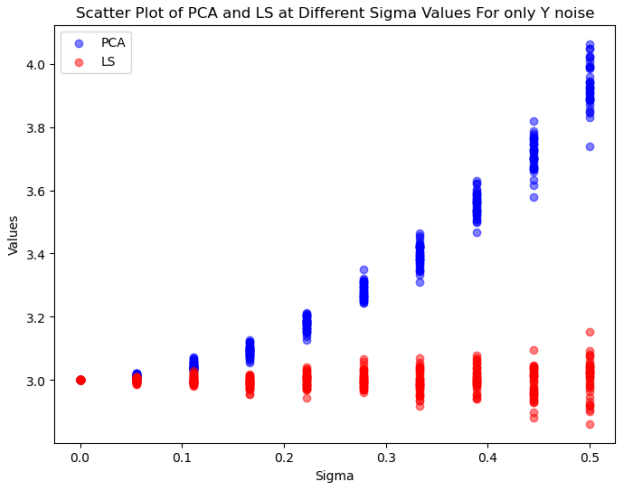
CSE Template

Rohan Mukherjee

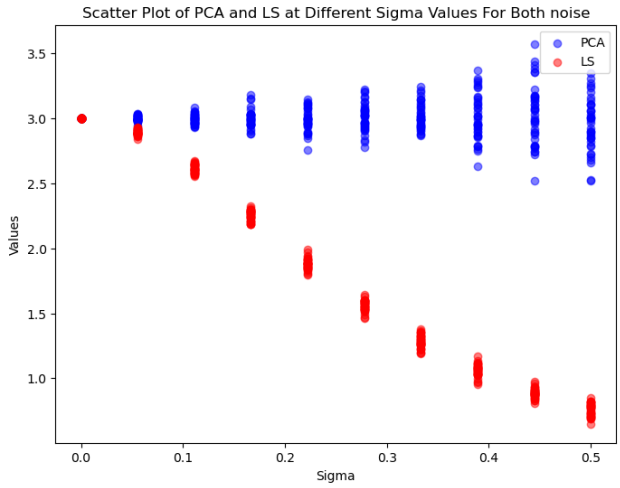
February 5, 2025

1. (a) For a bunch of different values of n and σ, LS usually returns something around ± 0.02, and PCA returns something around ± 1. For true independence, you would expect $E((X - E(X))(Y - E(Y))) = 0$, and as that is what LS returns, that is what you would get normally. On the other hand, if we write our data matrix $Z = \begin{pmatrix} x & y \end{pmatrix}$, then $Z^T Z = \begin{pmatrix} x^T x & x^T y \\ x^T y & y^T y \end{pmatrix}$, and since $x^T x, y^T y \approx \sigma^2$, while $x^T y \approx 0$, we get the eigenvectors of this matrix that PCA returns is just: $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Since they have the same eigenvalue, PCA might return either of them, so we should get ± 1. However, I am not fully sure why it chooses these eigenvectors instead of the standard basis ones, I think its probably an artifact of how the algorithm is initialized (presumably, you would only have to do one iteration of the power method).

   (b) Here is the picture of that or only y noise:



(c) This is the scatter plot I got for both x and y noise:



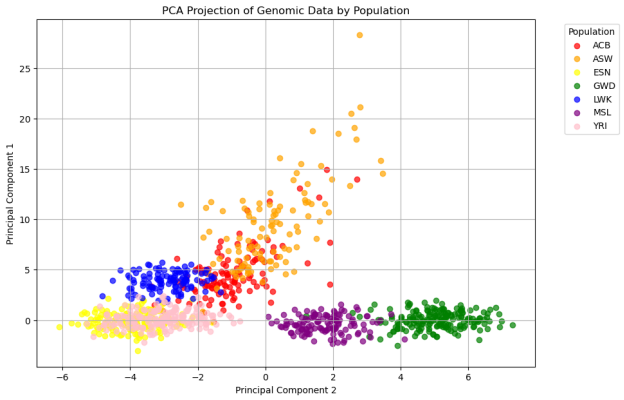
(d) From the matrix before, if $y = 3x$, we should $Z^T Z = \|x\|^2 \begin{pmatrix} 1 & 3 \\ 3 & 9 \end{pmatrix}$ which (after removing the $\|x\|^2$), has largest eigenvalue of 10 with corresponding eigenvector $\begin{pmatrix} 1 \\ 3 \end{pmatrix}$. However, if there is noise on only the y, with $y = 3x = \varepsilon$ where $\varepsilon \sim N(0, \sigma^2 I_n)$, then $x^T x$, stays the same while $x^T y = x^T(y + \varepsilon) = x^T y + x^T \varepsilon = 3x^T x + x^T \varepsilon$, where $x^T \varepsilon$is a $N(0, \sum x_i^2 \sigma^2)$ random variable. For large σ, this should be $3x^T x$ but in reality can be very different from it, and the last coordinate $y^T y = 9x^T x + 6x^T \varepsilon + \varepsilon^T \varepsilon$ also has giant variance. So the direction of PCA can be very spread out, which is what we observe for large σ.

However, if there is noise on both the x and y, recall that PCA is trying to find the direcdtion of maximum variance. Since we are making the noise in both x and y independently with same variance, the noise in the x direction and y direction equals proportionally. So, from the matrix, we would get something like: $\begin{pmatrix} x^T x + 2x^T \varepsilon_x + \varepsilon_x^T \varepsilon_x & 3x^T x + x^T \varepsilon_x \\ 3x^T x + x^T \varepsilon_y & 9x^T x + 6x^T \varepsilon_y + \varepsilon_y^T \varepsilon_y \end{pmatrix}$ The second order error terms become small, and now as we have equal spread in the x and y directions, PCA will be robust and return the correct eigenvalue of 3. Before, we only have spread in the y direction which confused the PCA.

For noise on both x and y, we are outputting for LS: $(x + \varepsilon_x - \mu_x \mathbf{1})^T (3x + \varepsilon_y - \mu_y \mathbf{1}) / \|x + \varepsilon_x - \mu_x \mathbf{1}\|$, since the mean of ε_x and ε_y are both 0. This gives has an error term of $\varepsilon_x^T (3x) + x^T \varepsilon_y$, which has a giant variance, being normally distributed with mean 0 and variance $\sigma^2 \sum_i (4x_i)^2$. This is like $4 \cdot 333$ times σ^2 in variance. So LS gets very confused and is super susceptible to noise, which gets crushed in this case. The error

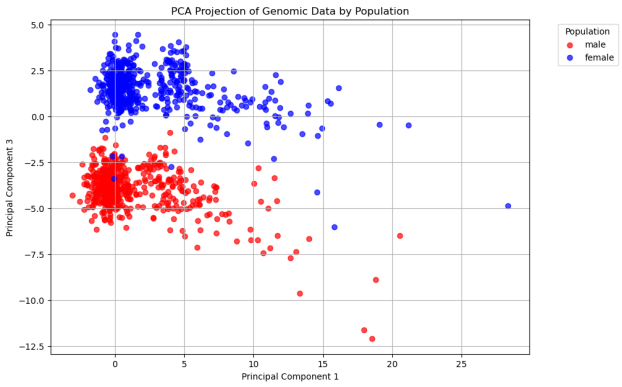
term $\varepsilon_x^T(3x)$ is dominant, and was not present before where LS worked well. It causes a lot of variance which didn't happen before.

2. (a) Here is the plot of the first 2 PCs:



(b) MSL and GWD, being Mende, Sierra Leone and Gambian, Western Division, Mandinka are extremely close geographically, while still being on the African continent. ESN and YRI, being Yoruba in Ibadan, Nigera, and Esan in Nigera, are also clearly both in Nigeria, which explains why they are close in the PCA. Nigeria is very far east from Sierra Leone and Gambia, which explains the distinct spread between the two groups. The blue group, being noticeably above ESN and YRI can be explained since it means kenya, which is far to the east of Nigeria. It is extremely far from the green and purple group because of the large geographical gap between west and east Africa. The orange and red group, being ACB and ASW, are African Caribbeans in Barbados and Americans of African ancestry are both located in the Americas, namely closer to North America, extremely far away from Africa. This is why there is such a big spread between them and everything else. So it seems that the second PC represents the east/west divide in Africa, and the first represents the America/Africa geographical divide.

(c) Here is the plot of the first vs 3rd PCs:



(d) I mean this is as clear as day. The third PC is repesentative of if the sample was a man or a woman.

(e) Let $x \in \mathbb{R}^n$ be a vector to be chosen later, and $X \in \mathbb{R}^{m \times n}$ be our data matrix, and consider

$$\tilde{X} = \begin{pmatrix} X \\ x^T \end{pmatrix}$$

Then,

$$\tilde{X}^T \tilde{X} = \begin{pmatrix} X^T & x \end{pmatrix} \begin{pmatrix} X \\ x^T \end{pmatrix} = X^T X + x x^T$$

Using eigendecomposition of $X^T X$, which is the right singular vectors, we get $X^T X = \sum_i \sigma_i^2 v_i v_i^T$. We have been given that the first 3 are unique. If we now strategically choose $x = \sqrt{\sigma_1^2 - \sigma_2^2} v_2$, we can see that $v_1 v_1^T$ and $v_2 v_2^T$ now have the same coefficients and hence the first principal component is no longer unique.

Similarly, if the first $k+1$ are unique, we can choose $x = \sqrt{\sigma_k^2 - \sigma_{k+1}^2} v_{k+1}$ and make the kth PC not unique.