STAT 215A Fall 2019 Week 4

Tiffany Tang

9/20/19

Announcements

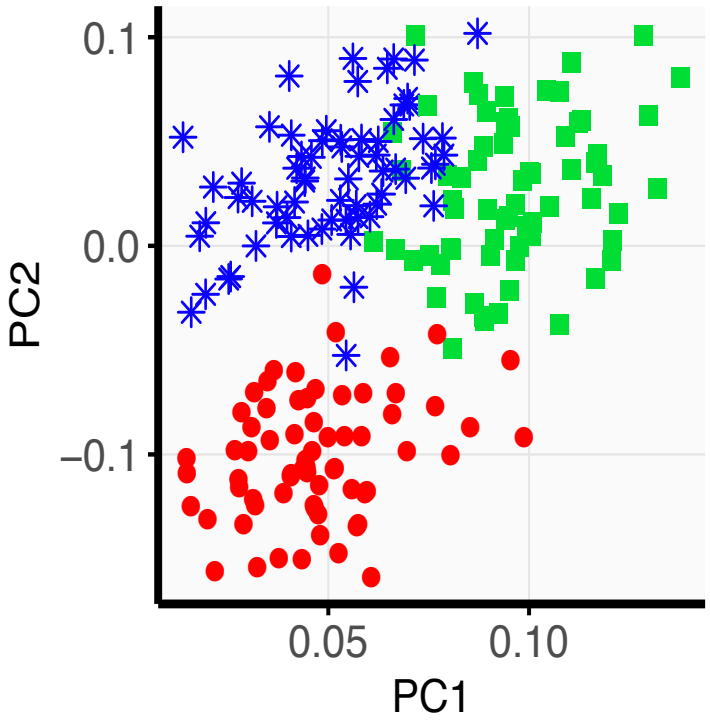
- Congrats on finishing lab 1!!!!
- I will send out instructions on how to do peer reviews later today
  - Completed peer reviews due in one week at **11:59pm Thursday Sept 26**
- Lab 2 + Homework 2 will be released next Friday

Plan for Today:

- PCS Documentation
- Brief review of PCA
- Alternatives to PCA
- End early due to climate strike

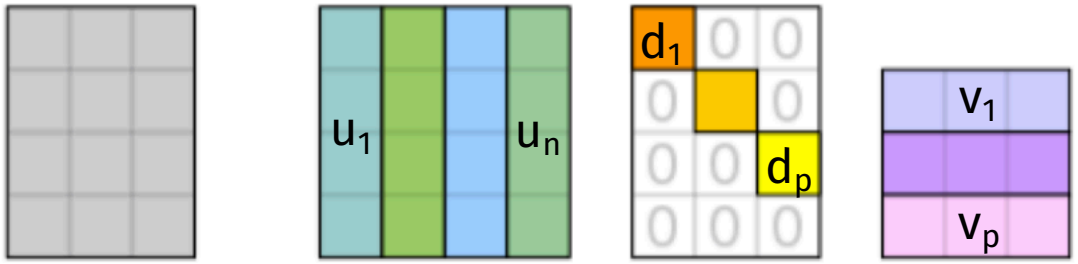
PCS Documentation

- Karl Kumbier, Bin Yu - Three principles of data science: predictability, computability, and stability (PCS)
- https://zenodo.org/record/1456199#.XYRzWJNKgWp
- Make your analysis transparent!
- Encourages reproducible research



Review of PCA

SVD



$$\mathbf{X} = \mathbf{U}\ \mathbf{D}\ \mathbf{V}^T$$

n x p n x n n x p p x p

$$d_1 \geq \cdots \geq d_p$$

$$\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}_{nxn}$$

$$\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_{pxp}$$

In R: svd()

PCA

PC directions: dominant feature patterns

$$\mathbf{v}_j = \underset{\mathbf{v}\in\mathbb{R}^p}{\operatorname{argmax}}\ \mathbf{v}^\top \mathbf{X}^\top \mathbf{X}\,\mathbf{v} \qquad \text{subject to} \quad \|\mathbf{v}\|_2^2 = 1, \quad \mathbf{v}^\top \mathbf{v}_i = 0 \ \ \forall\, i < j.$$

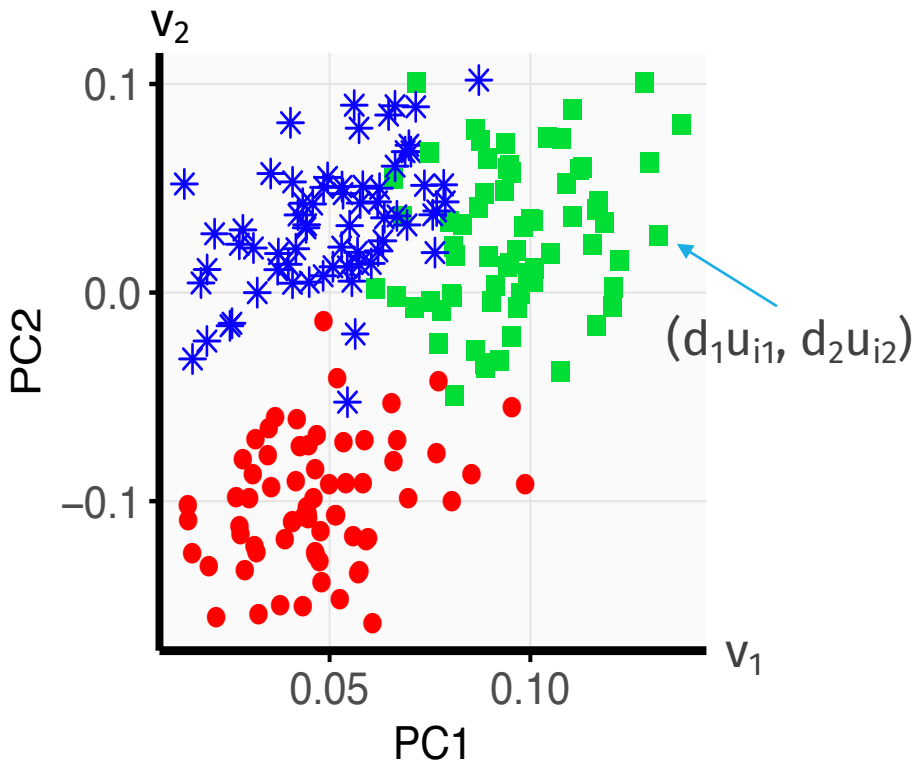
PC scores: dominant observation patterns

$$d_j\,\mathbf{u}_j = \mathbf{X}\,\mathbf{v}_j$$

(projection of data onto directions of maximizing variance)

Proportion of Variance Explained:

$$\frac{\mathbf{v}_j^\top \mathbf{X}^\top \mathbf{X}\,\mathbf{v}_j}{\operatorname{tr}(\mathbf{X}^\top \mathbf{X})} = \frac{d_j^2}{\sum_{i=1}^p d_i^2}$$



Practical Considerations with PCA

- PCA is optimal with Gaussian data, but can also work with non-Gaussian data in practice (but not always)
- What to do with categorical data?
  - One-hot encoding
- Only need to run PCA once to get all orthogonal, nested components

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

Image source: https://www.kaggle.com/ayushkaul/using-categorical-data-with-one-hot-encoding

Other Alternatives

- Modifications of PCA:
  - **Sparse PCA:** sparse, interpretable PCs
  - **Kernel PCA:** want non-linear PCs
  - **Functional PCA:** for functional/time series data
  - **Robust PCA:** for grossly corrupted observations
  - Downside: requires additional tuning parameters, which are difficult to tune
- Other methods for dimensionality reduction and pattern recognition
  - **NMF:** want non-negative components (e.g., bag of words example)
  - **t-SNE:** searches for a low-dimensional *manifold* representation (typically <4 dimensions)
  - Downside: non-nested components and non-unique solution