Permutation invariant matrix models and natural language data

Sanjaye Ramgoolam

Queen Mary, University of London

String Data, 2021
Wits University, Johannesburg

Based on

D. Kartsaklis, S. Ramgoolam, M. Sadrzadeh, "**Linguistic Matrix Theory**,"**2017** arXiv:1703.10252 [cs.CL], Annales de l'Institut Henri Poincare D, 2018
S. Ramgoolam, "**Permutation invariant Gaussian matrix Models**," **2018** arXiv:1809.07559 [hep-th], Nuclear Physics B, 2019
S. Ramgoolam, M. Sadrzadeh, L. Sword, "**Gaussianity and typicality in matrix distributional semantics**," **2019** arXiv:1912.10839 [hep-th] Annales de l'Institut Henri Poincare D (to appear).
G. Barnes, A. Padellaro, S. Ramgoolam, " **Permutation invariant Gaussian 2-Matrix models,**" **2021-04** arXiv:2104.03707
G. Barnes, A. Padellaro, S. Ramgoolam, "**Hidden symmetries and large N factorisation for permutation invariant matrix observables**," **2021-12** arXiv:2112.00498 [hep-th]
M. Accettulli Huber, A. Correa, S. Ramgoolam, M. Sadrzadeh, "**Permutation invariant matrix statistics and computational language tasks**," **2022-01** to appear arXiv:2201.XXXX [cs.CL]

Introduction

- We are studying an **ensemble of matrices** $\{X^A_{i,j}\}$. A is a label running over a finite set of objects, e.g. a finite set of words in a language, say $200 - 300$ words. The indices i, j are running over $i, j \in \{1, 2, \cdots, D\}$, D between 100 to 2000.

- **Compositional distributional semantics** (Coecke, Sadrzadeh,Clark, 2010) - in computational linguistics - provides algorithms for the construction of ensembles of matrices associated to certain classes of words, depending on their grammatical structure.

- The LMT programme develops **Permutation Invariant Random Matrix Theory** to study the statistics of these matrices.

Introduction

Theoretical developments:

- Enumeration and construction of permutation invariant matrix observables : PIMOs are associated to directed graphs (LMT-2017 paper and 2-Matrix paper in 04/2021).
- General permutation invariant Gaussian action : There is a 13-parameter family of Gaussian actions. Parameterisation of the couplings done using representation theory of S_D. Expectation values can be computed using Wick's theorem (2018 paper).
- Computer code available for these computations as part of the 2-Matrix paper 04/2021.
- Large N factorisation property for the PIMOs 12/2021.

Introduction

Statistical Physics and computational tasks with Natural Language Data:

- Evidence for approximate Gaussianity in matrix data from compositional distributional semantics (2019 paper).
- New geometrical and statistical tools, based on PIMOs for natural language tasks involving synonmyms, antonyms, hypernyms and hyponyms. (2022 paper).
- Remark : The distinction between synonyms and antonyms is an important challenge in distributional semantics.

OUTLINE

1. Compositional distributional semantics :
   - vectors, matrices M and tensors for words ..
   - Permutation invariance motivated from the extraction of meaning in CDS.
2. Gaussianity in permutation invariant sector
   - Based on RMT and particle physics.
   - 5-parameter theoretical model and tests :
3. Representation theory approach to the general 13-parameter model
   - Graph basis and Rep Basis for quadratic invariants.
   - Comparison to data of the 13-parameter model: approximate Gaussianity and quantifying departures from Gaussianity.
4. PIMOs and Language tasks.

Part 1 : Distributional Semantics and Word Vectors

The idea is that the **meaning of a word** is captured by the **contexts** in which it occurs (Harris 1954, Firth 1957).

One constructs **vectors** for words, from the frequencies with which they occur in the vicinity of a specified set of context words, using a collection of text (corpus).

For example if we choose context words (pet , feed), then the co-occurrence frequencies of a word W

$$\begin{aligned} a(W) &= \text{co-occurrence frequency of } W \text{ with pet} \\ b(W) &= \text{co-occurrence frequency of } W \text{ with feed} \end{aligned}$$

are used to define a vector $Vector(W)$

$$\text{Vector } (W) = \begin{pmatrix} a(W) \\ b(W) \end{pmatrix}$$

For example, the frequencies of Cat , Dog and Baby in a collection of books might be

$$\begin{aligned} Cat &= (103, 92) \\ Dog &= (99, 102) \\ Baby &= (2, 80) \end{aligned}$$

The similarity in meaning of Cat and Dog is reflected in the fact that they are vectors pointing in approximately the same direction in the space spanned by pet and feed.

$$v = a(pet) + b(feed)$$

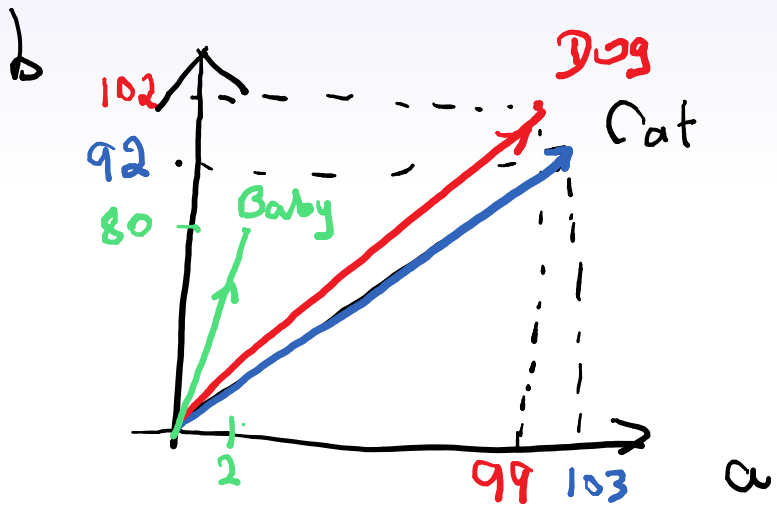


Figure: Example of Vectors from frequencies

Word vectors have been used in tasks such as computing word similarity.

Compositional Distributional Semantics

In Coecke, Sadrzadeh, Clark (2010), formal models of grammatical composition were combined with distributional semantics, to develop compositional distributional semantics

This allows the construction of the meaning of sentences (and composite expressions) from the meaning of the constituent words.

In this framework, words are associated with vectors, matrices, and higher rank tensors, depending on their grammatical type.

Part 2: Linguistic Matrix Theory : Permutation invariance and Gaussianity

The focus of LMT was to study the statistics of matrices associated to adjectives (and separately to verbs).

It is not practical to make sense of the distributions of the individual matrix elements.

We use the ideas of Random Matrix Theory (RMT) : We want to model aspects of a complex system (in this case adjective matrices in natural language) using probability distributions over large matrices.

Following Wigner and Dyson, RMT has been applied to complex nuclei, molecules, chaotic systems, financial correlators, biological networks etc.

Random Matrix Theory - Matrix Integrals

The simplest RMTs are defined by

$$\mathcal{Z}(M) = \int dM\, e^{-\mathrm{tr}\ \mathrm{M}^2}$$

M is a hermitian matrix (the Hamiltonian in Wigner's original applications).

The weight is invariant under $U(D)$ symmetry

$$M \to UMU^\dagger$$
$$\mathrm{tr}\ \mathrm{M}^2 \to \mathrm{tr}\ \mathrm{M}^2$$

Symmetry

The unitary invariance corresponds to base changes in the Hilbert space.

In the context of the Linguistic Application, some of the observables considered have continuous symmetries.

E.g. Given two vectors p_i and q_i, the inner product

$$\sum_{i=1}^{D} p_i q_i$$

is used to measure word similarity.

Symmetry

But in general, we don't expect all the information in the word vectors/tensors to be invariant under these continuous symmetries.

The symmetry of permutations S_D is still expected - as it corresponds to re-ordering the context words.

Indeed if we consider the quantity (Kullback-Leibler divergence)

$$\sum_i p_i \log p_i - p_i \log q_i$$

which is used to measure entailment - this is S_D invariant but not $O(D)$ invariant.

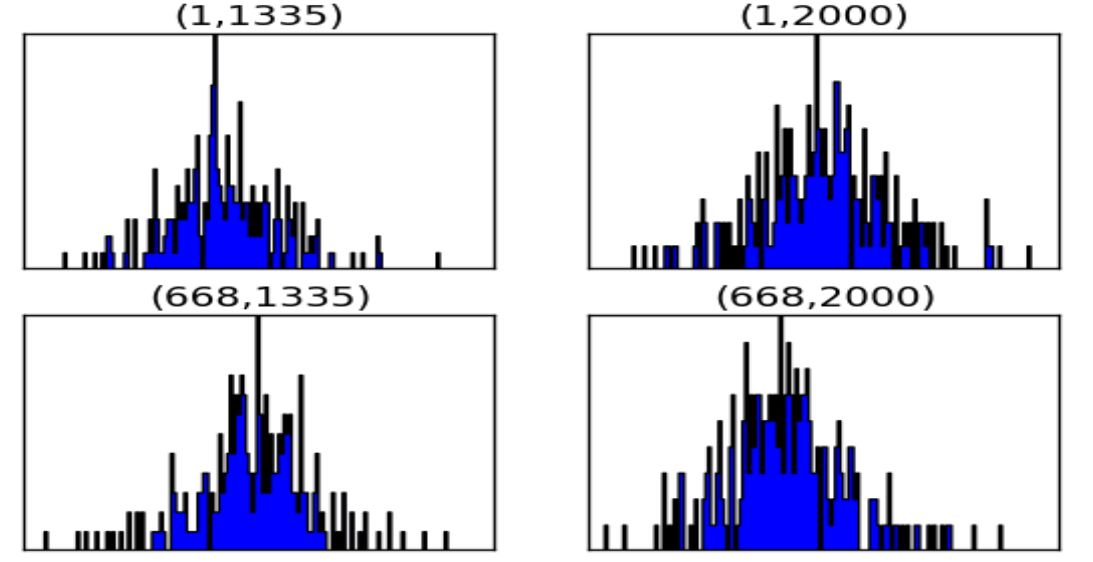
LMT-2017 : Building matrices.

In the paper, LMT-2017, the CS experts constructed an ensemble of matrices for adjectives. Build vectors for nouns such as “car” using frequency counts, processed according to traditional Distributional Semantics methods ; and also noun phrases such as “black car”. Then use linear regression to find matrix for “black”, optimised to ensure that the matrix “black” applied to noun-vector “X” gives the vector for “black X”.

This follows earlier work in CDS, e.g. (Baroni, Bernardi, Zamparelli, 2014 ; Baroni, Zamparelli, 2010, Grefenstette, Dinu, Zhang, Sadrzadeh, Baroni, 2013)

Other methods involve machine learning (Sadrzadeh, Wijnholds, Clark 2020)

Qualitative inspection of random matrix elements is supportive of the idea of Gaussians for the LR models.



Recap : One-variable Gaussian (normal) distributions.

The shape is described by a function of the height x, and parameters μ, σ

$$f(x : \mu, \sigma) = \frac{e^{\frac{-\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}}$$

The mean and standard deviation are calculated by doing integrals

$$< x > = \int_{-\infty}^{\infty} dx\ f(x : \mu, \sigma)\ x$$
$$< x^2 > = \int_{-\infty}^{\infty} dx\ f(x : \mu, \sigma)\ x^2$$

$$< x >= \mu$$
$$< x^2 >= \mu^2 + \sigma^2$$

Intro - Gaussian (normal) distributions.

All the higher moments $< x^n >$ (for $n \geq 3$))

$$< x^n >= \int_{-\infty}^{\infty} dx\ f(x : \mu, \sigma)\ x^n$$

are functions of μ, σ, hence are determined by $< x >, < x^2 >$.

In modeling a data set by the Gaussian model, calculate $< x >_{expt}, < x^2 >_{expt}$ to determine μ, σ. Then predict and compare.

LMT : Permutation Invariant Gaussian Matrix theory for NLP data

The partition function of the 5-parameter model in LMT-2017 is

$$\mathcal{Z}(\Lambda, a, b, J^0, J^S) = \int dM\, e^{-\frac{\Lambda}{2}\sum_{i=1}^{D} M_{ii}^2 - \frac{1}{4}(a+b)\sum_{i<j}(M_{ij}^2+M_{ji}^2)} e^{-\frac{1}{2}(a-b)\sum_{i<j} M_{ij}M_{ji} + J^0\sum_i M_{ii} + J^S\sum_{i<j}(M_{ij}+M_{ji})}$$

The observables of the model are S_D invariant polynomials in the matrix variables:

$$f(M_{i,j}) = f(M_{\sigma(i),\sigma(j)})$$

Expectation values of $f(M)$ are computed as

$$\langle f(M)\rangle \equiv \frac{1}{\mathcal{Z}}\int dM f(M)\mathrm{EXP}$$

as functions of J_0, J_S, Λ, a, b.

Computations of expectation values in a Gaussian theory use the formula

$$\mathcal{Z} = \int dx \exp\left(-\frac{1}{2}\sum_{i,j=1}^{N} x_i A_{ij} x_j + \sum_i s_i x_i\right) = \sqrt{\frac{(2\pi)^N}{\det A}} \exp\left(\frac{1}{2} s_i (A^{-1})_{ij} s_j\right) .$$

which is fundamental to perturbative QFT.

In the above, we have a product of D integrals for M_{ii} and there are $D(D-1)/2$ decoupled integrals involving, for each $1 < i < j < D$, a quadratic form involving $M_{ij}^2, M_{ji}^2, M_{ij}M_{ji}$. The computation of the generating function for expectation values involves inverting these 2×2 quadratic forms.

Permutation Invariant Gaussian Matrix theory : Fixing parameters from data

The linear averages

$$\langle \sum_i M_{ii} \rangle \quad \langle \sum_{i \neq j} M_{ij} \rangle$$

and the quadratic averages

$$\langle \sum_i M_{ii}^2 \rangle \quad \langle \sum_{i \neq j} M_{ij}^2 \rangle \quad \langle \sum_{i \neq j} M_{ij} M_{ji} \rangle$$

are compared with corresponding experimental averages

$$\langle f(M) \rangle_{EXPT} = \frac{1}{N_w} \sum_w f(M^w)$$

obtained by summing over words. $f(M)$ involves sum over matrix elements.

The comparison determines the 5 parameters of the model J_0, J_S, Λ, a, b.

Comparing cubic and quartic averages

With the parameters of the Gaussians thus fixed, we compare theory and experiment.

$$M_{d:3} \equiv \sum_i \langle M_{ii}^3 \rangle$$

$$M_{o:3,1} \equiv \sum_{i \neq j} \langle M_{ij}^3 \rangle$$

$$M_{o:3,2} = \sum_{i \neq j \neq k} \langle M_{ij} M_{jk} M_{ki} \rangle$$

For simpler observables, theory and expt within 40 percent

$$\frac{(M_{d:3})_{THRY}}{(M_{d:3})_{EXPT}} = 0.57$$

near $D = 2000$.

Within realm of perturbation theory.

For $M_{o:3,2}$ large difference. Ratio is 0.013

PART 3: General permutation invariant Gaussian matrix model.

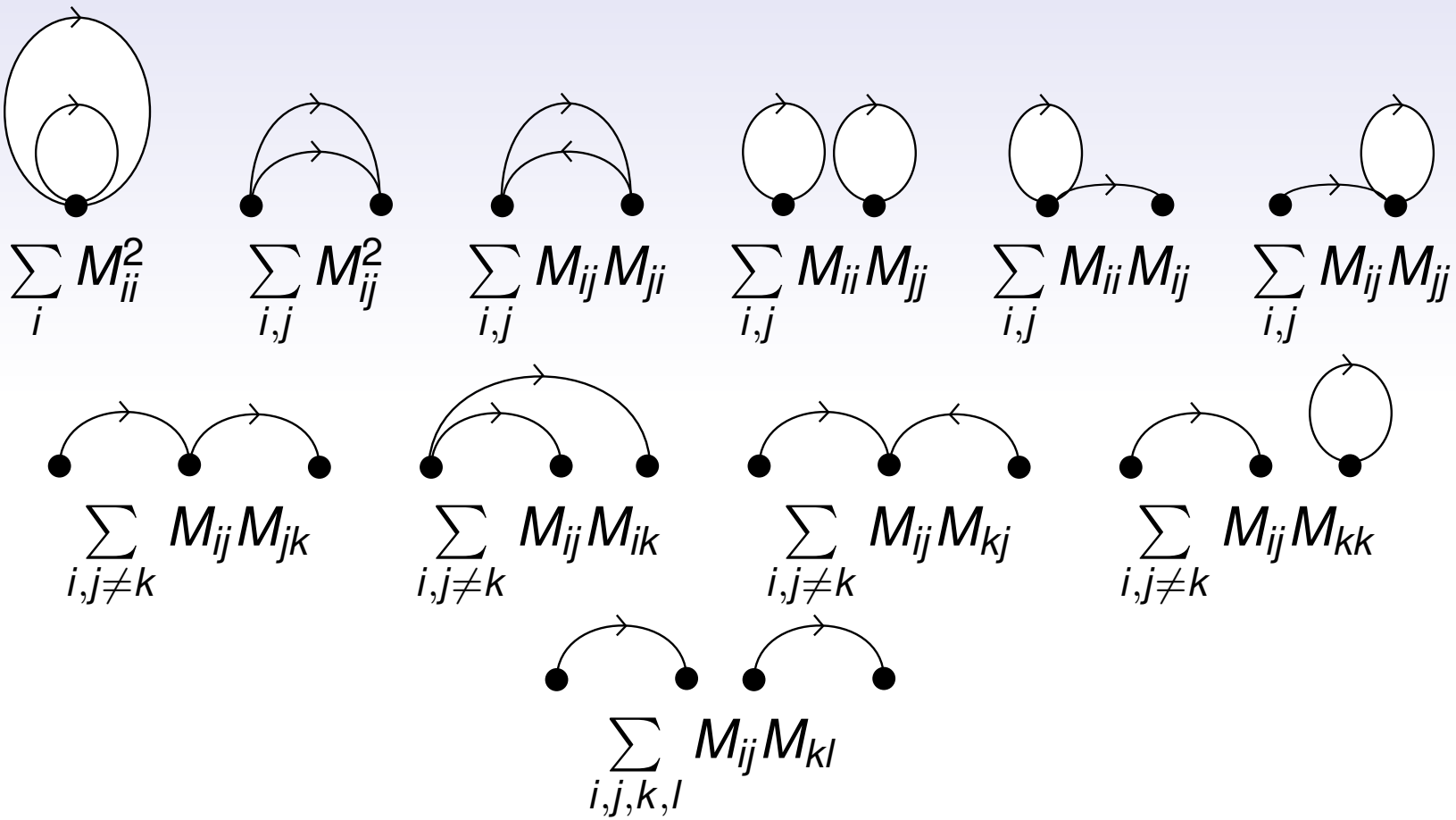


Figure: S_D invariant functions and corresponding graphs illustrated for the 11 quadratic invariants

The most general permutation invariant Gaussian will have 13 parameters. 2 for the linear invariants, and 11 for the quadratic invariants.

$$e^{-\sum_{a=1}^{11} \lambda_a M_{ij} B_a^{ijkl} M_{kl}}$$

To solve this Gaussian model, we want to invert this quadratic form.

$$\sum_a \lambda_a B_a^{ijkl} = \lambda_1 \delta^{ij}\delta^{jk}\delta^{kl} + \lambda_2 \delta^{ik}\delta^{jl} + \cdots + \lambda_{11}$$

RECALL :

$$\mathcal{Z} = \int dx \exp\left(-\frac{1}{2}\sum_{i,j=1}^{N} x_i A_{ij} x_j + \sum_i s_i x_i\right) = \sqrt{\frac{(2\pi)^N}{\det A}} \exp\left(\frac{1}{2} s_i (A^{-1})_{ij} s_j\right) .$$

Using the graph basis description of the quadratic form, not clear how to invert it, and not clear how to choose λ_a to ensure convergence of the measure.

We can use symmetry - representation theory of S_D in order to bring the quadratic form to a nearly diagonal form. When we make full use of symmetry, we reduce the problem to the inversion of two numbers, and a symmetric 2 by 2 matrix and a symmetric 3 by 3 matrix.

Invariance under $M \to U_\sigma M U_\sigma^\dagger$, for matrices U_σ corresponding to permutations $\sigma \in S_D$ amounts to imposing invariance $M_{ij} \to M_{\sigma(i)\sigma(j)}$

Linear combinations of M_{ij} form a representation of S_D which is $V_D \otimes V_D$. V_D is spanned by e_i for $i \in \{1, 2, \cdots, D\}$.

$$\sigma : e_i \to e_{\sigma(i)}$$
$$\sigma : M_{ij} \to M_{\sigma(i)\sigma(j)}$$

$$V_D = V_0 \oplus V_H$$

V_0 spanned by $e_1 + e_2 + \cdots + e_D$. V_H is spanned by $e_i - e_j$.

$$\begin{aligned} V_D \otimes V_D &= 2V_0 \oplus 3V_{[D-1,1]} \oplus V_{[D-2,2]} \oplus V_{[D-2,1,1]} \\ &= 2V_0 \oplus 3V_H \oplus V_2 \oplus V_3 \end{aligned}$$

$$\text{Span}\,\{M_{ij} : 1 \le i, j \le D\} = 2V_0 \oplus 3V_H \oplus V_2 \oplus V_3$$

$$= \bigoplus_{\alpha=1}^{2} V_0^{(\alpha)} \oplus \bigoplus_{\alpha=1}^{3} V_H^{(\alpha)} \oplus V_2 \oplus V_3$$

$$M_{ij} = \sum_{\Lambda, m_\Lambda, \alpha} C_{ij}^{\Lambda, m_\Lambda, \alpha} S_{\Lambda, m_\Lambda, \alpha}$$

$$\begin{aligned} \text{Symmetric } (V_H \otimes V_H) &= V_0 \oplus V_H \oplus V_3 \\ \text{Anti-symmetric } (V_H \otimes V_H) &= V_2 \end{aligned}$$

The matrix model computations will not need the explicit forms of the Clebsch for V_2, V_3 – will need projectors – which are actually simple.

Because of the tensor product decompositions above, projectors for V_2, V_3 in $V_D \otimes V_D$ can be written in terms of polynomials $F(i,j)$.

$$F(i,j) = \sum_{a=1}^{D-1} C_{a,i} C_{a,j} = \delta_{ij} - \frac{1}{N}$$

Computations of expectation values reduce to F-sums.
For example

$$< k,l|P_{V_2}|i,j >= \frac{1}{2}(F_{k,i}F_{l,j} - F_{l,i}F_{k,j})$$

The model is defined by integration. The partition function is

$$\mathcal{Z}(\mu_1, \mu_2; \Lambda_{V_0}, \Lambda_H, \Lambda_{V_2}, \Lambda_{V_3}) = \int dM e^{-\mathcal{S}}$$

where the action is a combination of linear and quadratic functions.

$$\mathcal{S} = -\sum_{\alpha=1}^{2} \mu_\alpha^{V_0} S^{V_0;\alpha} + \frac{1}{2} \sum_{\alpha,\beta=1}^{2} S^{V_0;\alpha} (\Lambda_{V_0})_{\alpha\beta} S^{V_0;\beta} + \frac{1}{2} \sum_{a=1}^{D-1} \sum_{\alpha,\beta=1}^{3} S_a^{H;\alpha} (\Lambda_H)_{\alpha\beta} S_a^{H;\beta}$$
$$+ \frac{1}{2} \Lambda_{V_2} \sum_{a=1}^{(D-1)(D-2)/2} S_a^{V_2} S_a^{V_2} + \frac{1}{2} \Lambda_{V_3} \sum_{a=1}^{D(D-3)/2} S_a^{V_3} S_a^{V_3} .$$

Tensor product of irreps $R \otimes S$ contains the trivial V_0 only if $R = S$ and $V_0 \subset Sym^2(R)$.
The expectation values of permutation invariant polynomials $f(M)$ are defined by

$$\langle f(M) \rangle = \frac{1}{\mathcal{Z}} \int dM e^{-\mathcal{S}} f(M) .$$

$$11 = 1 + 1 + (2)(3)/2 + 3(4)/2 = 1 + 1 + 3 + 6 = 11$$

The quadratic form already partially diagonalised. Change from euclidean measure in M to euclidean measures in S is orthogonal so Jacobian for change of measure is just 1.

$$\langle S^{V_0;\alpha} \rangle = \sum_\beta (\Lambda^{-1})_{\alpha\beta}\ \mu_\beta\ .$$

We introduce the definition

$$\tilde{\mu}_\alpha \equiv \sum_\beta (\Lambda^{-1})_{\alpha\beta}\ \mu_\beta\ .$$

We have defined variables $\widetilde{\mu}_1$, $\widetilde{\mu}_2$ for convenience. The variables transforming according to V_H, V_2, V_3 have vanishing expectation values

$$\begin{aligned}\langle S_a^{H;\alpha} \rangle &= 0 \\ \langle S_a^{V_2} \rangle &= 0 \\ \langle S_a^{V_3} \rangle &= 0\ .\end{aligned}$$

The quadratic expectation values are

$$\langle S^{V_i;\alpha} S^{V_j;\beta} \rangle = \langle S^{V_i;\alpha} S^{V_j;\beta} \rangle_{\rm conn} + \langle S^{V_i;\alpha} \rangle\ \langle S^{V_j;\beta} \rangle\ .$$

where

$$\langle S_a^{V_i;\alpha} S_b^{V_j;\beta} \rangle_{\rm conn} = \delta(V_i, V_j)(\Lambda_{V_i}^{-1})_{\alpha\beta}\delta_{ab}\ .$$

$$\sum_i \langle M_{ii} \rangle = \widetilde{\mu}_1 + \sqrt{(D-1)}\widetilde{\mu}_2$$

$$\sum_{i,j} \langle M_{ij} \rangle = D\widetilde{\mu}_1$$

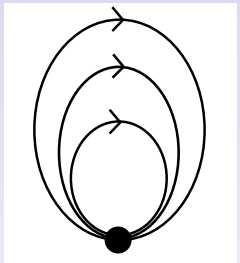
$$\sum_{i,j} \langle M_{ij} M_{ij} \rangle = \widetilde{\mu}_1^2 + \widetilde{\mu}_2^2 + (\Lambda_{V_0}^{-1})_{11} + (\Lambda_{V_0}^{-1})_{22} + (D-1)(\Lambda_H^{-1})_{22} + (D-1)(\Lambda_H^{-1})_{33}$$
$$+ (D-1)(\Lambda_H^{-1})_{11} + \frac{D(D-3)}{2}(\Lambda_{V_2})^{-1} + \frac{(D-1)(D-2)}{2}(\Lambda_{V_3})^{-1}$$

Similar equations for the other 10 quadratic graph-basis observables. These are identified with word averages and used to determine the μ, Λ parameters.

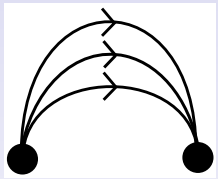
Given all the linear and quadratic expectation values, we can use Wick's theorem to calculate theoretical cubic and quartic expectation values.

$$\sum_i \langle M_{ii}^3 \rangle = 3\left(\frac{1}{D}\widetilde{\mu}_1 + \frac{\sqrt{(D-1)}}{D}\widetilde{\mu}_2\right) \times \left(\frac{1}{D}(\Lambda_{V_0}^{-1})_{11} + \frac{(D-1)}{D}(\Lambda_{V_0}^{-1})_{22} + 2\frac{\sqrt{(D-1)}}{D}(\Lambda_{V_0}^{-1})_{12}\right.$$
$$+ \frac{(D-1)}{D}(\Lambda_H^{-1})_{11} + \frac{(D-1)}{D}(\Lambda_H^{-1})_{22} + \frac{(D-1)(D-2)}{D}(\Lambda_H^{-1})_{33} + 2\frac{(D-1)}{D}(\Lambda_H^{-1})_{12}$$
$$\left. + 2\frac{(D-1)}{D}\sqrt{(D-2)}(\Lambda_H^{-1})_{13} + 2\frac{(D-1)}{D}\sqrt{(D-2)}(\Lambda_H^{-1})_{23} + \frac{1}{3D}\left(\widetilde{\mu}_1 + \sqrt{(D-1)}\widetilde{\mu}_2\right)^2\right).$$

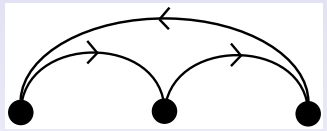
$$\sum_{i_1,i_2,i_3,i_4,i_5,i_6,i_7} \langle M_{i_1 i_2} M_{i_3 i_4} M_{i_5 i_6} M_{i_7 i_7} \rangle = 3D^3(\Lambda_{V_0}^{-1})_{11}^2 + 3D^3\sqrt{(D-1)}(\Lambda_{V_0}^{-1})_{11}(\Lambda_{V_0}^{-1})_{12}$$
$$+ 6D^3\widetilde{\mu}_1^2(\Lambda_{V_0}^{-1})_{11} + 3D^3\sqrt{(D-1)}\widetilde{\mu}_1\widetilde{\mu}_2(\Lambda_{V_0}^{-1})_{11} + D^3\sqrt{(D-1)}\widetilde{\mu}_1^2(\Lambda_{V_0}^{-1})_{12} +$$
$$+ D^3\widetilde{\mu}_1^4 + D^3\sqrt{(D-1)}\widetilde{\mu}_1^3\widetilde{\mu}_2.$$



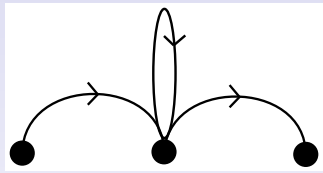
$\sum_i M_{ii}^3$



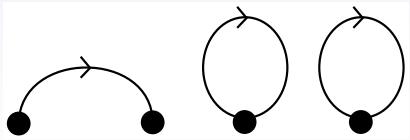
$\sum_{i,j} M_{ij}^3$



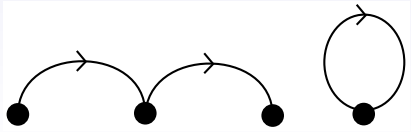
$\sum_{i,j,k} M_{ij} M_{jk} M_{ki}$



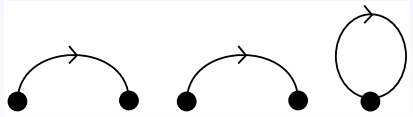
$\sum_{i,j,k} M_{ij} M_{jj} M_{jk}$



$\sum_{i,j,k,l} M_{ij} M_{kk} M_{ll}$

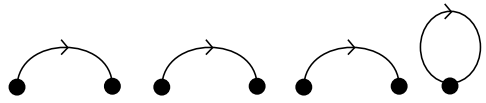


$\sum_{i,j,k,l} M_{ij} M_{jk} M_{ll}$



$\sum_{i,j,k,l,m} M_{ij} M_{kl} M_{mm}$

$\sum_{i,j,k,l,m,n} M_{ij} M_{kl} M_{mn}$



$\sum_{i_1,\ldots,i_7} M_{i_1 i_2} M_{i_3 i_4} M_{i_5 i_6} M_{i_7 i_7}$

$\sum_{i_1,\ldots,i_8} M_{i_1 i_2} M_{i_3 i_4} M_{i_5 i_6} M_{i_7 i_8}$

Figure: The 10 higher order observable graph diagrams labelled with the associated sum.

Adjectives at D = 2000 :

Graph	Expectation value	Theoretical val.	Experimental val.	Ratio
1	$\sum_i \langle (M_{ii})^3 \rangle$	1.44×10^{-1}	2.52×10^{-1}	0.57
2	$\sum_{i,j} \langle (M_{ij})^3 \rangle$	8.43×10^{-1}	3.65	0.23
3	$\sum_{i,j,k} \langle M_{ij} M_{jk} M_{ki} \rangle$	1.68	10.6	0.16
4	$\sum_{i,j,k} \langle M_{ij} M_{jj} M_{jk} \rangle$	53.8	80.1	0.67
5	$\sum_{i,j,k,l} \langle M_{ij} M_{kk} M_{ll} \rangle$	2.94×10^6	3.03×10^6	0.97
6	$\sum_{i,j,k,l} \langle M_{ij} M_{jk} M_{ll} \rangle$	4.83×10^4	5.04×10^4	0.96
7	$\sum_{i,j,k,l,m} \langle M_{ij} M_{kl} M_{mm} \rangle$	5.93×10^7	6.01×10^7	0.99
8	$\sum_{i,j,k,l,m,n} \langle M_{ij} M_{kl} M_{mn} \rangle$	1.38×10^9	1.40×10^9	0.98
9	$\sum_{i_1 \ldots i_7} \langle M_{i_1 i_2} M_{i_3 i_4} M_{i_5 i_6} M_{i_7 i_7} \rangle$	7.83×10^{10}	8.14×10^{10}	0.96
10	$\sum_{i_1 \ldots i_8} \langle M_{i_1 i_2} M_{i_3 i_4} M_{i_5 i_6} M_{i_7 i_8} \rangle$	1.86×10^{12}	1.96×10^{12}	0.95

There are in fact 52 cubic observables (graphs between one and 6 nodes) and 296 quartic observables (graphs between one and 8 nodes)

The above ratios give strong evidence that Matrix constructions in compositional distributional semantics can provide another arena for the application of approximate gaussianity and symmetry (here permutation symmetry) to real world data.

The widespread applicability of traditional RMT a la Wigner in data sciences beyond physics (e.g. financial correlations, gene networks etc.) can be taken as evidence for Gaussianity, for observables invariant under continuous symmetries (e.g. trM^3 , $trM^2 trM$). The results above are an indication that the Gaussianity extends to more general permutation invariant observables.

Approximate Gaussianity - generalized to higher dimensional QFT - underlies the applicability of perturbative QFT to particle physics, cosmology (e.g. CMB fluctuations) and elsewhere in condensed matter physics. The applicability of gaussian matrix models (zero dim QFT), with continuous or discrete symmetries, to matrix data could be an avenue for importing further insights from QFT to data.

In the paper **2022-01** we have done further Gaussianity tests with datasets of machine-learned matrices for verbs. We think a more robust measure of Gaussianity is

$$\left| \frac{\langle \mathcal{O}_a(M) \rangle_{expt} - \langle \mathcal{O}_a(M) \rangle_{theor}}{\langle \delta \mathcal{O}_a(M) \rangle_{expt}} \right|$$

Adjectives normalized differences (consistent with the picture with ratios)

jj 300	jj 700	jj 1300	jj 2000
0.238	0.229	0.273	0.307
0.192	0.653	0.884	1.039
0.020	0.765	1.160	1.375
0.054	0.242	0.307	0.366
0.011	0.024	0.026	0.022
0.052	0.002	0.041	0.040
0.014	0.016	0.017	0.011
0.024	0.021	0.022	0.015
0.037	0.037	0.038	0.025
0.047	0.042	0.043	0.031

Extends to machine-learned verbs :

obj	01sub09obj	02sub08obj	subj
0.205935	0.205601	0.200036	0.0648322
0.053190	0.042145	0.017755	0.101824
0.023944	0.041406	0.063718	0.0181965
0.067072	0.062928	0.032055	0.139924
0.054866	0.044782	0.031757	0.073066
0.020466	0.044131	0.060853	0.039253
0.079102	0.083503	0.084444	0.015464
0.086337	0.064538	0.023994	0.089363
0.068266	0.051903	0.020784	0.043241
0.092624	0.086513	0.059468	0.085318
0.169206	0.166239	0.16634	0.130891
0.173461	0.175396	0.181464	0.154943
0.132151	0.142041	0.166086	0.048161
0.011549	0.023646	0.048226	0.051810
0.067946	0.032733	0.021534	0.040914

Part 4 : Natural Language tasks.

A dataset of matrices constructed by adapting word2vec in

J. Maillard, S. Clark, E. Grefenstette, **"A type-driven tensor-based semantics for CCG,"** Proceedings of the Type Theory and Natural Language Semantics Workshop, EACL 2014.

G. Wijnholds, M. Sadrzadeh, S. Clark, **"Representation learning for type-driven composition,"** (**WSC2020**) Proceedings of the 24th Conference on Computational Natural Language Learning, pgs. 313-324, 2020.

(**WSC2020**) construct an ensemble of matrices of size 100×100, for a dataset of verbs in SimVerb3500.

SimVerb3500 contains pairs of verbs, organised by labels : synonym-pairs, antonym-pairs, hypernym-hyponym pairs, co-hyponym pairs, no-relation pairs.

Observable-deviation vectors for verbs

$$v_a^A = \mathcal{O}_a(M^A) - \langle \mathcal{O}_a(M) \rangle_{expt}$$

where

$$\langle \mathcal{O}_a(M) \rangle_{expt} = \frac{1}{N_{verbs}} \sum_A \mathcal{O}_a(M^A)$$

List of observables used

#	observable
1	M_{ii}
2	M_{ij}
3	$M_{ij}M_{ij}$
4	$M_{ij}M_{ji}$
5	$M_{ii}M_{ij}$
6	$M_{ii}M_{ji}$
7	$M_{ij}M_{ik}$
8	$M_{ij}M_{kj}$
9	$M_{ij}M_{jk}$
10	$M_{ij}M_{kl}$
11	$(M^2)_{ii}$
12	$M_{ii}M_{jj}$
13	$M_{ii}M_{jk}$
14	$(M^3)_{ii}$
15	$(M^3)_{ij}$
16	$M_{ij}M_{jk}M_{ki}$
17	$M_{ij}M_{jj}M_{jk}$
18	$M_{ij}M_{kk}M_{ll}$
19	$M_{ij}M_{jk}M_{ll}$
20	$M_{ij}M_{kl}M_{mm}$
21	$M_{ij}M_{kl}M_{mn}$
22	$M_{ij}M_{kl}M_{mn}M_{oo}$
23	$M_{ij}M_{kl}M_{mn}M_{op}$
24	$(M^4)_{ii}$
25	$(M^4)_{ij}$
26	$M_{ij}M_{jk}M_{pq}M_{qr}$
27	$M_{ij}M_{jk}M_{kl}$
28	$M_{jk}M_{kl}M_{lm}$

Table: A complete list of the observables used in this paper. Summations over every index are understood. The horizontal lines mark the different subsets used to build the observable vectors. These subsets are made out of 13, 23, 28 and 15 (=28-13) observables.

A physics motivated choice of inner product for the vectors

$$g_{dev}(v^A, v^B) = \sum_a \frac{v_a^A v_a^B}{\langle (\delta \mathcal{O}_a(M))^2 \rangle} ,$$

$$\langle (\delta \mathcal{O}_a(M))^2 \rangle = \frac{1}{N_{verbs}} \sum_A (\mathcal{O}_a(M^A) - \langle \mathcal{O}_a(M) \rangle_{expt})^2$$

This orthogonal choice motivated by large N factorisation : distinct combinatoric structures associated with invariant are orthogonal at large N - trace structures for continuous symmetries, graphs for S_N.

Digression : Large N factorisation for permutation invariant matrix observables (PIMOs)

In the paper **2021-12** we prove a large N factorisation result for PIMOS.
At a special point in the 13-parameter space of permutation invariant models, we have just the simplest $O(N)$ invariant action.

$$S = \mathrm{trMM}^{\mathrm{T}}$$

With this action the 2-point function defines an inner product for observables

$$\langle : \mathcal{O}_a(M) :: \mathcal{O}_b(M) : \rangle$$

In this inner product, distinct graphs are orthogonal at large N.
Reasonable to expect this should generalize.

A second choice of inner product for the vectors

A choice from statistics (Mahalanobis metric)

$$g_{dev}^{\mathrm{Maha}}(v^A, v^B) = \sum_{a,b} v_a^A K_{ab} v_b^B$$

K_{ab} is the inverse of the correlation matrix $\langle \delta\mathcal{O}_a \delta\mathcal{O}_b \rangle_{EXPT}$.
Notice that if the observables are uncorrelated then
$K_{aa} = 1/\langle (\delta\mathcal{O}_a)^2 \rangle$, and we recover the orthogonal metric.

Cosine of angle : A familiar measure in a new embedding space

A commonly used measure of “semantic similarity” in distributional semantics is the “cosine of angle ” between vectors for words.

In traditional distributional semantics applications, we are looking at vectors where the different components correspond to different context words (or linear combinations of context words).

Here we adapt to the setting of Observable deviation vectors equipped with a metric (one of the two above).

$$Cos(v^A, v^B) = \frac{g_{val}(v^A, v^B)}{\sqrt{g_{val}(v^A, v^A) g_{val}(v^B, v^B)}}$$

Statistical comparison of semantic relations

We compare the cosines, averaged over synonym-pairs, antonym-pairs and no-relation pairs.

Separation of SYN/NONE/ANT

Means of cosines of synonym, none and antonym pairs

	ANTONYM	NONE	SYNONYM
obj	0.093 ± 0.439	0.168 ± 0.506	0.284 ± 0.480
01sub09obj	0.088 ± 0.441	0.169 ± 0.504	0.288 ± 0.477
02sub08obj	0.092 ± 0.445	0.172 ± 0.502	0.295 ± 0.473
subj	0.148 ± 0.450	0.196 ± 0.517	0.314 ± 0.478

The means for synonyms,none, antonyms are consistently separated. 4 rows are discrete choices in the ML construction. This is for 13-dimensional vectors. SImilar results for 28- or 10-dimensional vectors.

Geometrical divide for distinguishing SYN/ANT

$$\underbrace{D_{\mathrm{A}} + \frac{\Delta_{\mathrm{A}}}{\Delta_{\mathrm{S}}} \frac{(D_{\mathrm{S}} - D_{\mathrm{A}})}{2}}$$

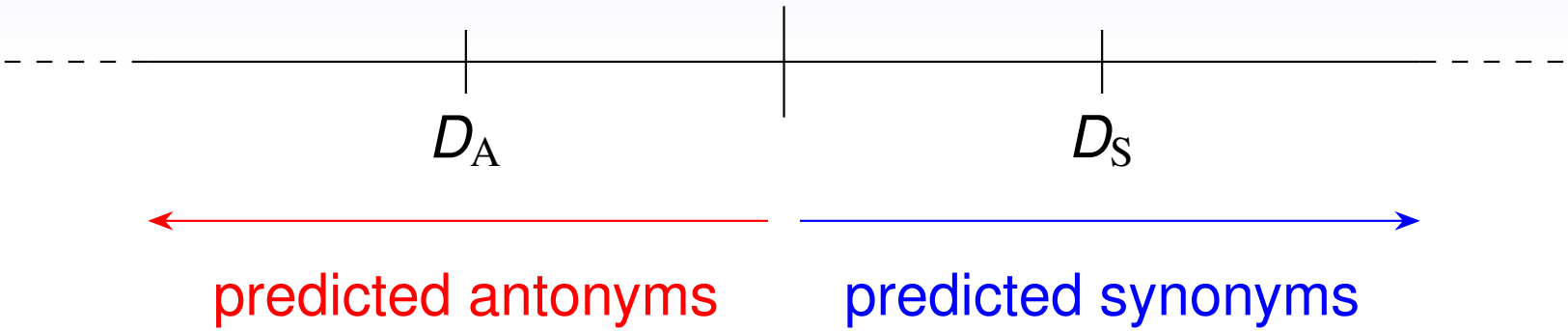


Figure: Figure illustrating antonym-synonym separation criterion

Success rates

With the orthogonal metric (results with the Mahalanobis metric are similar)

$$\frac{1}{N_{tot}} \sum_{i \in \text{cat}} \frac{N_{correct}^{(i)}}{N_{tot}^{(i)}}$$

	obj	01sub09obj	02sub08obj	subj
13 obs	0.560	0.555	0.557	0.523
15 obs	0.579	0.568	0.575	0.554
28 obs	0.556	0.561	0.578	0.564

Lengths for distintiguishing hypernym from hyponym in a hyper/hypo pair

Hyper/Hypo Example : Animal /Dog.

Hypernym is a more general class. Hyponym is an example within the class.

Statistical physics concepts (e.g. entropy) have been applied to hyper/hypo distinction in conventional word vectors.

Adapting these ideas to the present case, one can argue that the lengths of the observable deviation vectors should be longer for the hypernym in a pair than the hyponym pair.

"Chasing hypernyms in vector spaces with entropy," E Santus, A Lenci, Q Lu, SS Im Walde, 2014, Proc. 14th conference of the European Chapter of Assoc. Comp. Ling.

We apply in (the 2022-01 paper) this expectation to predict - based purely on the ODVs - whether a word in a hyper/hypo pair is a hypernym or hyponym.

Using that to predict, using orthogonal metric, we find the following success rates :

$$\text{ratio} = \frac{\#\text{ of pairs for which hyper} > \text{hypo}}{\text{all pairs of hyper/hyponyms}}$$

28 obs	high node	low node
0.627	0.602	0.629

With Mahalanobis metric, we get better performance on this task :

28 obs	high node	low node
0.677	0.652	0.653

Summary and Outlook

Permutation invariant matrix observables (PIMOs) show evidence for approximate Gaussianity in natural language data.

They show promising results in natural language tasks.

A future direction is to use Machine Learning techniques to explore the parameter spaces in these tasks (e.g. parameters in the metric ; divides) to get better performance in specific tasks.

These investigations of Gaussianity and statistics can be applied very generally – whenever there is a collection of real world entities (e.g. genes, proteins) such that each entity can be associated to a matrix, leading to a coherent ensemble of matrices.

Theoretical directions : Large N factorization, should have generalizations ; 2-matrix and tensor models ...

Supplement 1: The parameters of the 5$-parameter model.

Comparisons done for a range $300 \leq D \leq 2000$ in steps of 100.

$$\frac{J_0}{D} = 1.31 \times 10^{-2} \quad , \quad \frac{\Lambda}{D^2} = 2.86 \times 10^{-3}$$

$$\frac{J_s}{D} = 4.51 \times 10^{-4} \quad , \quad \frac{a}{D^2} = 1.95 \times 10^{-3} \quad , \quad \frac{b}{D^2} = 2.01 \times 10^{-3}$$

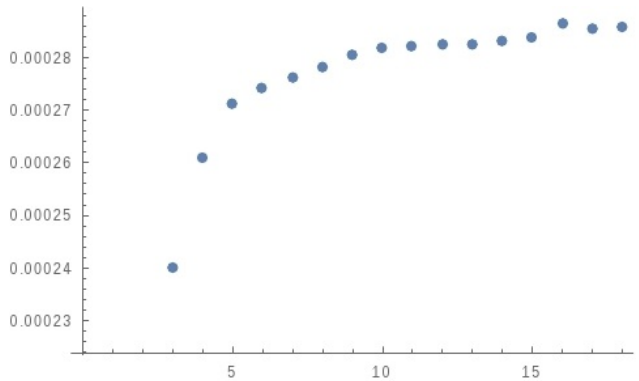


Figure: The ratio $\frac{\Lambda}{D^2}$ stabilizing at large D.

Supplement 2: The parameters of the 13-parameter model.

To 3 significant figures, the parameter values for $D = 2000$ are given below.

Parameter	Value
$\widetilde{\mu_1}$	4.84×10^{-1}
$\widetilde{\mu_2}$	1.01
$(\Lambda_{V_0}^{-1})_{11}$	4.00×10^{-2}
$(\Lambda_{V_0}^{-1})_{12}$	5.10×10^{-2}
$(\Lambda_{V_0}^{-1})_{22}$	2.49×10^{-1}
$(\Lambda_{H}^{-1})_{11}$	1.45×10^{-2}
$(\Lambda_{H}^{-1})_{12}$	1.02×10^{-4}
$(\Lambda_{H}^{-1})_{13}$	2.28×10^{-4}
$(\Lambda_{H}^{-1})_{22}$	2.91×10^{-4}
$(\Lambda_{H}^{-1})_{23}$	1.22×10^{-4}
$(\Lambda_{H}^{-1})_{33}$	7.27×10^{-4}
$(\Lambda_{V_2}^{-1})$	2.49×10^{-4}
$(\Lambda_{V_3}^{-1})$	2.41×10^{-4}

From : 1912** paper ; using equations for 13-parameter model from the 1810** paper and data from 1703** paper.

Supplement 3: Object and subject construction in ML method for verbs

To test whether large matrix representations of words can be effectively modelled with Gaussianity operators, we use the verb matrix representations obtained in (Clark,Sadrzadeh,Wijnholt-2020). These matrices were obtained using an extension of the *skipgram* model. In this model, predictions are made for the representation $\mathbf{n}$ of a specified target word, for the representation $\mathbf{c}$ of the word that follows or precedes it, the context word, and for the representation $\bar{\mathbf{c}}$ of words that never apear as context of the target. These predictions are updated by minimizing the following objective function

$$\sum_{c \in C} \log \sigma(\mathbf{n} \cdot \mathbf{c}) + \sum_{\bar{\mathbf{c}} \in \bar{C}} \log \sigma(-\mathbf{n} \cdot \bar{\mathbf{c}}),$$

until convergence is reached. Here σ is a sigmoid function that rescales its input to values between 0 and 1.
In the case where the targets are transitive verbs, a matrix representing the verb is multiplied with a vector representating either an object or a subject, resulting in $\mathbf{n}$, and Eq. 3 is minimized using a verb's subject or object vector representation, respectively, as contexts $\mathbf{c}$ and $\bar{\mathbf{c}}$.