Supplementary Material: Two Step CCA: A new spectral method for estimating vector models of words

Paramveer S. Dhillon DHILLON@CIS.UPENN.EDU
Computer & Information Science, University of Pennsylvania, Philadelphia, PA 19104 U.S.A

Jordan Rodu JRODU@WHARTON.UPENN.EDU
Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 U.S.A

Dean P. Foster FOSTER@WHARTON.UPENN.EDU
Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 U.S.A

Lyle H. Ungar UNGAR@CIS.UPENN.EDU
Computer & Information Science, University of Pennsylvania, Philadelphia, PA 19104 U.S.A

1. Proof of Theorem 1

The key to the proof is that CCA can be understood using the same machinery as is used for analyzing linear regression. The equivalent regression problem is that we want to recover the word *type* of length v given its context. For a more in-depth discussion of how CCA relates to regression, see (Glahn, 1968), for example. Thus, consider the case of predicting a vector $\mathbf{y}$ of length v (the word *type*) from a vector $\mathbf{x}$ (the context, which is of dimension $2hv$ in the one step CCA case and dimension $2k$ in the two step CCA). Consider the linear model

$$y = \mathbf{x}\beta + \epsilon$$

Note that, we are predicting only one dimension of our v-dimensional vector $\mathbf{y}$ at a time.

We want to understand the variance of our prediction of a word given the context. As is typical in regression, we calculate a standard error for each coefficient in our contexts, $\approx O(\frac{1}{\sqrt{n}})$. For one step CCA (OSCCA), we have $\mathbf{X} = [\mathbf{L}\ \ \mathbf{R}]$, and running a regression we will get a prediction error on order of $\frac{hv}{n}$, but since we have v such y's we get a total prediction error on the order of $\frac{hv^2}{n}$.

For the two-step case (TSCCA) we have $\mathbf{X} = [\mathbf{L}\mathbf{\Phi_L}\ \ \mathbf{R}\mathbf{\Phi_R}]$. As mentioned earlier, note that now we are working with about $2k$ predictors instead of $2hv$ predictors. If we knew the true $\mathbf{\Phi_L}$ and $\mathbf{\Phi_R}$, and thus the true subspace covered by our predictors, the regression error would be on the order of $\frac{kv}{n}$ (again, since there are v entries in our vector $\mathbf{y}$). Instead, we have an estimation of $\mathbf{\Phi_L}$ and $\mathbf{\Phi_R}$. If these were computed on infinite amount of data, then we would be arbitrarily close to the true subspace and we would be done. However since they come from a sample, we are using $\widehat{\mathbf{\Phi_L}}$ and $\widehat{\mathbf{\Phi_R}}$ which are approximation to the ideal $\mathbf{\Phi_L}$ and $\mathbf{\Phi_R}$. So our task is to understand the error introduced by this sample approximation of the true CCA. First, we develop some notation and concepts found in (Stewart, 1990).

Consider two subspaces $\mathcal{V}$ and $\hat{\mathcal{V}}$ and the respective matrices containing an orthonormal basis for these subspaces $\mathbf{V}$ and $\hat{\mathbf{V}}$. Let $\gamma_1, \gamma_2, \ldots$ be the singular values of the matrix $\mathbf{V}^\top\hat{\mathbf{V}}$, then define

$$\theta_i = \cos^{-1}\gamma_i$$

and define the canonical angle matrix $\mathbf{\Theta} = \text{diag}(\theta_\mathbf{1}, \ldots, \theta_\mathbf{k})$.

These values of $\mathbf{\Theta}$ capture the effect of using estimated singular vectors, $\hat{\mathbf{V}}$ to form an underlying subspace, as compared to the true subspace formed by the true singular vectors $\mathbf{V}$ stemming from infinite data. The largest canonical angle captures the largest angle between any two vectors– one from the perturbed subspace and one from the true subspace. The second largest canonical angle captures the second largest angle between any two vectors given that they are orthogonal to the original two, and so on. In this proof we will only make use of the largest canonical angle to provide a loose upper bound on the error stemming from the imperfect estimation of the true subspace.

Now, consider a matrix $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ and take the thin singular value decomposition of $\mathbf{A}$ and $\hat{\mathbf{A}}$ (and here

we take the liberty of applying diag in a block matrix sense)

$$\mathbf{A} = [\mathbf{U_1}\mathbf{U_2}]\text{diag}(\mathbf{\Lambda_1}, \mathbf{\Lambda_2})[\mathbf{V_1}\mathbf{V_2}]^\top$$
$$\hat{\mathbf{A}} = [\hat{\mathbf{U}}_\mathbf{1}\hat{\mathbf{U}}_\mathbf{2}]\text{diag}(\hat{\mathbf{\Lambda}}_\mathbf{1}, \hat{\mathbf{\Lambda}}_\mathbf{2})[\mathbf{V_1}\mathbf{V_2}]^\top$$

In our case we have that $\lambda_i = 0$ for all $\lambda_i \in \mathbf{\Lambda_2}$.

From (Stewart & guang Sun, 1990), we have that

$$\max\{||\sin\mathbf{\Theta}||_2, ||\sin\mathbf{\Psi}||_2\} \leq c||E||_2 \tag{1}$$

for some constant c where here $\mathbf{\Theta}$ is the matrix of canonical angles formed from the subspaces of $\mathbf{U}$ and $\hat{\mathbf{U}}$, and $\mathbf{\Psi}$ is the matrix of canonical angles formed between the subspaces of $\mathbf{V}$ and $\hat{\mathbf{V}}$. Note that since $\mathbf{\Theta}$ and $\mathbf{\Psi}$ are diagonal matrices the induced norms $||\cdot||_2$ recover the largest canonical angle of each subspace, and hence we can simultaneously derive an upper bound for the largest canonical angle of either subspace.

We have now developed the machinery we need to analyze the two step CCA.

Without loss of generality, assume that $\mathbf{L}^\top\mathbf{L} = \mathbf{R}^\top\mathbf{R} = \mathbf{I}$, then ultimately we are interested in projection onto the subspace spanned by $\mathbf{B} = [\mathbf{LU_1} \quad \mathbf{RV_1}]$. Note that because of our assumption the projection onto $\mathbf{LU_1}$ is $\mathbf{LU_1U_1^\top L^\top}$ and similarly for $\mathbf{RV_1}$. Furthermore, note from our assumptions that $\mathbf{LU_1}$ forms an orthonormal basis for the space spanned by $\mathbf{LU_1}$ (since

$$(\mathbf{LU_1})^\top(\mathbf{LU_1}) = \mathbf{U_1^\top L^\top LU_1} = \mathbf{I}$$

and similarly for $\mathbf{L\hat{U}_1}$, $\mathbf{RV_1}$, and $\mathbf{R\hat{V}_1}$).

Lastly, and critically, the singular values of $\mathbf{U_1^\top L^\top L\hat{U}_1}$ are identical to those of $\mathbf{U_1^\top\hat{U}_1}$ (similarly for $\mathbf{RV_1}$ etc.) and so from above we have that the matrix of canonical angles between the subspaces $\mathbf{LU_1}$ and $\mathbf{L\hat{U}_1}$ are identical to $\mathbf{\Theta}$ (the matrix of canonical angles between $\mathbf{U_1}$ and $\hat{\mathbf{U}}_\mathbf{1}$), and likewise the matrix of canonical angles between the subspaces $\mathbf{RV_1}$ and $\mathbf{R\hat{V}_1}$ are identical to $\mathbf{\Psi}$ (the matrix of canonical angles between $\mathbf{V_1}$ and $\hat{\mathbf{V}}_\mathbf{1}$), and thus the maximal angle enjoys the same bound derived above. If we can get a handle on the spectral norm of $\mathbf{E}$, which will come directly from random matrix theory, then we can bound the largest canonical angle of our two subspaces.

We know that $\mathbf{E}$ is a random matrix of i.i.d Gaussian entries with variance $\frac{1}{n}$, and that the largest singular value of a matrix is the spectral norm of the matrix. From random matrix theory we know that the square of the spectral norm of $\mathbf{E}$ is $O(\frac{\sqrt{hv}}{\sqrt{n}})$ (Rudelson & Vershynin, 2010).

The strategy will be to divide the variance in the prediction of $\mathbf{y}$ into two separate parts i.e. the variance that comes from predicting using the incorrect subspace, and then the variance from regression (as stated above) if we had the correct subspace.

Let $\hat{\mathbf{X}} = [\mathbf{L\hat{\Phi}_L} \quad \mathbf{R\hat{\Phi}_R}]$ (the incorrect subspace) and $\mathbf{X} = [\mathbf{L\Phi_L} \quad \mathbf{R\Phi_R}]$ (the true version). To get a handle on predicting with the incorrect subspace (we will consider the subspaces $\mathbf{L\Phi_L}$ and $\mathbf{R\Phi_R}$ separately here, but note that from (1) the angles between the subspaces and their respective perturbed subspaces are bounded by a common bound) we note that, for the regression of $\mathbf{Y}$ on $\mathbf{X}$ we have

$$\beta|\hat{\mathbf{X}} = \frac{\text{Cov}(\mathbf{Y}, \hat{\mathbf{X}})}{\text{Var}(\hat{\mathbf{X}})}$$

and

$$\beta|\mathbf{X} = \frac{\text{Cov}(\mathbf{Y}, \mathbf{X})}{\text{Var}(\mathbf{X})}$$

and

$$\text{Cov}(\mathbf{Y}, \mathbf{X}) = \text{Cov}(\mathbf{Y}, \hat{\mathbf{X}})$$

so trivially

$$\begin{aligned}\beta|\hat{\mathbf{X}} &= \beta|\mathbf{X} * \frac{\text{Var}(\mathbf{X})}{\text{Var}\hat{\mathbf{X}}} \\ &= \beta|\mathbf{X} * \frac{\text{Var}(\mathbf{X})}{\text{Var}(\mathbf{X}) + \text{Var}(\mathbf{X} - \hat{\mathbf{X}})}\end{aligned}$$

Let $\hat{y}$ be the the estimate of y from the true subspace, and $\hat{\hat{y}}$ be the estimate from the perturbed subspace. For the first part of our proof, bounding the error that comes from predicting with the incorrect subspace, we want to bound $\mathbb{E}(\hat{y} - \hat{\hat{y}})^2$.

We have

$$\begin{aligned}\mathbb{E}\left[\hat{y} - \hat{\hat{y}}\right]^2 &= \mathbb{E}\left[\beta|\mathbf{X} * \mathbf{x} - \beta|\hat{\mathbf{X}} * \mathbf{x}\right]^2 \\ &= \mathbb{E}\left[(\beta|\mathbf{X} - \beta|\hat{\mathbf{X}}) * \mathbf{x}\right]^2 \\ &= \mathbb{E}\left[\left(\beta|\mathbf{X} - \beta|\mathbf{X}\frac{\text{Var}(\mathbf{X})}{\text{Var}(\mathbf{X}) + \text{Var}(\mathbf{X} - \hat{\mathbf{X}})}\right) * \mathbf{x}\right]^2 \\ &= \mathbb{E}\left[\beta|\mathbf{X}\left(\mathbf{1} - \frac{\text{Var}(\mathbf{X})}{\text{Var}(\mathbf{X}) + \text{Var}(\mathbf{X} - \hat{\mathbf{X}})}\right) * \mathbf{x}\right]^2 \\ &= \mathbb{E}\left[\beta|\mathbf{X} * \mathbf{x}\left(\frac{\text{Var}(\mathbf{X} - \hat{\mathbf{X}})}{\text{Var}(\mathbf{X}) + \text{Var}(\mathbf{X} - \hat{\mathbf{X}})}\right)\right]^2 \\ &= \mathbb{E}\left[\hat{y} * \left(\frac{\text{Var}(\mathbf{X} - \hat{\mathbf{X}})}{\text{Var}(\mathbf{X}) + \text{Var}(\mathbf{X} - \hat{\mathbf{X}})}\right)\right]^2\end{aligned} \tag{2}$$

Because we are working with a ratio of variances instead of actual variances, then without loss of generality we can set $\text{Var}(\hat{\mathbf{X}}) = 1$ for all predictors.

Now, we don't really care what the exact 'true' $\mathbf{X}$'s are (formed with the true singular vectors), because we only care about predicting y and not actually recovering the true β's associated with our SVD. This means we do not suffer from the usual constraints imposed on the erratic behavior of singular vectors. Usually one must handle this kind of error with respect to the entire subspace since singular vectors are highly unstable. In our case, however, we are free to compare to any 'true' vectors we like from the correct subspace, as long as they span the entire true subspace (and nothing more).

We will define a theoretical set of predictors to compare with, then. We are doing this to obtain an upper bound for the total possible variance of $\text{Var}(x - \hat{x})$ for any acceptable set of x's in the true underlying subspace (where we take acceptable to mean that the x's span the true subspace and nothing more).

We handle each subspace $\mathbf{L}\hat{\mathbf{U}}_\mathbf{1}$ and $\mathbf{R}\hat{\mathbf{V}}_\mathbf{1}$ separately. The construction is to take our first vector and 'choose' a vector from the true subspace that lies such that the angle between the two vectors is the maximal canonical angle between the true and perturbed subspaces.

We proceed to our second predictor and choose a vector from the true subspace such the second 'true' predictor is orthogonal to the first. Note that the angle between our second observed $\hat{x}$ and the second chosen x is at most the maximal canonical angle by assumption. Again, because we don't care about the β's associated with our true singular vectors, but only about prediction quality of our perturbed subspace, we need not be worried that our 'chosen' vectors might not be the true singular vectors. We continue in this manner until we have expired all of our predictors from both sets of spaces.

We know from above that the sine of the maximal angle of of both sets of subspaces is $O\left(\frac{\sqrt{hv}}{\sqrt{n}}\right)$ and so we have that the maximal variation

$$\frac{\text{Var}(\mathbf{X} - \hat{\mathbf{X}})}{\text{Var}(\hat{\mathbf{X}})} \sim O\left(\frac{\sqrt{hv}}{\sqrt{n}}\right)$$

and so from 2 we have

$$\begin{aligned}\mathbb{E}(\hat{y} - \hat{\hat{y}})^2 &= \mathbb{E}\left[\hat{y} * O\left(\frac{\sqrt{hv}}{\sqrt{n}}\right)\right]^2 \\ &\approx O\left(\frac{hv}{n} * \frac{1}{v}\right) = O\left(\frac{h}{n}\right)\end{aligned}$$

We have v of these to predict, so we have a total error attributable to subspace estimation on the order of $\frac{hv}{n}$. Adding regression error as we did earlier (which is on the order of $\frac{kv}{n}$), we get a total error of $\frac{(h+k)v}{n}$. We recall that the error from the one step CCA (OSCCA) is on the order of $\frac{hv^2}{n}$ which yields an error ratio of $\frac{h+k}{hv}$.

References

Glahn, Harry R. Canonical Correlation and Its Relationship to Discriminant Analysis and Multiple Regression. *Journal of the Atmospheric Sciences*, 25 (1):23–31, January 1968.

Rudelson, Mark and Vershynin, Roman. Non-asymptotic theory of random matrices: extreme singular values, 2010. URL http://www.citebase.org/abstract?id=oai:arXiv.org:1003.2990.

Stewart, G. W. Perturbation theory for the singular value decomposition. In *IN SVD AND SIGNAL PROCESSING, II: ALGORITHMS, ANALYSIS AND APPLICATIONS*, pp. 99–109. Elsevier, 1990.

Stewart, G.W. and guang Sun, Ji. *Matrix perturbation theory*. Computer science and scientific computing. Academic Press, 1990. ISBN 9780126702309. URL http://books.google.com/books?id=l78PAQAAMAAJ.