Linguistic Matrix Theory

Dimitrios Kartsaklis[1] Sanjaye Ramgoolam[2,3] Mehrnoosh Sadrzadeh[4]

[1]Department of Theoretical and Applied Linguistics, University of Cambridge
[2]School of Physics and Astronomy, Queen Mary University of London
[3]School of Physics and Centre for Theoretical Physics, University of the Witwatersrand
[4]School of Electronic Engineering and Computer Science, Queen Mary University of London
dk426@cam.ac.uk, s.ramgoolam@qmul.ac.uk, m.sadrzadeh@qmul.ac.uk

We propose a Matrix Theory approach to tensor-based models of meaning, based on permutation symmetry along with Gaussian weights and their perturbations. A simple Gaussian model is tested against word matrices created from a large corpus of text. We characterize the cubic and quartic departures from the model, which we propose, alongside the Gaussian parameters, as signatures for comparison of linguistic corpora. We propose that perturbed Gaussian models with permutation symmetry provide a promising framework for characterizing the nature of universality in the statistical properties of word matrices. The matrix theory framework developed here exploits the view of statistics as zero dimensional perturbative quantum field theory. It perceives language as a physical system realizing a universality class of matrix statistics characterized by permutation symmetry.[1]

1 Introduction

Meaning representation is a task at the core of Computational Linguistics research. At the word level, models based on the so-called *distributional hypothesis* (the meaning of a word is determined by the contexts in which it occurs) [8] associate meaning with vectors of statistics reflecting the co-occurrence of the word with a set of contexts. While distributional models of this form have been proved very useful in evaluating the semantic similarity of words by application of vector algebra tools, their statistical nature do not allow them to scale up to the level of multi-word phrases or sentences. Recent methods [4, 6, 10] address this problem by adopting a compositional approach: the meaning of relational words such as verbs and matrices is associated with matrices or higher order tensors, and composition with the noun vectors takes the form of tensor contraction. In tensor-based models of this form, the grammatical type of each word determines the vector space in which the word lives: take N to be the noun space and S the sentence space, then an adjective becomes a linear map $N \to N$ living in $N \otimes N^*$, an intransitive verb a map $N \to S$ in $N^* \otimes S$, and a transitive verb a tensor of order 3 in $N^* \otimes S \otimes N^*$. Hence, given a transitive sentence of the form "John likes Mary", vectors $\overrightarrow{John}$, $\overrightarrow{Mary}$ representing the meaning of the noun arguments and an order-3 tensor M_{likes} for the verb, the meaning of the sentence is a vector in S computed as $\overrightarrow{John} M_{\text{likes}} \overrightarrow{Mary}$.

The association of relational words such as adjectives and verbs in a corpus with matrices produces a large amount of matrix data, and raises the question of characterising the information present in this data. Matrix distributions have been studied in a variety of areas of applied and theoretical physics. Wigner and Dyson studied the energy levels of complex nuclei, which are eigenvalues of Hamiltonians which describe the complex quantum interactions of the constituent protons and neutrons. The techniques they developed have been applied to complex atoms, molecules, subsequently to scattering matrices, chaotic

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

Submitted to:
QPL 2017

systems amd financial correlations (see for example [12, 7, 3, 5]). The spectral studies of Wigner and Dyson focused on systems with continuous symmetries, described by unitary, orthogonal or symplectic groups. Matrix theory has also seen a flurry of applications in fundamental physics, an important impetus coming from the AdS/CFT correspondence [11], which gives an equivalence between four dimensional quantum field theories and ten dimensional string theory. Important observables in this correspondence can be computed using reduced matrix models where the quantum field theory path integrals simplify to ordinary matrix integrals (for reviews of these directions in AdS/CFT see [1, 14, 13]). This sets us back to the world of matrix distributions.

In the linguistic matrix theory we develop here, we study the matrices coming from linguistic data using Gaussian Matrix distributions. The matrices we use are not hermitian or real symmetric; they are general real matrices. Hence, a distribution of eigenvalues is not the natural way to study their statistics. Another important property of the linguistic application at hand is that while it is natural to consider matrices of a fixed size $D \times D$, there is no reason to expect the linguistic or statistical properties of these matrices to be invariant under a continuous symmetry. It is true that dot products of vectors (which are used in measuring word similarity in distributional semantics) are invariant under the continuous orthogonal group $O(D)$. However, in general we may expect no more than an invariance under the discrete symmetric group S_D of $D!$ permutations of the basis vectors. The general framework for our investigations will therefore be Gaussian matrix integrals of the form:

$$\int dM e^{L(M)+Q(M)+perturbations} \tag{1}$$

where $L(M)$ is a linear function of the matrix M, invariant under S_D, and $Q(M)$ is a quadratic function invariant under S_D. Allowing linear terms in the Gaussian action means that the matrix model can accommodate data which has non-zero expectation value. The quadratic terms are eleven in number for $D \geq 4$, but we will focus on a simple solvable subspace which involves three of these quadratic invariants along with two linear ones.

2 Creating matrices for representing meaning

We work on a dataset of 171 verbs and 273 adjectives. The selection process was designed to put emphasis on words with a sufficient number of relatively frequent noun arguments in our training corpus, since this is very important for creating reliable matrices representing their meaning. Our goal is to use vectors representing the distributional meaning of the argument nouns in order to create appropriate matrices representing the meaning of the verbs and adjectives in a compositional setting. For example, given an adjective-noun compound such as "red car", our goal is to produce a matrix M_{red} such that $M_{red}\overrightarrow{car} = \overrightarrow{y}$, where $\overrightarrow{car}$ is the distributional vector of "car" and $\overrightarrow{y}$ a vector reflecting the distributional behaviour of the compound "red car". Note that a non-compositional solution for creating such a vector $\overrightarrow{y}$ would be to treat the compound "red car" as a single word and apply the same process we used for creating the vectors of nouns above [2]. This would allow us to create a dataset of the form $\{(\overrightarrow{car}, \overrightarrow{red\ car}), (\overrightarrow{door}, \overrightarrow{red\ door}), \cdots\}$ based on all the argument nouns of the specific adjective (or verb for that matter); the problem of finding a matrix which, when contracted by the vector of a noun, will approximate the distributional vector of the whole compound, can be solved by applying multi-linear regression on this dataset. We apply this method to produce matrices for all verbs and adjectives in our dataset, based on their argument nouns. For each word we create $D \times D$ matrices for various Ds, ranging from 300 to 2000 dimensions in steps of 100.

3 The 5-parameter Gaussian Model

We consider a simple S_D invariant Gaussian matrix model. The measure dM is a standard measure on the D^2 matrix variables (see [9] for the explicit formula).This is multiplied by an exponential of a quadratic function of the matrices. The parameters J^0, J^S are coefficients of terms linear in the diagonal and off-diagonal matrix elements respectively. The parameter Λ is the coefficient of the square of the diagonal elements, while a, b are coefficients for off-diagonal elements. The partition function of the model is

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

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

$$f(M_{i,j}) = f(M_{\sigma(i),\sigma(j)}) \tag{3}$$

At quadratic order there are 11 polynomials, which are listed in the Appendix B of [9]. We have only used three of these invariants in the model above. The most general matrix model compatible with S_D symmetry would consider all the eleven parameters and allow coefficients for each of them. In this paper, we restrict attention to the simple 5-parameter model, where the integral factorizes into D integrals for the diagonal matrix elements and $D(D-1)/2$ integrals for the off-diagonal elements. Each integral for a diagonal element is a 1-variable integral. For each (i, j) with $i < j$, we have an integral over 2 variables. Expectation values of $f(M)$ are computed as

$$\langle f(M)\rangle \equiv \frac{1}{\mathscr{Z}}\int dM f(M)\text{EXP} \tag{4}$$

where EXP is the exponential term in the partition function. In the following we give expressions for a set of linear, quadratic, cubic and quartic expectation values computed from theory. The computation follows standard techniques from the path integral approach to quantum field theory. This involves introducing sources J_{ij} for all the matrix elements and computing the general Gaussian integrals as function of all these sources. Taking appropriate derivatives of the result gives the expectation values of the observables.

Since the theory is Gaussian, all the correlators can be given by Wick's theorem in terms of the linear and quadratic expectation values. We are thus able to compute the expectation values for permutation invariant linear and quadratic averages. Comparison of these with the data allows us to fix the 5 parameters of the model. The model is then used to compute cubic and quartic averages such as $M_{d:3}, M_{o:3,1}, M_{o:3;2}$ defined below. These are treated as predictions to be compared to the data.

$$M_{d:3} = \sum_i \langle M_{ii}^3\rangle, \;\; M_{o:3,1} = \sum_{i\neq j}\langle M_{ij}^3\rangle, \;\; M_{o:3,2} = \sum_{i\neq j\neq k}\langle M_{ij}M_{jk}M_{ki}\rangle \tag{5}$$

4 Results

We find that for cubic averages such as $M_{d:3}, M_{o:3,1}$, the 5-parameter theory gives a good approximation to the data, which suggests that for these averages, the theory can be a reasonable starting point for perturbation theory. For more general cubic invariants, the 5-parameter Gaussian model is well off. THis is in fact not too surprising. A better starting point will be the Gaussian model with all eleven quadratic parameters turned on. Analytic and computational work on these more general models is planned for the near future. These Gaussian characteristics, as well as the cubic and quadratic departures from Gaussianity can serve as signatures of linguistic corpora, which can be used as a tool for comparative lingustics.

References

[1] O. Aharony, S. Gubser, J. Maldacena, H. Ooguri & Y. Oz (2000): *Large N field theories, string theory and gravity*. Physics Reports 323(3), pp. 183–386.

[2] M. Baroni, R. Bernardi & R. Zamparelli (2014): *Frege in Space: A Program of Compositional Distributional Semantics*. Linguistic Issues in Language Technology 9.

[3] C. WJ Beenakker (1997): *Random-matrix theory of quantum transport*. Reviews of modern physics 69(3), p. 731.

[4] B. Coecke, M. Sadrzadeh & S. Clark (2010): *Mathematical Foundations for a Compositional Distributional Model of Meaning. Lambek Festschrift*. Linguistic Analysis 36, pp. 345–384.

[5] A. Edelman & Y. Wang (2013): *Random matrix theory and its innovative applications*. In: Advances in Applied Mathematics, Modeling, and Computational Science, Springer, pp. 91–116.

[6] E. Grefenstette & M. Sadrzadeh (2015): *Concrete models and empirical evaluations for acategorical compositional distributional model of meaning*. Computational Linguistics 41, pp. 71–118.

[7] T. Guhr, A. Müller-Groeling & H.A. Weidenmüller (1998): *Random-matrix theories in quantum physics: common concepts*. Physics Reports 299(4), pp. 189–425.

[8] Z. Harris (1968): *Mathematical Structures of Language*. Wiley.

[9] D. Kartsaklis, S. Ramgoolam & M. Sadrzadeh (2017): *Linguistic Matrix Theory*. arXiv preprint arXiv:1703.10252.

[10] D. Kartsaklis, M. Sadrzadeh & S. Pulman (2012): *A Unified Sentence Space for Categorical Distributional-Compositional Semantics: Theory and Experiments*. In: Proceedings of 24th International Conference on Computational Linguistics (COLING 2012): Posters, Mumbai, India, pp. 549–558.

[11] J. M. Maldacena (1999): *The Large N limit of superconformal field theories and supergravity*. Int. J. Theor. Phys. 38, pp. 1113–1133, doi:10.1023/A:1026654312961. [Adv. Theor. Math. Phys.2,231(1998)].

[12] M. L. Mehta (2004): *Random matrices*. 142, Academic press.

[13] S. Ramgoolam (2016): *Permutations and the combinatorics of gauge invariants for general N*. PoS CORFU2015, p. 107.

[14] Sanjaye Ramgoolam (2008): *Schur-Weyl duality as an instrument of Gauge-String duality*. AIP Conf. Proc. 1031, pp. 255–265, doi:10.1063/1.2972012.