^{p}(\mathbf{WH})_{ij}^2$$

Let $\mathbf{x} = \text{vec}(\mathbf{X})$ and $\hat{\mathbf{x}} = \text{vec}(\mathbf{WH})$, then

$$\sum_{i=1}^{N}\sum_{j=1}^{p}\{x_{ij} - (\mathbf{WH})_{ij}\}^2 = \mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\hat{\mathbf{x}} + \hat{\mathbf{x}}^T\hat{\mathbf{x}}$$

In general: $\mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\hat{\mathbf{x}} + \hat{\mathbf{x}}^T\hat{\mathbf{x}} \geq 0$

Proportion of explained variance: $\dfrac{2\mathbf{x}^T\hat{\mathbf{x}} - \hat{\mathbf{x}}^T\hat{\mathbf{x}}}{\mathbf{x}^T\mathbf{x}} \leq 1$

Text mining example (cats and cars)

A document-term matrix

	lion	tiger	cheetah	jaguar	porsche	ferrari
document 1	2	2	1	2	0	0
document 2	2	3	3	3	0	0
document 3	1	1	1	1	0	0
document 4	2	2	2	3	1	1
document 5	0	0	0	1	1	1
document 6	0	0	0	2	1	2

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

Text mining example (cats and cars)

Rank 1 factorization

$$\mathbf{WH} = \begin{bmatrix} 1.5 \\ 2.3 \\ 0.8 \\ 2.3 \\ 0.6 \\ 1.1 \end{bmatrix} [0.8\ 0.9\ 0.8\ 1.4\ 0.3\ 0.5] = \begin{bmatrix} 1.2 & 1.4 & 1.2 & 2.0 & 0.5 & 0.7 \\ 1.9 & 2.1 & 1.9 & 3.2 & 0.8 & 1.1 \\ 0.7 & 0.8 & 0.7 & 1.2 & 0.3 & 0.4 \\ 1.9 & 2.1 & 1.9 & 3.2 & 0.8 & 1.1 \\ 0.5 & 0.6 & 0.5 & 0.9 & 0.2 & 0.3 \\ 0.9 & 1.0 & 0.9 & 1.5 & 0.4 & 0.5 \end{bmatrix}$$

Explained variance: 0.83

Text mining example (cats and cars)

Rank 2 factorization

$$\mathbf{WH} = \begin{bmatrix} 0.0 & 2.4 \\ 0.0 & 3.8 \\ 0.0 & 1.4 \\ 1.8 & 2.8 \\ 1.8 & 0.0 \\ 3.0 & 0.0 \end{bmatrix} \begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.6 & 0.5 & 0.6 \\ 0.7 & 0.8 & 0.7 & 0.8 & 0.0 & 0.0 \end{bmatrix} = \begin{bmatrix} 1.6 & 1.9 & 1.6 & 1.9 & 0.0 & 0.0 \\ 2.6 & 2.9 & 2.6 & 2.9 & 0.0 & 0.0 \\ 0.9 & 1.1 & 0.9 & 1.1 & 0.0 & 0.0 \\ 1.9 & 2.1 & 1.9 & 3.2 & 0.8 & 1.1 \\ 0.0 & 0.0 & 0.0 & 1.1 & 0.8 & 1.1 \\ 0.0 & 0.0 & 0.0 & 1.8 & 1.4 & 1.8 \end{bmatrix}$$

Explained variance: 0.98

Probabilistic latent semantic analysis

In text mining, $\mathbf{X}$ is a document-term matrix of counts

Let $x_{11}, \ldots, x_{Np}$ have a multinomial distribution, that is,

$$P(x_{11}, \ldots, x_{Np}) = \frac{m}{\prod_{i=1}^{N} \prod_{j=1}^{p} x_{ij}!} \prod_{i=1}^{N} \prod_{j=1}^{p} \{P(d_i, t_j)\}^{x_{ij}}$$

where $m = \sum_{i=1}^{N} \sum_{j=1}^{p} x_{ij}$ and $P(d_i, t_j)$ is the joint probability of document i and term j

The number of parameters of this saturated model is $Np - 1$

So if $N = 20$ and $p = 1000$, then $Np - 1 = 19999$

Probabilistic latent semantic analysis

Under the saturated model the maximum likelihood estimates of $P(d_1, t_1), \ldots, P(d_N, t_p)$ are

$$\begin{aligned} \tilde{P}(d_1, t_1) &= \frac{x_{11}}{m} \\ \tilde{P}(d_1, t_2) &= \frac{x_{12}}{m} \\ \vdots & \quad \vdots \\ \tilde{P}(d_N, t_p) &= \frac{x_{Np}}{m} \end{aligned}$$

Probabilistic latent semantic analysis

Let $c_1, \ldots, c_r$ be r latent classes, then

$$\begin{aligned} P(d_i, t_j) &= \sum_{k=1}^{r} P(d_i, t_j, c_k) \\ &= \sum_{k=1}^{r} P(d_i, t_j \,|\, c_k) P(c_k) \\ &= \sum_{k=1}^{r} \underbrace{P(d_i \,|\, c_k) P(t_j \,|\, c_k)}_{\text{local independence}} P(c_k) \end{aligned}$$

The number of parameters is reduced to $r(N-1) + r(p-1) + r - 1$, which equals $3 \cdot 19 + 3 \cdot 999 + 2 = 3056$ if $N = 20$, $p = 1000$, and $r = 3$

Probabilistic latent semantic analysis

Maximum likelihood estimation of $P(d_i \,|\, c_k)$, $P(t_j \,|\, c_k)$, and $P(c_k)$, for all i, j, and k

To find the maximum likelihood estimates the function

$$\sum_{i=1}^{N} \sum_{j=1}^{p} x_{ij} \ln \left\{ \sum_{k=1}^{r} P(d_i \,|\, c_k) P(t_j \,|\, c_k) P(c_k) \right\}$$

can be maximized with respect to $P(d_i \,|\, c_k)$, $P(t_j \,|\, c_k)$, and $P(c_k)$, for all i, j, and k, subject to

- $P(d_i \,|\, c_k) \geq 0$, $P(t_j \,|\, c_k) \geq 0$, $P(c_k) \geq 0$, for all i, j, and k
- $\sum_{i=1}^{N} P(d_i \,|\, c_k) = 1$, $\sum_{j=1}^{p} P(t_j \,|\, c_k) = 1$, $\sum_{k=1}^{r} P(c_k) = 1$, for all i, j, k

using the expectation-maximization (EM) algorithm

Probabilistic latent semantic analysis

The probability matrix

$$\mathbf{P} = \begin{bmatrix} P(d_1, t_1) & P(d_1, t_2) & \dots & P(d_1, t_p) \\ P(d_2, t_1) & P(d_2, t_2) & \dots & P(d_2, t_p) \\ \vdots & \vdots & & \vdots \\ P(d_N, t_1) & P(d_N, t_2) & \dots & P(d_N, t_p) \end{bmatrix}$$

is approximated by the three-way factorization

$$\begin{aligned} \hat{\mathbf{P}} &= \begin{bmatrix} \hat{P}(d_1 \,|\, c_1) & \dots & \hat{P}(d_1 \,|\, c_r) \\ \vdots & & \vdots \\ \hat{P}(d_N \,|\, c_1) & \dots & \hat{P}(d_N \,|\, c_r) \end{bmatrix} \begin{bmatrix} \hat{P}(c_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \hat{P}(c_r) \end{bmatrix} \begin{bmatrix} \hat{P}(t_1 \,|\, c_1) & \dots & \hat{P}(t_p \,|\, c_1) \\ \vdots & & \vdots \\ \hat{P}(t_1 \,|\, c_r) & \dots & \hat{P}(t_p \,|\, c_r) \end{bmatrix} \\ &= \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^T \end{aligned}$$

where $\mathbf{U} \geq 0$, $\boldsymbol{\Sigma} \geq 0$, and $\mathbf{V} \geq 0$

Probabilistic latent semantic analysis

Two goodness of fit statistics

$H_0\colon\ r = r_0$ and $H_A\colon\ r \neq r_0$

The Pearson chi-square statistic

$$X^2 = \sum_{i=1}^{N}\sum_{j=1}^{p} \frac{\{x_{ij} - m\hat{P}(d_i, t_j)\}^2}{m\hat{P}(d_i, t_j)}$$

The likelihood ratio statistic

$$G^2 = 2\sum_{i=1}^{N}\sum_{j=1}^{p} x_{ij}\left\{\ln\left(\frac{x_{ij}}{m}\right) - \ln\hat{P}(d_i, t_j)\right\}$$

For both $df = Np - r(N + p - 1)$

Probabilistic latent semantic analysis

The reconstruction error is given by

$$\sum_{i=1}^{N}\sum_{j=1}^{p}\{x_{ij} - m(\hat{\mathbf{P}})_{ij}\}^2 = \sum_{i=1}^{N}\sum_{j=1}^{p} x_{ij}^2 - 2\sum_{i=1}^{N}\sum_{j=1}^{p} x_{ij} m(\hat{\mathbf{P}})_{ij} + \sum_{i=1}^{N}\sum_{j=1}^{p} m^2(\hat{\mathbf{P}})_{ij}^2$$

Let $\mathbf{x} = \text{vec}(\mathbf{X})$ and $\hat{\mathbf{x}} = \text{vec}(m\hat{\mathbf{P}})$, then

$$\sum_{i=1}^{N}\sum_{j=1}^{p}\{x_{ij} - m(\hat{\mathbf{P}})_{ij}\}^2 = \mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\hat{\mathbf{x}} + \hat{\mathbf{x}}^T\hat{\mathbf{x}}$$

In general: $\mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\hat{\mathbf{x}} + \hat{\mathbf{x}}^T\hat{\mathbf{x}} \geq 0$

Proportion of explained variance: $\dfrac{2\mathbf{x}^T\hat{\mathbf{x}} - \hat{\mathbf{x}}^T\hat{\mathbf{x}}}{\mathbf{x}^T\mathbf{x}} \leq 1$

Text mining example (cats and cars)

A document-term matrix

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

Let $m = \sum_{i=1}^{N} \sum_{j=1}^{p} x_{ij} = 41$, then the estimate of $\mathbf{P}$ under the saturated model is

$$\tilde{\mathbf{P}} = \mathbf{X}/m = \begin{bmatrix} .049 & .049 & .024 & .049 & .000 & .000 \\ .049 & .073 & .073 & .073 & .000 & .000 \\ .024 & .024 & .024 & .024 & .000 & .000 \\ .049 & .049 & .049 & .073 & .024 & .024 \\ .000 & .000 & .000 & .024 & .024 & .024 \\ .000 & .000 & .000 & .049 & .024 & .049 \end{bmatrix}$$

If $r = 6$, then $\mathbf{X}/m$ is exactly reconstructed by $\hat{\mathbf{P}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$

Text mining example (cats and cars)

For $r = 2$,

$$\hat{\mathbf{P}} = \underbrace{\begin{bmatrix} .234 & .000 \\ .367 & .000 \\ .133 & .000 \\ .266 & .275 \\ .000 & .272 \\ .000 & .453 \end{bmatrix}}_{\mathbf{U}} \underbrace{\begin{bmatrix} .731 & .000 \\ .000 & .269 \end{bmatrix}}_{\mathbf{\Sigma}} \underbrace{\begin{bmatrix} .234 & .267 & .234 & .266 & .000 & .000 \\ .000 & .000 & .000 & .365 & .272 & .363 \end{bmatrix}}_{\mathbf{V}^T}$$

$$= \begin{bmatrix} .040 & .046 & .040 & .045 & .000 & .000 \\ .063 & .072 & .063 & .071 & .000 & .000 \\ .023 & .026 & .023 & .026 & .000 & .000 \\ .045 & .052 & .045 & .079 & .020 & .027 \\ .000 & .000 & .000 & .027 & .020 & .027 \\ .000 & .000 & .000 & .045 & .033 & .044 \end{bmatrix}$$

Text mining example (cats and cars)

H_0: $r = 2$ and H_A: $r \neq 2$

Pearson's statistic: $X^2 = .8452 \Rightarrow P(\chi^2 > .8452 \,|\, df = 14) > .05$

Likelihood ratio statistic: $G^2 = .8902 \Rightarrow P(\chi^2 > .8902 \,|\, df = 14) > .05$

Not statistically significant, so H_0 cannot be rejected

Proportion of explained variance: .98