# The Role of Dimensionality Reduction in Distributional Semantics

## or: having fun with matrix algebra

Stefan Evert

Technische Universität Darmstadt, Germany
evert@linglit.tu-darmstadt.de

Leuven Statistics Days
8 June 2012

TECHNISCHE UNIVERSITÄT DARMSTADT





## Outline







## Outline







## General definition of DSMs

A **distributional semantic model** (DSM) is a scaled and/or transformed co-occurrence matrix **M**, such that each row **m** represents the distribution of a **target term** across contexts.

| | get | see | use | hear | eat | kill |
|---|---|---|---|---|---|---|
| knife | 0.027 | -0.024 | 0.206 | -0.022 | -0.044 | -0.042 |
| cat | 0.031 | 0.143 | -0.243 | -0.015 | -0.009 | 0.131 |
| dog | -0.026 | 0.021 | -0.212 | 0.064 | 0.013 | 0.014 |
| boat | -0.022 | 0.009 | -0.044 | -0.040 | -0.074 | -0.042 |
| cup | -0.014 | -0.173 | -0.249 | -0.099 | -0.119 | -0.042 |
| pig | -0.069 | 0.094 | -0.158 | 0.000 | 0.094 | 0.265 |
| banana | 0.047 | -0.139 | -0.104 | -0.022 | 0.267 | -0.042 |

**Term** = word form, lemma, phrase, morpheme, word pair, ...
**Targets** = rows (terms whose distribution is represented)
**Features** = columns (individual contexts or collocates)

## Notation: term-context matrix

Frequency matrix $\mathbf{F} \in \mathbb{R}^{k \cdot n}$ (**term-context** row vectors $\mathbf{f}_i \in \mathbb{R}^n$)

$$\mathbf{F} = \begin{bmatrix} \cdots & \mathbf{f}_1^T & \cdots \\ \cdots & \mathbf{f}_2^T & \cdots \\ & \vdots & \\ & \vdots & \\ \cdots & \mathbf{f}_k^T & \cdots \end{bmatrix}$$

| | Felidae | Pet | Feral | Bloat | Philosophy | Kant | Back pain |
|---|---|---|---|---|---|---|---|
| cat | 10 | 10 | 7 | – | – | – | – |
| dog | – | 10 | 4 | 11 | – | – | – |
| animal | 2 | 15 | 10 | 2 | – | – | – |
| time | 1 | – | – | – | 2 | 1 | – |
| reason | – | 1 | – | – | 1 | 4 | 1 |
| cause | – | – | – | 2 | 1 | 2 | 6 |
| effect | – | – | – | 1 | – | 1 | – |

Interpretation as **collection of row vectors**:

- $\mathbf{F} = (f_{ij})$, where $f_{ij} = (\mathbf{f}_i)_j$ = frequency count of target term $t_i$ in context $c_j$ (wrt. **context tokens**, here: Wikipedia articles)





## Notation: term-term matrix

Cooccurrence matrix $\mathbf{M} \in \mathbb{R}^{k \cdot n}$ (**term-term** row vectors $\mathbf{m}_i \in \mathbb{R}^n$)

$$\mathbf{M} = \begin{bmatrix} \cdots & \mathbf{m}_1^T & \cdots \\ \cdots & \mathbf{m}_2^T & \cdots \\ & \vdots & \\ & \vdots & \\ \cdots & \mathbf{m}_k^T & \cdots \end{bmatrix}$$

| | breed | tail | feed | kill | important | explain | likely |
|---|---|---|---|---|---|---|---|
| cat | 83 | 17 | 7 | 37 | – | 1 | – |
| dog | 561 | 13 | 30 | 60 | 1 | 2 | 4 |
| animal | 42 | 10 | 109 | 134 | 13 | 5 | 5 |
| time | 19 | 9 | 29 | 117 | 81 | 34 | 109 |
| reason | 1 | – | 2 | 14 | 68 | 140 | 47 |
| cause | – | 1 | – | 4 | 55 | 34 | 55 |
| effect | – | – | 1 | 6 | 60 | 35 | 17 |

Interpretation as **collection of row vectors**:

- $\mathbf{M} = (m_{ij})$, where $m_{ij} = (\mathbf{m}_i)_j$ = cooccurrence frequency of target term $t_i$ with feature term $\tau_j$ (a **collocate** of $t_i$)





## Document vectors as centroids

$$\mathbf{F} = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \phi_1 & \phi_2 & \cdots & \phi_n \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \end{bmatrix} \qquad \mathbf{d}_j = \frac{1}{n_j} \sum_{i=1}^{k} \phi_{ji} \cdot \mathbf{m}_i = \frac{1}{n_j} \mathbf{M}^T \cdot \phi_j$$

- column vector $\phi_j \in \mathbb{R}^k$ = bag-of-words representation of context $c_j$, i.e. a term frequency vector with $\phi_{ji} = f_{ij}$
- context $c_j$ can be represented in term space by **document vector** $\mathbf{d}_j \in \mathbb{R}^n$ = weighted centroid of the corresponding term vectors (Schütze 1998)
- $n_j = \sum_i \phi_{ji}$ = document size of $c_j$





## DSM parameters

Corpus with linguistic annotation

⇓

Term-context vs. term-term matrix

⇓

Type & size of context

⇓

Feature scaling

⇓

Similarity/**distance measure** & normalisation

⇓

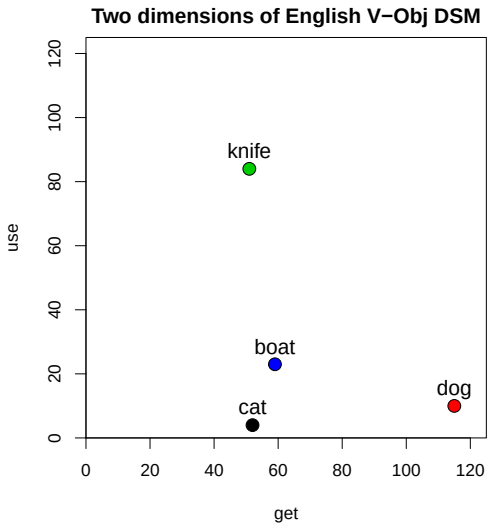
**Dimensionality reduction**

⇓

Semantic distance, nearest neighbours, semantic maps, ...

## Geometric interpretation and semantic distance

- row vector $\mathbf{m}_{dog}$ describes usage of word *dog* in the corpus
- can be seen as coordinates of point in $n$-dimensional Euclidean space $\mathbb{R}^n$
- illustrated for two dimensions: *get* and *use*
- $\mathbf{m}_{dog} = (115, 10)$

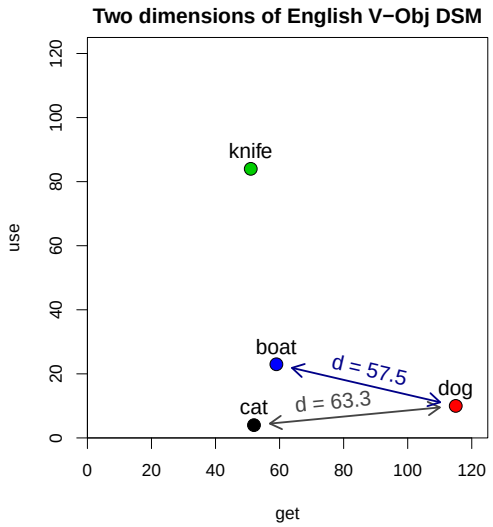







## Geometric interpretation and semantic distance

- similarity = spatial proximity (Euclidean metric)
- location depends on frequency of noun ($f_{dog} \approx 2.7 \cdot f_{cat}$)

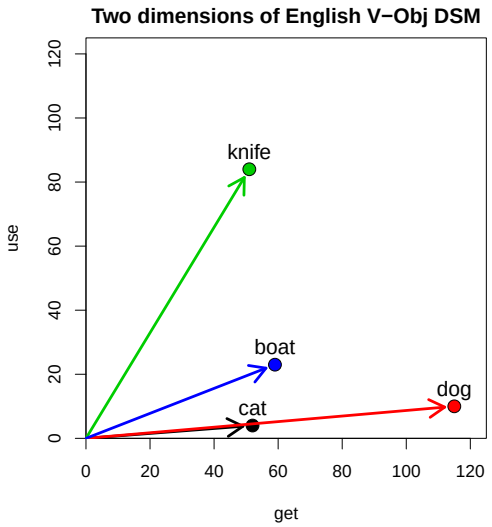







## Geometric interpretation and semantic distance

- similarity = spatial proximity (Euclidean metric)
- location depends on frequency of noun ($f_{dog} \approx 2.7 \cdot f_{cat}$)
- direction more important than location

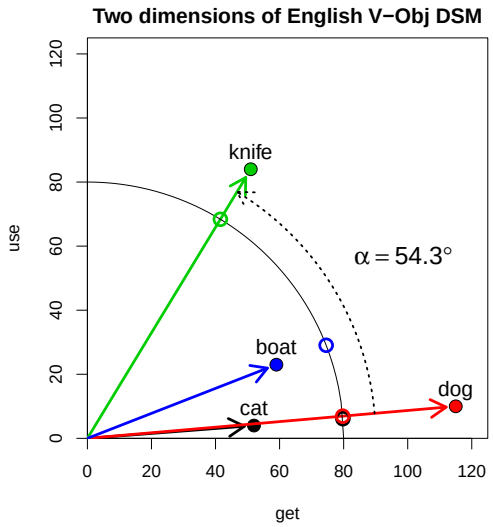







## Geometric interpretation and semantic distance

- similarity = spatial proximity (Euclidean metric)
- location depends on frequency of noun ($f_{dog} \approx 2.7 \cdot f_{cat}$)
- direction more important than location
- normalise "length" $\|\mathbf{m}_{dog}\|$ of vector
- or use angle $\alpha$ as distance measure



## Metric & norm

- **metric** $d(\mathbf{x}, \mathbf{y})$ as measure of semantic (dis)similarity
- **norm** $\|\mathbf{x}\|$ = measure of vector length induces a homogeneous, translation-invariant metric $d(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\|$
- family of Minkowski *p*-**norms** for $p \in [1, \infty]$:

$$\|\mathbf{x}\|_p := \left(|x_1|^p + \cdots + |x_n|^p\right)^{1/p}$$

(for $p < 1$, the triangle inequality is not satisfied)

- includes the intuitive **Euclidean** norm for $p = 2$:

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$$

- from now on $\|\mathbf{x}\| := \|\mathbf{x}\|_2$ unless specified otherwise





## Euclidean norm & inner product

- Euclidean norm $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ is special because it is induced by an **inner product**:

$$\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x}^T \mathbf{y} = x_1 y_1 + \cdots + x_n y_n$$

- **angle** $\varphi$ between vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

$$\cos \varphi := \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}$$

☞ **cosine similarity** is popular "distance" measure for DSM

- $\mathbf{x}$ and $\mathbf{y}$ are **orthogonal** iff $\langle \mathbf{x}, \mathbf{y} \rangle = 0$
  - the **shortest connection** between a point $\mathbf{x}$ and a subspace $A$ is orthogonal to all vectors $\mathbf{y} \in A$





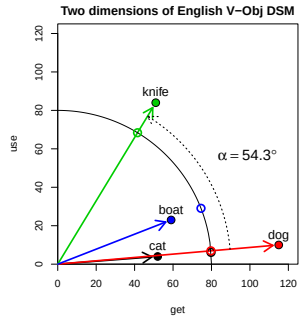
## Euclidean distance or cosine similarity?

Which is better, Euclidean distance or cosine similarity?

Equivalent if vectors are normalised ($\|\mathbf{x}\|_2 = 1$),
i.e. same ranking of distances between different points

$$\cos \varphi = \langle \mathbf{x}, \mathbf{y} \rangle$$

$$\begin{aligned} \|\mathbf{x} - \mathbf{y}\|^2 &= \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle \\ &= \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle - 2 \langle \mathbf{x}, \mathbf{y} \rangle \\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2 \langle \mathbf{x}, \mathbf{y} \rangle \\ &= 2 - 2 \cos \varphi \end{aligned}$$








## An exercise in matrix algebra

Task: compute distances (or similarities) between all target terms $t_i$ in row-normalised matrix $\mathbf{M}$ as quickly as possible.

$$\cos \varphi_{ij} = \langle \mathbf{m}_i, \mathbf{m}_j \rangle = \mathbf{m}_i^T \mathbf{m}_j \quad \text{for } i, j \in \{1, \ldots, k\}$$

$$\cos \begin{bmatrix} \cdots & \cdots & \vdots & \cdots \\ & & \varphi_{ij} & \\ & & \vdots & \end{bmatrix} = \begin{bmatrix} \cdots & \mathbf{m}_i^T & \cdots \end{bmatrix} \cdot \begin{bmatrix} & \vdots & \\ & \mathbf{m}_j & \\ & \vdots & \end{bmatrix}$$

$$\cos \varphi = \mathbf{M} \cdot \mathbf{M}^T$$

## Outline







## Distributional Memory (Baroni and Lenci 2010)

- Tensor of (word, link, word) triples, e.g. (*book*, OBJ, *read*)
  - also (*sharp*, AS ADJ AS, *knife*); (*geek*, USE, *computer*); ...
- TypeDM: feature scores = local MI (Evert 2004) based on number of distinct surface realisations of the link pattern
  - 30,686 target terms × 25,336 link types × 30,686 collocates
- $W_1 \times LW_2$ matricization yields state-of-the-art DSM
  - very **high-dimensional**: 30,686 × 3,127,436 matrix
  - extremely **sparse**: 131 million nonzero cells = 0.137%

➥ Dimensionality reduction to make data set manageable
  - e.g. 1.25 M uninformative features with single nonzero entry





## Goals of dimensionality reduction

- Numerical convenience
- Noise reduction (Landauer and Dumais 1997)
- Latent meaning dimensions (Schütze 1992, 1998)

A simple approach: **feature selection**
- drop least frequent, variable, informative, ... features
- convenient, but no noise reduction & latent dimensions

General form: map data points into **low-dimensional subspace**
- exploit correlations between features ➜ less information loss





## Approaches to dimensionality reduction

excerpt from verb-object DSM based on British National Corpus

| | buy | purchase | sell | write | read | draft |
|---|---|---|---|---|---|---|
| company | 81 | 17 | 50 | 1 | 2 | 2 |
| ticket | 178 | 9 | 98 | 7 | 0 | 0 |
| coffee | 21 | 0 | 9 | 0 | 0 | 0 |
| electricity | 2 | 1 | 15 | 1 | 0 | 0 |
| chocolate | 19 | 0 | 2 | 0 | 0 | 0 |
| letter | 4 | 0 | 3 | 950 | 223 | 25 |
| note | 1 | 0 | 2 | 167 | 70 | 4 |
| statement | 0 | 0 | 1 | 18 | 58 | 7 |
| agreement | 0 | **45** | 0 | 3 | 2 | 13 |

**feature selection** (2 dimensions)

# Approaches to dimensionality reduction

excerpt from verb-object DSM based on British National Corpus

| | buy | purchase | sell | write | read | draft |
|---|---|---|---|---|---|---|
| company | 81 | 17 | 50 | 1 | 2 | 2 |
| ticket | 178 | 9 | 98 | 7 | 0 | 0 |
| coffee | 21 | 0 | 9 | 0 | 0 | 0 |
| electricity | 2 | 1 | 15 | 1 | 0 | 0 |
| chocolate | 19 | 0 | 2 | 0 | 0 | 0 |
| letter | 4 | 0 | 3 | 950 | 223 | 25 |
| note | 1 | 0 | 2 | 167 | 70 | 4 |
| statement | 0 | 0 | 1 | 18 | 58 | 7 |
| agreement | 0 | **45** | 0 | 3 | 2 | 13 |

aggregate **meaningful feature combinations**





# Approaches to dimensionality reduction

excerpt from verb-object DSM based on British National Corpus

| | buy | purchase | sell | write | read | draft |
|---|---|---|---|---|---|---|
| company | 84 | 7 | 47 | 2 | 0 | 0 |
| ticket | 177 | 14 | 99 | 7 | 1 | 1 |
| coffee | 20 | 2 | 11 | 0 | 0 | 0 |
| electricity | 8 | 1 | 4 | 1 | 0 | 0 |
| chocolate | 15 | 1 | 9 | 0 | 0 | 0 |
| letter | 4 | 0 | 3 | 948 | 230 | 25 |
| note | 1 | 0 | 1 | 174 | 42 | 5 |
| statement | 0 | 0 | 0 | 30 | 7 | 1 |
| agreement | 3 | **0** | 2 | 4 | 1 | 0 |

by **regression** into 2-dimensional subspace





# Approaches to dimensionality reduction

excerpt from verb-object DSM based on British National Corpus

| | buy | purchase | sell | write | read | draft |
|---|---|---|---|---|---|---|
| company | 84 | 7 | 47 | $2-1$ | 1 | 0 |
| ticket | 177 | 14 | 99 | $8-1$ | $2-1$ | 1 |
| coffee | 20 | 2 | 11 | 0 | 0 | 0 |
| electricity | 8 | 1 | 4 | 1 | 0 | 0 |
| chocolate | 15 | 1 | 9 | 0 | 0 | 0 |
| letter | $6-2$ | 0 | $4-1$ | 948 | 230 | 25 |
| note | 1 | 0 | 1 | 174 | 42 | 5 |
| statement | 0 | 0 | 0 | 30 | 7 | 1 |
| agreement | 3 | 0 | 2 | 4 | 1 | 0 |

by **regression** into 2-dimensional subspace





# Approaches to dimensionality reduction

excerpt from verb-object DSM based on British National Corpus

| | buy | purchase | sell | write | read | draft | dim1 | dim2 |
|---|---|---|---|---|---|---|---|---|
| company | 84 | 7 | 47 | $2-1$ | 1 | 0 | 2 | 96 |
| ticket | 177 | 14 | 99 | $8-1$ | $2-1$ | 1 | 8 | 203 |
| coffee | 20 | 2 | 11 | 0 | 0 | 0 | 0 | 23 |
| electricity | 8 | 1 | 4 | 1 | 0 | 0 | 1 | 9 |
| chocolate | 15 | 1 | 9 | 0 | 0 | 0 | 0 | 18 |
| letter | $6-2$ | 0 | $4-1$ | 948 | 230 | 25 | 976 | $-2$ |
| note | 1 | 0 | 1 | 174 | 42 | 5 | 179 | 0 |
| statement | 0 | 0 | 0 | 30 | 7 | 1 | 31 | 0 |
| agreement | 3 | 0 | 2 | 4 | 1 | 0 | 4 | 3 |

by **regression** into 2-dimensional subspace
first dimension: written material
second dimension: commodities

## Outline



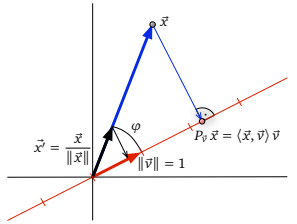




## Dimensionality reduction by orthogonal projection

- Approach: map data points into linear subspace with $d \ll n$ dimensions, shifting their positions as little as possible
  - same intuition as for linear regression: residuals = "noise"
  - i.e. minimise displacement $\tilde{\mathbf{x}} - \mathbf{x}$ between original data point $\mathbf{x}$ and mapped point $\tilde{\mathbf{x}}$ in low-dimensional subspace

- For each data point $\mathbf{x}$, best possible mapping is orthogonal projection $\tilde{\mathbf{x}} = \mathbf{P}_A\mathbf{x}$ into a given subspace $A$
- $\|\mathbf{x}\|^2 = \|\mathbf{P}_A\mathbf{x}\|^2 + \underbrace{\|\mathbf{x} - \mathbf{P}_A\mathbf{x}\|^2}_{\text{displacement}}$
- Based on Euclidean distance








## Dimensionality reduction by orthogonal projection

- $d$-dimensional subspace $A$ spanned by basis vectors $\mathbf{b}_1, \ldots, \mathbf{b}_d$ with $\langle \mathbf{b}_i, \mathbf{b}_j \rangle = \delta_{ij}$, forming an orthogonal $n \times d$ matrix:

$$\mathbf{Q} = \begin{bmatrix} \vdots & & \vdots \\ \vdots & & \vdots \\ \mathbf{b}_1 & \cdots & \mathbf{b}_d \\ \vdots & & \vdots \\ \vdots & & \vdots \end{bmatrix} \qquad \mathbf{P}_A\mathbf{x} = \sum_{i=1}^{d} \mathbf{b}_i(\mathbf{b}_i^T\mathbf{x}) = \mathbf{Q}\mathbf{Q}^T\mathbf{x}$$

- $\mathbf{P}_A\mathbf{x} = \mathbf{Q}\mathbf{Q}^T\mathbf{x}$ = projection into subspace $A \subseteq \mathbb{R}^n$
- $\mathbf{Q}^T\mathbf{x}$ = projection into internal Cartesian coordinates of $A$
  - $\|\mathbf{Q}^T\mathbf{x}\| = \|\mathbf{P}_A\mathbf{x}\|$ ($\mathbf{Q}$ is isometric embedding)
  - $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_d$ (identity matrix)





## Dimensionality reduction by orthogonal projection

- Project row vectors $\mathbf{m}$ of co-occurrence matrix $\mathbf{M}$ by matrix multiplication ➜ row vectors $\tilde{\mathbf{m}}$ of matrix $\tilde{\mathbf{M}}$

$$\tilde{\mathbf{M}} = \mathbf{M}\mathbf{P}_A = \mathbf{M}\mathbf{Q}\mathbf{Q}^T$$

- Total displacement given by **Frobenius norm** $\|\tilde{\mathbf{M}} - \mathbf{M}\|_2$

$$\sum_{i=1}^{k} \|\mathbf{P}_A\mathbf{m}_i - \mathbf{m}_i\|^2 = \|\mathbf{M}\mathbf{P}_A - \mathbf{M}\|^2 = \|\mathbf{M}\|^2 - \|\mathbf{M}\mathbf{P}_\mathbf{A}\|^2$$

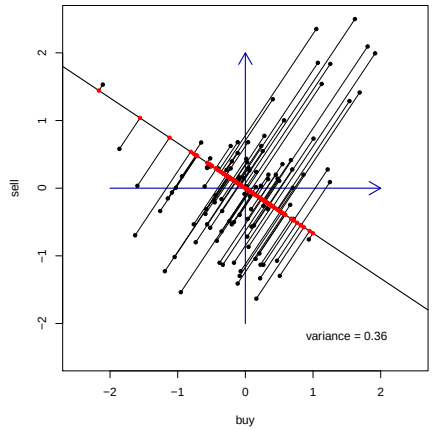
☞ Goal: find subspace $A$ that maximises

$$\|\mathbf{M}\mathbf{P}_\mathbf{A}\|^2 = \|\mathbf{M}\mathbf{Q}\mathbf{Q}^T\|^2 = \|\mathbf{M}\mathbf{Q}\|^2$$

## Dimensionality reduction by orthogonal projection

- For one-dimensional subspace: $\mathbf{P}_A = \mathbf{b}\mathbf{b}^T$, so maximise

$$\|\mathbf{Mb}\|^2 = \langle \mathbf{Mb}, \mathbf{Mb} \rangle = (\mathbf{b}^T\mathbf{M}^T)(\mathbf{Mb}) = \mathbf{b}^T(\mathbf{M}^T\mathbf{M})\mathbf{b}$$



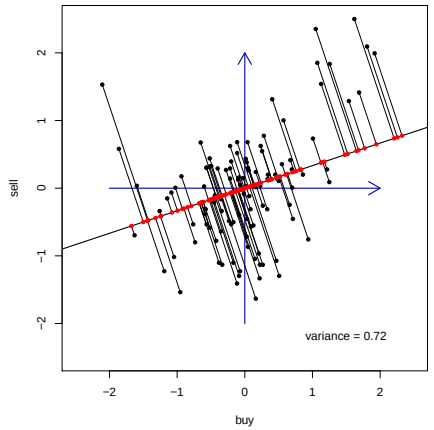



Dimensionality reduction | Singular value decomposition (SVD)

Dimensionality reduction by orthogonal projection

- For one-dimensional subspace: $\mathbf{P}_A = \mathbf{b}\mathbf{b}^T$, so maximise

$$\|\mathbf{Mb}\|^2 = \langle \mathbf{Mb}, \mathbf{Mb} \rangle = (\mathbf{b}^T\mathbf{M}^T)(\mathbf{Mb}) = \mathbf{b}^T(\mathbf{M}^T\mathbf{M})\mathbf{b}$$



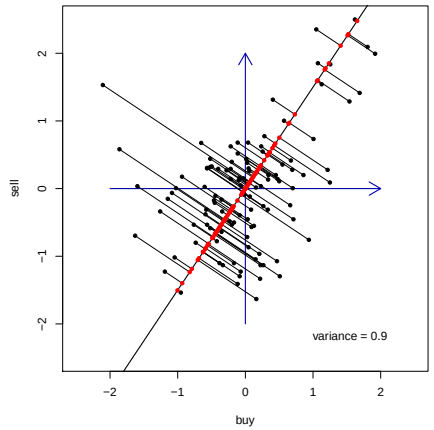

Stefan Evert (TU Darmstadt) | Dimensionality Reduction for DSM | wordspace.collocations.de | 29 / 50

Dimensionality reduction | Singular value decomposition (SVD)

Dimensionality reduction by orthogonal projection

- For one-dimensional subspace: $\mathbf{P}_A = \mathbf{b}\mathbf{b}^T$, so maximise

$$\|\mathbf{Mb}\|^2 = \langle \mathbf{Mb}, \mathbf{Mb} \rangle = (\mathbf{b}^T\mathbf{M}^T)(\mathbf{Mb}) = \mathbf{b}^T(\mathbf{M}^T\mathbf{M})\mathbf{b}$$



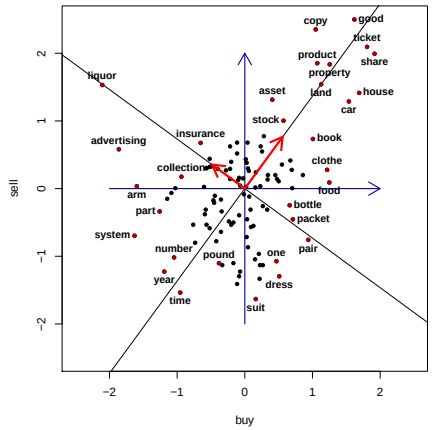

Stefan Evert (TU Darmstadt) | Dimensionality Reduction for DSM | wordspace.collocations.de | 29 / 50

Dimensionality reduction | Singular value decomposition (SVD)

Dimensionality reduction by orthogonal projection

- For one-dimensional subspace: $\mathbf{P}_A = \mathbf{b}\mathbf{b}^T$, so maximise

$$\|\mathbf{Mb}\|^2 = \langle \mathbf{Mb}, \mathbf{Mb} \rangle = (\mathbf{b}^T\mathbf{M}^T)(\mathbf{Mb}) = \mathbf{b}^T(\mathbf{M}^T\mathbf{M})\mathbf{b}$$




Stefan Evert (TU Darmstadt) | Dimensionality Reduction for DSM | wordspace.collocations.de | 29 / 50

## Dimensionality reduction by orthogonal projection

- For one-dimensional subspace: $\mathbf{P}_A = \mathbf{b}\mathbf{b}^T$, so maximise

$$\|\mathbf{M}\mathbf{b}\| = \langle \mathbf{M}\mathbf{b}, \mathbf{M}\mathbf{b} \rangle = (\mathbf{b}^T\mathbf{M}^T)(\mathbf{M}\mathbf{b}) = \mathbf{b}^T(\mathbf{M}^T\mathbf{M})\mathbf{b}$$

- Solution: $\mathbf{b}$ = eigenvector for largest eigenvalue of the symmetric, positive semi-definite **covariance matrix** $\mathbf{M}^T\mathbf{M}$
- Best $d$-dimensional subspace is given by orthogonal eigenvectors $\mathbf{b}_1, \ldots, \mathbf{b}_d$ corresponding to the $d$ largest eigenvalues $s_1 \geq s_2 \geq \ldots \geq s_d \geq 0$ of $\mathbf{M}^T\mathbf{M}$
- Quality of the approximation
  - $\|\mathbf{M}\mathbf{Q}_d\| = s_1 + \cdots + s_d$ vs. $\|\mathbf{M}\| = \sum_{i=1}^{n} s_i$
  - relative "importance" of dimension $\mathbf{b}_i$ given by $s_i / \|\mathbf{M}\|$





## Eigenvalue decomposition

- Symmetric, positive semi-definite matrix $\mathbf{M}^T\mathbf{M}$ has eigenvalue decomposition

$$\mathbf{M}^T\mathbf{M} = \mathbf{V} \cdot \mathbf{S} \cdot \mathbf{V}^T$$

where $\mathbf{V}$ is an orthogonal matrix of eigenvectors (columns) and $\mathbf{S} = \mathrm{Diag}(s_1, \ldots, s_n)$ a diagonal matrix of eigenvalues

$$\mathbf{V} = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} s_1 & & & & \\ & s_2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & s_n \end{bmatrix}$$

- Best $d$-dimensional subspace: $\mathbf{b}_1 = \mathbf{v}_1, \ldots, \mathbf{b}_d = \mathbf{v}_d$
- Dimensionality reduction: $\mathbf{M}_d = \mathbf{M}\mathbf{V}_d$





## Singular value decomposition (SVD)

- The idea of eigenvalue decomposition can be generalised to an arbitrary (non-symmetric, non-square) matrix $\mathbf{M}$
  - ☞ such a matrix need not have any eigenvalues
- **Singular value decomposition** (**SVD**) factorises $\mathbf{B}$ into

$$\mathbf{M} = \mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^T$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal coordinate transformations and $\mathbf{\Sigma}$ is a rectangular-diagonal matrix of **singular values** (with customary ordering $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$)

- Truncated SVD only computes first $d$ nonzero singular values
  - ☞ $\mathbf{\Sigma}$ is a square $d \times d$ matrix





## Truncated SVD illustration

$$\mathbf{M} \approx \tilde{\mathbf{M}}_d = \mathbf{U}_d \cdot \mathbf{\Sigma}_d \cdot \mathbf{V}_d^T$$

$$\begin{bmatrix} & n & \\ k & \tilde{\mathbf{M}} & \\ & & \end{bmatrix} = \begin{bmatrix} & d & \\ k & \mathbf{U} & \\ & & \end{bmatrix} \cdot \begin{bmatrix} \sigma_1 & d & \\ d & \ddots & \\ \mathbf{\Sigma} & & \sigma_d \end{bmatrix} \cdot \begin{bmatrix} & n & \\ d & \mathbf{V}^T & \\ & & \end{bmatrix}$$

## Dimensionality reduction by SVD

$$\mathbf{M}^T\mathbf{M} = (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)^T(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T) = \mathbf{V}\mathbf{\Sigma}\underbrace{\mathbf{U}^T\mathbf{U}}_{=\mathbf{I_d}}\mathbf{\Sigma}\mathbf{V}^T$$

$$= \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T$$

- Eigenvectors of $\mathbf{M}^T\mathbf{M}$ = right singular vectors of $\mathbf{M}$ (columns of $\mathbf{V}$) with eigenvalues $s_i = \sigma_i^2$, i.e. $\mathbf{S} = \mathbf{\Sigma}^2$
- Dimensionality reduction by SVD:

$$\tilde{\mathbf{M}}_d = \mathbf{M}\mathbf{V}_d = \mathbf{U}_d\mathbf{\Sigma}_d \qquad \text{(in } \mathbb{R}^d\text{)}$$

$$\tilde{\mathbf{M}}_d = \mathbf{M}\mathbf{V}_d\mathbf{V}_d^T = \mathbf{U}_d\mathbf{\Sigma}_d\mathbf{V}_d^T \qquad \text{(in original space)}$$

☞ "importance" of dimension $\mathbf{v}_i$ given by $\sigma_i^2$





## SVD dimensionality reduction example

| | buy | purchase | sell | write | read | draft | dim1 | dim2 |
|---|---|---|---|---|---|---|---|---|
| company | 84 | 7 | 47 | 2 − 1 | 1 | 0 | 2 | 96 |
| ticket | 177 | 14 | 99 | 8 − 1 | 2 − 1 | 1 | 8 | 203 |
| coffee | 20 | 2 | 11 | 0 | 0 | 0 | 0 | 23 |
| electricity | 8 | 1 | 4 | 1 | 0 | 0 | 1 | 9 |
| chocolate | 15 | 1 | 9 | 0 | 0 | 0 | 0 | 18 |
| letter | 6 − 2 | 0 | 4 − 1 | 948 | 230 | 25 | 976 | −2 |
| note | 1 | 0 | 1 | 174 | 42 | 5 | 179 | 0 |
| statement | 0 | 0 | 0 | 30 | 7 | 1 | 31 | 0 |
| agreement | 3 | 0 | 2 | 4 | 1 | 0 | 4 | 3 |

$$\tilde{\mathbf{M}}_2 = \mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^T = \mathbf{u}_1\sigma_1\mathbf{v}_1^T + \mathbf{u}_2\sigma_2\mathbf{v}_2^T \qquad \mathbf{M}\mathbf{V}_2 = \mathbf{U}_2\mathbf{\Sigma}_2$$





## Outline







## Interpretations of SVD

- "Noise reduction": projection into $d$-dimensional subspace
  - ☞ minimise cost = diplacement of points (Euclidean distance)
- Matrix approximation: $\tilde{\mathbf{M}}_d$ is best rank-$d$ approximation of $\mathbf{M}$
  - ☞ minimise Frobenius norm $\|\tilde{\mathbf{M}}_d - \mathbf{M}\|_2^2 = \sum_{i=1}^k \|\tilde{\mathbf{m}}_i - \mathbf{m}_i\|^2$
- Distance-preserving embedding into $d$-dimensional space
  - ☞ minimise $\sum_{i=1}^k \sum_{j=1}^k \left| \|\mathbf{m}_i - \mathbf{m}_j\|^2 - \|\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j\|^2 \right|$
  - ☞ principal component analysis (**PCA**) is best distance-preserving projection = SVD for column-centered $\mathbf{M}$ (i.e. $\sum_i \mathbf{m}_i = \mathbf{0}$)
- Latent class model (➜ latent meaning dimensions)
  - ☞ $\tilde{\mathbf{M}}_d = \sum_{i=1}^d \mathbf{u}_i\sigma_i\mathbf{v}_i^T$ (conditional independence given class $i$)

## SVD as a topic model

- Truncated SVD decomposition of term-document matrix:

$$\mathbf{F} \approx \tilde{\mathbf{F}} = \sum_{i=1}^{d} \mathbf{u}_i \sigma_i \mathbf{v}_i^T$$

  - $\sigma_i$ = prior frequency of topic $i$
  - $\mathbf{u}_i$ = word frequency distribution for topic $i$
  - $\mathbf{v}_i$ = contribution of topic $i$ to each document
  - ☞ assumes unscaled frequency counts $\mathbf{F}$

- This topic model is known as latent semantic indexing (LSI)
- Latent semantic analysis (LSA, Landauer and Dumais 1997) interprets topics as meaning components





## Interpretations of SVD

- "Noise reduction": projection into $d$-dimensional subspace
  - ☞ minimise cost = diplacement of points (Euclidean distance)
- Matrix approximation: $\tilde{\mathbf{M}}_d$ is best rank-$d$ approximation of $\mathbf{M}$
  - ☞ minimise Frobenius norm $\|\tilde{\mathbf{M}}_d - \mathbf{M}\|_2^2 = \sum_{i=1}^{k} \|\tilde{\mathbf{m}}_i - \mathbf{m}_i\|^2$
- Distance-preserving embedding into $d$-dimensional space
  - ☞ minimise $\sum_{i=1}^{k} \sum_{j=1}^{k} \left| \|\mathbf{m}_i - \mathbf{m}_j\|^2 - \|\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j\|^2 \right|$
  - ☞ principal component analysis (**PCA**) is best distance-preserving projection = SVD for column-centered $\mathbf{M}$ (i.e. $\sum_i \mathbf{m}_i = \mathbf{0}$)
- Latent class model (➜ latent meaning dimensions)
  - ☞ $\tilde{\mathbf{M}}_d = \sum_{i=1}^{d} \mathbf{u}_i \sigma_i \mathbf{v}_i^T$ (conditional independence given class $i$)
- Matrix factorization $\tilde{\mathbf{M}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ with $\mathbf{\Sigma} = \mathrm{Diag}(\sigma_1, \ldots, \sigma_d)$
  - ☞ SVD: Frobenius cost $\|\tilde{\mathbf{M}} - \mathbf{M}\|_2$, $\mathbf{U}, \mathbf{V}$ orthogonal, $\mathbf{\Sigma} \geq 0$
  - ☞ always implies a latent class model
  - ☞ $\mathbf{\Sigma}$ can be absorbed into $\mathbf{U}, \mathbf{V}$ under relaxed constraints





## Is SVD really a distance-preserving embedding?

- SVD equivalent to PCA only for column-centered matrix
  - centering destroys sparseness and non-negativity of $\mathbf{M}$
  - does not seem appropriate for highly skewed frequency data
  - PCA preserves Euclidean distance, but DSMs often use cosine

☞ SVD preserves **inner products** = cosine for normalised $\mathbf{M}$

- recall that $\cos\varphi = \mathbf{M}\mathbf{M}^T$ if $\|\mathbf{m}_i\| = 1 \; \forall i$

$$\mathbf{M}\mathbf{M}^T = \mathbf{U}\mathbf{\Sigma}\underbrace{\mathbf{V}^T\mathbf{V}}_{\mathbf{I}}\mathbf{\Sigma}\mathbf{U}^T = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^T$$

- since $\mathbf{U}$ is isometric, best rank-$d$ approximation to $\mathbf{M}\mathbf{M}^T$ is given by first singular values $\mathbf{U}_d\mathbf{\Sigma}_d^2\mathbf{U}_d^T = (\mathbf{U}_d\mathbf{\Sigma}_d)(\mathbf{U}_d\mathbf{\Sigma}_d)^T$

➥ $\tilde{\mathbf{M}}_d = \mathbf{U}_d\mathbf{\Sigma}_d$ preserves inner products $\langle \tilde{\mathbf{m}}_i, \tilde{\mathbf{m}}_j \rangle$
(and hence cosines computed without renormalization of $\tilde{\mathbf{M}}_d$)





## Outline

## Alternative dimensionality reduction techniques

different methods available depending on interpretation of SVD

- SVD as orthogonal projection
  - **random indexing** (RI) projects into random subspace
  - randomly generated unit basis vectors $\mathbf{b}_i$ (sparse or Gaussian) are approximately orthogonal, i.e. $\langle \mathbf{b}_i, \mathbf{b}_j \rangle \approx \delta_{ij}$
  - Johnson-Lindenstrauss lemma: distances are preserved well if $d$ is sufficiently high (cf. Papadimitriou *et al.* 1998)
  - ☞ no "noise reduction" effect (correlations not exploited)
- SVD as rank-$d$ matrix approximation
  - wrt. other cost function, e.g. $\|\tilde{\mathbf{M}} - \mathbf{M}\|_1$
  - I am not aware of any standard algorithm / implementation
- SVD as decorrelation
  - independent component analysis (**ICA**) has been applied to separation of word senses (Rapp 2003)
  - ☞ does not seem useful for dimensionality reduction





## Alternative dimensionality reduction techniques

different methods available depending on interpretation of SVD

- SVD as distance-preserving embedding
  - non-linear and non-metric embeddings: **kernel PCA**, (non-metric) **multidimensional scaling** (MDS), ...
- SVD as matrix factorization
  - **non-negative matrix factorization** (Lee and Seung 2001)
  - $\mathbf{M} \approx \mathbf{WH}$ with $\mathbf{W}, \mathbf{H} \geq 0$
  - cost function: Frobenius $\|\mathbf{M} - \mathbf{WH}\|_2$, cross-entropy, ...
  - ☞ expensive iterative algorithm, non-unique solution
- SVD as latent class (topic) model
  - probabilistic topic models are more plausible for frequency data, e.g. **PLSA** (Hoffmann 1999)
  - PLSA is equivalent to NMF with cross-entropy cost function
  - latent Dirichlet allocation (LDA) and other Bayesian models





## Outline





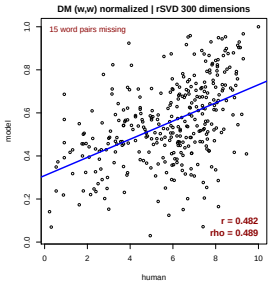


## A case study on the usefulness of dimensionality reduction

- Distributional Memory with $W_1 \times LW_2$ matricization
  - $k = 30{,}686$ target terms
  - $n = 3{,}127{,}436$ feature dimensions
- Two standard evaluation tasks
  - TOEFL synonym test (Landauer and Dumais 1997)
  - WordSim-353 semantic similarity ratings for 353 noun pairs (Finkelstein *et al.* 2002), with Spearman rank correlation $\rho$
- Dimensionality reduction techniques
  - feature selection (based on number of nonzero entries)
  - random indexing (RI) with sparse random vectors
  - RI + singular value decomposition (using randomized SVD)
  - aggregation: collapse DM tensor into $W_1 \times W_2$ matrix (yields $30{,}686 \times 30{,}686$ matrix with 6.41% nonzero cells)
- Caveat: no parameter optimization

## A case study on the usefulness of dimensionality reduction

| | TOEFL | WordSim |
|---|---|---|
| full 3.1M | 76.3% | .430 |
| top 1M | 76.3% | .430 |
| top 100k | 77.5% | .430 |
| top 5k | 71.3% | .400 |
| RI 5k | 76.3% | .439 |
| RI 1k | 78.8% | .400 |
| RI 6k + SVD 300 | 67.5% | .426 |
| $W_1 \times W_2$ full 30k | 76.3% | .461 |
| $W_1 \times W_2$ SVD 300 | 70.0% | .489 |








## Outline







## Things I love to talk about . . .

- Analysis of PLSA as matrix factorization
- Term-document vs. term-term matrix, higher-order models
  - ☞ can be illustrated nicely for sentence context
- Composition and dimensionality reduction
  - ☞ is vector multiplication etc. compatible with SVD?
- Sentence and document vectors
  - ☞ centroid? compositional?
- Non-linear dimensionality reduction techniques
  - ☞ useful for sparse high-dimensional vectors?
- Broad-scale evaluation and parameter optimization of DSM
  - ☞ single evaluation tasks give skewed picture
- Extension to tensor factorization
  - ☞ Tucker decomposition, non-negative tensor factorization





## References I

Baroni, Marco and Lenci, Alessandro (2010). Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, **36**(4), 673–712.

Evert, Stefan (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714. Available from `http://www.collocations.de/phd.html`.

Finkelstein, Lev; Gabrilovich, Evgeniy; Matias, Yossi; Rivlin, Ehud; Solan, Zach; Wolfman, Gadi; Ruppin, Eytan (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, **20**(1), 116–131.

Hoffmann, Thomas (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99)*.

Landauer, Thomas K. and Dumais, Susan T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, **104**(2), 211–240.

Lee, Daniel D. and Seung, H. Sebastian (2001). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13: Proceedings of the NIPS 2000 Conference*, pages 556–562. MIT Press.

# References II

Papadimitriou, Christos H.; Raghavan, Prabhakar; Tamaki, Hisao; Vempala, Santosh (1998). Latent semantic indexing: A probabilistic analysis. In *Proceedings of the 17th ACM Symposium on the Principles of Database Systems*, pages 159–168.

Rapp, Reinhard (2003). Die Erkennung semantischer Mehrdeutigkeiten mittels Unabhängigkeitsanalyse. In *Proceedings of the GLDV-Frühjahrstagung 2003*, Köthen, Germany.

Schütze, Hinrich (1992). Dimensions of meaning. In *Proceedings of Supercomputing '92*, pages 787–796, Minneapolis, MN.

Schütze, Hinrich (1998). Automatic word sense discrimination. *Computational Linguistics*, **24**(1), 97–123.