Principal component analysis

Representation learning

Useful representations of data

Representation learning:

- **Given**: raw feature vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$.
- **Goal**: learn a "useful" feature transformation $\boldsymbol{\phi}\colon \mathbb{R}^d \to \mathbb{R}^k$.
  (Often $k \ll d$—i.e., *dimensionality reduction*—but not always.)

  Can then use $\boldsymbol{\phi}$ as a feature map for supervised learning.

Some previously encontered examples:

- Feature maps corresponding to pos. def. kernels (+approximations).
  (Usually *data-oblivious*—feature map doesn't depend on the data.)
- Centering $\boldsymbol{x} \mapsto \boldsymbol{x} - \boldsymbol{\mu}$
  (Effect: resulting features have mean $\mathbf{0}$.)
- Standardization $\boldsymbol{x} \mapsto \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$.
  (Effect: resulting features have mean $\mathbf{0}$ and unit variance.)

What other properties of a feature representation may be desirable?

Principal component analysis

Dimensionality reduction via projections

Projections

- **Input**: $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$, target dimensionality $k \in \mathbb{N}$.
- **Output**: a k-dimensional subspace, represented by an orthonormal basis $\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_k \in \mathbb{R}^d$.
- **(Orthogonal) projection**: projection of $\boldsymbol{x} \in \mathbb{R}^d$ to $\operatorname{span}(\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_k)$ is

$$\underbrace{\left(\sum_{i=1}^{k} \boldsymbol{q}_i \boldsymbol{q}_i^\top\right)}_{\boldsymbol{\Pi}} \boldsymbol{x} \;=\; \sum_{i=1}^{k} \langle \boldsymbol{q}_i, \boldsymbol{x} \rangle \boldsymbol{q}_i \;\in\; \mathbb{R}^d .$$

Can also represent the projection of $\boldsymbol{x}$ in terms of its coefficients w.r.t. the orthonormal basis $\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_k$:

$$\boldsymbol{\phi}(\boldsymbol{x}) \;:=\; \begin{bmatrix} \langle \boldsymbol{q}_1, \boldsymbol{x} \rangle \\ \langle \boldsymbol{q}_2, \boldsymbol{x} \rangle \\ \vdots \\ \langle \boldsymbol{q}_k, \boldsymbol{x} \rangle \end{bmatrix} \;\in\; \mathbb{R}^k .$$

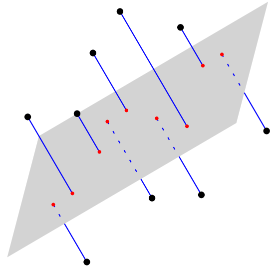
Projection of minimum residual squared error

Minimize residual squared error

Objective: find k-dimensional projector $\boldsymbol{\Pi} \colon \mathbb{R}^d \to \mathbb{R}^d$ such that the average residual squared error

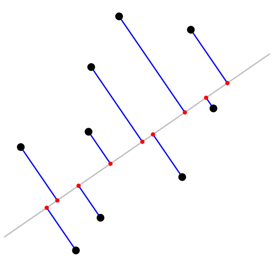
$$\frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\Pi} \boldsymbol{x}_i\|_2^2$$

is as small as possible.



Projection of minimum residual squared error

$k = 1$ **case** $(\boldsymbol{\Pi} = \boldsymbol{q}\boldsymbol{q}^\top)$



Objective: find unit vector $\boldsymbol{q} \in \mathbb{R}^d$ to minimize

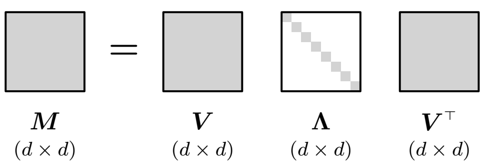
$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} \left\| \boldsymbol{x}_i - \boldsymbol{q}\boldsymbol{q}^\top \boldsymbol{x}_i \right\|_2^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i\|_2^2 - \boldsymbol{q}^\top \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right) \boldsymbol{q} \\ &= \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i\|_2^2 - \boldsymbol{q}^\top \left(\frac{1}{n} \boldsymbol{A}^\top \boldsymbol{A} \right) \boldsymbol{q} \end{aligned}$$

(where $\boldsymbol{x}_i^\top$ is i-th row of $\boldsymbol{A} \in \mathbb{R}^{n \times d}$).

$$\operatorname*{arg\,min}_{\boldsymbol{q} \in \mathbb{R}^d : \|\boldsymbol{q}\|_2 = 1} \frac{1}{n} \sum_{i=1}^{n} \left\| \boldsymbol{x}_i - \boldsymbol{q}\boldsymbol{q}^\top \boldsymbol{x}_i \right\|_2^2 \;\equiv\; \operatorname*{arg\,max}_{\boldsymbol{q} \in \mathbb{R}^d : \|\boldsymbol{q}\|_2 = 1} \boldsymbol{q}^\top \left(\frac{1}{n} \boldsymbol{A}^\top \boldsymbol{A} \right) \boldsymbol{q} .$$

Aside: Eigendecompositions

Every symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ guaranteed to have eigendecomposition with real eigenvalues:



$$= \sum_{i=1}^{d} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$$

real **eigenvalues**: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \quad (\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_d))$;
corresponding orthonormal **eigenvectors**: $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_d \quad (\boldsymbol{V} = [\boldsymbol{v}_1 | \boldsymbol{v}_2 | \cdots | \boldsymbol{v}_d])$.

Fixed-point characterization of eigenvectors:

$$\boldsymbol{M} \boldsymbol{v}_i \;=\; \lambda_i \boldsymbol{v}_i .$$

Eigendecompositions

Variational characterization of eigenvectors:

$$\begin{aligned} \max_{\boldsymbol{q} \in \mathbb{R}^d} \ & \boldsymbol{q}^\top \boldsymbol{M} \boldsymbol{q} \\ \text{s.t. } \|\boldsymbol{q}\|_2 &= 1 \end{aligned}$$

- Maximum value: λ_1 (top eigenvalue)
- Maximizer: $\boldsymbol{v}_1$ (top eigenvector)

For $i > 1$,

$$\begin{aligned} \max_{\boldsymbol{q} \in \mathbb{R}^d} \ & \boldsymbol{q}^\top \boldsymbol{M} \boldsymbol{q} \\ \text{s.t. } \|\boldsymbol{q}\|_2 &= 1 \\ \langle \boldsymbol{q}, \boldsymbol{v}_j \rangle &= 0 \ \forall j < i \end{aligned}$$

- Maximum value: λ_i (i-th largest eigenvalue)
- Maximizer: $\boldsymbol{v}_i$ (i-th eigenvector)

Principal component analysis ($k = 1$)

$k = 1$ **case** ($\boldsymbol{\Pi} = \boldsymbol{q}\boldsymbol{q}^\top$)

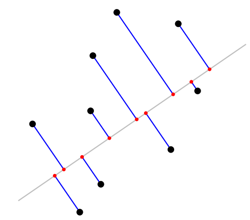
$$\operatorname*{arg\,min}_{\boldsymbol{q} \in \mathbb{R}^d : \|\boldsymbol{q}\|_2 = 1} \frac{1}{n} \sum_{i=1}^n \left\| \boldsymbol{x}_i - \boldsymbol{q}\boldsymbol{q}^\top \boldsymbol{x}_i \right\|_2^2 \quad \equiv \quad \operatorname*{arg\,max}_{\boldsymbol{q} \in \mathbb{R}^d : \|\boldsymbol{q}\|_2 = 1} \boldsymbol{q}^\top \left(\frac{1}{n} \boldsymbol{A}^\top \boldsymbol{A} \right) \boldsymbol{q} \,.$$

Solution: eigenvector of $\boldsymbol{A}^\top \boldsymbol{A}$ corresponding to largest eigenvalue (i.e., the top eigenvector $\boldsymbol{v}_1$).

$$\boldsymbol{q}^\top \left(\frac{1}{n} \boldsymbol{A}^\top \boldsymbol{A} \right) \boldsymbol{q} = \frac{1}{n} \sum_{i=1}^n \langle \boldsymbol{q}, \boldsymbol{x}_i \rangle^2$$

(variance in direction $\boldsymbol{q}$, assuming $\frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i = \boldsymbol{0}$).

top eigenvector $\equiv$ direction of maximum variance



Principal component analysis (general k)

General k case ($\boldsymbol{\Pi} = \boldsymbol{Q}\boldsymbol{Q}^\top$)

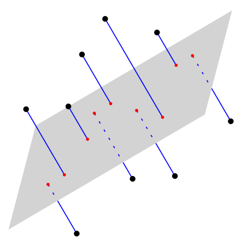
$$\operatorname*{arg\,min}_{\substack{\boldsymbol{Q} \in \mathbb{R}^{d \times k}: \\ \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}}} \frac{1}{n} \sum_{i=1}^n \left\| \boldsymbol{x}_i - \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{x}_i \right\|_2^2 \quad \equiv \quad \operatorname*{arg\,max}_{\substack{\boldsymbol{Q} \in \mathbb{R}^{d \times k}: \\ \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}}} \sum_{i=1}^k \boldsymbol{q}_i^\top \left(\frac{1}{n} \boldsymbol{A}^\top \boldsymbol{A} \right) \boldsymbol{q}_i \,.$$

Solution: k eigenvectors of $\boldsymbol{A}^\top \boldsymbol{A}$ corresponding to k largest eigenvalue

$$\sum_{i=1}^k \boldsymbol{q}_i^\top \left(\frac{1}{n} \boldsymbol{A}^\top \boldsymbol{A} \right) \boldsymbol{q}_i = \sum_{i=1}^k \frac{1}{n} \sum_{j=1}^n \langle \boldsymbol{q}_i, \boldsymbol{x}_j \rangle^2$$

(sum of variances in $\boldsymbol{q}_i$ directions, assuming $\frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i = \boldsymbol{0}$).

top k eigenvectors $\equiv$ k-dim. subspace of maximum variance



Principal component analysis (PCA)

Data matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$

Rank k PCA (k dimensional linear subspace)

- Get top k eigenvectors $\widehat{\boldsymbol{V}}_k := [\boldsymbol{v}_1 | \boldsymbol{v}_2 | \dots | \boldsymbol{v}_k]$ of

$$\frac{1}{n} \boldsymbol{A}^\top \boldsymbol{A} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top \,.$$

- *Feature map*: $\boldsymbol{\phi}(\boldsymbol{x}) := (\langle \boldsymbol{v}_1, \boldsymbol{x} \rangle, \langle \boldsymbol{v}_2, \boldsymbol{x} \rangle, \dots, \langle \boldsymbol{v}_k, \boldsymbol{x} \rangle) \in \mathbb{R}^k$.
- *Decorrelating property*:

$$\frac{1}{n} \sum_{i=1}^n \boldsymbol{\phi}(\boldsymbol{x}_i) \boldsymbol{\phi}(\boldsymbol{x}_i)^\top = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_k) \,.$$

- *Approx. reconstruction*: $\boldsymbol{x} \mapsto \widehat{\boldsymbol{V}}_k \boldsymbol{\phi}(\boldsymbol{x})$.

Principal component analysis (PCA)

Data matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$

Rank k PCA with centering (k dimensional affine subspace)

- Get top k eigenvectors $\widehat{\boldsymbol{V}}_k := [\boldsymbol{v}_1 | \boldsymbol{v}_2 | \dots | \boldsymbol{v}_k]$ of

$$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^{\top}$$

  where $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$.

- *Feature map*: $\boldsymbol{\phi}(\boldsymbol{x}) := (\langle \boldsymbol{v}_1, \boldsymbol{x} - \boldsymbol{\mu}\rangle, \langle \boldsymbol{v}_2, \boldsymbol{x} - \boldsymbol{\mu}\rangle, \dots, \langle \boldsymbol{v}_k, \boldsymbol{x} - \boldsymbol{\mu}\rangle) \in \mathbb{R}^k$.

- *Decorrelating property*:

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\phi}(\boldsymbol{x}_i) = \boldsymbol{0}$$

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_i)^{\top} = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_k).$$

- *Approx. reconstruction*: $\boldsymbol{x} \mapsto \boldsymbol{\mu} + \widehat{\boldsymbol{V}}_k \boldsymbol{\phi}(\boldsymbol{x})$.

Example: PCA on OCR digits data

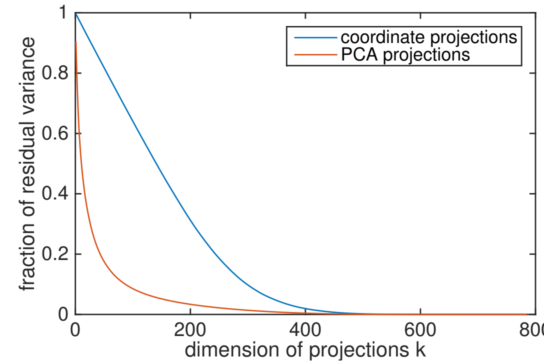
Data $\{\boldsymbol{x}_i\}_{i=1}^{n}$ from $\mathbb{R}^{784}$.

- Fraction of residual variance left by rank-k PCA projection:

$$1 - \frac{\sum_{j=1}^{k} \text{variance in direction } \boldsymbol{v}_j}{\text{total variance}}.$$

- Fraction of residual variance left by best k *coordinate* projections:

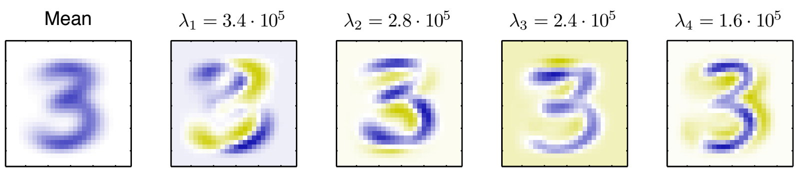
$$1 - \frac{\sum_{j=1}^{k} \text{variance in direction } \boldsymbol{e}_j}{\text{total variance}}.$$



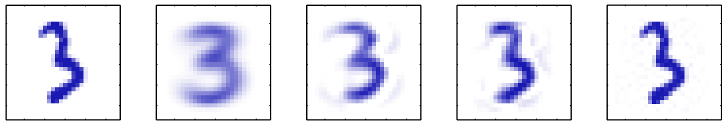
Example: compressing digits images

16×16 pixel images of handwritten 3s (as vectors in $\mathbb{R}^{256}$)

Mean $\boldsymbol{\mu}$ and eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4$



Reconstructions:



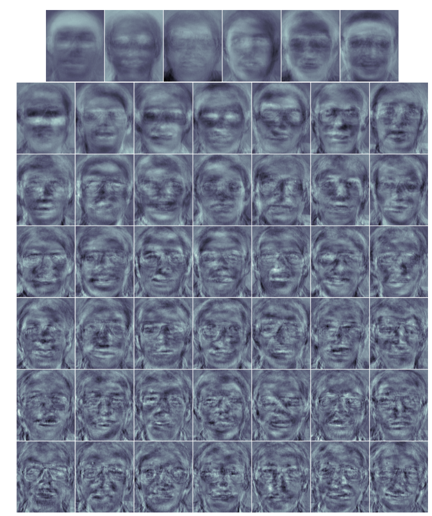
$\boldsymbol{x}$ $\quad k = 1 \quad k = 10 \quad k = 50 \quad k = 200$

Only have to store k numbers per image,
along with the mean $\boldsymbol{\mu}$ and k eigenvectors ($256(k + 1)$ numbers).

Example: eigenfaces

92×112 pixel images of faces (as vectors in $\mathbb{R}^{10304}$)

100 example images



top $k = 48$ eigenvectors

Other examples

- $\boldsymbol{x} \in \mathbb{R}^d$: movement of stock prices for d different stocks in one day.

  Principal component: combination of stocks that account for the most variation in stock price movement.

- $\boldsymbol{x} \in \{1, 2, \ldots, 5\}^d$: levels at which various terms describe an individual (e.g., "jolly", "impulsive", "outgoing", "conceited", "meddlesome")

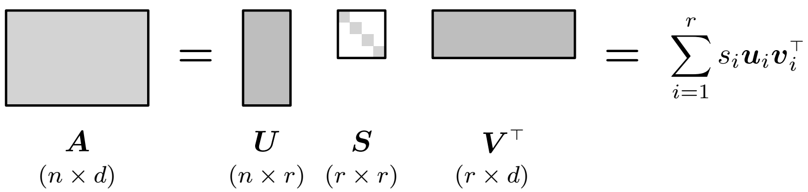
  Principal components: major personality axes in a population (e.g., "extroversion", "agreeableness", "conscientiousness")

- ...

Singular value decomposition

Singular value decomposition

Every matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ has a **singular value decomposition (SVD)**



where

- $r = \operatorname{rank}(\boldsymbol{A}) \quad (r \leq \min\{n, d\})$;
- $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$ (i.e., $\boldsymbol{U} = [\boldsymbol{u}_1 | \boldsymbol{u}_2 | \cdots | \boldsymbol{u}_r]$ has orthonormal columns)
  left singular vectors;
- $\boldsymbol{S} = \operatorname{diag}(s_1, s_2, \ldots, s_r)$ where $s_1 \geq s_2 \geq \cdots \geq s_r > 0$
  singular values;
- $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$ (i.e., $\boldsymbol{V} = [\boldsymbol{v}_1 | \boldsymbol{v}_2 | \cdots | \boldsymbol{v}_r]$ has orthonormal columns)
  right singular vectors.

SVD vs PCA

If SVD of $\boldsymbol{A}$ is $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top = \sum_{i=1}^{r} s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, then:

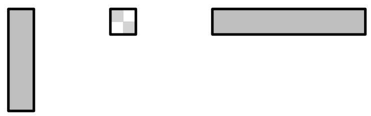
- non-zero eigenvalues of $\boldsymbol{A}^\top \boldsymbol{A}$ are $s_1^2, s_2^2, \ldots, s_r^2$, (squares of singular values of $\boldsymbol{A}$);
- corresponding eigenvectors are $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r \in \mathbb{R}^d$ (right singular vectors of $\boldsymbol{A}$).

By symmetry, also have:

- non-zero eigenvalues of $\boldsymbol{A}\boldsymbol{A}^\top$ are $s_1^2, s_2^2, \ldots, s_r^2$, (squares of singular values of $\boldsymbol{A}$);
- corresponding eigenvectors are $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r \in \mathbb{R}^d$ (left singular vectors of $\boldsymbol{A}$).

Low-rank SVD

For any $k \leq \text{rank}(\boldsymbol{A})$, **rank-$k$ SVD approximation**:



$$= \sum_{i=1}^{k} s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$$

$\widehat{\boldsymbol{U}}_k$ $(n \times k)$ $\quad \widehat{\boldsymbol{S}}_k$ $(k \times k)$ $\quad \widehat{\boldsymbol{V}}_k^\top$ $(k \times d)$

(Just retain top k left/right singular vectors and singular values from SVD.)

Best rank-k approximation:

$$\widehat{\boldsymbol{A}} := \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{S}}_k \widehat{\boldsymbol{V}}_k^\top = \mathop{\arg\min}_{\substack{\boldsymbol{M} \in \mathbb{R}^{n \times d}: \\ \text{rank}(\boldsymbol{M}) \leq k}} \sum_{i=1}^{n} \sum_{j=1}^{d} (A_{i,j} - M_{i,j})^2.$$

Minimum value is simply given by

$$\sum_{i=1}^{n} \sum_{j=1}^{d} (A_{i,j} - \widehat{A}_{i,j})^2 = \sum_{t>k} s_t^2.$$

Example: latent semantic analysis

Represent corpus of documents by counts of words they contain:

	aardvark	abacus	abalone	$\cdots$
document 1	3	0	0	$\cdots$
document 2	7	0	4	$\cdots$
document 3	2	4	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

- One column per vocabulary word in $\boldsymbol{A} \in \mathbb{R}^{n \times d}$
- One row per document in $\boldsymbol{A} \in \mathbb{R}^{n \times d}$
- $A_{i,j}$ = numbers of times word j appears in document i.

Example: latent semantic analysis

Statistical model for document-word count matrix.

Parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_k, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots, \boldsymbol{\pi}_n, \ell_1, \ell_2, \dots, \ell_n)$.

- $k \ll \min\{n, d\}$ "topics", each represented by a distributions over vocabulary words:
  $$\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_k \in \mathbb{R}_+^d.$$
  Each $\boldsymbol{\beta}_t = (\beta_{t,1}, \beta_{t,2}, \dots, \beta_{t,d})$ is a probability vector, so $\sum_{j=1}^{d} \beta_{t,j} = 1$.
- Each document i is associated with a probability distribution $\boldsymbol{\pi}_i = (\pi_{i,1}, \pi_{i,2}, \dots, \pi_{i,k})$ over topics, so $\sum_{t=1}^{k} \pi_{i,t} = 1$.

Model posits that document i's count vector (i-th row in $\boldsymbol{A}$) follows a multinomial distribution with probabilities given by $\sum_{t=1}^{k} \pi_{i,t} \boldsymbol{\beta}_t$:

$$\begin{bmatrix} A_{i,1} & A_{i,2} & \dots & A_{i,d} \end{bmatrix} \sim \text{Multinomial}\left(\ell_i, \sum_{t=1}^{k} \pi_{i,t} \boldsymbol{\beta}_t^\top\right).$$

Expected value is $\ell_i \sum_{t=1}^{k} \pi_{i,t} \boldsymbol{\beta}_t^\top$.

Example: latent semantic analysis

Suppose $\boldsymbol{A} \sim P_{\boldsymbol{\theta}}$.

In expectation, $\boldsymbol{A}$ has rank $\leq k$:

$$\mathbb{E}(\boldsymbol{A}) = \underbrace{\begin{bmatrix} \leftarrow & \ell_1 \boldsymbol{\pi}_1^\top & \rightarrow \\ \leftarrow & \ell_2 \boldsymbol{\pi}_2^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \ell_n \boldsymbol{\pi}_n^\top & \rightarrow \end{bmatrix}}_{n \times k} \underbrace{\begin{bmatrix} \longleftarrow & \boldsymbol{\beta}_1^\top & \longrightarrow \\ \longleftarrow & \boldsymbol{\beta}_2^\top & \longrightarrow \\ & \vdots & \\ \longleftarrow & \boldsymbol{\beta}_k^\top & \longrightarrow \end{bmatrix}}_{k \times d}.$$

Observed matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \mathbb{E}(\boldsymbol{A}) + \textbf{Zero mean noise}$$

so $\boldsymbol{A}$ is generally of rank $\min\{n, d\} \gg k$.

Example: latent semantic analysis

Using SVD: rank-k SVD $\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{S}}_k \widehat{\boldsymbol{V}}_k^\top$ of $\boldsymbol{A}$ gives approximation to $\boldsymbol{L}\boldsymbol{B}^\top$:

$$\widehat{\boldsymbol{A}} := \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{S}}_k \widehat{\boldsymbol{V}}_k^\top \approx \mathbb{E}(\boldsymbol{A}).$$

(SVD helps remove some of the effect of the noise.)

- Each of the n documents can be summarized by k numbers:

$$\widehat{\boldsymbol{A}} \widehat{\boldsymbol{V}}_k = \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{S}}_k \in \mathbb{R}^{n \times k}.$$

- New document *feature representation* very useful for information retrieval.

  (Example: cosine similarities between documents become faster to compute and possibly less noisy.)

- Actually estimating $\boldsymbol{\pi}_i$ and $\boldsymbol{\beta}_t$ takes a bit more work.

Recap

- **PCA**: directions of maximum variance in data $\equiv$ subspace that minimizes residual squared error.
- **SVD**: general decomposition for arbitrary matrices

  Low-rank SVD: best low-rank approximation of a matrix in terms of average squared errors
- **PCA/SVD**: often useful when low-rank structure is expected (e.g., probabilistic modeling).