Supplemental Materials of Getting in Shape: Word Embedding SubSpaces

Tianyuan Zhou[1], **João Sedoc**[2] and **Jordan Rodu**[1*]
[1]Department of Statistics, University of Virginia
[2]Department of Computer and Information Science, University of Pennsylvania
tz8hu@virginia.edu, joao@cis.upenn.edu, jsr6q@virginia.edu

1 Supplementary Theory

Restatement of Lemma 1 with proof

Lemma 1. *$\sigma(\hat{Y}) = \sigma(U_X^T U_Y \Sigma_Y)$, where U_X and U_Y are left singular vectors of X and Y, and Σ_Y is the diagonal matrix where the diagonal elements are singular values of Y. $\hat{Y}$ is the projection of Y on X by linear regression. (We'd better isolate this so that we can have several corollaries, such as when $\Sigma_Y = \sigma_Y I$ then $\sigma(\hat{Y}) = \sigma_Y \sigma(U_X^T U_Y)$)*

Proof. *By linear regression, we have $\hat{Y} = X(X^TX)^{-1}X^TY = U_X U_X^T Y = U_X U_X^T U_Y \Sigma_Y V_Y^T$. Note that the leftmost matrix U_X and the rightmost matrix V_Y^T doesn't change the singular vales of $\hat{Y}$, and only the middle part $U_X^T U_Y \Sigma_Y$ decides $\sigma(\hat{Y})$. So we have $\sigma(\hat{Y}) = \sigma(U_X^T U_Y \Sigma_Y)$.*

Restatement of propositions with proof

Proposition 1. *Suppose we have a set of i.i.d. r.v.s, denote as $X_1, X_2, \cdots$ where each $X_i = (x_{i,1}, \cdots, x_{i,p})$ is of a p-dim row vector whose elements are of mean 0, independent and have identical first to fourth moments, and suppose we have a stretching matrix $\Sigma = diag(\sqrt{\sigma_1}, \sqrt{\sigma_2}, \cdots, \sqrt{\sigma_p})$. Then we have the following two results:*

$$corr(||X_i - X_j||_2^2, ||X_i\Sigma - X_j\Sigma||_2^2) = \frac{||\Sigma^2||_1/\sqrt{p}}{||\Sigma^2||_2} = \frac{(\sigma_1 + \sigma_2 + \cdots + \sigma_p)/\sqrt{p}}{\sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2}}$$

and

$$corr(< X_i, X_j >, < X_i\Sigma, X_j\Sigma >) = \frac{||\Sigma^4||_1/\sqrt{p}}{||\Sigma^4||_2} = \frac{(\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2)/\sqrt{p}}{\sqrt{\sigma_1^2 4 + \sigma_2^4 + \cdots + \sigma_p^4}}$$

where $||A||_2$ and $||A||_1$ are entry-wise 2-norm (Frobenius norm) and 1-norm (absolute summation) respectively (not norm induced by vector norm).

In addition, denote I as a set of column index such as $\{1, 2\}$, and denote $X_{i,I}$ as the subspace of X_i indexed by I. For example, $X_{i,\{1\}} = [X_{i,1}, 0, \cdots, 0]$, namely only keep the first column of X_i but let all other columns be 0. The correlation of $||X_i\Sigma - X_j\Sigma||_2^2$ and $||X_{i,I} - X_{j,I}||_2^2$ is $\frac{\Sigma_{i\in I}\sigma_i/\sqrt{|I|}}{\sqrt{\sigma_1^2+\sigma_2^2+\cdots+\sigma_p^2}}$

Proof. *We only talk about $corr(||X_i - X_j||_2^2, ||X_i\Sigma - X_j\Sigma||_2^2)$ as the proof of other terms are the same as this. $corr(||X_i - X_j||_2^2, ||X_i\Sigma - X_j\Sigma||_2^2) = \frac{cov(||X_i-X_j||_2^2, ||X_i\Sigma-X_j\Sigma||_2^2)}{\sqrt{var(||X_i-X_j||_2^2) var(||X_i\Sigma-X_j\Sigma||_2^2)}}$. Denote each dimensions of X_i as $u_1, u_2, \cdots, u_p$, and of X_j as $v_1, \cdots, v_p$. Then $||X_i - X_j||_2^2 = (u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_p - v_p)^2$, $||X_i\Sigma - X_j\Sigma||_2^2 = (u_1\sqrt{\sigma_1} - v_1\sqrt{\sigma_1})^2 + (u_2\sqrt{\sigma_2} - v_2\sqrt{\sigma_2})^2 + \cdots + (u_p\sqrt{\sigma_p} - v_p\sqrt{\sigma_p})^2$. Since for different dimensions $u_i, v_i \perp u_j, v_j$, $cov(||X_i - X_j||_2^2, ||X_i\Sigma - X_j\Sigma||_2^2 = \sigma_1 var((u_1 - v_1)^2) + \sigma_2 var((u_2 - v_2)^2) + \cdots + \sigma_1 var((u_p - v_p)^2) = var((u_1 - v_1)^2)(\sigma_1 + \sigma_2 + \cdots + \sigma_p)$ because $var((u_1 - v_1)^2) = var((u_2 - v_2)^2) = \cdots = var((u_p - v_p)^2)$. Similarly, we have $var(||X_i - X_j||_2^2 = p\sigma_x^2$ and $var(||X_i\Sigma - X_j\Sigma||_2^2 = var((u_1 - v_1)^2)(\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2)$. Substitute these expressions into $corr(||X_i - X_j||_2^2, ||X_i\Sigma - X_j\Sigma||_2^2)$ and we have the proof.*

The proof is very straightforward. In this theorem, what we want to say is that if we stretch the isotropic matrix, than the distance and inner product structure will be destroyed, in the sense of its correlation with the original ones will be low if the stretching matrix Σ is very imbalanced. We have the following corollary: in an extreme case, which is nearly the case in the word vectors, if $\Sigma = diag(1, 0, 0, \cdots, 0)$, then $corr(||X_i - X_j||_2^2, ||X_i\Sigma - X_j\Sigma||_2^2) = corr(< X_i, X_j >, < X_i\Sigma, X_j\Sigma >) = 1/\sqrt{p}$. Or more generally, if $\Sigma = diag(1, 1, \cdots, 1, 0, 0, \cdots, 0)$ where there are m terms of 1 and $p - m$ terms of 0, then $\Sigma = diag(1, 0, 0, \cdots, 0)$, then $corr(||X_i - X_j||_2^2, ||X_i\Sigma - X_j\Sigma||_2^2) = corr(< X_i, X_j >, < X_i\Sigma, X_j\Sigma >) = \sqrt{\frac{m}{p}}$.

We have one more theorem on the image-kernel decomposition of this transformation if it is a random projection:

Proposition 2. *Suppose we have a set of i.i.d. r.v.s, denote as $X_1, X_2, \cdots$ where each $X_i = (x_{i,1}, \cdots, x_{i,p})$ is of a p-dim row vector following multivariate normal distribution*

*Contact Author

$N((0,0,\cdots,0),\sigma_x^2 I_p)$. *Also we have a random-projection matrix* $\Sigma = diag(1,1,\cdots,1,0,0,\cdots,0)$ *where the first* m *terms are* 1 *and the others are 0. Then* $\frac{||X_i\Sigma - X_j\Sigma||_2^2}{||X_i - X_j||_2^2}$, *namely the ratio between the distance of each pair of data points after transformation and of before transformation, follows* $Beta(\frac{m}{2},\frac{p-m}{2})$ *distribution. And this ratio is independent with* $||X_i - X_j||_2^2$, *which follows* $\sqrt{2\sigma_x^2}\chi_p^2$ *distribution.*

Proof. *Denote each component of* X_i *as* $\{u_1,u_2,\cdots,u_p\}$ *and each component of* X_j *as* $\{v_1,v_2,\cdots,v_p\}$ *Here we can write* $||X_i - X_j||_2^2 = \Sigma_{l=1}^p (u_l - v_l)^2 = \Sigma_{l=1}^m (u_l - v_l)^2 + \Sigma_{l=m+1}^p (u_l - v_l)^2$ *and* $||X_i\Sigma - X_j\Sigma||_2^2 = \Sigma_{l=1}^m (u_l - v_l)^2$. *Note that* $\forall l\ (u_l - v_l)^2 \sim \sqrt{2\sigma_x^2}\chi_1^2$. *So that* $\frac{||X_i\Sigma - X_j\Sigma||_2^2}{||X_i - X_j||_2^2} = \frac{\Sigma_{l=1}^m (u_l - v_l)^2}{\Sigma_{l=1}^m (u_l - v_l)^2 + \Sigma_{l=m+1}^p (u_l - v_l)^2}$ *follows* $Beta(\frac{m}{2},\frac{p-m}{2})$, *and this ratio is independent with* $||X_i - X_j||_2^2$ *due to Beta-Gamma relationship.*

In the case of X and Y are independent, we have the following series of results.

Lemma 2. *Suppose* $X_{n\times p}$ *and* $Y_{n\times p}$ *are two independent orthogonal matrices. The column vectors of* X *are Haar invariant which is obtained through performing the Gram–Schmidt procedure for a matrix whose elements are independent standard normals, or obtained through getting the left singular vectors of such a matrix. Then if* p *is given and* $n \to \infty$, *each element of* $\sqrt{n}X^TY$ *converges to* $N(0,1)$, *or* $\sqrt{X}^TY$ *converges to Gaussian random matrix in distribution.*

Proof. *The typical method of obtaining a Haar invariant orthogonal matrix is from performing the Gram–Schmidt procedure for a Gaussian random matrix [**?**]. Also, by performing SVD on a Gaussian random matrix, both the left and right singular vectors are Haar invariant [**?**]. Note that both the left singular vectors and the basis get from Gram–Schmidt procedure are the orthonormal basis of the columns spaces. That is to say, they are the same upon a unitary transformation. Therefore, we only need to prove the matrix obtained through performing the Gram–Schmidt procedure for a Gaussian random matrix.*

It is easy to show that the inner product of two independent n*-dim unit vectors, one of which is uniformly distributed on* S_{n-1}, *follows* $2 * Beta(\frac{n-1}{2},\frac{n-1}{2}) - 1$ *which will converges in distribution to* $N(0,\frac{1}{n})$ *[**?**].*

Now suppose we have two orthogonal matrices $X_{(n\times p)} = [X_{1(n\times 1)}, X_{2(n\times 1)},\cdots,X_{p(n\times 1)}]$ *and* $Y_{(n\times p)} = [Y_{1(n\times 1)}, Y_{2(n\times 1)},\cdots,Y_{p(n\times 1)}]$ *either of which has* p *orthogonal columns. All* X_i*'s are uniformly distributed on* S_{n-1}. *Let* $D = X^TY$. *Then each element of* D *is the inner product of two column vectors of* X *and* Y *i.e.* $D_{ij} = X_i^TY_j$.

Let's study the joint distribution of all elements of D. *At first, we study the joint distribution of* D_{11} *and* D_{12}, $f(D_{11},D_{12}) = f(D_{11})f(D_{12}|D_{11})$. *From previous argument, we know that* $\sqrt{n}D_{11} = \sqrt{n}X_1^TY_1$ *will converge to* $N(0,1)$ *as* $n \to \infty$. $f(D_{12}|D_{11})$ *depends on* D_{11} *because* X_1 *and* X_2 *are orthogonal to each other. Note that, given* X_1, X_2 *can be generated in the following process [**?**]: randomly generate a vector with elements follow* $i.i.d.$ $N(0,1)$ *and normalize it to make it uniformly on* S_{n-1} *denoted as* $\tilde{X_2}$, *and, Gram-Schmidt orthonormalize it. That is to say* $X_2 = \frac{\tilde{X_2} - (\tilde{X_2}^TX_1)X_1}{||\tilde{X_2} - (\tilde{X_2}^TX_1)X_1||_2}$. *Note that* $\tilde{X_2}^TX_1$ *will also converge to* $N(0,\frac{1}{n})$. *Therefore,* $D_{12}|D_{11}$ *has the same distribution of* $X_2^TY_1 = \frac{\tilde{X_2}^TY_1 - (\tilde{X_2}^TX_1)(X_1^TY_1)}{||\tilde{X_2} - (\tilde{X_2}^TX_1)X_1||_2}$. *We can find that* $\sqrt{n}X_2^TY_1 = \frac{\sqrt{n}\tilde{X_2}^TY_1 - (\tilde{X_2}^TX_1)(\sqrt{n}X_1^TY_1)}{||\tilde{X_2} - (\tilde{X_2}^TX_1)X_1||_2}$, *where* $||\tilde{X_2} - (\tilde{X_2}^TX_1)X_1||_2 \xrightarrow{P} 1$, $(\tilde{X_2}^TX_1)(\sqrt{n}X_1^TY_1) \xrightarrow{P} 0$ *and* $\sqrt{n}\tilde{X_2}^TY_1 \xrightarrow{d} N(0,1)$, *so by Slutsky's theorem* $\sqrt{n}X_2^TY_1 \xrightarrow{d} N(0,1)$, *which is independent of* D_{11}. *That is to say,* D_{11} *and* D_{12} *are asymptotically* $i.i.d.$.

With the same argument, if we already have $X_1,\cdots,X_l$ *and* $Y_1,\cdots,Y_p$, *then we can still generates* $\tilde{X_{l+1}}$ *and* $\tilde{Y_{m+1}}$ *and get* $X_{l+1} = \frac{\tilde{X_{l+1}} - \Sigma_{i=1}^l(\tilde{X_{l+1}}^TX_i)X_i}{||\tilde{X_{l+1}} - \Sigma_{i=1}^l(\tilde{X_{l+1}}^TX_i)X_i||_2}$. *Then for any* Y_m, *we have* $\sqrt{n}X_{l+1}^TY_m = \frac{\sqrt{n}\tilde{X_{l+1}}^TY_m - \Sigma_{i=1}^l(\sqrt{n}\tilde{X_{l+1}}^TX_i)(X_i^TY_m)}{||\tilde{X_{l+1}} - \Sigma_{i=1}^l(\tilde{X_{l+1}}^TX_i)X_i||_2}$. *Since* p *is fixed and* $l \le p$, *both* $\Sigma_{i=1}^l(\sqrt{n}\tilde{X_{l+1}}^TX_i)(X_i^TY_m)$ *and* $\Sigma_{i=1}^l(\tilde{X_{l+1}}^TX_i)X_i$ *are finite sum of terms which will converge to* 0 *in probability, and as a result, both two terms will converge to* 0 *in probability. Note that* $\sqrt{n}\tilde{X_{l+1}}^TY_m$ *will converge to* $N(0,1)$. *Then by Slutsky's Theorem,* $\sqrt{n}X_{l+1}^TY_m$ *will converge to* $N(0,1)$. *Similarly, if we already have* $Y_1,\cdots,Y_l$ *and* $X_1,\cdots,X_p$, *we can generate* Y_{l+1} *and get the same results.*

With this argument, we get the joint distribution of all elements of D, *all of which converge to* $i.i.d.$ $N(0,\frac{1}{n})$ *in distribution, namely* $\sqrt{n}D$ *converges to Gaussian random matrix in distribution.*

Corollary 1. $\sigma(D) \xrightarrow{d} \sigma(G(p))$ *if* $D \xrightarrow{d} G(p)$. *Moreover, the singular values of* $\hat{Y}$ *when* Y *is isotropic noise, distributed as* $\frac{1}{\sqrt{n}}\sigma(G(p))$.

Proof. *Since* $\sigma(\cdot)$ *is a continuous mapping, by continuous mapping theorem we have the result.*

Corollary 2. *If* X *and* Y *are independent matrices of shape* $n \times p$ *where* $n \to \infty$ *and* p *is given. Denote* $SVD(X) = U_X\Sigma_XV_X^\top$ *and* $SVD(Y) = U_Y\Sigma_YV_Y^\top$. *If either* U_X *or* U_Y *is Haar invariant that can be obtained from performing the Gram–Schmidt procedure for a matrix whose elements are independent standard normals, or obtained through getting the left singular vectors of such a matrix, then the probability of not existing a singular value of* $\widehat{Y}$ *smaller than a given number* s $\mathrm{P}(\sigma_1(\widehat{Y}) \le s) \le \Pi_{i=1}^p F_{\chi_p^2}(\frac{s^2}{n\sigma_i^2(Y)}) + \epsilon$, *where* $F_{\chi_p^2}(\cdot)$ *is the cumulative distribution function of* χ_p^2 *and* ϵ *is any given small positive number.*

Proof. *By previous theorem, we now that for all* $1 \le i \le p$, *there exists at least a singular value in* $\widehat{Y}$ *at least as large as* $||U_X^\top U_{yi}||_2\sigma_i(Y)$ *where* U_{yi} *is the corresponding left singular vector of* $\sigma_i(Y)$. *That is to say,* $\sigma_1(\widehat{Y}) \ge \max_i(||U_X^\top U_{yi}||_2\sigma_i(Y))$.

$U_X^\top U_{yi}$ is the i-th column of $U_X^\top U_Y$, and by previous lemma, $\sqrt{n}U_X^\top U_Y \xrightarrow{d} G(p)$. So that $||\sqrt{n}U_X^\top U_{yi}||_2^2$ is the summation of the square of the i-th column of $\sqrt{n}U_X^\top U_Y$ which converges in distribution to the χ_p^2 by continuous mapping theorem. That is to say, $\forall t \geq 0, \mathrm{P}(||\sqrt{n}U_X^\top U_{yi}||_2^2 \leq t) \to F_{\chi_p^2}(t)$. Also by continuous mapping theorem, $\forall s \geq 0, \mathrm{P}(\max_i(||U_X^\top U_{yi}||_2\sigma_i(Y)) \leq s) \to \mathrm{P}(\forall i \quad ||U_X^\top U_{yi}||_2\sigma_i(Y) \leq s) \to \Pi_{i=1}^p(||U_X^\top U_{yi}||_2\sigma_i(Y) \leq s) \to \Pi_{i=1}^p F_{\chi_p^2}(\frac{s^2}{n\sigma_i^2(Y)})$. By the definition of convergence, for any $\epsilon > 0$, $\mathrm{P}(\max_i(||U_X^\top U_{yi}||_2\sigma_i(Y)) \leq t) \leq \Pi_{i=1}^p F_{\chi_p^2}(\frac{s^2}{n\sigma_i^2(Y)}) + \epsilon$. Since $\sigma_1(\widehat{Y}) \geq \max_i(||U_X^\top U_{yi}||_2\sigma_i(Y))$, $\mathrm{P}(\sigma_1(\widehat{Y}) \leq s) \leq \mathrm{P}(\max_i(||U_X^\top U_{yi}||_2\sigma_i(Y)) \leq s) \leq \Pi_{i=1}^p F_{\chi_p^2}(\frac{s^2}{n\sigma_i^2(Y)}) + \epsilon$.

2 Supplementary Empirical Results

2.1 Sequential Comparisons of More Dependency Embeddings

Here we show the empirical results of 5 more kinds of Dependency Embeddings along with the one shown in the main body. The 6 kinds of Dependency Embeddings are of 250-dim or 500-dim, and of CBOW, Skig-Gram or GloVe. The overall trends of them are similar.

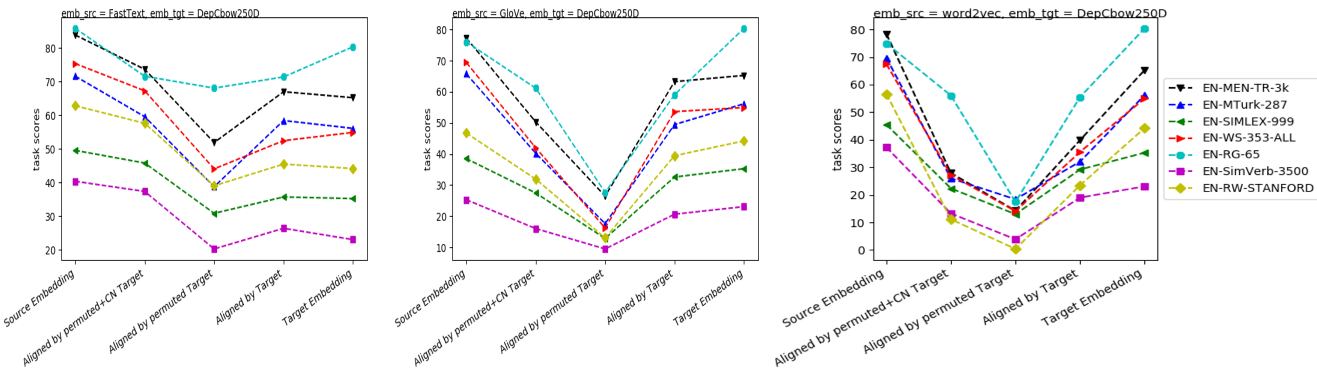

Figure 1: Sequential Comparison of FastText/GloVe/word2vec Aligned With Dependency Embedding (CBOW, 250-dim)

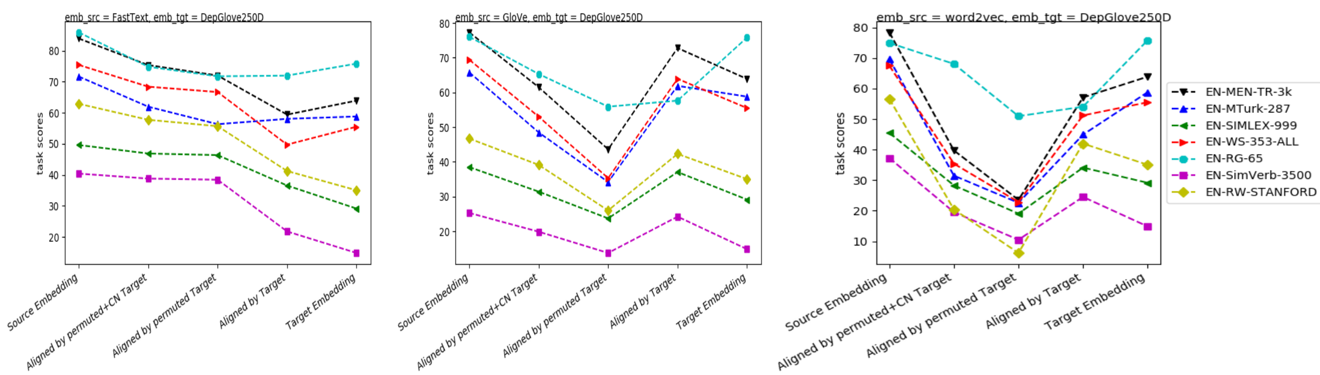

Figure 2: Sequential Comparison of FastText/GloVe/word2vec Aligned With Dependency Word Embedding (GloVe, 250D)