Common Principal Component Analysis

Benjamin Draves

October 22, 2020

Motivation - Multiplex Networks

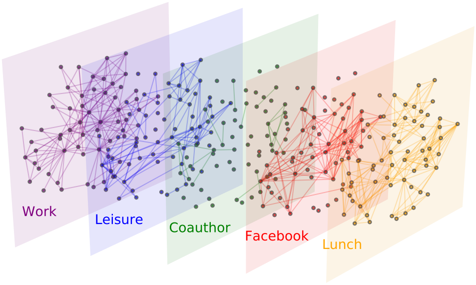


Figure: Multiplex Network of Aarhus Computer Science Department. Vertices are members of the department and each layer encodes a different type of interaction.[8]

- Multiplex networks - set of networks over a common vertex set.
- Adjacency matrices $A^{(g)} \in [0,1]^{n\times n}$ often low effective rank.
- Simultaneous dimensionality reduction of $\{A^{(g)}\}_{g=1}^m$.

Outline of Talk

1. Principal Component Analysis (PCA) Overview
   - Motivating example: Palmer Penguins
   - Derivation of principal components
   - Computational details and connection to SVD
2. Common Principal Component Analysis (CPCA) Overview
   - Common Principal Components definition
   - MLE & Spectral approaches to estimation
   - Computational details and connection to SVD
3. Randomized Algorithms for Truncated Singular Value Decompositions
   - Algorithm Sketch
   - Theoretical Performance
   - Application to Palmer Penguins

Palmer Penguins [5]

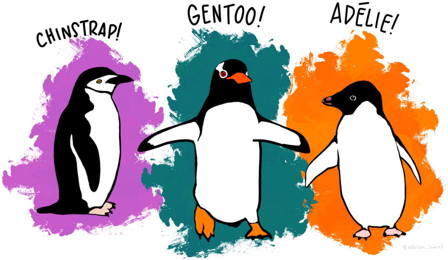


Figure: Artwork by @allison_horst.

- Four continuous variables
  - Flipper length
  - Bill length
  - Bill depth
  - Body mass

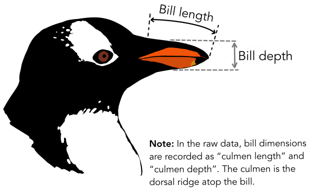
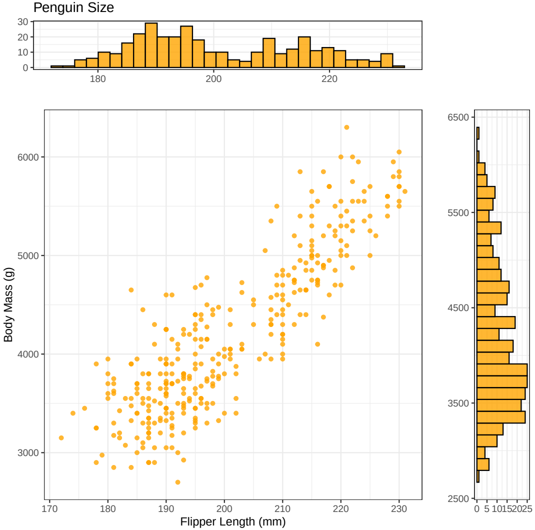
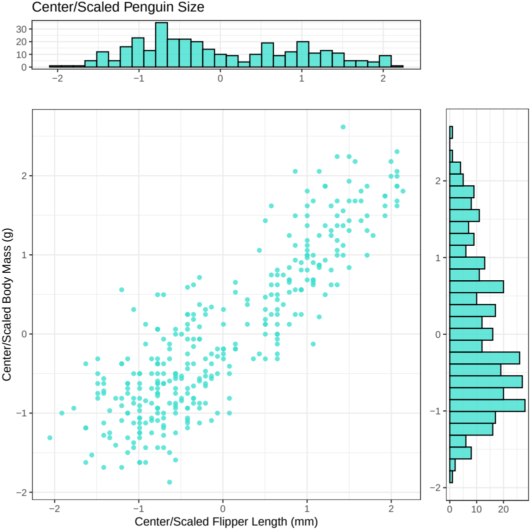


Figure: Artwork by @allison_horst.

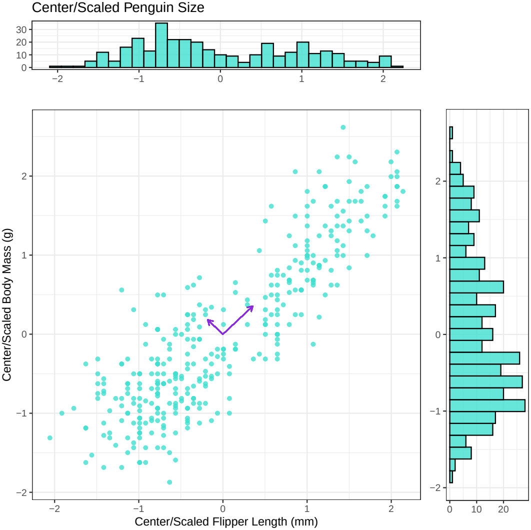
Dimensionality Reduction and PCA



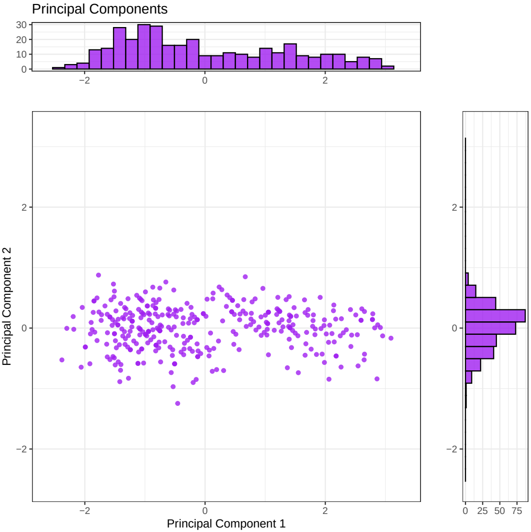
Dimensionality Reduction and PCA



Dimensionality Reduction and PCA



Dimensionality Reduction and PCA



Notation and Population Parameters

- Goal: Transform data to principal components that are uncorrelated and ordered by their contribution to the variance.
- Let $\mathrm{x} \in \mathbb{R}^p$ be a random vector with variance $\mathrm{Var}(\mathrm{x}) = \Sigma \in \mathbb{R}^{p \times p}$.
- Assume that Σ is positive definite with distinct eigenvalues

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$$

  and corresponding eigenvectors $\{\mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_p\} \subset \mathbb{R}^p$.
- Let F_x be the cumulative density function for x.

Defining Principal Directions

- The principal components (PC) are defined sequentially [6].
- Let $\{a_1, a_2, \dots, a_p\} \subset \mathbb{R}^p$ be the principal *directions*.
- a_1 solves the constrained optimization problem

$$a_1 = \underset{\|\alpha\|_2=1}{\arg\max} \operatorname{Var}(\alpha^T x).$$

- a_2 solves the constrained optimization problem

$$a_2 = \underset{\|\alpha\|_2=1}{\arg\max} \operatorname{Var}(\alpha^T x) \text{ subject to } \operatorname{Cov}(\alpha^T x, a_1^T x) = 0.$$

- a_k solves the constrained optimization problem

$$a_k = \underset{\|\alpha\|_2=1}{\arg\max} \operatorname{Var}(\alpha^T x) \text{ subject to } \operatorname{Cov}(\alpha^T x, a_i^T x) = 0 \text{ for } i \in [k-1]$$

Deriving a_1

- Notice $\text{Var}(a_1^T x) = a_1^T \Sigma a_1$ and $\|a_1\|_2 = a_1^T a_1$.
- Write the lagrangian $\mathcal{L}(\alpha, \lambda) = \alpha^T \Sigma \alpha - \lambda(\alpha^T \alpha - 1)$
- Differentiate with respect to α yields the eigenvalue-vector equation

$$\Sigma\alpha - \lambda\alpha = (\Sigma - \lambda \mathrm{I})\alpha = 0.$$

- (λ, α) is an eigenvalue-eigenvector pair of Σ.
- Since α is a unit eigenvector, to maximize $\text{Var}(\alpha^T x)$ notice

$$\text{Var}(\alpha^T x) = \alpha^T \Sigma \alpha = \alpha^T (\lambda \alpha) = \lambda \alpha^T \alpha = \lambda$$

- Thus, $\lambda = \lambda_1$ and $a_1 = v_1$ maximizes this equation.

Deriving a_2

- Recall a_2 has the added constraint $\text{Cov}(a_2^T x, a_1^T x) = 0$.
- Writing this constraint out

$$\begin{aligned}
\text{Cov}(a_2^T x, a_1^T x) &= \mathbb{E}[a_2^T x x^T a_1] - \mathbb{E}[a_2^T x]\mathbb{E}[x^T a_1] \\
&= a_2^T E[xx^T]a_1 - a_2^T \mathbb{E}[x]\mathbb{E}[x^T]a_1 \\
&= a_2^T \left(E[xx^T] - \mathbb{E}[x]\mathbb{E}[x]^T \right) a_1 \\
&= a_2^T \Sigma a_1
\end{aligned}$$

- Since a_1 is an eigenvector $a_2^T \Sigma a_1 = \lambda_1 a_2^T a_1 = 0$
- Implies orthogonality of a_1 and a_2.

Deriving a_2 (cont.)

- Write the lagrangian $\mathcal{L}(\alpha, \lambda, \phi) = \alpha^T \Sigma \alpha - \lambda(\alpha^T \alpha - 1) - \phi \alpha^T a_1$
- Differentiating with respect to α gives

$$\frac{\partial \mathcal{L}(\alpha, \lambda, \phi)}{\partial \alpha} = \Sigma \alpha - \lambda \alpha - \phi a_1$$

- Multiplying through by a_1 and setting equal to zero gives

$$\begin{aligned} a_1^T \Sigma a_2 - \lambda a_1^T a_2 - \phi a_1^T a_1 &= 0 \\ (\lambda_1 - \lambda) a_1^T a_2 - \phi a_1^T a_1 &= 0 \\ \phi &= 0 \end{aligned}$$

Deriving a_2 (cont.)

- Returning to the differentiated lagrangian, we have

$$\frac{\partial \mathcal{L}(\alpha, \lambda, \phi)}{\partial \alpha} = (\Sigma - \lambda \mathrm{I})a_2 = 0.$$

- (λ, a_2) is a eigenvalue-eigenvector pair of Σ.
- Therefore, setting $\lambda = \lambda_2$ and $a_2 = v_2$ maximize $\mathrm{Var}(a_2^T x)$ subject to $a_1^T a_2 = 0$.

Principal Directions & Principal Components

Principal Directions

Let $a_k \in \mathbb{R}^p$ be the solution to the optimization problem

$$a_k = \arg\max_{\|\alpha\|_2 = 1} \mathrm{Var}(\alpha^T x) \text{ subject to } \mathrm{Cov}(\alpha^T x, a_i^T x) = 0 \text{ for } i \in [k-1].$$

Then a_k is the k-th eigenvector of Σ.

Principal Components

The *principal components*, $z \in \mathbb{R}^p$, are the coordinates of the data in the transformed space. The *principal component* vector is given by

$$z = [v_1^T x, v_2^T x, \dots, v_p^T x]^T.$$

Principal Components Properties

- Let Σ have eigendecomposition $\Sigma = \mathrm{V}\Lambda\mathrm{V}^T$ so that $\mathrm{V} = [v_1, v_2, \dots, v_p]$ and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ [6].
- Then the principal components can be written as $z = \mathrm{V}^T x$.
  1. Uncorrelated components:
  $$\mathrm{Var}(z) = \mathrm{Var}(V^T x) = V^T \Sigma V = V^T V \Lambda V^T V = \Lambda$$
  2. Decreasing contribution to variance: for all $i < j$
  $$\mathrm{Var}(z_j) = v_j^T \Sigma v_j = \lambda_j > \lambda_i = v_i^T \Sigma v_i = \mathrm{Var}(z_i)$$
  3. Principal axes: The principal directions are the principal axes of the level curves formed by the quadratic form $x^T \Sigma^{-1} x = c$ for some constant $c > 0$.

Implementation of PCA

- In practice, Σ is unobserved and must be estimated to estimate the principal directions.
- Suppose $x_1, x_2, \dots, x_n \overset{i.i.d.}{\sim} F_x$ and *assume* $\mathbb{E}[x] = 0$.
- An unbiased estiamte of Σ is given by

$$\mathrm{S} = \frac{1}{n-1}\sum_{i=1}^{n} x_i x_i^T = \frac{1}{n-1}\mathrm{X}^T\mathrm{X}.$$

  where X has the observations $\{x_i\}_{i=1}^n$ in its rows.
- Estimate V by finding eigenvectors of S.

Implementation of PCA

- Alternatively, suppose X has SVD $X = \hat{U}\hat{D}\hat{V}^T$.
- S can be written as

$$S = \frac{1}{n-1}X^TX = \frac{1}{n-1}\hat{V}\hat{D}^2\hat{V}^T$$

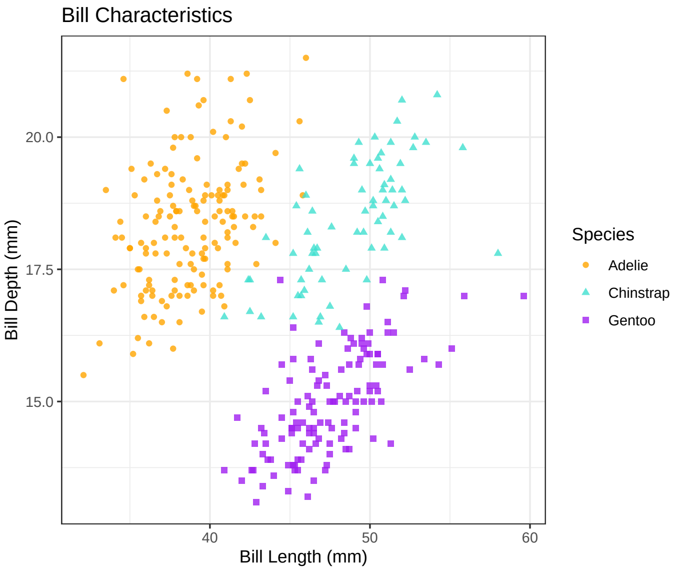
- Estimated principal components are $\{\hat{z}_i = \hat{V}^T x_i\}_{i=1}^n$. In matrix notation,

$$\hat{\mathrm{Z}} = \mathrm{X}\hat{V} = \hat{U}\hat{D}$$

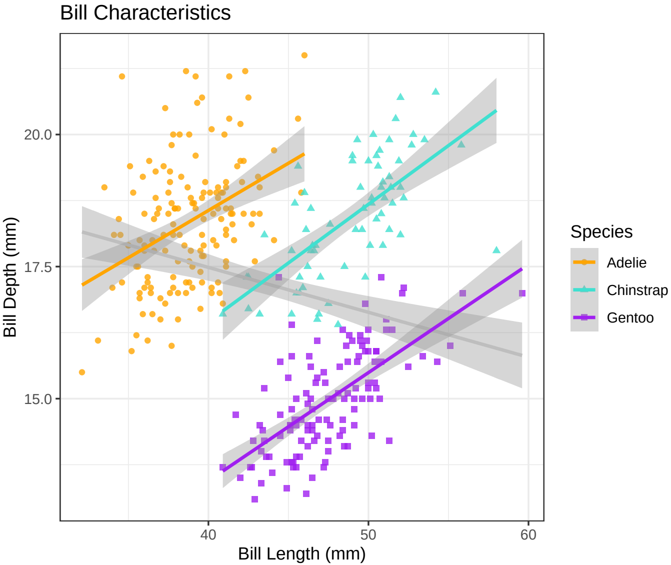
- **Principal components can be estimated by the (truncated) SVD of X.**

Part I: Questions?

Palmer Penguins



Palmer Penguins



CPC Hypothesis

- Common Principal Components (CPC) introduced by Flury 1984 [3]
- Suppose $x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)} \sim F_i$ where $\text{Var}(x_1^{(i)}) = \Sigma_i$.
- Assumed the Σ_i are co-diagonalizable by $\mathrm{V} \in \mathbb{R}^{p \times p}$

$$\Sigma_i = \mathrm{V}\Lambda_i\mathrm{V}^T$$

- Goal: Complete PCA on each population while *leveraging the fact each population shares common principal directions*.
- Estimation approach:
  1. Estimate V by pooling information across populations
  2. Estimate Λ_i independently in each population.

Estimation: MLE

- Assume $x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)} \overset{i.i.d.}{\sim} N(\mu_i, \Sigma_i)$
- Suppose population i has n_i samples and let S_i be the unbiased variance estimator of Σ_i.
- Then the likelihood function can be written as

$$\mathcal{L}(\Sigma_1, \ldots, \Sigma_m | S_1, \ldots, S_m) \propto \prod_{k=1}^{m} \exp\left[\mathrm{Tr}\left(-\frac{n_i}{2}\Sigma_i^{-1} S_i\right)\right] |\Sigma_i|^{-n_i/2}$$

- This likelihood is bounded and always has a solution but has stability issues due to singularities that arise in practice.
- Flury & Gautschi [2] developed an algorithm that can be applied to solve this problem.

Estimation: Spectral

- Unclear how MLE pools information across populations.
- Krzanowski [7] suggested estimating V by finding the eigenvectors of

$$\bar{S} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{S}_i = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{n_i - 1}\sum_{j=1}^{n_i}(\mathrm{x}_j^{(i)} - \bar{x}^{(i)})(\mathrm{x}_j^{(i)} - \bar{x}^{(i)})^T\right]$$

- Let $\bar{S}$ have eigendecomposition $\mathrm{S} = \hat{V}\bar{\hat{\Lambda}}\hat{V}^T$ then the common principal components are given by

$$\hat{\mathrm{Z}}^{(i)} = X^{(i)}\hat{V}$$

- The CPC parameter estimates are given by $(\hat{\mathrm{V}}, \{\mathrm{diag}(\hat{V}^T S_i \hat{V})\}_{i=1}^m)$
- Estimate V by finding eigenvectors of $\bar{\mathrm{S}}$.

Implementation of CPCA

- Assume $\mathbb{E}[\mathrm{x}_1^{(i)}] = 0$ for all $i \in [m]$ and $n_i = n$ for all $i \in [m]$.
- Let $\mathrm{X} = [X^{(1)T} X^{(2)T} \dots X^{(m)T}]^T \in \mathbb{R}^{nm \times p}$ have SVD $X = \hat{U}\hat{D}\hat{V}^T$.
- Then $\bar{S}$ can be written as

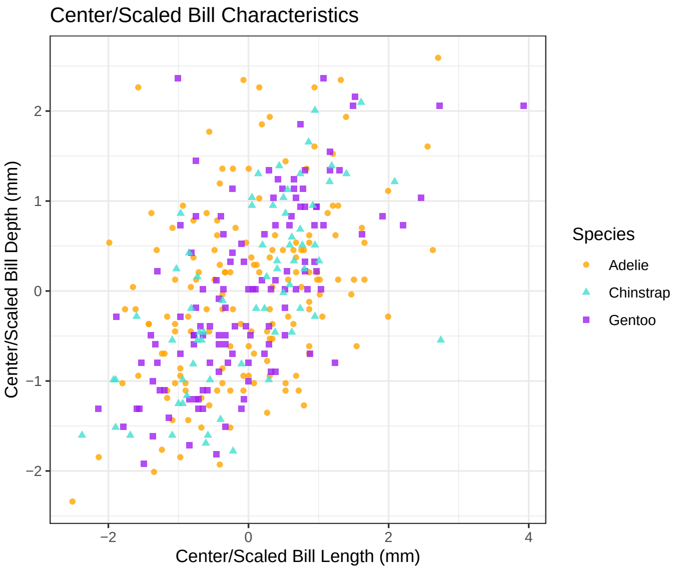
$$\bar{S} = \frac{1}{m(n-1)} X^T X = \frac{1}{m(n-1)} \hat{V}\hat{D}^2\hat{V}^T$$

- Further, notice the common principal components can be written

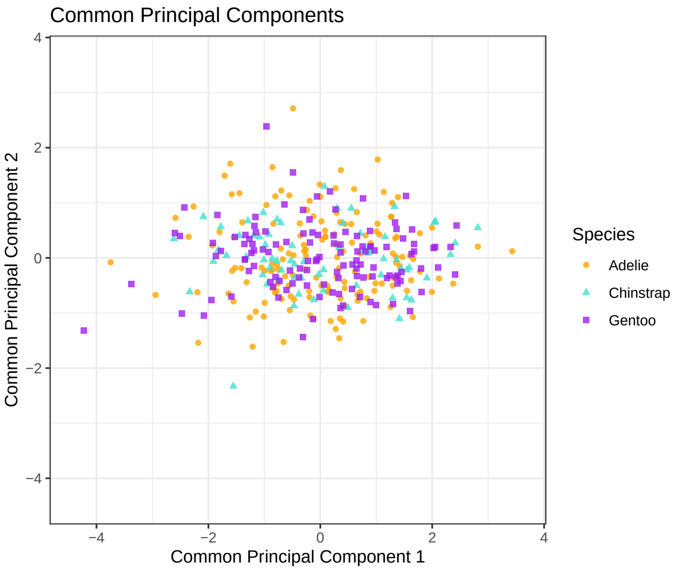
$$\begin{bmatrix} \hat{\mathrm{Z}}^{(1)} \\ \vdots \\ \hat{\mathrm{Z}}^{(m)} \end{bmatrix} = \begin{bmatrix} \mathrm{X}^{(1)}\hat{V} \\ \vdots \\ \mathrm{X}^{(m)}\hat{V} \end{bmatrix} = \begin{bmatrix} \mathrm{X}^{(1)} \\ \vdots \\ \mathrm{X}^{(m)} \end{bmatrix} \hat{V} = X\hat{V} = \hat{U}\hat{D}.$$

- **The CPCs can be estimated by the SVD of** X.

Palmer Penguins



Palmer Penguins



Palmer Penguins

- After estimating $\hat{V}$, estimate $\hat{\Lambda}_i = \text{diag}(\hat{V}^T S_i \hat{V})$.
- Estimate how much each CPC captures variance from each species by

$$\%\text{ variance explained by CPC}_i = \frac{\hat{\lambda}_i}{\sum_{j=1}^{p} \hat{\lambda}_j}$$

Species	CPC1	CPC2
Adelie	0.7	0.3
Chinstrap	0.83	0.17
Gentoo	0.82	0.18

Table: Variance explained by each CPC by species.

Part II: Questions?

Randomized Algorithms

- Randomized algorithms are growing in popularity for the computation of matrix decompositions [4].
- Most randomized algorithms follow a two step approach
  1. Introduce randomness that reduces the size/complexity of the problem
  2. Use deterministic algorithms to complete the matrix decomposition on the smaller subproblem
- Most work focuses on introducing the 'right type' of randomness that preserves the matrix's spectral properties.

Randomized Algorithm for Truncated SVD

rSVD algorithm [1]

1. Let $\Omega \in \mathbb{R}^{p \times k}$ have random Gaussian entries and k is the target rank.
2. Compute the QR decomposition of $X\Omega = QR$.
3. Let $Q^T X$ have SVD $Q^T X = \tilde{U}\hat{\Sigma}\hat{V}^T$.
4. Set the left singular vectors $\hat{U} = Q\tilde{U}$.
5. Return $\hat{X}_k = \hat{U}\hat{\Sigma}\hat{V}^T$.

- Q is first commuted to approximate $col(X)$, $X \approx QQ^T X$.
- Deterministic SVD only used on $k \times p$ matrix instead of $n \times p$ matrix.

Randomized SVD: Visualization

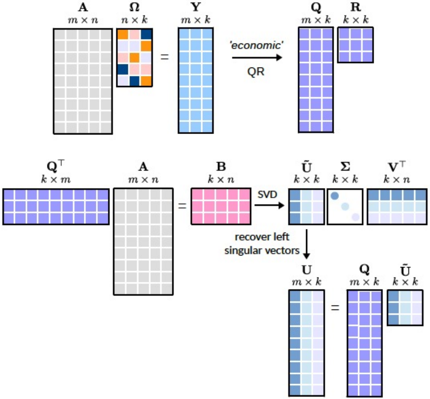


Figure: Figure from Figure 8 in Erichson et. al [1]

Error Analysis

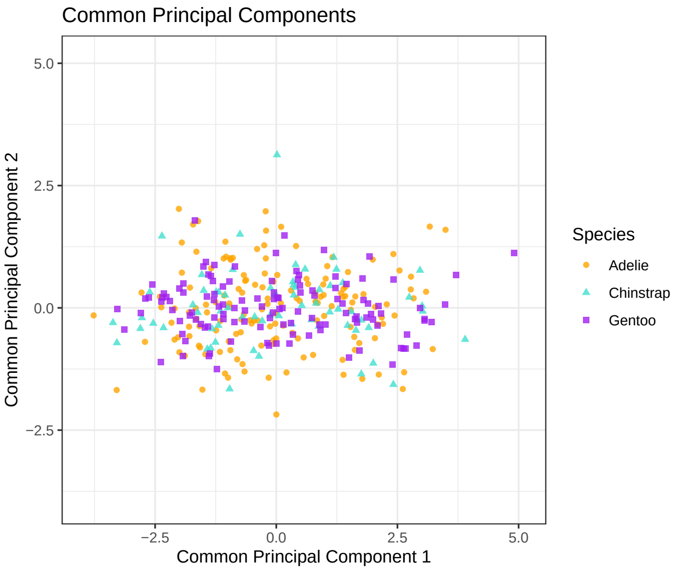
Theoretical Performance of rSVD

Let X_k be the k-rank truncated SVD of X computed by rSVD. Then expected spectral norm error is given by [9, 4, 1].

$$\mathbb{E}\|X - X_k\|_2 \leq \left[1 + \sqrt{\frac{k}{p-1}} + \frac{e\sqrt{k+p}}{p} \cdot \sqrt{\min\{m,n\} - k}\right]^{\frac{1}{2q+1}} \sigma_{k+1}(X)$$

- (p, q) are two parameters to rSVD that improve its application.
- Eckart and Young guarantee this error is bounded below by $\sigma_{k+1}(X)$.

Palmer Penguins



Palmer Penguins

- Estimate how much each CPC captures variance from each species by

$$\% \text{ variance explained by } \mathrm{CPC}_i = \frac{\hat{\lambda}_i}{\sum_{j=1}^{p} \hat{\lambda}_j}$$

Species	CPC1	CPC2	CPC3	CPC4
Adelie	0.58	0.17	0.16	0.09
Chinstrap	0.68	0.13	0.08	0.10
Gentoo	0.76	0.08	0.09	0.07

Table: Variance explained by each CPC by species.

Conclusion

- PCA is a powerful tool for dimensionality reduction in highly interrelated data.
- CPCA is an extension that allows for population specific variation along common principal direction.
- PCA and CPCA can both be carried out by using the truncated SVD.
- Randomized algorithms are making the computation of truncated SVDs more scalable to large datasets.

Part III: Questions?

References I

- N. Erichson et al. “Randomized Matrix Decompositions using R”. In: *Submitted to JSS* (Aug. 2016).
- Bernhard Flury and Walter Gautschi. “An Algorithm for Simultaneous Orthogonal Transformation of Several Positive Definite Symmetric Matrices to Nearly Diagonal Form”. In: *Siam Journal on Scientific and Statistical Computing* 7 (Jan. 1986). DOI: 10.1137/0907013.
- Bernhard N. Flury. “Common Principal Components in K Groups”. In: *Journal of the American Statistical Association* 79.388 (1984), pp. 892–898. ISSN: 01621459.
- N. Halko, P. G. Martinsson, and J. A. Tropp. “Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions”. In: *SIAM Review* 53.2 (2011), pp. 217–288.

References II

Allison Marie Horst, Alison Presmanes Hill, and Kristen B Gorman. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. R package version 0.1.0. 2020. URL: https://allisonhorst.github.io/palmerpenguins/.

Ian Jolliffe. "Principal Component Analysis". In: *International Encyclopedia of Statistical Science*. Ed. by Miodrag Lovric. Springer Berlin Heidelberg, 2011.

W. J. Krzanowski. "Principal Component Analysis in the Presence of Group Structure". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 33.2 (1984), pp. 164–168.

Matteo Magnani, Barbora Micenkova, and Luca Rossi. *Combinatorial Analysis of Multiple Networks*. 2013. arXiv: 1303.4986 [cs.SI].

References III

Per-Gunnar Martinsson. *Randomized methods for matrix computations*. 2019. arXiv: 1607.01649 [math.NA].