Unsupervised Dimensionality Reduction via PCA

David I. Inouye
Very high-dimensional data is becoming ubiquitous

- Images (1 million pixels)
- Text (100k unique words)
- Genetics (4 million SNPs)
- Business data (12 million products)
Why dimensionality reduction?
Visualization

- Allows 2D scatterplot visualizations even of high-dimensional data (2D projection of digits)

https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html
Why dimensionality reduction? Lower computation costs

- Suppose original dimension is large like $d = 100000$ (e.g., images, DNA sequencing, or text)

- If we reduce to $k = 100$ dimensions, the training algorithm can be sped up by $1000\times$

4-5 million SNPs in human genome. [https://www.diagnosticsolutionslab.com/tests/genomicinsight](https://www.diagnosticsolutionslab.com/tests/genomicinsight)
Why dimensionality reduction? Underlying phenomena is on lower dimensional space.
Outline of Principal Components Analysis (PCA)

1. Motivation for dimensionality reduction
2. Formal PCA problem: Min reconstruction
3. Derive PCA formulation for 1D
   ▶ Least error 1D projection is orthogonal
   ▶ Sum over all data points
4. Solution is based on truncated SVD
5. Equivalent problem: Max variance
Math: Principal Component Analysis (PCA) can be formalized as minimizing the linear reconstruction error of the data using only $k \leq d$ dimensions.

- PCA can be formalized as

$$\min_{Z \in \mathbb{R}^{n \times k}, W \in \mathbb{R}^{d \times k}} \|X_c - ZW^T\|_F^2 \quad \text{s.t.} \quad W^TW = I_k$$

- where

$$X_c = X - \mathbf{1}_n \mu_x^T \in \mathbb{R}^{n \times d} \quad \text{(centered input data)}$$
Review of linear algebra
and introduction to numpy Python library

- See Jupyter notebook, which can be opened and run in Google Colab
Math: Principal Component Analysis (PCA) can be formalized as minimizing the linear reconstruction error of the data using only $k \leq d$ dimensions.

- PCA can be formalized as

$$
\min_{Z \in \mathbb{R}^{n \times k}, W \in \mathbb{R}^{d \times k}} \|X_c - ZW^T\|_F^2 \quad \text{s.t. } W^T W = I_k
$$

- where

$$X_c = X - \mathbf{1}_n \mu_x^T \in \mathbb{R}^{n \times d} \quad \text{(centered input data)}$$
Math: Principal Component Analysis (PCA) can be formalized as minimizing the linear reconstruction error of the data using only \( k \leq d \) dimensions

\[
\min_{Z \in \mathbb{R}^{n \times k}, W \in \mathbb{R}^{d \times k}} \| X_c - Z W^T \|_F^2 \quad \text{s.t.} \quad W^T W = I_k
\]

- Let’s stare at this equation some more 😊
- Why is this dimensionality reduction?
- What does the orthogonal constraint mean?
- Why minimize the squared Frobenius norm?

\[
\| X_c - Z W^T \|_F^2 = \sum_{i=1}^{n} \| x_i^T - z_i^T W^T \|_2^2 = \sum_{i=1}^{n} \| x_i - W z_i \|_2^2
\]

- For analysis, let’s simplify to a single dimension (i.e., \( k = 1 \))
  \[
  \sum_{i=1}^{n} \| x_i - z_i w \|_2^2 \text{ where } z_i \text{ is a scalar}
  \]
What is the best projection given a fixed subspace (line in 1D case)?

- If we are given \( \mathbf{w} \), what is the best \( \mathbf{z} \) (i.e. minimum reconstruction error) for a given \( \mathbf{x} \)?
  - \[ \min_{\mathbf{z}} \| \mathbf{x} - \mathbf{z} \mathbf{w} \|_2^2 \]

The orthogonal projection!
- \( \mathbf{z} = \mathbf{x}^T \mathbf{w} = \| \mathbf{x} \| \| \mathbf{w} \| \cos \theta = \| \mathbf{x} \| \cos \theta \)
- \( \mathbf{z} = \| \mathbf{x} \| \cos \theta = \text{hyp} \cdot \frac{\text{adj}}{\text{hyp}} = \text{adj} \)
- \( \mathbf{z} \mathbf{w} \) is a scaled vector along the line defined by \( \mathbf{w} \)
Thus, we can simplify to only minimizing over $W$

$$\min_{z,w: \|w\|_2=1} \sum_{i=1}^{n} \|x_i - z_i w\|_2^2 = \min_{w: \|w\|_2=1} \sum_{i=1}^{n} \|x_i - (x_i^T w) w\|_2^2$$

- Now we can return to the Frobenius norm:

$$\min_{w: \|w\|_2=1} \|X_c - zw^T\|_F^2 \quad \text{where} \quad z = X_c w$$

- What is $zw^T$? Have we seen something like this before?

- This is the best rank-1 approximation to $X_c$, which is given by the SVD!
  - $w = v_1$ and $z = \sigma_1 u_1$, where $\sigma_1$, $u_1$, $v_1$ are the first singular value, left singular vector and right singular vector respectively.
For $k \geq 1$, the PCA solution is the top $k$ right singular vectors.

- If $X_c = USV^T$, then the general solution is $W^* = V_{1:k}$

- Remember: SVD is best $k$ dim. approximation
Check: The solution reveals the truncated SVD as best approximation

\[
\min_{W} \|X_c - (X_c W)W^T\|_F^2 \quad \text{s.t. } W^TW = I
\]

\[
X_c = USV^T
\]

\[
W = V_{1:k}
\]
Check: The solution reveals the truncated SVD as best approximation

\[
\min_{W} \left\| X_c - (X_c W) W^T \right\|_F^2 \quad \text{s.t.} \quad W^T W = I
\]
Intuition: Principal component analysis finds the best linear projection onto a lower-dimensional space.

\[
\min_{w: \|w\|_2=1} \|X_c - zw^T\|_F^2
\]

where \( z = X_c w \)

2D to 1D projection: Red lines show the projection error onto 1D lines. PCA finds the line that has the smallest projection error (in this example, when it aligns with the purple).

https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues
Minimizing reconstruction error (red lines) is equivalent to maximizing the variance of projection (spread of red points)

Max reconstruction error
Min variance

Min reconstruction error
Max variance

\[
\text{argmax}_{w: \|w\|_2 = 1} \|X_c - zw^T\|_F^2
\]

\[
= \text{argmin}_{w: \|w\|_2 = 1} \|z\|_2^2 = \text{argmin}_{w: \|w\|_2 = 1} \sigma_z^2
\]

where \( z = X_c w \)

\[
\text{argmin}_{w: \|w\|_2 = 1} \|X_c - zw^T\|_F^2
\]

\[
= \text{argmax}_{w: \|w\|_2 = 1} \sigma_z^2
\]
Derivation of equivalence will require 3 facts

1. Squared Frobenius norm is trace of matrix product (generalizes $\|x\|_2^2 = x^T x$):
   - $\|A\|_F^2 = \text{Tr}(A^T A)$

2. If $A \in \mathbb{R}^{n \times d}$ is a centered data matrix, then the Frobenius norm is the scaled sum of 1D variances:
   - $\|A\|_F^2 = \text{Tr}(A^T A) = n \text{Tr}(\hat{\Sigma}) = n \sum_j \sigma_j^2$
   - where $\hat{\Sigma}$ is the empirical covariance matrix and $\sigma_j^2$ is variance for the $j$-th dimension

3. Optimization solutions are invariant when the objective is multiplied by positive constant or a constant is added,
   - $\arg\min_W f(W) = \arg\min_W af(W) + b$, $\forall a > 0, b \in \mathbb{R}$
The PCA objective can be decomposed into the original variance minus the variance of projection

- Minimize reconstruction error
  \[ \min_{W:W^TW = I_K} \|X_c - (X_c W)W^T\|_F^2 \]
- \[ \|X_c - X_c W W^T\|_F^2 \]
- \[ = \text{Tr}[(X_c - X_c W W^T)^T (X_c - X_c W W^T)] \]
- \[ = \text{Tr}[(X_c^T - W W^T X_c^T)(X_c - X_c W W^T)] \]
- \[ = \text{Tr}[X_c^T X_c - W W^T X_c^T X_c - X_c^T X_c W W^T + W W^T X_c^T X_c W W^T] \]
- \[ = \text{Tr}[X_c^T X_c] - \text{Tr}[W W^T X_c^T X_c] - \text{Tr}[X_c^T X_c W W^T] + \text{Tr}[W W^T X_c^T X_c W W^T] \]
- \[ = \text{Tr}[X_c^T X_c] - \text{Tr}[W^T X_c^T X_c W] - \text{Tr}[W^T X_c^T X_c W] + \text{Tr}[W^T X_c^T X_c W W^T W] \]
- \[ = \text{Tr}[X_c^T X_c] - \text{Tr}[W^T X_c^T X_c W] - \text{Tr}[W^T X_c^T X_c W] + \text{Tr}[W^T X_c^T X_c W] \]
- \[ = \text{Tr}[X_c^T X_c] - \text{Tr}[(X_c W)W^T X_c W] \]
- \[ = \text{Tr}[X_c^T X_c] - \text{Tr}[Z^T Z] \]
- \[ = n \sum_{j=1}^d \sigma_{x,j}^2 - n \sum_{j=1}^k \sigma_{z,j}^2 \]
Equivalence is derived by manipulating optimization problem

- $\text{argmin } \|X_c - (X_c W)W^T\|^2_F$
- $w : w^T w = I_k$
- $= \text{argmin } n \sum_{j=1}^{d} \sigma_{x,j}^2 - n \sum_{j=1}^{k} \sigma_{z,j}^2$
- $w : w^T w = I_k$
- $= \text{argmin } - \sum_{j=1}^{k} \sigma_{z,j}^2$
- $w : w^T w = I_k$
- $= \text{argmax } \sum_{j=1}^{k} \sigma_{z,j}^2$
- $w : w^T w = I_k$
- This last one is exactly maximizing the variance along the projected dimensions of $z$
Equivalent solutions: The solution to both problems is the top $k$ right singular vectors of $X_c$

- Minimize reconstruction error
  \[
  \min_{W:WW^T=I_k} \|X_c - (X_cW)W^T\|_F^2
  \]
  - Singular value decomposition (SVD) of $X_c = USV^T$
  - Solution: $W^* = V_{1:k}$

- Maximize variance of latent projection (equivalent solution)
  \[
  \max_{W:WW^T=I_k} \sum_{j=1}^k \sigma_{Z,j}^2
  \]
  - Equivalent solution is the eigenvectors of $X_c^T X_c = n \hat{\Sigma}$
    - $X_c^T X_c = (USV^T)^T(USV^T) = (VSU^T)(USV^T) = VS(U^T U)SV^T = VS^2V^T = Q\Lambda Q^T$
  - Solution: $W^* = Q_{1:k} \equiv V_{1:k}$
Recap: Principal Components Analysis (PCA)

1. Motivation for dimensionality reduction
2. Formal PCA problem: Min reconstruction
3. Derive PCA formulation for 1D
   ▶ Least error 1D projection is orthogonal
   ▶ Sum over all data points
4. Solution is based on truncated SVD
5. Alternative viewpoint: Max variance
   ▶ Derive equivalence
   ▶ Derive equivalent solutions
Demo of PCA via sklearn (time permitting)

- Random projections vs PCA projections
- Visualizations of
  - Minimum reconstruction error
  - Maximum variance
  - Explained variance based on $k$
- Code examples
  - Digits
  - Eigenfaces
Questions?