Dimensionality Reduction Using Principal Components Analysis

University of Nairobi
School of Mathematics
Dr. James Katende

June 11, 2020

Dimensionality Reduction

- We have a data set X with m rows (cases) and n measurements (or variables). $(m \geq n)$
- We wish to visualise the data but n is large and we can't plot the data in 2D or 3D
- We can throw away some variables but we may be losing important information
- How do we work around this?
- We transform the data into a new set of uncorrelated variables which can be ranked in order of importance.
- The technique used is called principal components analysis

Principal Components Analysis

- $T = XV$ is the magic transformation
- V is an $n \times n$ orthogonal matrix and its coloumns are the eigen vectors of the matrix X^tX
- X^tX is the covariance matrix assuming X is centered.
- T is exactly the same size as X and the entries of T are called the scores.

$$X^tX = VDV^t \tag{1}$$

  Where D is a diagonal matrix and the entries in the main diagonal are the eigenvalues of the matrix X^tX in descending order

- We label the eigenvalues of X^tX as $\lambda_1, \lambda_2 \ldots, \lambda_n$ with $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n$ and corresponding eigenvectors $v_1, v_2, \ldots, v_n$
- The eigen vectors $v_1, v_2, \ldots, v_n$ are called principal components and its entries are called loadings.

Singular Value Decomposition

- There is a more computationally efficient way to obtain T
- $X = U\Sigma V^t$, U is an $m \times m$ matrix, V is an $n \times n$,matrix, and Σ is an $m \times n$ matrix with an $n \times n$ diagonal block matrix at the top.
- The entries in the diagonal block are $\sigma_1 = \sqrt{\lambda_1}$,$\sigma_2 = \sqrt{\lambda_2}$, ...,$\sigma_n = \sqrt{\lambda_n}$ and are called the singular values of X.
- And the vectors $v_1, v_2, \ldots, v_n$ are the coloumns of the matrix V and are called the right singular vectors and $\sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_n$.
- U and V are orthogonal matrices $(V^{-1} = V^t, U^{-1} = U^t)$

$$T = XV = U\Sigma. \tag{2}$$

Reducing the Dimension

- Since the singular vectors are in some order of importance we can drop a few and know exactly how much information is retained.
- We only retain the ones corresponding to the large singular values.
- If we pick only the first two eigen vectors then we have

$$T_2 = XV_2 \tag{3}$$

  where V_2 contains the first two coloumns of V and now T_2 is an $m \times 2$ matrix with only two coloumns (transformed measurments) instead of the original n measurments.
- We still have all the m cases but only two variables.
- We can then plot these two variables in $2D$ and be able to visualize the data and retain a reasonable amount of information

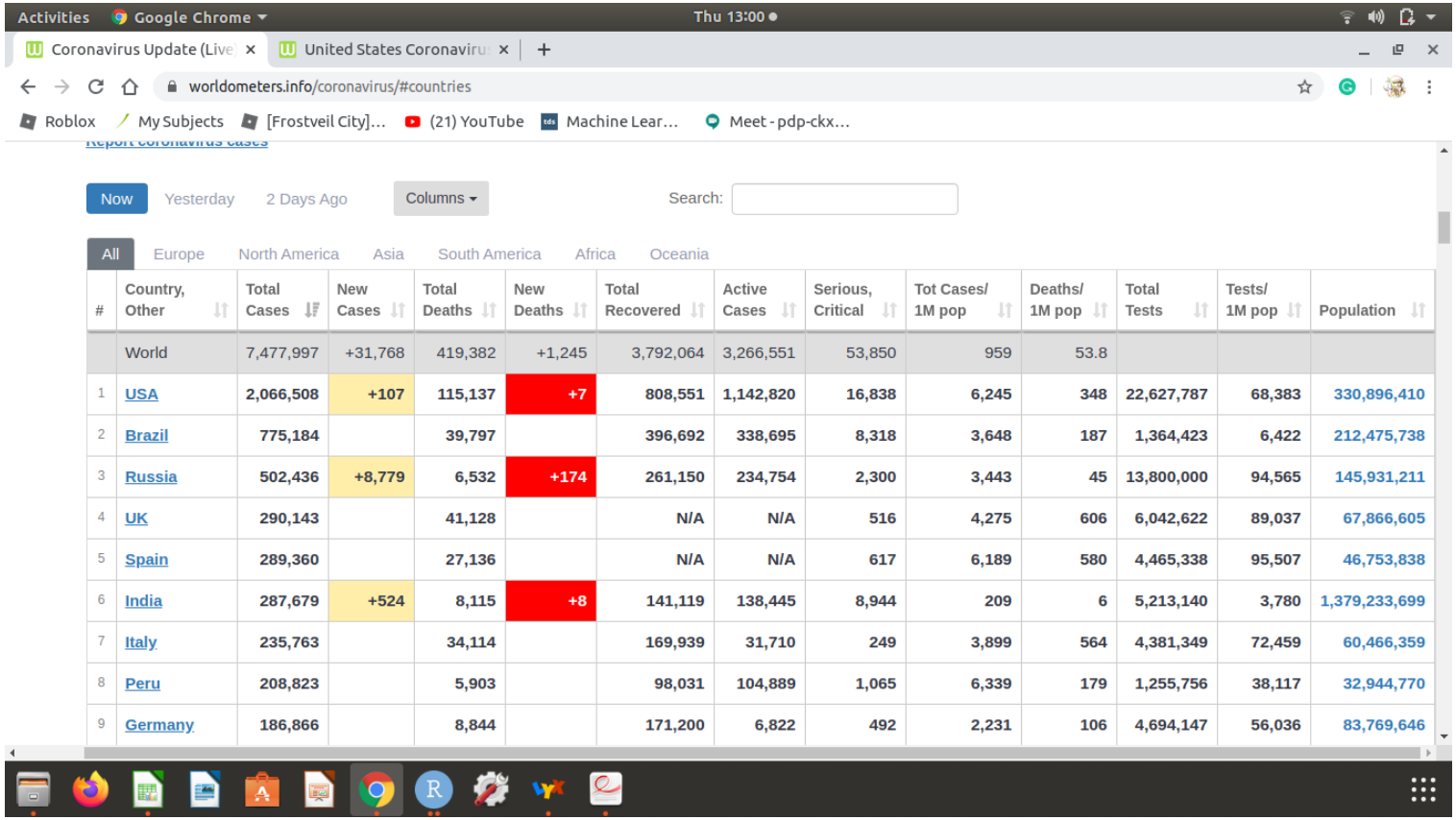
How Much Information is Retained?

- If we choose to retain a total of r principal components, then the information retained is given by

$$\frac{\sigma_1 + \sigma_2 + \cdots + \sigma_r}{\sigma_1 + \sigma_2 + \cdots + \sigma_n} \times 100 \tag{4}$$

- 95% is good but we are often able to get good insights with lower values
- Just use your judgement since this is mostly an exploratory venture.
- Many times we are able to work with only two principal components with fairly good results.
- With only two components we can graph the data and obtain a good visualization of the data.

Corona Virus Data



#	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Population
	World	7,477,997	+31,768	419,382	+1,245	3,792,064	3,266,551	53,850	959	53.8			
1	USA	2,066,508	+107	115,137	+7	808,551	1,142,820	16,838	6,245	348	22,627,787	68,383	330,896,410
2	Brazil	775,184		39,797		396,692	338,695	8,318	3,648	187	1,364,423	6,422	212,475,738
3	Russia	502,436	+8,779	6,532	+174	261,150	234,754	2,300	3,443	45	13,800,000	94,565	145,931,211
4	UK	290,143		41,128		N/A	N/A	516	4,275	606	6,042,622	89,037	67,866,605
5	Spain	289,360		27,136		N/A	N/A	617	6,189	580	4,465,338	95,507	46,753,838
6	India	287,679	+524	8,115	+8	141,119	138,445	8,944	209	6	5,213,140	3,780	1,379,233,699
7	Italy	235,763		34,114		169,939	31,710	249	3,899	564	4,381,349	72,459	60,466,359
8	Peru	208,823		5,903		98,031	104,889	1,065	6,339	179	1,255,756	38,117	32,944,770
9	Germany	186,866		8,844		171,200	6,822	492	2,231	106	4,694,147	56,036	83,769,646

Figure: Source: https://www.worldometers.info/coronavirus/#countries

Acquiring and Analysing the Data

- Obtained the data from https://www.worldometers.info/coronavirus/#countries using Libreoffice calc
- Did some processing in Libreoffice Calc
- Performed the analysis using R software.
- Libreoffice and R are open source softwares and can be downloaded for free.
- Search youtube for videos on how to download and install the software.
- We now go to a live demo of the Data Analysis.