Ling 5801: Lecture Notes 18
Dimensionality Reduction

Contents

Imagine we have the following data for noun phrase head words expanding to modifier words:

$$\mathbf{G} = \begin{array}{c} \\ \text{orzo} \\ \text{penne} \\ \text{ziti} \\ \text{pici} \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} & \text{with} & & \text{in} & & \text{have} \end{array} \\ \begin{pmatrix} \cdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots \\ \cdots & 0 & \cdots & 1 & \cdots & 2 & \cdots \\ \cdots & 1 & \cdots & 2 & \cdots & 3 & \cdots \\ \cdots & 3 & \cdots & 3 & \cdots & 6 & \cdots \\ \cdots & 0 & \cdots & 0 & \cdots & 1 & \cdots \\ \cdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots \end{pmatrix} \end{array}$$

This is pretty sparse, e.g. no instances of *orzo* modified by *with*.

We can generalize over limited data if we blur or 'smooth' it by removing dimensions of variance.

18.1 Center

First we center our data $\mathbf{G} \in \mathbb{R}^{N \times V}$:

$$\mathbf{X} \overset{\text{def}}{=} (\mathbf{G} - \frac{\overbrace{\mathbf{1}\mathbf{1}^\top \mathbf{G}}^{\text{broadcasted means}}}{N})$$

18.2 Best-fit line

Then we find a line $\mathbf{r}_\mathbf{X}^{(I)} \in \mathbb{R}^V$ capturing the most variance in centered data $\mathbf{X}$.

Start with random initial line $\mathbf{r}_\mathbf{X}^{(0)}$, then iteratively project it through variance $\mathbf{X}^\top\mathbf{X}$ and renormalize:

$$\mathbf{r}_\mathbf{X}^{(i)} = \frac{\mathbf{X}^\top\mathbf{X}\,\mathbf{r}_\mathbf{X}^{(i-1)}}{\|\mathbf{X}^\top\mathbf{X}\,\mathbf{r}_\mathbf{X}^{(i-1)}\|_2} \tag{1}$$

(Weight all data points by similarity to $\mathbf{r}^{(i-1)}$, then average coordinates, then move to unit circle.)

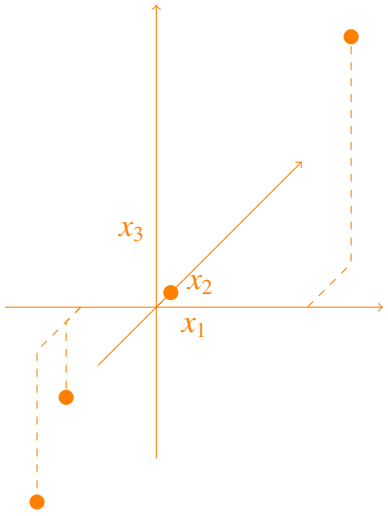
This proceeds until i converges ($i = I$).

For example:

$$\mathbf{X} = \mathbf{G} - \frac{\overbrace{\mathbf{1}^{N\times N}\,\mathbf{G}}^{\text{column means}}}{N} = \begin{bmatrix} -1 & -.5 & -1 \\ 0 & .5 & 0 \\ 2 & 1.5 & 3 \\ -1 & -1.5 & -2 \end{bmatrix} \quad \text{(centered)}$$

$$\mathbf{r}^{(0)} = \begin{bmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix}$$

$$\mathbf{X}^\top\mathbf{X}\,\mathbf{r}^{(0)} = \begin{bmatrix} -1 & 0 & 2 & -1 \\ -.5 & .5 & 1.5 & -1.5 \\ -1 & 0 & 3 & -2 \end{bmatrix} \begin{bmatrix} -1 & -.5 & -1 \\ 0 & .5 & 0 \\ 2 & 1.5 & 3 \\ -1 & -1.5 & -2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix} = \begin{bmatrix} \frac{20}{\sqrt{3}} \\ \frac{18}{\sqrt{3}} \\ \frac{31}{\sqrt{3}} \end{bmatrix}$$

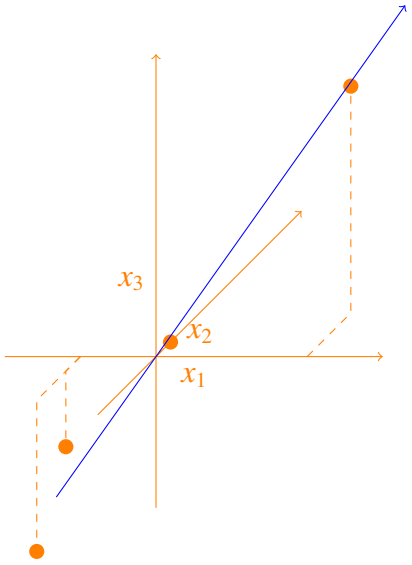


$$\mathbf{r}^{(1)} = \begin{bmatrix} 0.48722554 \\ 0.43850298 \\ 0.75519958 \end{bmatrix}$$

$$\mathbf{r}^{(2)} = \begin{bmatrix} 0.48765374 \\ 0.43679415 \\ 0.75591316 \end{bmatrix}$$

$$\mathbf{r}^{(3)} = \begin{bmatrix} 0.48767114 \\ 0.4367649 \\ 0.75591884 \end{bmatrix}$$

$$\mathbf{r}^{(4)} = \begin{bmatrix} 0.48767151 \\ 0.43676433 \\ 0.75591892 \end{bmatrix}$$



18.3 Principal Components Analysis

Next we collapse the space of the data along this line $\mathbf{r}$ of greatest variance.

Done by projecting remaining variance $\mathbf{X}^{(k-1)}$ onto $\mathbf{r}$, then back using $\mathbf{r}^\top$, and subtracting from $\mathbf{X}^{(k-1)}$.

Each time we do this makes a simpler, lower-dimensional space $\mathbf{X}^{(k)}$ of the remaining variance:

$$\mathbf{X}^{(0)} = \mathbf{X} \tag{2}$$

$$\mathbf{X}^{(k)} = \mathbf{X}^{(k-1)} - \mathbf{X}^{(k-1)}\mathbf{r}^{(k)(I)}\mathbf{r}^{(k)(I)\top} \tag{3}$$

where $\mathbf{r}^{(k)(i)}$ is a linear regression on $\mathbf{X}^{(k-1)}$, as before:

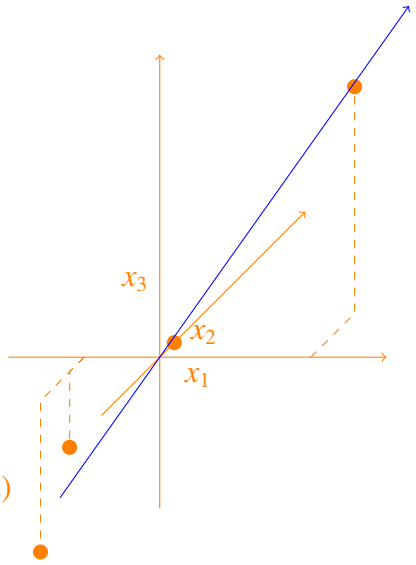
$$\mathbf{r}^{(k)(i)} = \frac{\mathbf{X}^{(k-1)\top}\mathbf{X}^{(k-1)}\mathbf{r}^{(k)(i-1)}}{||\mathbf{X}^{(k-1)\top}\mathbf{X}^{(k-1)}\mathbf{r}^{(k)(i-1)}||_2} \tag{4}$$

We keep doing this until we have a set of K lines (principal components) that approximate the data.

For example:

$$\mathbf{G} = \begin{array}{c} \\ \text{orzo} \\ \text{penne} \\ \text{ziti} \\ \text{pici} \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} \cdots & \text{with} & \cdots & \text{in} & \cdots & \text{have} & \cdots \end{array} \\ \begin{pmatrix} \cdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots \\ \cdots & 0 & \cdots & 1 & \cdots & 2 & \cdots \\ \cdots & 1 & \cdots & 2 & \cdots & 3 & \cdots \\ \cdots & 3 & \cdots & 3 & \cdots & 6 & \cdots \\ \cdots & 0 & \cdots & 0 & \cdots & 1 & \cdots \\ \cdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots \end{pmatrix} \end{array}$$

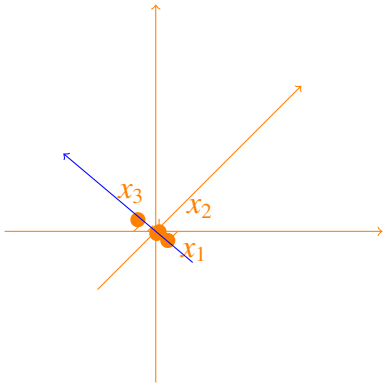
$$\mathbf{X}^{(0)} = \mathbf{G} - \frac{\mathbf{1}^{N\times N}\,\mathbf{G}}{N} = \begin{bmatrix} -1 & -.5 & -1 \\ 0 & .5 & 0 \\ 2 & 1.5 & 3 \\ -1 & -1.5 & -2 \end{bmatrix} \quad \text{(centered)}$$



$$\mathbf{r}^{(1)(I)} = \begin{bmatrix} .49 \\ .44 \\ .76 \end{bmatrix}$$

Now let's add another component:

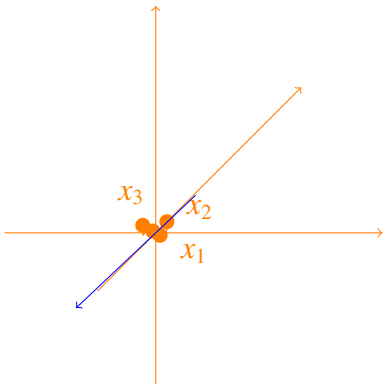
$$\mathbf{X}^{(1)} = \begin{bmatrix} -0.2870376 & 0.13853747 & 0.10513275 \\ -0.10649876 & 0.40461846 & -0.16507921 \\ 0.0989363 & -0.20261489 & 0.05324187 \\ 0.29460005 & -0.34054104 & 0.00670458 \end{bmatrix}$$



$$\mathbf{r}^{(2)(I)} = \begin{bmatrix} -.56 \\ .82 \\ -.11 \end{bmatrix}$$

Now we could add another component (but this wouldn't be *reduced* anymore):

$$\mathbf{X}^{(2)} = \begin{bmatrix} -0.14021386 & -0.07691476 & 0.13489797 \\ 0.12318608 & 0.06757412 & -0.11851576 \\ -0.0284893 & -0.01562789 & 0.02740919 \\ 0.04551707 & 0.02496854 & -0.0437914 \end{bmatrix}$$



$$\mathbf{r}^{(3)(I)} = \begin{bmatrix} .67 \\ .37 \\ -.64 \end{bmatrix}$$

Now define a 'smoothed' matrix $\hat{\mathbf{X}} \in \mathbb{R}^{N \times D}$ by projecting $\mathbf{X}$ into this reduced space, then back:

$$\hat{\mathbf{X}} = \underbrace{\mathbf{X}\left[\mathbf{r}^{(1)} \cdots \mathbf{r}^{(K)}\right]}_{\text{data points in } K\text{-space}} \begin{bmatrix} \mathbf{r}^{(1)\top} \\ \vdots \\ \mathbf{r}^{(K)\top} \end{bmatrix} \tag{5}$$

Then un-center it to get $\hat{\mathbf{G}}$ – a 'smoothed' version of $\mathbf{G}$:

$$\hat{\mathbf{G}} = \hat{\mathbf{X}} + \frac{\mathbf{1}^{N \times N}\,\mathbf{G}}{N} \tag{6}$$

Here's what the reconstruction looks like using the first two principal components:

$$\hat{\mathbf{G}} = \begin{bmatrix} -1 & -.5 & -1 \\ 0 & .5 & 0 \\ 2 & 1.5 & 3 \\ -1 & -1.5 & -2 \end{bmatrix} \begin{bmatrix} .49 & -.56 \\ .44 & .82 \\ .76 & -.11 \end{bmatrix} \begin{bmatrix} .49 & .44 & .76 \\ -.56 & .82 & -.11 \end{bmatrix} + \begin{bmatrix} 1 & 1.5 & 3 \\ 1 & 1.5 & 3 \\ 1 & 1.5 & 3 \\ 1 & 1.5 & 3 \end{bmatrix}$$

$$= \begin{array}{r} \\ \text{orzo} \\ \text{penne} \\ \text{ziti} \\ \text{pici} \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccc} \cdots & \text{with} & \cdots & \text{in} & \cdots & \text{have} & \cdots \end{array} \\ \begin{pmatrix} \cdots & 0.13 & \cdots & 1.07 & \cdots & 1.85 & \cdots \\ \cdots & 0.87 & \cdots & 1.93 & \cdots & 3.12 & \cdots \\ \cdots & 3.04 & \cdots & 3.04 & \cdots & 6.00 & \cdots \\ \cdots & -0.05 & \cdots & -0.04 & \cdots & 1.02 & \cdots \\ \cdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots \end{pmatrix} \end{array}$$

That solved our zero-count problem for *orzo*!

Not so much for *pici* though (it has negative counts!)... That's a problem with linear regression.

We might fix this by *not* centering first, or by using other techniques, like neural nets (next course)!

Reduced dimensionality vectors are also associated with words ('word embeddings').

- Data dimensionality D is very large, e.g. set of co-occurring words at various offset distances.
- Reduced dimensionality K is usually about 100 to 1000.
- Dimensionality reduction uses recurrent neural networks.