# Factorization of Latent Variables in Distributional Semantic Models

**Arvid Österlund** and **David Ödling**
KTH Royal Institute of Technology, Sweden
arvidos|dodling@kth.se

**Magnus Sahlgren**
Gavagai, Sweden
mange@gavagai.se

## Abstract

This paper discusses the use of factorization techniques in distributional semantic models. We focus on a method for redistributing the weight of latent variables, which has previously been shown to improve the performance of distributional semantic models. However, this result has not been replicated and remains poorly understood. We refine the method, and provide additional theoretical justification, as well as empirical results that demonstrate the viability of the proposed approach.

## 1 Introduction

Distributional Semantic Models (DSMs) have become standard paraphernalia in the natural language processing toolbox, and even though there is a wide variety of models available, the basic parameters of DSMs (context type and size, frequency weighting, and dimension reduction) are now well understood. This is demonstrated by the recent convergence of state-of-the-art results (Baroni et al., 2014; Levy and Goldberg, 2014).

However, there are a few notable exceptions. One is the performance improvements demonstrated in two different papers using a method for redistributing the weight of principal components (PCs) in factorized DSMs (Caron, 2001; Bullinaria and Levy, 2012). In the latter of these papers, the factorization of latent variables in DSMs is used to reach a perfect score of 100% correct answers on the TOEFL synonym test. This result is somewhat surprising, since the factorization method is *the inverse of what is normally used*.

Neither the result nor the method has been replicated, and therefore remains poorly understood. The goal of this paper is to replicate and explain the result. In the following sections, we first provide a brief review of DSMs and factorization, and review the method for redistributing the weight of latent variables. We then replicate the 100% score on the TOEFL test and provide additional state-of-the-art scores for the BLESS test. We also provide a more principled reformulation of the factorization method that is better suited for practical applications.

## 2 Distributional Semantics

Consider a set of words $W = \{w_1, \ldots, w_n\}$ and a set of context words $C = \{c_1, \ldots, c_m\}$. The DSM representation is created by registering an occurrence of a word $w_i$ with a set of context words $c_j, \ldots, c_k$ with a corresponding increment of the projection of $w_i$ on the $c_j, \ldots, c_k$ bases. In other words, each cell $f_{ij}$ in the matrix representation $F$ represents a co-occurrence count between a word $w_i$ and a context $c_j$. In the following, we use $W = C$, making the co-occurrence matrix symmetric $F_{n \times n}$. We also adhere to standard practice of weighting the co-occurrence counts with Positive Pointwise Mutual Information (PPMI) (Niwa and Nitta, 1994), which is a variation of the standard PMI weighting,[1] which simply discards non-positive PMI values.

## 3 Singular Value Decomposition

The high dimensionality of the co-occurrence matrix makes it necessary in most practical applications to apply some form of dimensionality reduction to $F$, with the goal of finding a basis $\{\hat{x_j}, ..., \hat{x_k}\}$ that restates the original basis $\{x_k, ...\}$ in a lower-dimensional space $\hat{F}$, where $\hat{F}$ denotes the rank-$k$ approximation of $F$:

$$\min_{\hat{F} \in R^n \times R^k} |F - \hat{F}| \quad (1)$$

[1] $\text{PMI}(f_{ij}) = \log \frac{f_{ij}(\sum_{ij} f_{ij})^2}{\sum_i f_{ij} \sum_j f_{ij} \sum_{ij} f_{ij}}$.



*Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 227–231,
Lisbon, Portugal, 17-21 September 2015. 

Assuming Gaussian-like distributions,[2] a canonical way of achieving this is to maximize the variance of the data in the new basis. This enables ordering of the new basis according to how much of the variance in the original data each component describes.

A standard co-occurrence matrix is positive and symmetric and thus has, by the spectral theorem, a spectral decomposition of an ordered set of positive eigenvalues and an orthogonal set of eigenvalues:

$$F = U\Sigma V^T \tag{2}$$

where $U$ holds the eigenvectors of $F$, $\Sigma$ holds the eigenvalues, and $V \in U(w)$ is a unitary matrix mapping the original basis of $F$ into its eigenbasis. Hence, by simply choosing the first $k$ eigenvalues and their respective eigenvectors we have the central result:

$$\min_k |F - \hat{F}| \rightarrow \hat{F} \approx U_k \Sigma_k V_k^T \tag{3}$$

where $\hat{F}$ is the best rank-$k$ approximation in the Frobenius-norm. This is commonly referred to as *truncated* Singular Value Decomposition (SVD).

Finally, using cosine similarity,[3] $V$ is redundant due to invariance under unitary transformations, which means we can represent the principal components of $\hat{F}$ in its most compact form $\hat{F} \equiv U\Sigma$ without any further comment.

This projection onto the eigenbasis does not only provide an efficient compression of the sparse co-occurrence data, but has also been shown to improve the performance and noise tolerance of DSMs (Schütze, 1992; Landauer and Dumais, 1997; Bullinaria and Levy, 2012).

## 4 The Caron $p$-transform

Caron (2001) introduce a method for renormalization of the latent variables through an exponent factor $p \in R$:

$$U\Sigma \rightarrow U\Sigma^p \tag{4}$$

which is shown to improve the results of factorized models using both information retrieval and question answering test collections. We refer to this renormalization as the *Caron $p$-transform*. Bullinaria and Levy (2012) further corroborate Caron's result, and show that the optimum exponent parameter $p$ for DSMs is with strong statistical significance $p < 1$. Moreover, due to the redistribution of weight to the lower variance PCs, Bullinaria and Levy (2012) show that similar performance improvements can be achieved by simply *removing* the first PCs. We refer to this as *PC-removal*. A highlight of their results is a perfect score of 100% on the TOEFL synonym test.

Apart form the perfect score on the TOEFL test, it is noteworthy that the PC-removal scheme is the inverse of how SVD is normally used in DSMs; instead of retaining only the first PCs – which is the standard way of using the SVD in DSMs – the PC-removal scheme *deletes* them, and instead retains all the rest.

## 5 Experiments

We replicate the experiment setup of Bullinaria and Levy (2012) by removing punctuation and decapitalizing the ukWaC corpus (Baroni et al., 2009). The DSM includes the 50,000 most frequent words along with the remaining 23 TOEFL words and is populated using a $\pm 2$-sized context window. Co-occurrence counts are weighted with PPMI, and SVD is applied to the resulting matrix, reducing the dimensionality to 5,000. The results of removing the first PCs versus applying the Caron $p$-transform are shown in Figure 1, which replicates the results from Bullinaria and Levy (2012).

In order to better understand what influence the transform has on the representations, we also provide results on the BLESS test (Baroni and Lenci, 2011), which lists a number of related terms to 200 target terms. The related terms represent 8 different kinds of semantic relations (co-hyponymy, hypernymy, meronymy, attribute, event, and three random classes corresponding to randomly selected nouns, adjectives and verbs), and it is thus possible to use the BLESS test to determine what type of semantic relation a model favors. Since our primary interest here is in *paradigmatic* relations, we focus on the hypernymy and co-hyponymy relations, and require that the model scores one of the related terms from these classes higher than the related terms from the other classes. The corpus was split into different sizes to test the statistical significance of the weight redistribution effect. Furthermore, it shows that the optimal weight distribution depends on the size of the data.

[2] It is well known that the Gaussian assumption does not hold in reality, and consequently there are also other approaches to dimensionality reduction based on multinomial distributions, which we will not consider in this paper.

[3] $\cos(w_i, w_j) = \frac{w_i \cdot w_j}{|w_i||w_j|}$

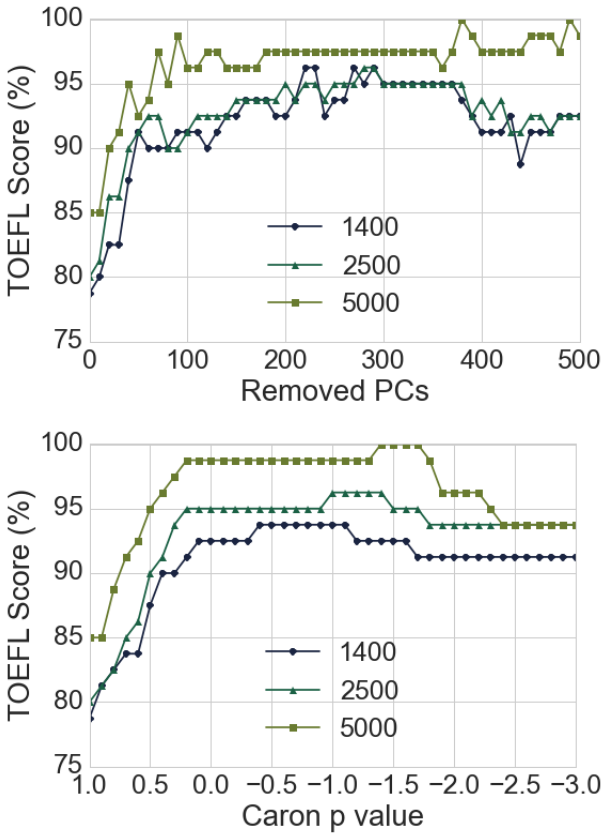



Figure 1: TOEFL score for the PC-removal scheme and the Caron $p$-transform for the span of PCs.

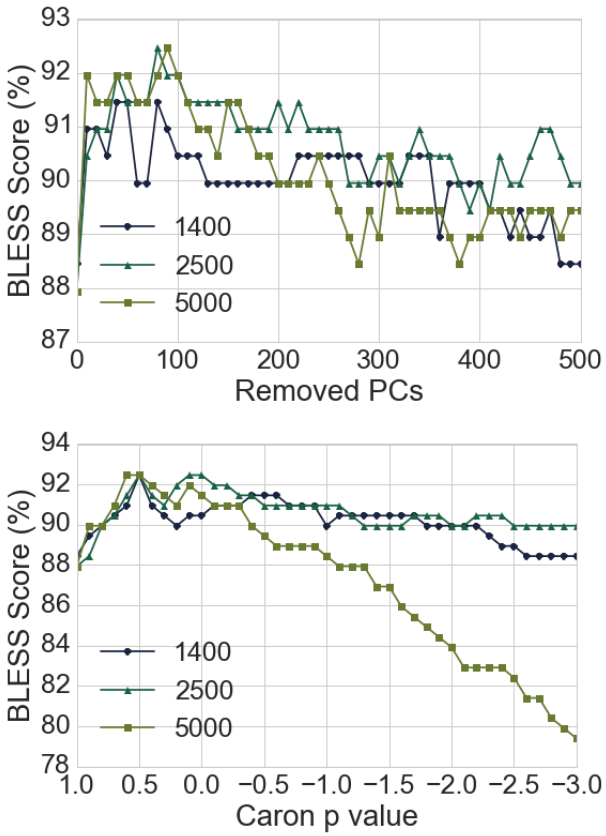



Figure 2: BLESS score for the PC-removal scheme and the Caron $p$-transform for the span of PCs.

Figure 2 shows the BLESS results for both the PC removal scheme and the Caron $p$-transform for different sizes of the corpus. The best score is 92.96% for the PC removal, and 92.46% for the Caron $p$-transform, both using the full data set. Similarly to the TOEFL results, we see better results for a larger number of removed PCs. Interestingly, there is clearly a larger improvement in performance of the Caron $p$-transform than for the PC removal scheme.

Figure 3 shows how the redistribution affects the different relations in the BLESS test. The violin plots are based on the maximum values of each relation, and the width of the violin represents the normalized probability density of cosine measures. The cosine distributions, $\Theta_i$, are based on the best matches for each category $i$, and normalized by the total mean and variance amongst all categories $\hat{\Theta}_i = \frac{\Theta_i - \mu}{\sigma}$. Thus, the figure illustrates how well each category is separated from each other, the larger separation the better.

The results in Figure 3 indicate that the top 120 PCs contain a higher level of co-hyponymy relations than the lower; removing the top 120 PCs gives a violin shape that resembles the inverse of the plot for the top 120 PCs. Although neither part of the PC span is significantly better in separating the categories, it is clear that removing the first 120 PCs increases the variance within the categories

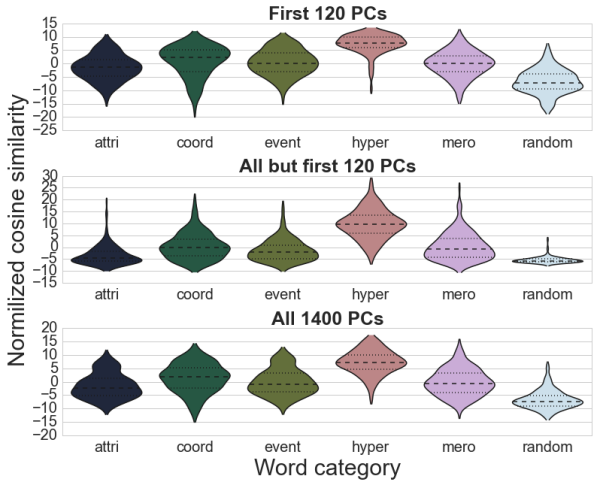



Figure 3: BLESS targets versus categories from 1,400 PCs representation of the entire corpus.

and especially amongst the coord-category. This is an interesting result, since it seems to contradict the hypothesis that removing the first PCs improves the semantic quality of the representations – there is obviously valuable semantic information in the first PCs.

Table 1 summarizes our top results on the TOEFL, BLESS, and also the SimLex-999 similarity test (Hill et al., 2014), and compares them to a baseline score from the Skipgram model (Mikolov et al., 2013a), trained on the same data using a window size of 2, negative samples, and 400-dimensional vectors.

| | TOEFL | BLESS | SimLex-999 |
|---|---|---|---|
| PC removal | 100 | 92.96 | 46.52 |
| Caron $p$ | 100 | 92.46 | 46.66 |
| Skipgram | 83.75 | 83.00 | 39.91 |

Table 1: Top results for the PC removal and Caron $p$ on each test compared to the Skipgram model.

Unfortunately, the optimal redistribution of weight on the PCs for the respective top scores differ between the experiments. For the PC removal the optimal number of removed PCs is 379 for TOEFL, 15 for BLESS and 128 for SimLex-999, while the optimal number for the Caron $p$-transform is -1.4 for TOEFL, 0.5 for BLESS and -0.40 for SimLex-999. Hence, there is likely no easy way to find a general expression of the optimal redistribution of weight on the PCs for a given application.

## 6 The Pareto Principle

It is common practice to reduce the dimensionality of an $n$-dimensional space to as many PCs as it takes to cover 80% of the total eigenvalue mass. This convention is known as the *Pareto principle* (or 80/20-rule), and generally gives a good trade-off between compression and precision. The results presented in the previous section suggest a type of inversion of this principle in the case of DSMs.

Given a computational and practical limit of the number of PCs $m$ with weights $\Sigma = \{\sigma_1, ..., \sigma_m\}$, the optimal redistribution of weight on these components is such that the first $l - m$ components $\sigma_1, ..., \sigma_{m-l}$ is transformed such that they constitute 20% of the new total mass. Where $l - m$ is the number of components representing the last 20 % of the original mass. In other words, the function $f : \Sigma \rightarrow \hat{\Sigma}$ performing this redistribution is such that:

$$\frac{\sum_{i=1}^{m-l} \hat{\sigma}_i}{\sum_{i=1}^{m} \hat{\sigma}_i} \approx 20\% \quad (5)$$

In this formulation, we can consider the Caron $p$-transform and the PC-removal scheme as special cases, where the Caron $p$-transform is given by:

$$f(\sigma_i) = \sigma_i^p \quad \forall i,\ p \in R \quad (6)$$

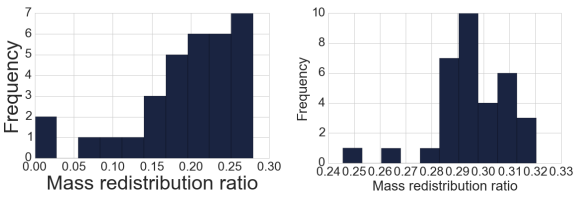
and the PC-removal scheme by:

$$f(\sigma_i) = (1 - \delta(F))\sigma_i \quad \forall i,\ F = \{1....l\} \quad (7)$$

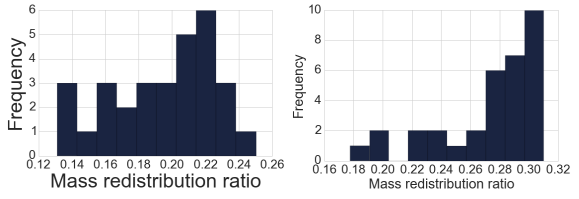
where $\delta(F)$ denotes the generalized Kronecker delta function.

To test this claim, we form this quotient for the distributions of weights at the optimal parameters for the Caron $p$-transform and the PC-removal scheme for both the BLESS and TOEFL tests for each of the 40 sub-corpora.

Even though the results are not as optimal for the BLESS test as for the TOEFL, the results in Figure 4 point in favor of this measure. The optimal mass distributions for the Caron $p$-transform and the PC removal are all around 20 %.




(a) TOEFL - PC removal (b) BLESS - PC removal




(c) TOEFL - Caron $p$ (d) BLESS - Caron $p$

Figure 4: The mass redistribution ratio for the best results on the 1,400 PC models.

This result does not only apply for 1,400 PCs, but has also been verified on a smaller set of matrices with sizes of 2,500 PCs, 4,000 PCs and 5,000 PCs. The results for 1,400 PCs and 5,000 PCs are shown in Table 2. As can be seen in this table, the rule of thumb yields reasonable good guesses for both Caron $p$ and PC removal, over the different tests and for various number of PCs.

**5,000 PC representation**

| | PC removal | Caron $p$ |
|---|---|---|
| Parameter | 493 | -0.80 |
| TOEFL | 100 | 98.75 |
| BLESS | 89.45 | 88.95 |
| SimLex | 44.82 | 45.74 |

**1,400 PC representation**

| | PC removal | Caron $p$ |
|---|---|---|
| Parameter | 204 | -0.80 |
| TOEFL | 93.75 | 93.75 |
| BLESS | 89.95 | 90.95 |
| SimLex | 45.45 | 46.47 |

Table 2: Results for the PC removal and Caron $p$ using the 80/20 rule

## 7 Conclusions and future work

This paper has discussed the method of redistributing the weight of the first PCs in factorized DSMs. We have replicated previously published results, and provided additional empirical justification for the method. The method significantly outperforms the baseline Skipgram model on all tests used in the experiments. Our results also suggest a slight refinement of the method, for which we have provided both theoretical and empirical justification. The resulting rule of thumb method leads to stable results that may be useful in practice.

Although the experiments in this paper has provided further evidence for the usefulness of redistributing the weight in factorized models, it also raises additional interesting research questions. For example, does the method also improve models that have been trained on smaller data sets? Does it also hold for non-Gaussian factorization like Non-negative Matrix Factorization? How does the method affect the (local) structural properties of the representations; do factorized models display the same type of structural regularities as has been observed in word embeddings (Mikolov et al., 2013b), and would it be possible to use methods such as relative neighborhood graphs (Gyllensten and Sahlgren, 2015) to explore the local effects of the transformation?

## References


Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. In *Proceedings of GEMS*, pages 1–10.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pages 238–247.

John Bullinaria and Joseph Levy. 2012. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods*, 44(3):890–907.

John Caron. 2001. Experiments with LSA scoring: Optimal rank and basis. In Michael Berry, editor, *Computational Information Retrieval*, pages 157–169.

Amaru Cuba Gyllensten and Magnus Sahlgren. 2015. Navigating the semantic horizon using relative neighborhood graph. In *Proceedings of EMNLP*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *CoRR*, abs/1408.3456.

Thomas Landauer and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Proceedings of NIPS*, pages 2177–2185.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings NAACL-HLT*, pages 746–751.

Yoshiki Niwa and Yoshihiko Nitta. 1994. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of COLING*, pages 304–309.

Hinrich Schütze. 1992. Dimensions of meaning. In *Proceedings of Supercomputing*, pages 787–796.