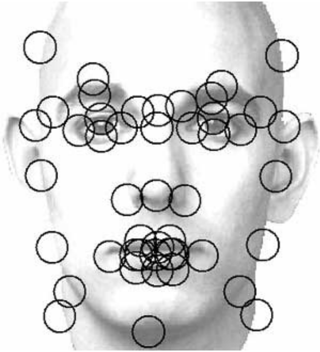
Non-identifiability of word embeddings, and connections to shape analysis

Simon Preston, Rachel Carrington, Karthik Bharath
simon.preston@nottingham.ac.uk

University of Nottingham

December 14, 2019

Shape and invariances

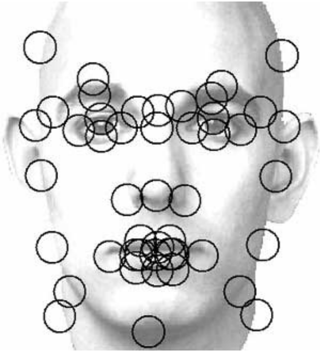


"Configuration"

$$\rightarrow \quad \overset{m\times k}{\mathbf{V}} = \begin{pmatrix} x_1 & x_2 & \cdots & x_k \\ y_1 & y_2 & \cdots & y_k \\ z_1 & z_2 & \cdots & z_k \end{pmatrix}$$

- "Shape": info invariant to translation, scaling, rotation (+ reflection)

Shape and invariances



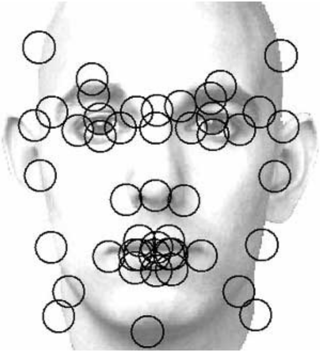
$\rightarrow \quad \overset{m\times k}{\mathbf{V}} = \begin{pmatrix} x_1 & x_2 & \cdots & x_k \\ y_1 & y_2 & \cdots & y_k \\ z_1 & z_2 & \cdots & z_k \end{pmatrix}$

"Configuration"

(Assume $\mathbf{V}$ is centred: $\mathbf{V1} = \mathbf{0}$)

- "Shape": info invariant to translation, scaling, rotation (+ reflection)

Shape and invariances



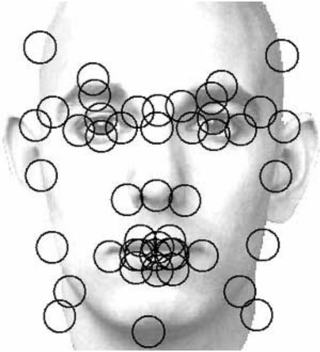
$\rightarrow \quad \overset{m\times k}{\mathbf{V}} = \begin{pmatrix} x_1 & x_2 & \cdots & x_k \\ y_1 & y_2 & \cdots & y_k \\ z_1 & z_2 & \cdots & z_k \end{pmatrix}$

"Configuration"

(Assume $\mathbf{V}$ is centred: $\mathbf{V}\mathbf{1} = \mathbf{0}$)

- "Shape": info invariant to translation, scaling, rotation (+ reflection)
- Can identify shape as $[\mathbf{V}] = \{c\mathbf{Q}\mathbf{V} : c \in \mathbb{R}^+; \mathbf{Q} \in O(m)\}$

Shape and invariances



"Configuration"

$$\rightarrow \quad \overset{m\times k}{\mathbf{V}} = \begin{pmatrix} x_1 & x_2 & \cdots & x_k \\ y_1 & y_2 & \cdots & y_k \\ z_1 & z_2 & \cdots & z_k \end{pmatrix}$$

(Assume $\mathbf{V}$ is centred: $\mathbf{V1} = \mathbf{0}$)

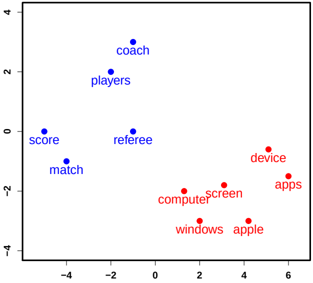
- "Shape": info invariant to translation, scaling, rotation (+ reflection)
- Can identify shape as $[\mathbf{V}] = \{c\mathbf{QV} : c \in \mathbb{R}^+; \mathbf{Q} \in O(m)\}$
- "Shape function": $g(\cdot)$ such that $g(\mathbf{V}) = g(c\mathbf{QV})$.

Text data

Representation, $\mathbf{X}$

$$
\begin{array}{cccccc|l}
\text{abide} & \text{absence} & \text{absent} & \text{abundant} & \text{abyss} & \cdots & \\
\hline
1 & 2 & 3 & 1 & 1 & \cdots & \text{doc } 1 \\
1 & 0 & 2 & 0 & 0 & \cdots & \text{doc } 2 \\
0 & 1 & 1 & 0 & 0 & \cdots & \text{doc } 3 \\
0 & 2 & 0 & 0 & 0 & \cdots & \text{doc } 4 \\
0 & 3 & 1 & 0 & 0 & \cdots & \text{doc } 5 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots &
\end{array}
$$

Word embedding, $\mathbf{V}$



Rise of Word Embedding Models

TITLE	CITED BY	YEAR
Distributed representations of words and phrases and their compositionality T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean Neural information processing systems	16538	2013
Efficient estimation of word representations in vector space T Mikolov, K Chen, G Corrado, J Dean arXiv preprint arXiv:1301.3781	13414	2013
Glove: Global vectors for word representation J Pennington, R Socher, C Manning Proceedings of the 2014 conference on empirical methods in natural language ...	10790	2014

Rise of Word Embedding Models

TITLE	CITED BY	YEAR
Distributed representations of words and phrases and their compositionality T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean Neural information processing systems	16538	2013
Efficient estimation of word representations in vector space T Mikolov, K Chen, G Corrado, J Dean arXiv preprint arXiv:1301.3781	13414	2013
Glove: Global vectors for word representation J Pennington, R Socher, C Manning Proceedings of the 2014 conference on empirical methods in natural language ...	10790	2014

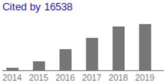
Cited by 16538

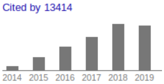
Cited by 13414

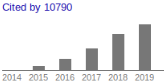
Cited by 10790

Rise of Word Embedding Models

TITLE	CITED BY	YEAR
Distributed representations of words and phrases and their compositionality T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean Neural information processing systems	16538	2013
Efficient estimation of word representations in vector space T Mikolov, K Chen, G Corrado, J Dean arXiv preprint arXiv:1301.3781	13414	2013
Glove: Global vectors for word representation J Pennington, R Socher, C Manning Proceedings of the 2014 conference on empirical methods in natural language …	10790	2014







Invariance and identifiability issues for word embeddings R Carrington, K Bharath, S Preston Advances in Neural Information Processing Systems, 15114-15123	2019

Why word embeddings?

- Word embedding $\mathbf{V}$ encodes word meaning.
- Used for, and evaluated on, word tasks.
- Word similarity: "given word A, how similar is word B?"
- Word association: "A is to B as C is to what?", e.g. Paris is to France as Madrid is to ...?
- Task performance measured by $g(\text{data}, \mathbf{V})$.

Word embeddings as matrix factorisation

Simple word embedding model: take $\mathbf{V}$ as minimiser of

$$\|\mathbf{X} - \mathbf{U}\mathbf{V}\|^2 = \sum_{ij} \left(x_{ij} - \mathbf{u}_i^\top \mathbf{v}_j \right)^2$$

Word embeddings as matrix factorisation

Simple word embedding model: take **V** as minimiser of

$$\|\mathbf{X} - \mathbf{U}\mathbf{V}\|^2 = \sum_{ij} \left(x_{ij} - \mathbf{u}_i^\top \mathbf{v}_j \right)^2$$

Solution for **V**: write SVD of $\mathbf{X} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^\top$. Then $\|\mathbf{X} - \mathbf{X}_d\|$ is minimised by $\mathbf{X}_d = \mathbf{A}_d \boldsymbol{\Sigma}_d \mathbf{B}_d^T$, so take

$$\mathbf{U}^* = \mathbf{A}_d, \quad \mathbf{V}^* = \boldsymbol{\Sigma}_d \mathbf{B}_d^\top.$$

Word embeddings as matrix factorisation

Simple word embedding model: take $\mathbf{V}$ as minimiser of

$$\|\mathbf{X} - \mathbf{U}\mathbf{V}\|^2 = \sum_{ij} \left(x_{ij} - \mathbf{u}_i^\top \mathbf{v}_j \right)^2$$

Solution for $\mathbf{V}$: write SVD of $\mathbf{X} = \mathbf{A}\mathbf{\Sigma}\mathbf{B}^\top$. Then $\|\mathbf{X} - \mathbf{X}_d\|$ is minimised by $\mathbf{X}_d = \mathbf{A}_d \mathbf{\Sigma}_d \mathbf{B}_d^T$, so take

$$\mathbf{U}^* = \mathbf{A}_d, \quad \mathbf{V}^* = \mathbf{\Sigma}_d \mathbf{B}_d^\top.$$

... or $\mathbf{V}^* = \mathbf{\Sigma}_d^{1-\alpha} \mathbf{B}_d^\top$?

Word embeddings as matrix factorisation

Simple word embedding model: take $\mathbf{V}$ as minimiser of

$$\|\mathbf{X} - \mathbf{U}\mathbf{V}\|^2 = \sum_{ij} \left(x_{ij} - \mathbf{u}_i^\top \mathbf{v}_j \right)^2$$

Solution for $\mathbf{V}$: write SVD of $\mathbf{X} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^\top$. Then $\|\mathbf{X} - \mathbf{X}_d\|$ is minimised by $\mathbf{X}_d = \mathbf{A}_d \boldsymbol{\Sigma}_d \mathbf{B}_d^T$, so take

$$\mathbf{U}^* = \mathbf{A}_d, \quad \mathbf{V}^* = \boldsymbol{\Sigma}_d \mathbf{B}_d^\top.$$

... or $\mathbf{V}^* = \boldsymbol{\Sigma}_d^{1-\alpha} \mathbf{B}_d^\top$?

... or $\mathbf{V}^* = \mathbf{C}\mathbf{B}_d^\top$ for any $\mathbf{C} \in GL(d)$?!

Word embeddings as matrix factorisation

Simple word embedding model: take $\mathbf{V}$ as minimiser of

$$\|\mathbf{X} - \mathbf{U}\mathbf{V}\|^2 = \sum_{ij} \left(x_{ij} - \mathbf{u}_i^\top \mathbf{v}_j \right)^2 = f(\mathbf{X}, \mathbf{U}\mathbf{V})$$

Solution for $\mathbf{V}$: write SVD of $\mathbf{X} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^\top$. Then $\|\mathbf{X} - \mathbf{X}_d\|$ is minimised by $\mathbf{X}_d = \mathbf{A}_d \boldsymbol{\Sigma}_d \mathbf{B}_d^T$, so take

$$\mathbf{U}^* = \mathbf{A}_d, \quad \mathbf{V}^* = \boldsymbol{\Sigma}_d \mathbf{B}_d^\top.$$

... or $\mathbf{V}^* = \boldsymbol{\Sigma}_d^{1-\alpha} \mathbf{B}_d^\top$?

... or $\mathbf{V}^* = \mathbf{C}\mathbf{B}_d^\top$ for any $\mathbf{C} \in GL(d)$?!

Non-identifiability: $f(\mathbf{X}, \mathbf{U}\mathbf{V}) = f(\mathbf{X}, \mathbf{U}\mathbf{C}^{-1}\mathbf{C}\mathbf{V})$

Non-identifiability of different embedding models

$$\textbf{LSA:}\ \sum_{ij} \left(x_{ij} - \mathbf{u}_i^\top \mathbf{v}_j\right)^2$$

$$\textbf{word2vec:}\ \sum_{ij} \log\left(\sigma\left(\mathbf{u}_i^T \mathbf{v}_j\right)\right) + k \cdot \frac{\sum_l x_{il} \sum_m x_{mj}}{\sum_{ij} x_{ij}} \log\left(\sigma\left(-\mathbf{u}_i^T \mathbf{v}_j\right)\right)$$

$$\textbf{GloVe:}\ \sum_{ij} h\left(x_{ij}\right)\left(\mathbf{u}_i^T \mathbf{v}_j - h_1\left(x_{ij}\right)\right)^2$$

Non-identifiability of different embedding models

$$\textbf{LSA:} \ \sum_{ij} \left(x_{ij} - \mathbf{u}_i^\top \mathbf{v}_j\right)^2$$

$$\textbf{word2vec:} \ \sum_{ij} \log\left(\sigma\left(\mathbf{u}_i^T \mathbf{v}_j\right)\right) + k \cdot \frac{\sum_l x_{il} \sum_m x_{mj}}{\sum_{ij} x_{ij}} \log\left(\sigma\left(-\mathbf{u}_i^T \mathbf{v}_j\right)\right)$$

$$\textbf{GloVe:} \ \sum_{ij} h\left(x_{ij}\right)\left(\mathbf{u}_i^T \mathbf{v}_j - h_1\left(x_{ij}\right)\right)^2$$

Each is such that for any particular solution

$$\mathbf{V}^* = \arg\min_{\mathbf{V}} \ f(\mathbf{X}, \mathbf{UV})$$

a general solution set is

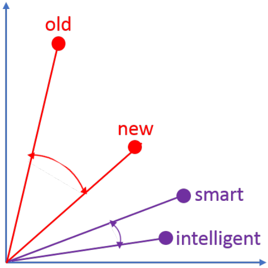
$$\{\mathbf{V} : \mathbf{V} = \mathbf{CV}^*, \mathbf{C} \in \mathrm{GL}(r)\}$$

Non-identifiability - implications? (1)

Word similarity task:

data

Word 1	Word 2	Human Score (y_i)	Embedding Score (z_i)
old	new	1.58	$\cos(\mathbf{v}_{\text{"old"}}, \mathbf{v}_{\text{"new"}})$
smart	intelligent	9.2	*etc*
hard	difficult	8.77	...
happy	cheerful	9.55	...
hard	easy	0.95	...
fast	rapid	8.75	...



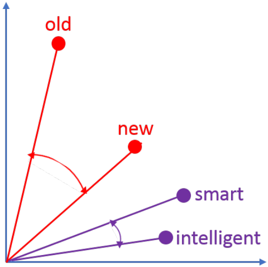
$$\cos(\mathbf{v}_i, \mathbf{v}_j) = \mathbf{v}_i^\top \mathbf{v}_j / (\|\mathbf{v}_i\| \cdot \|\mathbf{v}_j\|)$$

Non-identifiability - implications? (1)

Word similarity task:

data

Word 1	Word 2	Human Score (y_i)	Embedding Score (z_i)
old	new	1.58	$\cos(\mathbf{v}_{\text{“old”}}, \mathbf{v}_{\text{“new”}})$
smart	intelligent	9.2	*etc*
hard	difficult	8.77	...
happy	cheerful	9.55	...
hard	easy	0.95	...
fast	rapid	8.75	...



$$\cos(\mathbf{v}_i, \mathbf{v}_j) = \mathbf{v}_i^\top \mathbf{v}_j / \left(\|\mathbf{v}_i\| \cdot \|\mathbf{v}_j\|\right)$$

...then measure “embedding performance” by

$$g(\text{data}, \mathbf{V}) = \text{corr}(\{y_i\}, \{z_i\})$$

Non-identifiability - implications? (2)

Word analogy task:

- Paris is to France as Madrid is to . . . ? Given $\mathbf{V}$ solve

$$\arg\max_i \left[\cos(\mathbf{v}_i, \mathbf{v}_{\text{“France”}}) - \cos(\mathbf{v}_i, \mathbf{v}_{\text{“Paris”}}) + \cos(\mathbf{v}_i, \mathbf{v}_{\text{“Madrid”}})\right]$$

- The data are a set of human-chosen analogies.
- Performance metric $g(\text{data}, \mathbf{V})$ is the proportion correct.

Non-identifiability - implications? (2)

Word analogy task:

- Paris is to France as Madrid is to ...? Given $\mathbf{V}$ solve

$$\arg\max_i \left[\cos(\mathbf{v}_i, \mathbf{v}_{\text{“France”}}) - \cos(\mathbf{v}_i, \mathbf{v}_{\text{“Paris”}}) + \cos(\mathbf{v}_i, \mathbf{v}_{\text{“Madrid”}})\right]$$

- The data are a set of human-chosen analogies.
- Performance metric $g(\text{data}, \mathbf{V})$ is the proportion correct.

For both similarity and analogy, g depends on $\mathbf{V}$ only via $\cos(\mathbf{v}_i, \mathbf{v}_j)$.

Hence $g(\text{data}, \mathbf{V}) = g(\text{data}, c\mathbf{Q}\mathbf{V}) \Rightarrow g$ is a shape function.

Mis-match of invariances

Training objective $f(\mathbf{X}, \mathbf{UV})$ invariant to $\mathbf{V} \mapsto \mathbf{CV}$.

Test objective $g(\text{data}, \mathbf{V})$ invariant to $\mathbf{V} \mapsto c\mathbf{QV}$

$$\begin{aligned} f\left(\boldsymbol{X}, \boldsymbol{UV}\right) &= f\left(\boldsymbol{X}, \boldsymbol{UC}^{-1}\boldsymbol{CV}\right), \qquad && \boldsymbol{C} \in \mathrm{GL}(r) \\ g\left(\boldsymbol{D}, \boldsymbol{V}\right) &= g\left(\boldsymbol{D}, c\boldsymbol{QV}\right), \qquad && \boldsymbol{Q} \in \mathrm{O}(d), c \in \mathbb{R} \end{aligned}$$

What is the set $\mathcal{F}_d \subset \mathrm{GL}(d)$ which leaves f invariant but not g?

Mis-match of invariances

Training objective $f(\mathbf{X}, \mathbf{UV})$ invariant to $\mathbf{V} \mapsto \mathbf{CV}$.

Test objective $g(\text{data}, \mathbf{V})$ invariant to $\mathbf{V} \mapsto c\mathbf{QV}$

$$\begin{aligned} f(\boldsymbol{X}, \boldsymbol{UV}) &= f\left(\boldsymbol{X}, \boldsymbol{UC}^{-1}\boldsymbol{CV}\right), \qquad && \boldsymbol{C} \in \mathrm{GL}(r) \\ g(\boldsymbol{D}, \boldsymbol{V}) &= g(\boldsymbol{D}, c\boldsymbol{QV}), \qquad && \boldsymbol{Q} \in \mathrm{O}(d), c \in \mathbb{R} \end{aligned}$$

What is the set $\mathcal{F}_d \subset \mathrm{GL}(d)$ which leaves f invariant but not g?

Write $\mathcal{F}_d = \tilde{\mathcal{F}}_d - c\mathcal{I}$, where

- $\tilde{\mathcal{F}}_d = \mathrm{GL}(d) \backslash \mathrm{O}(d)$, and can be identified with $\mathrm{UT}(d)$, upper tringular matrices with +ve diag elements. (Intuition: QR decomposition of C)
- $c\mathcal{I} = \{cI_d : c \in \mathbb{R}\}$ is set of scale transformations
- dimension of $\mathcal{F}_d$ is $d(d-1)/2 - 1$.

Is the mis-match a problem?

- "When all methods are allowed to tune a similar set of hyperparameters their performance is largely comparable"[1]

[1] Levy, Goldberg, Dagan, *Trans. Assoc. Comput. Ling.*, 2015

Is the mis-match a problem?

- "When all methods are allowed to tune a similar set of hyperparameters their performance is largely comparable"[1]
- Some hyperparameters index different elements of solution set f, chosen for performance in g, e.g. $\mathbf{V}^* = \mathbf{\Sigma}_d^{1-\alpha}\mathbf{B}_d^\top$
- f typically optimised by Monte Carlo (different solns explained as local optima - but also due to non-identifiabilty) then soln chosen for g.

[1] Levy, Goldberg, Dagan, *Trans. Assoc. Comput. Ling.*, 2015

Is the mis-match a problem?

- "When all methods are allowed to tune a similar set of hyperparameters their performance is largely comparable"[1]
- Some hyperparameters index different elements of solution set f, chosen for performance in g, e.g. $\mathbf{V}^* = \mathbf{\Sigma}_d^{1-\alpha}\mathbf{B}_d^\top$
- f typically optimised by Monte Carlo (different solns explained as local optima - but also due to non-identifiabilty) then soln chosen for g.
- Both are (implicitly) *supervised* approaches.

[1] Levy, Goldberg, Dagan, *Trans. Assoc. Comput. Ling.*, 2015

Resolving mis-match

$$\arg\min_{\mathbf{V}} \; f(\mathbf{X}, \mathbf{UV})$$

Identifying a solution unique up to orthogonal transformations:

- Impose constraint $\mathbf{V}\mathbf{V}^\top = \mathbf{I}$, then for any solution $\mathbf{V}^*$ any other solution $\mathbf{C}\mathbf{V}^*$ for $\mathbf{C} \in \mathrm{GL(d)}$ satisfies $g(\mathrm{data}, \mathbf{C}\mathbf{V}^*) = g(\mathrm{data}, \mathbf{V}^*)$.

Resolving mis-match

$$\arg\min_{\mathbf{V}} f(\mathbf{X}, \mathbf{UV})$$

Identifying a solution unique up to orthogonal transformations:

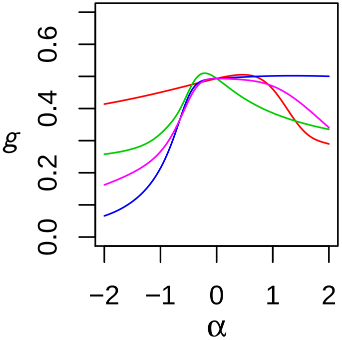
- Impose constraint $\mathbf{V}\mathbf{V}^\top = \mathbf{I}$, then for any solution $\mathbf{V}^*$ any other solution $\mathbf{C}\mathbf{V}^*$ for $\mathbf{C} \in \mathrm{GL(d)}$ satisfies $g(\mathrm{data}, \mathbf{C}\mathbf{V}^*) = g(\mathrm{data}, \mathbf{V}^*)$.

Identifying a unique solution:

- Additionally impose: (i) $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$, (ii) $\mathrm{diag}(\mathbf{U}^\top\mathbf{U})$ decreasing, (iii) positive first non-zero elements of each col of $\mathbf{U}$.

Sensitivity to particular solution

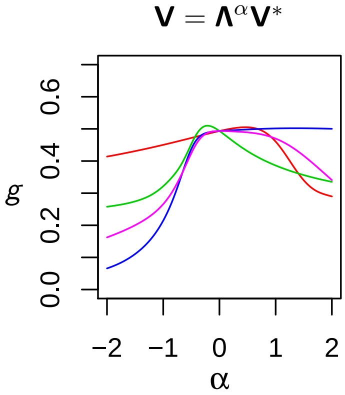
$$\mathbf{V} = \mathbf{\Lambda}^{\alpha} \mathbf{V}^{*}$$



$\mathbf{\Lambda} = \mathbf{\Sigma}_d$ $\quad (\mathbf{\Lambda})_i \sim |N(0,1)|$

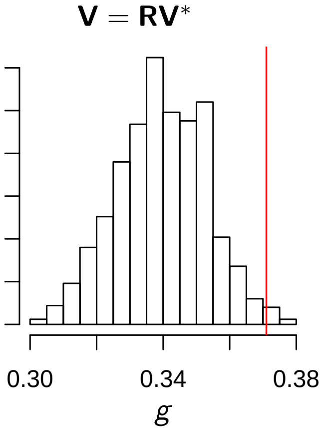
$(\mathbf{\Lambda})_i = i$ $\quad (\mathbf{\Lambda})_i \sim U(0,1)$

Sensitivity to particular solution



$\mathbf{\Lambda} = \mathbf{\Sigma}_d$ $\quad (\mathbf{\Lambda})_i \sim |N(0,1)|$

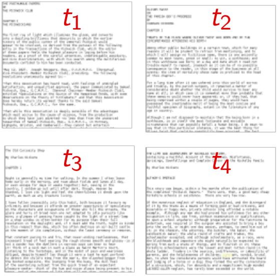
$(\mathbf{\Lambda})_i = i$ $\quad (\mathbf{\Lambda})_i \sim U(0,1)$



$(\mathbf{R})_{ij} \sim |N(0,1)|;\ j \geq i$

Outlook: dynamic embedding

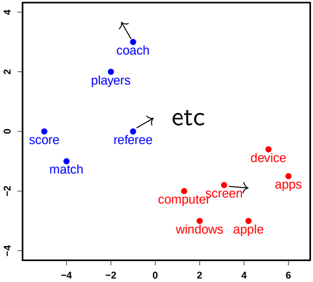
Text data



Representation, $\mathbf{X}$

$$
\begin{array}{c}
\begin{array}{cccccc} \text{abide} & \text{absence} & \text{absent} & \text{abundant} & \text{abyss} & \cdots \end{array} \\
\begin{pmatrix}
1 & 2 & 3 & 1 & 1 & \cdots \\
1 & 0 & 2 & 0 & 0 & \cdots \\
0 & 1 & 1 & 0 & 0 & \cdots \\
0 & 2 & 0 & 0 & 0 & \cdots \\
0 & 3 & 1 & 0 & 0 & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots
\end{pmatrix}
\end{array}
\begin{array}{l}
\\
\text{doc } 1 \quad t_1 \\
\text{doc } 2 \quad t_2 \\
\text{doc } 3 \quad t_3 \\
\text{doc } 4 \quad t_4 \\
\text{doc } 5 \quad t_5 \\
\end{array}
$$

Word embedding, $\mathbf{V}(t)$



Acknowledgements

Invariance and identifiability issues for word embeddings. Rachel Carrington, Karthik Bharath & Simon Preston. *NeurIPS, 2019*.

EPSRC Engineering and Physical Sciences Research Council

Bloomberg