Ling 5702: Lecture Notes 7
Dimensionality Reduction

Neurons or neural clusters do not correspond to labeled features like 'is married' or 'has a dog.' There are an unbounded number of such features, and only a few billion neurons (and only a few thousand cortical columns, if these are our units!). Properties such as these must therefore be *distributed* over a *reduced dimensional space* of representational units. This section will describe of model of dimensionality reduction known as Principal Components Analysis (PCA). We could use it to build a tractable neural-like model of linguistic semantic representations.

7.1 Linear Regression

First, we can find a line that captures most of the variance in a set of data using linear regression. Like the logistic regression described in the previous lecture notes, linear regression can be performed iteratively, starting with iteration $i = 0$ with random coefficients and then repeatedly refining it in the direction of greatest variance, and then normalizing the resulting vector using the two-norm:

$$r^{(0)} \in \mathbb{R}^L \tag{1}$$

$$r^{(i)} = \frac{X^\top X r^{(i-1)}}{\|X^\top X r^{(i-1)}\|_2} \tag{2}$$

where the vector $X^\top X r$ multiplies a set of (previous regression) line coefficients by the variance in X in a process similar to logistic regression, but without the logistic (σ) function:

$$(X^\top X r)_l = \sum_j \left(\sum_{l'} X_{j,l'} \cdot r_{l'} \right) \cdot X_{j,l} \tag{3}$$

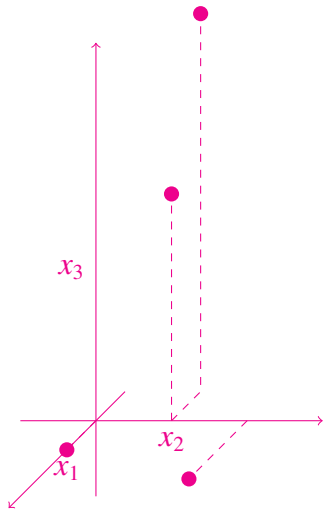
This proceeds until i converges ($i = I$).

For example:

$$X = \begin{bmatrix} -1 & 1 & 5 \\ 0 & 1 & 3 \\ 1 & 0 & 0 \\ 2 & 2 & 0 \end{bmatrix}$$

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

$$X^\top X r^{(0)} = \begin{bmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 0 & 2 \\ 5 & 3 & 0 & 0 \end{bmatrix} \begin{bmatrix} -1 & 1 & 5 \\ 0 & 1 & 3 \\ 1 & 0 & 0 \\ 2 & 2 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix} = \begin{bmatrix} \frac{4}{\sqrt{3}} \\ \frac{17}{\sqrt{3}} \\ \frac{37}{\sqrt{3}} \end{bmatrix}$$



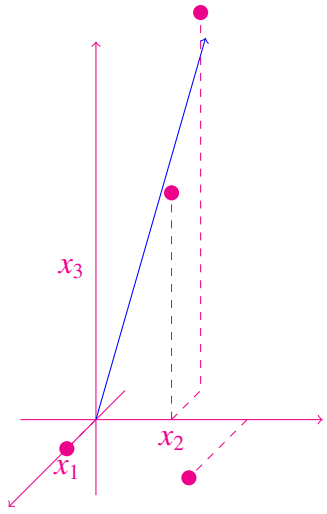
$$r^{(1)} = \begin{bmatrix} \frac{4}{\sqrt{1674}} \\ \frac{17}{\sqrt{1674}} \\ \frac{37}{\sqrt{1674}} \end{bmatrix}$$

$$r^{(2)} = \begin{bmatrix} 0.09776474 \\ 0.41550016 \\ 0.90432388 \end{bmatrix}$$

$$r^{(3)} = \begin{bmatrix} -0.07649084 \\ 0.28510223 \\ 0.95544015 \end{bmatrix}$$

$$r^{(4)} = \begin{bmatrix} -0.11977154 \\ 0.249467 \\ 0.96094797 \end{bmatrix}$$

$$r^{(5)} = \begin{bmatrix} -0.13022705 \\ 0.24068495 \\ 0.96182726 \end{bmatrix}$$



7.2 Principal Components Analysis

Principal Components Analysis (PCA) finds a reduced K-dimensional space (or 'hyper-plane') of unlabeled features or *components*, suspended in some original L-dimensional space (or 'hyper-space') of labeled features, which captures as much of the variation as possible of some initial set $X^{(0)}$ of J data points in this L-dimensional space. It does this by repeatedly subtracting out dimensional components of data points that lie along optimal linear regression lines $r^{(k)}$ which each capture as much of the variance as possible of the remaining effectively $(L-k)$-dimensional space $X^{(k-1)}$.

Each principal component k can be found by performing a linear regression on the data points in a partially-reduced (effectively $(L-(k-1))$-dimensional) space. Like the logistic regression described in the previous lecture notes, each of these k linear regressions can be performed iteratively, starting the iteration at $i = 0$ with random coefficients and then repeatedly refining it in the direction of greatest variance, and then normalizing the resulting vector using the two-norm:

$$r^{(k)(0)} \in \mathbb{R}^L \tag{4}$$

$$r^{(k)(i)} = \frac{X^{(k)\top} X^{(k)} r^{(k)(i-1)}}{\|X^{(k)\top} X^{(k)} r^{(k)(i-1)}\|_2} \tag{5}$$

where the vector $X^\top Xr$ multiplies a set of (previous regression) line coefficients by the variance in X in a process similar to logistic regression, but without the logistic (σ) function:

$$(X^\top Xr)_l = \sum_j \left(\sum_{l'} X_{j,l'} \cdot r_{l'} \right) \cdot X_{j,l} \tag{6}$$

This proceeds until i converges ($i = I$). The dimensional component defined by each linear regression is then subtracted from the L-dimensional vector of each data point in each partially

dimensionality-reduced data set $X^{(k-1)}$, starting with the original data set X at $k = 0$, further flattening each such data set into an even lower-dimensional (effectively $(L-k)$-dimensional) hyperplane $X^{(k)}$ which is orthogonal to (at right angles to) the regression line $r^{(k)}$:

$$X^{(0)} = X \tag{7}$$

$$X^{(k)} = X^{(k-1)} - X^{(k-1)} r^{(k)(I)} r^{(k)(I)\top} \tag{8}$$

where the matrix $Xrr^\top$ projects data points in X orthogonally onto a (normalized) regression line r:

$$(Xrr^\top)_{j,l} = \left(\sum_{l'} X_{j,l'} \cdot r_{l'} \right) \cdot r_l \tag{9}$$

This proceeds until a maximal reduced dimensionality is reached ($k = K$).

An explicit K-dimensional translation Y of the original L-dimensional data set X can now be obtained by multiplying X by a matrix consisting of these regression vectors concatenated together as columns:

$$Y = X \left[r^{(1)} \cdots r^{(K)} \right] \tag{10}$$

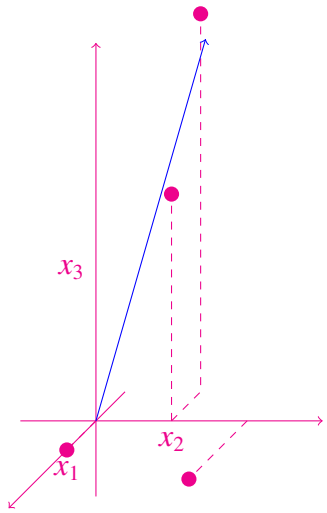
This K-dimensional data set Y can then be translated back into (a noisy copy of) X by multiplying Y by a matrix consisting of these regression vectors concatenated together as rows:

$$X \approx Y \begin{bmatrix} r^{(1)\top} \\ \vdots \\ r^{(K)\top} \end{bmatrix} \tag{11}$$

For example:

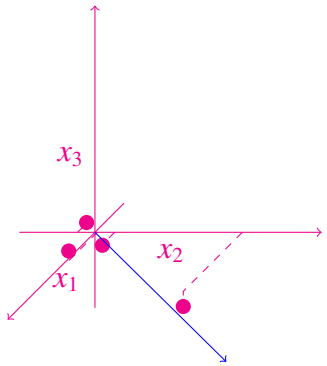
$$X^{(0)} = \begin{bmatrix} -1 & 1 & 5 \\ 0 & 1 & 3 \\ 1 & 0 & 0 \\ 2 & 2 & 0 \end{bmatrix}$$

$$r^{(1)(I)} = \begin{bmatrix} -0.13356538 \\ 0.23786699 \\ 0.96207047 \end{bmatrix}$$



$$X^{(1)} = \begin{bmatrix} -0.308 & -0.233 & 0.015 \\ 0.417 & 0.257 & -0.006 \\ 0.982 & 0.032 & 0.128 \\ 2.028 & 1.950 & -0.2001 \end{bmatrix}$$

$$r^{(2)(I)} = \begin{bmatrix} 0.76267961 \\ 0.64456156 \\ -0.05348083 \end{bmatrix}$$



$$X^{(2)} = \begin{bmatrix} -0.014 & 0.016 & -0.006 \\ 0.048 & -0.055 & 0.020 \\ 0.400 & -0.460 & 0.169 \\ -0.119 & 0.136 & -0.050 \end{bmatrix}$$

$$r^{(3)(I)} = \begin{bmatrix} 0.63283497 \\ -0.72660834 \\ 0.26750742 \end{bmatrix}$$