Gaussianity and typicality in matrix distributional semantics

Sanjaye Ramgoolam[1,2], Mehrnoosh Sadrzadeh[3], Lewis Sword[1]

[1]School of Physics and Astronomy, Queen Mary University of London
[2]School of Physics and Centre for Theoretical Physics, University of the Witwatersrand
[3]Department of Computer Science, University College London
s.ramgoolam@qmul.ac.uk, m.sadrzadeh@ucl.ac.uk,l.sword@se15.qmul.ac.uk

Constructions in type-driven compositional distributional semantics associate large collections of matrices of size D to linguistic corpora. We develop the proposal of analysing the statistical characteristics of this data in the framework of permutation invariant matrix models. The observables in this framework are permutation invariant polynomial functions of the matrix entries, which correspond to directed graphs. Using the general 13-parameter permutation invariant Gaussian matrix models recently solved, we find, using a dataset of matrices constructed via standard techniques in distributional semantics, that the expectation values of a large class of cubic and quartic observables show high gaussianity at levels between 90 to 99 percent. We find evidence that observables with similar matrix model characteristics of gaussianity also have high degrees of correlation between the ranked lists of words associated to these observables. [1]

1 Permutation invariant Gaussian matrix models and prediction of moments of matrices from compositional distributional semantics

A research programme "Linguistic Matrix Theory" of understanding the characteristics of randomness in natural language, specifically in matrix/tensor datasets arising from type-driven compositional distributional semantics [1, 2], using the framework of random matrix/tensor theories was initiated in [3]. This programme draws on and generalizes the application of random matrix theories to the energy level distribtions of complex nuclei [4]. The categorical compositional distributional semantics, and specifically the use of tensors therein, was inspired by the categorical foundations of quantum mechanics[5]. The setting has also been recasted as a "quantisation" functor in [6], very similar to that of TQFT, assigning semantic vector spaces and their tensors to natual language grammatical types, for an overview see [7].

In the Linguistic Matrix Theory (LMT) programme of [3], one of the first steps was to identify the appropriate type of symmetry. Here it was useful to consider the kinds of mathematical expressions which are used in distributional semantics to extract the meaning encoded in words. For vector, matrix and tensor data in D dimensions, some of these expressions are invariant under the orthogonal group of all rotations in D dimensions, but the generic expressions are only invariant under the smaller symmetry of all permutations of D objects, the symmetric group S_D. This motivated us to consider matrix models with S_D symmetry. The polynomial functions of matrix variables M_{ij} which are S_D invariant have an elegant classification in terms of polynomials labelled by directed graphs. The degree of the polynomial is the number of edges in the graph : the number of nodes is unconstrained. There are two graphs at linear order, each associated with a permutation invariant polynomial. A general permutation invariant linear function is a sum of these two polynomials with arbitrary coefficients. We restrict these linear coefficients to be real numbers numbers μ_1, μ_2. There are eleven independent quadratic functions. As a simple toy model

[1]This paper is an abstract based on previous work, for the full account please see [9].

Submitted to:
QPL 2020

we considered three quadratic polynomials with three associated coefficients $\Lambda_1, \Lambda_2, \Lambda_3$. We defined a function $S(\mu_1, \mu_2, \Lambda_1, \Lambda_2, \Lambda_3)$ and considered a probability distribution defined by the partition function

$$Z = \int dM e^{-S(\mu_1, \mu_2, \Lambda_1, \Lambda_2, \Lambda_3)} \tag{1}$$

Given any permutation invariant polynomial, which we will henceforth refer to as observables and denote $\mathscr{O}(M)$, we can calculate a theoretical expectation value

$$\langle \mathscr{O}(M) \rangle_{THEO} = \frac{1}{Z} \int dM e^{-S(\mu_1, \mu_2, \Lambda_1, \Lambda_2, \Lambda_3)} \mathscr{O}(M) \tag{2}$$

The expectation values of the linear and quadratic observables $\langle \mathscr{O} \rangle$ are expressible as simple functions of the μ_a, Λ_i. In order to match these probability distributions with experimental data, the experimental expectation values for these five observables were computed as averages over the words in the dataset

$$\frac{1}{N_{words}} \sum_A \mathscr{O}(M^A) \tag{3}$$

A is a label for the words in the dataset and N_{words} is the number of words in the dataset. Equating these to the theoretical expectation values, we determined the μ_a, Λ_i parameters of the model, for a given dataset.

The theoretical model was also used to calculate the expectation values of a number of cubic and quadratic observables. These theoretical values, using the input of μ_a, Λ_i determined as above, give the predictions of the 5-parameter Gaussian model for these observables. We calculated the ratios of the theoretical to experimental values, with a ratio close to 1 being good agreement between theory and experiment. The best ratios were approximately 60 %, but for a number of observables the ratios were very low, the lowest being around 0.6 % . We argued that a more complete treatment with a general Gaussian model that includes all the eleven parameters would likely give better ratios.

The theoretical model with all eleven quadratic parameters was solved in [8]. It was useful to employ a representation theoretic approach to the space of quadratic permutation invariant functions. The eleven parameters were organised according to four irreducible representations V_0, V_H, V_2, V_3 of S_D. Λ^{V_0} is a symmetric 2×2 matrix with three real parameters, Λ^{V_H} is a symmetric 3×3 real matrix with 6 parameters and $\Lambda^{V_2}, \Lambda^{V_3}$ are each real numbers. We have an action

$$S(\mu_a, \Lambda^{V_0}, \Lambda^{V_H}, \Lambda^{V_2}, \Lambda^{V_3}) \equiv S(\mu_a, \Lambda^V) \tag{4}$$

which defines a probability distribution and associated partition function

$$Z = \int dM e^{-S(\mu_a, \Lambda^V)} \tag{5}$$

Convergence of the measure requires that $\Lambda^{V_0}, \Lambda^{V_H}$ are positive semi-definite matrices, and $\Lambda^{V_2} \geq 0, \Lambda^{V_3} \geq 0$. The first main goal of this paper is to report on the application of this 13-parameter Gaussian model from [8] to the same dataset constructed in [3], to test its effectiveness at predicting cubic and quartic expectation values along the lines of the approach in [3].

We find that low order permutation invariant polynomials, and specifically the 13-parameter Gaussian permutation invariant matrix models, are indeed the right objects to detect strong evidence of Gaussianity. While the best theory/expt ratios achieved by the 5-parameter model are near 60%, the best ratios now are near 99% and indeed for a number of cubic and quartic observables, these ratios are above 90%. The lowest ratio is 16%, so that the Gaussian model still predicts the right order of magnitude of the expectation value even in the worst case. In all the experiments studied, we find that the linear and quadratic expectation values lead to theoretical parameters μ, Λ consistent with the convergence criteria.

1.1 Theory/Experiment comparisons for adjective matrices

In detail, the results for the Cubic and Quartic ratios for 13 parameter model are given below. The first table is for the matrices associated with adjectives, while the second is for verbs.
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

Our tables of theory/experiment ratios for $\langle \mathscr{O}(M) \rangle$ show that some pairs of observables have distinctly similar characteristics whether we are looking at expectation values or standard deviations. Each observable can also be used to rank the words in the dataset, starting from the word with the lowest $\mathscr{O}(M)$ to the one with the highest. Since ranked lists of words form a standard tool in distributional semantics, it is natural to ask whether observables which have very similar matrix model characteristics also produce similar ranked lists. We find evidence for a positive answer. For example, Figure 1 give the correlation plots for the ranked lists coming from graphs G2 and G3. Figure 2 gives the analogous plots for the pair $G3$ and $G10$. The pairs $G2, G3$ which have very similar ratios in the table above, also produce very well-correlated rankes word-lists. This behaviour is also seen using the Pearson rank corelation coefficient.

In the future, we will investigate multi-matrix and tensor generalizations of the work presented here.

References

[1] B. Coecke, M. Sadrzadeh, and S. Clark, "Mathematical Foundations for a Compositional Distributional Model of Meaning, " Lambek Festschrift. Linguistic Analysis, 36:345–384, 2010.

[2] M.Baroni, R. Zamparelli, "Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space," Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 1183–1193, 2010.

[3] D. Kartsaklis, S. Ramgoolam and M. Sadrzadeh, "Linguistic Matrix Theory," arXiv:1703.10252 [cs.CL].

[4] Wigner, E. (1955). "Characteristic vectors of bordered matrices with infinite dimensions". Annals of Mathematics. 62 (3): 548–564.

[5] S. Abramsky, B. Coecke. "A Categorical Semantics of Quantum Protocols". Proceedings of the 19th Annual IEEE Symposium on Logic in Computer Science. 415–425. 2004.

[6] B. Coecke, E. Grefenstette, M. Sadrzadeh. "Lambek vs. Lambek: Functorial vector space semantics and string diagrams for Lambek calculus". Annals of Pure and Applied Logic. 164 (11). 1079-1100. 2013.

[7] M. Sadrzadeh, "Quantization, Frobenius and Bi Algebras from the Categorical Framework of Quantum Mechanics to Natural Language Semantics". Front. Phys. 5, 18 pages, 2017.

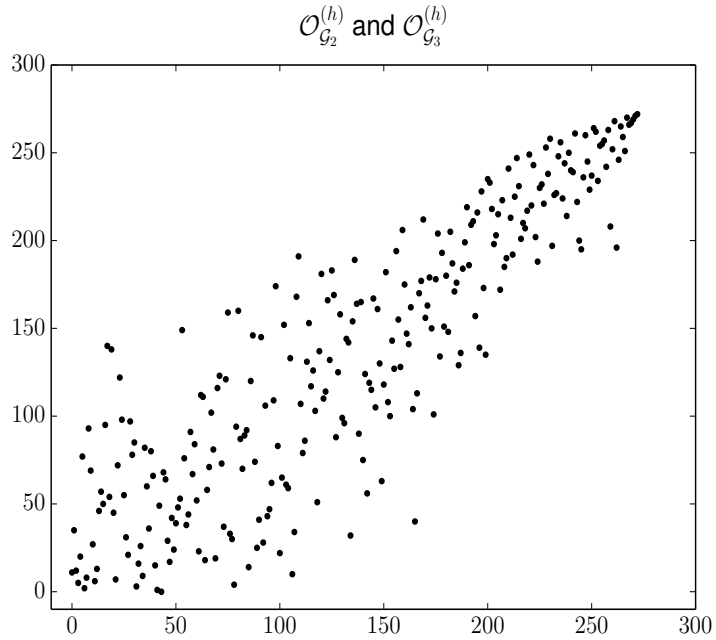

Figure 1: Rank correlation plot corresponding to graph 2 and 3 observables

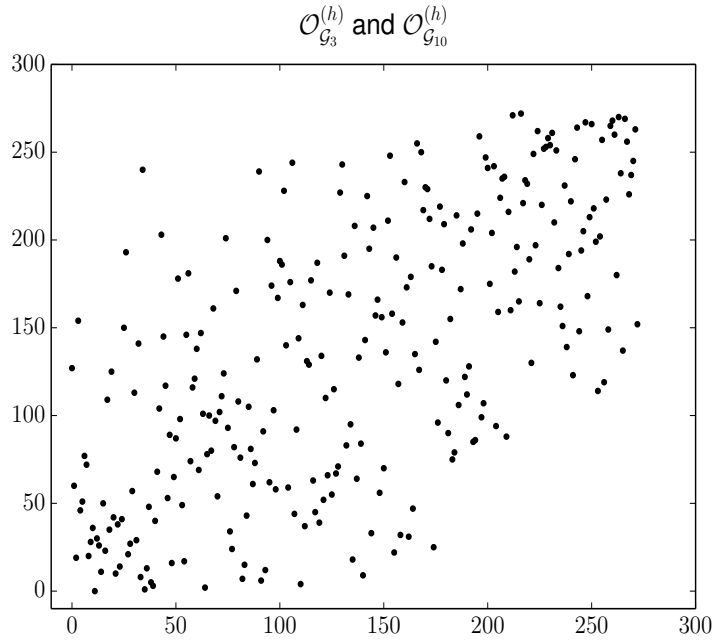

Figure 2: Rank correlation plot corresponding to graph 3 and 10 observables

[8] S. Ramgoolam, "Permutation invariant Gaussian matrix models," Nucl. Phys. B **945** (2019) 114682 [arXiv:1809.07559 [hep-th]].

[9] S. Ramgoolam, M. Sadrzadeh and L. Sword, "Gaussianity and typicality in matrix distributional semantics," arXiv:1912.10839 [hep-th].