Matrix Factorization

DS-GA 1013 / MATH-GA 2824 Optimization-based Data Analysis

http://www.cims.nyu.edu/~cfgranda/pages/OBDA_fall17/index.html

Carlos Fernandez-Granda

Low-rank models

Matrix completion

Structured low-rank models

Motivation

Quantity $y[i, j]$ depends on indices i and j

We observe examples and want to predict new instances

In collaborative filtering, $y[i, j]$ is rating given to a movie i by a user j

Collaborative filtering

$$Y := \begin{pmatrix} 1 & 1 & 5 & 4 \\ 2 & 1 & 4 & 5 \\ 4 & 5 & 2 & 1 \\ 5 & 4 & 2 & 1 \\ 4 & 5 & 1 & 2 \\ 1 & 2 & 5 & 5 \end{pmatrix}$$

	Bob	Molly	Mary	Larry
The Dark Knight	1	1	5	4
Spiderman 3	2	1	4	5
Love Actually	4	5	2	1
Bridget Jones's Diary	5	4	2	1
Pretty Woman	4	5	1	2
Superman 2	1	2	5	5

Simple model

Assumptions:

- Some movies are more popular in general
- Some users are more generous in general

$$y[i,j] \approx a[i]b[j]$$

- $a[i]$ quantifies popularity of movie i
- $b[j]$ quantifies generosity of user j

Simple model

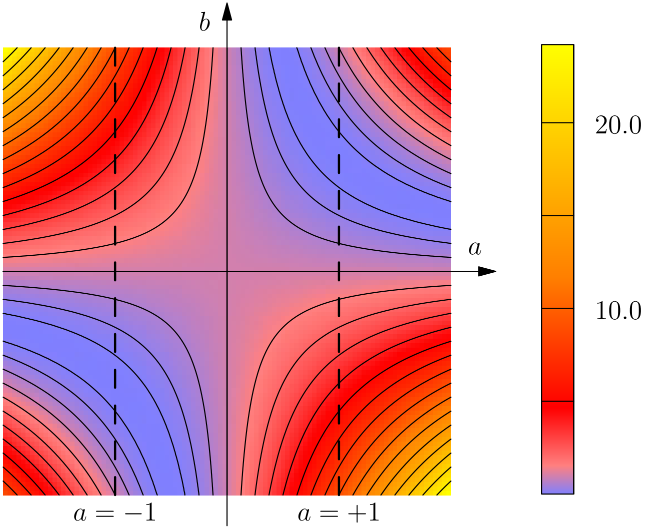
Problem: Fitting a and b to the data yields nonconvex problem

Example: 1 movie, 1 user, rating 1 yields cost function

$$(1 - ab)^2$$

To fix scale set $|a| = 1$

$(1 - ab)^2$



Rank-1 model

Assume m movies are all rated by n users

Model becomes

$$Y \approx \vec{a}\,\vec{b}^{\,T}$$

We can fit it by solving

$$\min_{\vec{a}\in\mathbb{R}^m,\,\vec{b}\in\mathbb{R}^n} \left|\left| Y - \vec{a}\,\vec{b}^{\,T} \right|\right|_{\mathrm{F}} \qquad \text{subject to} \quad ||\vec{a}||_2 = 1$$

Equivalent to

Rank-1 model

Assume m movies are all rated by n users

Model becomes

$$Y \approx \vec{a}\,\vec{b}^{\,T}$$

We can fit it by solving

$$\min_{\vec{a}\in\mathbb{R}^m,\,\vec{b}\in\mathbb{R}^n} \left|\left| Y - \vec{a}\,\vec{b}^{\,T} \right|\right|_{\mathrm{F}} \qquad \text{subject to} \quad \left|\left|\vec{a}\right|\right|_2 = 1$$

Equivalent to

$$\min_{X\in\mathbb{R}^{m\times n}} \left|\left| Y - X \right|\right|_{\mathrm{F}} \qquad \text{subject to} \quad \operatorname{rank}(X) = 1$$

Best rank-k approximation

Let USV^T be the SVD of a matrix $A \in \mathbb{R}^{m \times n}$

The truncated SVD $U_{:,1:k} S_{1:k,1:k} V_{:,1:k}^T$ is the best rank-k approximation

$$U_{:,1:k} S_{1:k,1:k} V_{:,1:k}^T = \arg\min_{\{\widetilde{A} \,\mid\, \text{rank}(\tilde{A})=k\}} \left|\left|A - \widetilde{A}\right|\right|_{\text{F}}$$

Rank-1 model

$$\sigma_1 \vec{u}_1 \vec{v}_1^{\,T} = \arg \min_{X \in \mathbb{R}^{m \times n}} ||Y - X||_{\mathrm{F}} \qquad \text{subject to} \quad \operatorname{rank}(X) = 1$$

The solution to

$$\min_{\vec{a} \in \mathbb{R}^m,\, \vec{b} \in \mathbb{R}^n} \left|\left| Y - \vec{a}\,\vec{b}^{\,T} \right|\right|_{\mathrm{F}} \qquad \text{subject to} \quad ||\vec{a}||_2 = 1$$

is

$$\vec{a}_{\min} =$$
$$\vec{b}_{\min} =$$

Rank-1 model

$$\sigma_1 \vec{u}_1 \vec{v}_1^{\,T} = \arg \min_{X \in \mathbb{R}^{m \times n}} ||Y - X||_{\text{F}} \qquad \text{subject to} \quad \text{rank}\,(X) = 1$$

The solution to

$$\min_{\vec{a} \in \mathbb{R}^m,\, \vec{b} \in \mathbb{R}^n} \left|\left| Y - \vec{a}\,\vec{b}^{\,T} \right|\right|_{\text{F}} \qquad \text{subject to} \quad ||\vec{a}||_2 = 1$$

is

$$\vec{a}_{\text{min}} = \vec{u}_1$$
$$\vec{b}_{\text{min}} = \sigma_1 \vec{v}_1$$

Rank-r model

Certain people like certain movies: r factors

$$y[i,j] \approx \sum_{l=1}^{r} a_l[i] b_l[j]$$

For each factor l

- $a_l[i]$: movie i is positively (> 0), negatively (< 0) or not (≈ 0) associated to factor l
- $b_l[j]$: user j likes (> 0), hates (< 0) or is indifferent (≈ 0) to factor l

Rank-r model

Equivalent to

$$Y \approx AB, \qquad A \in \mathbb{R}^{m \times r}, \quad B \in \mathbb{R}^{r \times n}$$

SVD solves

$$\min_{A \in \mathbb{R}^{m \times r}, B \in \mathbb{R}^{r \times n}} \left|\left| Y - A\,B \right|\right|_{\text{F}} \qquad \text{subject to} \quad \left|\left| \vec{a}_1 \right|\right|_2 = 1, \ldots, \left|\left| \vec{a}_r \right|\right|_2 = 1$$

Problem: Many possible ways of choosing $\vec{a}_1, \ldots, \vec{a}_r, \vec{b}_1, \ldots, \vec{b}_r$

SVD constrains them to be orthogonal

Collaborative filtering

$Y :=$

	Bob	Molly	Mary	Larry
The Dark Knight	1	1	5	4
Spiderman 3	2	1	4	5
Love Actually	4	5	2	1
Bridget Jones's Diary	5	4	2	1
Pretty Woman	4	5	1	2
Superman 2	1	2	5	5

SVD

$$A - \mu\vec{1}\,\vec{1}^{\,T} = USV^T = U \begin{bmatrix} 7.79 & 0 & 0 & 0 \\ 0 & 1.62 & 0 & 0 \\ 0 & 0 & 1.55 & 0 \\ 0 & 0 & 0 & 0.62 \end{bmatrix} V^T$$

$$\mu := \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}$$

Rank 1 model

$$\bar{A} + \sigma_1 \vec{u}_1 \vec{v}_1^T =$$

	Bob	Molly	Mary	Larry	
	1.34 (1)	1.19 (1)	4.66 (5)	4.81 (4)	The Dark Knight
	1.55 (2)	1.42 (1)	4.45 (4)	4.58 (5)	Spiderman 3
	4.45 (4)	4.58 (5)	1.55 (2)	1.42 (1)	Love Actually
	4.43 (5)	4.56 (4)	1.57 (2)	1.44 (1)	B.J.'s Diary
	4.43 (4)	4.56 (5)	1.57 (1)	1.44 (2)	Pretty Woman
	1.34 (1)	1.19 (2)	4.66 (5)	4.81 (5)	Superman 2

Movies

	D. Knight	Sp. 3	Love Act.	B.J.'s Diary	P. Woman	Sup. 2
$\vec{a}_1 =$	$(-0.45$	-0.39	0.39	0.39	0.39	$-0.45)$

Coefficients cluster movies into action (+) and romantic (-)

Users

$$\begin{array}{ccccc} & \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \\ \vec{b}_1 = & (3.74 & 4.05 & -3.74 & -4.05) \end{array}$$

Coefficients cluster people into action (-) and romantic (+)

Low-rank models

Matrix completion

Structured low-rank models

Netflix Prize

Matrix completion

Bob	Molly	Mary	Larry	
1	?	5	4	The Dark Knight
?	1	4	5	Spiderman 3
4	5	2	?	Love Actually
5	4	2	1	Bridget Jones's Diary
4	5	1	2	Pretty Woman
1	2	?	5	Superman 2

Matrix completion as an inverse problem

$$\begin{bmatrix} 1 & ? & 5 \\ ? & 3 & 2 \end{bmatrix}$$

For a fixed sampling pattern, underdetermined system of equations

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Y_{11} \\ Y_{21} \\ Y_{12} \\ Y_{22} \\ Y_{13} \\ Y_{23} \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 5 \\ 2 \end{bmatrix}$$

Isn't this completely ill posed?

Assumption: Matrix is low rank, depends on $\approx r\,(m+n)$ parameters

As long as data $>$ parameters recovery is possible (in principle)

$$\begin{bmatrix} 1 & 1 & 1 & 1 & ? & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ ? & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Matrix cannot be sparse

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 23 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Singular vectors cannot be sparse

$$\begin{bmatrix}1\\1\\1\\1\end{bmatrix}\begin{bmatrix}1 & 1 & 1 & 1\end{bmatrix} + \begin{bmatrix}0\\0\\0\\1\end{bmatrix}\begin{bmatrix}1 & 2 & 3 & 4\end{bmatrix} = \begin{bmatrix}1 & 1 & 1 & 1\\1 & 1 & 1 & 1\\1 & 1 & 1 & 1\\2 & 3 & 4 & 5\end{bmatrix}$$

Incoherence

The matrix must be incoherent: its singular vectors must be spread out

For $1/\sqrt{n} \leq \mu \leq 1$

$$\max_{1 \leq i \leq r, 1 \leq j \leq m} |U_{ij}| \leq \mu$$

$$\max_{1 \leq i \leq r, 1 \leq j \leq n} |V_{ij}| \leq \mu$$

for the left $U_1, \ldots, U_r$ and right $V_1, \ldots, V_r$ singular vectors

Measurements

We must see an entry in each row/column at least

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ ? & ? & ? & ? \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 \\ ? \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$$

Assumption: Random sampling (usually does not hold in practice!)

Low-rank matrix estimation

First idea:

$$\min_{X \in \mathbb{R}^{m \times n}} \operatorname{rank}(X) \quad \text{such that } X_\Omega = y$$

Ω: indices of revealed entries
y: revealed entries

Computationally intractable because of missing entries

Tractable alternative:

$$\min_{X \in \mathbb{R}^{m \times n}} ||X||_* \quad \text{such that } X_\Omega = y$$

Exact recovery

Guarantees by Gross 2011, Candès and Recht 2008, Candès and Tao 2009

$$\min_{X \in \mathbb{R}^{m \times n}} ||X||_* \quad \text{such that } X_\Omega = y$$

achieves exact recovery with high probability as long as the number of samples is proportional to $r\,(n+m)$ up to log terms

The proof is based on the construction of a dual certificate

Low-rank matrix estimation

If data are noisy

$$\min_{X \in \mathbb{R}^{m \times n}} \left|\left|X_\Omega - \vec{y}\right|\right|_2^2 + \lambda \left|\left|X\right|\right|_*$$

where $\lambda > 0$ is a regularization parameter

Matrix completion via nuclear-norm minimization

Bob	Molly	Mary	Larry	
1	2 (1)	5	4	The Dark Knight
2 (2)	1	4	5	Spiderman 3
4	5	2	2 (1)	Love Actually
5	4	2	1	Bridget Jones's Diary
4	5	1	2	Pretty Woman
1	2	5 (5)	5	Superman 2

Proximal gradient method

Method to solve the optimization problem

$$\text{minimize} \quad f(\vec{x}) + h(\vec{x}),$$

where f is differentiable and prox_h is tractable

Proximal-gradient iteration:

$$\vec{x}^{(0)} = \text{arbitrary initialization}$$
$$\vec{x}^{(k+1)} = \text{prox}_{\alpha_k h}\left(\vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)\right)$$

Proximal operator of nuclear norm

The solution X to

$$\min_{X \in \mathbb{R}^{m \times n}} \frac{1}{2} \left|\left| Y - X \right|\right|_{\mathrm{F}}^2 + \tau \left|\left| X \right|\right|_*$$

is obtained by soft-thresholding the SVD of Y

$$X_{\text{prox}} = \mathcal{D}_\tau \left(Y \right)$$

$$\mathcal{D}_\tau \left(M \right) := U \, \mathcal{S}_\tau \left(S \right) V^T \qquad \text{where } M = U \, S V^T$$

$$\mathcal{S}_\tau \left(S \right)_{ii} := \begin{cases} S_{ii} - \tau & \text{if } S_{ii} > \tau \\ 0 & \text{otherwise} \end{cases}$$

Subdifferential of the nuclear norm

Let $X \in \mathbb{R}^{m \times n}$ be a rank-r matrix with SVD USV^T, where $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$ and $S \in \mathbb{R}^{r \times r}$

A matrix G is a subgradient of the nuclear norm at X if and only if

$$G := UV^T + W$$

where W satisfies

$$\begin{aligned} ||W|| &\leq 1 \\ U^T W &= 0 \\ W\,V &= 0 \end{aligned}$$

Proximal operator of nuclear norm

The subgradients of

$$\frac{1}{2} \left|\left| Y - X \right|\right|_{\mathsf{F}}^2 + \tau \left|\left| X \right|\right|_*$$

are of the form

$$Y - X + \tau G$$

where G is a subgradient of the nuclear norm at X

$\mathcal{D}_\tau (Y)$ is a minimizer if and only if

$$G = \frac{1}{\tau} (Y - D_\tau (Y))$$

is a subgradient of the nuclear norm at $\mathcal{D}_\tau (Y)$

Proximal operator of nuclear norm

Separate SVD of Y into singular values greater or smaller than τ

$$
\begin{aligned}
Y &= U S V^T \\
&= \begin{bmatrix} U_0 & U_1 \end{bmatrix} \begin{bmatrix} S_0 & 0 \\ 0 & S_1 \end{bmatrix} \begin{bmatrix} V_0 & V_1 \end{bmatrix}^T
\end{aligned}
$$

$\mathcal{D}_\tau(Y) = U_0 (S_0 - \tau I) V_0^T$, so

$$
\frac{1}{\tau}(Y - \mathcal{D}_\tau(Y)) = U_0 V_0^T + \frac{1}{\tau} U_1 S_1 V_1^T
$$

Proximal gradient method

Proximal gradient method for the problem

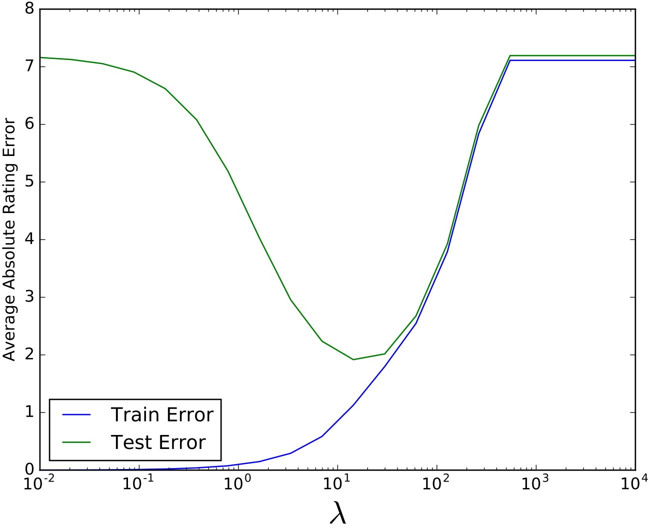
$$\min_{X \in \mathbb{R}^{m \times n}} \left|\left|X_{\Omega} - \vec{y}\right|\right|_2^2 + \lambda \left|\left|X\right|\right|_*$$

$$X^{(0)} = \text{arbitrary initialization}$$
$$M^{(k)} = X^{(k)} - \alpha_k \left(X_{\Omega}^{(k)} - \vec{y} \right)$$
$$X^{(k+1)} = \mathcal{D}_{\alpha_k \lambda} \left(M^{(k)} \right)$$

Real data

- Movielens database
- 671 users
- 300 movies
- Training set: 9 135 ratings
- Test set: 1 016

Real data



Average Absolute Rating Error
8
7
6
5
4
3
2
1
0
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
λ
Train Error
Test Error

Low-rank matrix completion

Intractable problem

$$\min_{X \in \mathbb{R}^{m \times n}} \operatorname{rank}(X) \quad \text{such that } X_\Omega \approx \vec{y}$$

Nuclear norm: convex but computationally expensive
due to SVD computations

Alternative

- Fix rank k beforehand
- Parametrize the matrix as AB where $A \in \mathbb{R}^{m \times r}$ and $B \in \mathbb{R}^{r \times n}$
- Solve

$$\min_{\widetilde{A} \in \mathbb{R}^{m \times r}, \widetilde{B} \in \mathbb{R}^{r \times n}} \left| \left| \left(\widetilde{A}\widetilde{B} \right)_{\Omega} - \vec{y} \right| \right|_2$$

by alternating minimization

Alternating minimization

Sequence of least-squares problems (much faster than computing SVDs)

- To compute $A^{(k)}$ fix $B^{(k-1)}$ and solve

$$\min_{\widetilde{A} \in \mathbb{R}^{m \times r}} \left|\left| \left(\widetilde{A} B^{(k-1)} \right)_{\Omega} - \vec{y} \right|\right|_2$$

- To compute $B^{(k)}$ fix $A^{(k)}$ and solve

$$\min_{\widetilde{B} \in \mathbb{R}^{r \times n}} \left|\left| \left(A^{(k)} \widetilde{B} \right)_{\Omega} - \vec{y} \right|\right|_2$$

Theoretical guarantees: Jain, Netrapalli, Sanghavi 2013

Low-rank models

Matrix completion

Structured low-rank models

Nonnegative matrix factorization

Nonnegative atoms/coefficients can make results easier to interpret

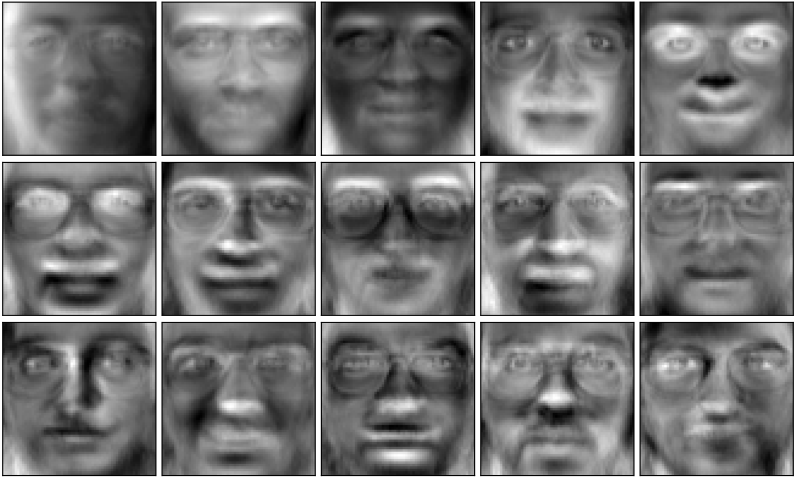
$$X \approx A\, B, \quad A_{i,j} \geq 0,\; B_{i,j} \geq 0, \text{ for all } i, j$$

Nonconvex optimization problem:

$$\begin{aligned} \text{minimize} \quad & \left|\left| X - \tilde{A}\, \tilde{B} \right|\right|_{\text{F}}^2 \\ \text{subject to} \quad & \tilde{A}_{i,j} \geq 0, \\ & \tilde{B}_{i,j} \geq 0, \qquad \text{for all } i, j \end{aligned}$$

$\tilde{A} \in \mathbb{R}^{m \times r}$ and $\tilde{B} \in \mathbb{R}^{r \times n}$

Faces dataset: PCA



Faces dataset: NMF

Topic modeling

$A :=$

singer	GDP	senate	election	vote	stock	bass	market	band	Articles
6	1	1	0	0	1	9	0	8	a
1	0	9	5	8	1	0	1	0	b
8	1	0	1	0	0	9	1	7	c
0	7	1	0	0	9	1	7	0	d
0	5	6	7	5	6	0	7	2	e
1	0	8	5	9	2	0	0	1	f

SVD

$$A = USV^T = U \begin{bmatrix} 23.64 & 0 & 0 & 0 & & \\ 0 & 18.82 & 0 & 0 & 0 & 0 \\ 0 & 0 & 14.23 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.63 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.03 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.36 \end{bmatrix} V^T$$

Left singular vectors

		a	b	c	d	e	f
U_1	$=$	$(-0.24$	-0.47	-0.24	-0.32	-0.58	$-0.47)$
U_2	$=$	$(\ 0.64$	-0.23	0.67	-0.03	-0.18	$-0.21)$
U_3	$=$	$(-0.08$	-0.39	-0.08	0.77	0.28	$-0.40)$

Right singular vectors

	singer	GDP	senate	election	vote	stock	bass	market	band
$V_1 =$	-0.18	-0.24	-0.51	-0.38	-0.46	-0.34	-0.2	-0.3	-0.22
$V_2 =$	0.47	0.01	-0.22	-0.15	-0.25	-0.07	0.63	-0.05	0.49
$V_3 =$	-0.13	0.47	-0.3	-0.14	-0.37	0.52	-0.04	0.49	-0.07

Nonnegative matrix factorization

$$X \approx W\ H$$

$$W_{i,j} \geq 0,\ H_{i,j} \geq 0,\ \text{for all } i, j$$

Right nonnegative factors

	singer	GDP	senate	election	vote	stock	bass	market	band
$H_1 =$	(0.34	0	3.73	2.54	3.67	0.52	0	0.35	0.35)
$H_2 =$	(0	2.21	0.21	0.45	0	2.64	0.21	2.43	0.22)
$H_3 =$	(3.22	0.37	0.19	0.2	0	0.12	4.13	0.13	3.43)

Interpretations:

- Count atom: Counts for each doc are weighted sum of H_1, H_2, H_3
- Coefficients: They cluster words into politics, music and economics

Left nonnegative factors

		a	b	c	d	e	f
W_1	=	(0.03	2.23	0	0	1.59	2.24)
W_2	=	(0.1	0	0.08	3.13	2.32	0)
W_3	=	(2.13	0	2.22	0	0	0.03)

Interpretations:

- Count atom: Counts for each word are weighted sum of W_1, W_2, W_3
- Coefficients: They cluster docs into politics, music and economics

Sparse PCA

Sparse atoms can make results easier to interpret

$$X \approx A\,B, \quad A \text{ sparse}$$

Nonconvex optimization problem:

$$\begin{aligned} \text{minimize} \quad & \left\|X - \tilde{A}\,\tilde{B}\right\|_2^2 + \lambda \sum_{i=1}^{k} \left\|\tilde{A}_i\right\|_1 \\ \text{subject to} \quad & \left\|\tilde{A}_i\right\|_2 = 1, \qquad 1 \le i \le k \end{aligned}$$

$\tilde{A} \in \mathbb{R}^{m \times r}$ and $\tilde{B} \in \mathbb{R}^{r \times n}$

Faces dataset