COMPSCI 589
Lecture 19: Principal Components Analysis

Benjamin M. Marlin

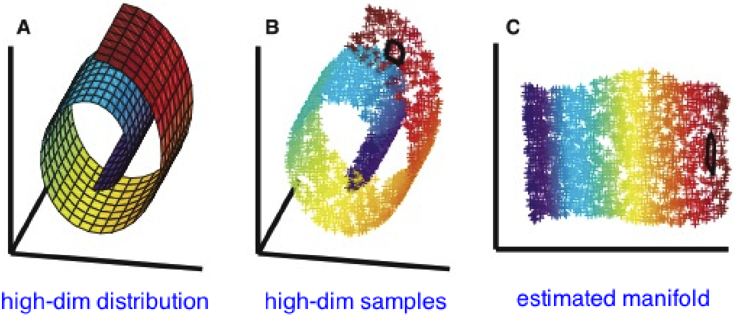
College of Information and Computer Sciences
University of Massachusetts Amherst

Slides by Benjamin M. Marlin (marlin@cs.umass.edu).
Created with support from National Science Foundation Award# IIS-1350522.

The Dimensionality Reduction Task

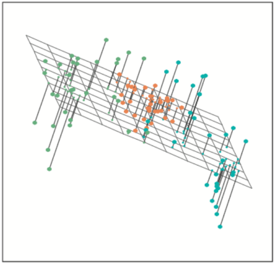
Definition: The Dimensionality Reduction Task

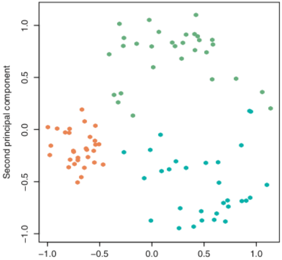
Given a collection of feature vectors $\mathbf{x}_i \in \mathbb{R}^D$, map the feature vectors into a lower dimensional space $\mathbf{z}_i \in \mathbb{R}^K$ where $K < D$ while preserving certain properties of the data.



Linear Dimensionality Reduction

- The simplest dimensionality reduction methods assume that the observed high dimensional data vectors $\mathbf{x}_i \in \mathbb{R}^D$ lie on a K-dimensional linear manifold within $\mathbb{R}^D$.
- Mathematically, the linear sub-space assumption can be written as $\mathbf{X} = \mathbf{Z} \times \mathbf{B}$





Learning

- The learning problem for linear dimensionality reduction is to estimate values for both $\mathbf{Z}$ and $\mathbf{B}$ given only the noisy observations $\mathbf{X}$.
- One possible learning criteria is to minimize the sum of squared errors when reconstructing $\mathbf{X}$ from $\mathbf{Z}$ and $\mathbf{B}$. This leads to:

$$\arg\min_{\mathbf{Z},\mathbf{B}} ||\mathbf{X} - \mathbf{ZB}||_F$$

where $||\mathbf{A}||_F$ is the Frobenius norm of matrix $\mathbf{A}$ (the sum of the squares of all matrix entries).

Singular Value Decomposition

- We can pick a unique representation for the subspace by specifying additional criteria. Classical Rank-K Singular Value Decomposition (K-SVD) corresponds to the following restriction:

$$\arg\min_{\mathbf{U},\mathbf{S},\mathbf{V}} ||\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}^T||_F$$

  where S is a $K \times K$ diagonal matrix with positive elements, $\mathbf{U}$ is an $N \times K$ matrix such that $\mathbf{U}^T\mathbf{U} = I$, and V is a DxK matrix such that $\mathbf{V}^T\mathbf{V} = I$.

- The matrix product $\mathbf{Z} = \mathbf{US}$ gives the optimal rank-K representation of $\mathbf{X}$ with respect to Frobenius norm minimization, with $\mathbf{V}^T$ acting as the basis for the space.

Eigenvectors

- Let $\mathbf{A} \in \mathbb{R}^{DxD}$ be a matrix, $\mathbf{v} \in \mathbb{R}^{D}$ be a vector, and λ be scalar.
- If $\mathbf{Av} = \lambda\mathbf{v}$ then $\mathbf{v}$ is a right eigenvector of A with eigenvalue λ.
- If $\mathbf{A}^T\mathbf{v} = \lambda\mathbf{v}$ then $\mathbf{v}$ is a left eigenvector of A with eigenvalue λ (equivalently $\mathbf{v}^T\mathbf{A} = \lambda\mathbf{v}^T$).
- If $\mathbf{A}$ is symmetric so that $\mathbf{A} = \mathbf{A}^T$, then the left and right eigenvectors of $\mathbf{A}$ are the same with the same eigenvalues.

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = 3 \begin{bmatrix} 1 & 1 \end{bmatrix}$$

- A full-rank (invertible) matrix $\mathbf{A} \in \mathbb{R}^{DxD}$ will have D linearly independent eigenvectors.

Eigendecomposition

- Let $\mathbf{V} \in \mathbb{R}^{DxD}$ be a matrix whose columns $\mathbf{v}_d$ are D linearly independent eigenvectors of $\mathbf{A}$ with Λ the corresponding diagonal matrix of eigenvalues such that $\Lambda_{dd} = \lambda_d$.Then:

$$\mathbf{AV} = \mathbf{V}\Lambda$$

$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^{-1}$$

$$\mathbf{V}^{-1}\mathbf{AV} = \Lambda$$

- Without loss of generality, we can assume that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_D$.

Eigendecomposition of a Symmetric Matrix

- If $\mathbf{A}$ is symmetric, we can choose D orthonormal eigenvectors so that $||\mathbf{v}_d||_2 = 1$, $\mathbf{v}_d^T \mathbf{v}_{d'} = 0$ and D real eigenvalues $\lambda_d \in \mathbb{R}$. This representation of $\mathbf{A}$ is unique. As a result, we have:

$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^T = \sum_{d=1}^{D} \lambda_d \mathbf{v}_d \mathbf{v}_d^T$$

$$\mathbf{V}^T\mathbf{A}\mathbf{V} = \Lambda$$

Representation of a Vector in the Eigen Basis

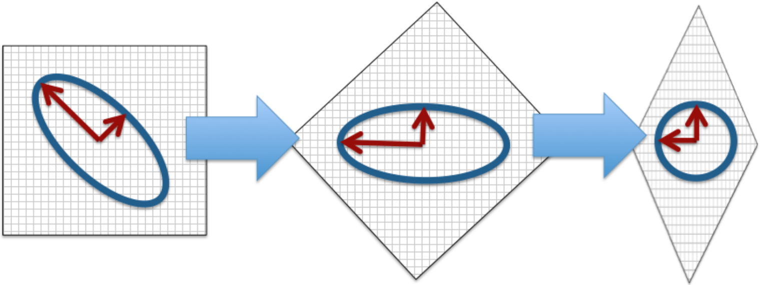
- Similarly, if **a** is an arbitrary vector, then we can also represent **a** using the basis provided by the eigevectors **V** of a real symmetric matrix **A**. We obtain:

$$\mathbf{a} = \sum_{d=1}^{D} \alpha_d \mathbf{v}_d \tag{1}$$

$$\alpha_d = \mathbf{a}^T \mathbf{v}_d \tag{2}$$

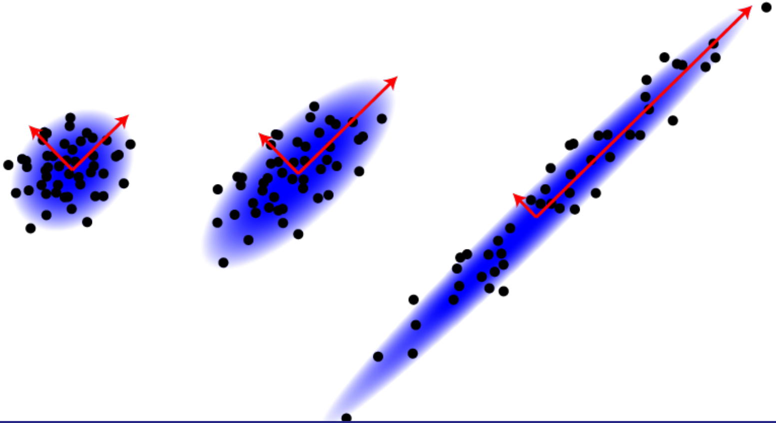
Geometry

- If $\mathbf{A}$ is a real symmetric matrix with positive eigenvalues, then the quadratic equation $\mathbf{x}^T\mathbf{A}\mathbf{x} = 0$ defines an ellipsoid in a D-dimensional space, which provides a different way of thinking about these operations:



Principal Component Analysis

- Given a data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$, the goal of Principal Component Analysis (PCA) is to identify the directions of maximum variance contained in the data.



Sample Variance in a Given Direction

- Let $\mathbf{w} \in \mathbb{R}^D$ such that $||\mathbf{w}||_2 = \sqrt{\mathbf{w}^T\mathbf{w}} = 1$.
- The sample estimate of the variance in the direction $\mathbf{w}$ given the data set $\mathbf{X}$ is given by the expression:

$$\frac{1}{N}\sum_{i=1}^{N}(\mathbf{X}_i\mathbf{w} - \mu)^2 \quad \text{where} \quad \mu = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{w}$$

Pre-Centering

- Under the assumption that the data are pre-centered so that $\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i = 0$, this expression simplifies to:

$$\frac{1}{N}\sum_{i=1}^{N}(\mathbf{X}_i\mathbf{w})^2 = (\mathbf{X}\mathbf{w})^T(\mathbf{X}\mathbf{w}) = \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}$$

The Direction of Maximum Variance

- Suppose we want to identify the direction $\mathbf{w}_1$ of maximum variance given the data matrix $\mathbf{X}$. We can formulate this optimization problem as follows:

$$\mathbf{w}_1 = \max_{\mathbf{w}} \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \; ... \; \text{st} \; ||\mathbf{w}||_2 = 1$$

- How can we solve this problem?

The Direction of Maximum Variance

- Let $\Sigma = \mathbf{X}^T\mathbf{X}$.
- Σ is real and symmetric, so it admits an eigendecomposition of the form:

$$\Sigma = \sum_{d=1}^{D} \sigma_d \mathbf{V}_d \mathbf{V}_d^T$$

- $\sigma_1 \geq \sigma_2 \geq, ..., \geq \sigma_D \geq 0$ are the eigenvalues of Σ.
- $\mathbf{V}_d \in \mathbb{R}^D$ are the eigenvectors of Σ. They satisfy:

$$||\mathbf{V}_d||_2 = \sqrt{\mathbf{V}_d^T \mathbf{V}_d} = 1 \text{ ... for all } d$$

$$\mathbf{V}_d^T \mathbf{V}_{d'} = 0 \text{ ... for all } d \neq d'$$

The Direction of Maximum Variance

- Using this result, we can write the optimization problem as:

$$\max_{\mathbf{w}} \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} \text{ ... st } ||\mathbf{w}||_2 = 1$$

$$\max_{\mathbf{w}} \mathbf{w}^T \left(\sum_{d=1}^{D} \sigma_d \mathbf{V}_d \mathbf{V}_d^T\right) \mathbf{w} \text{ ... st } ||\mathbf{w}||_2 = 1$$

$$\max_{\mathbf{w}} \sum_{d=1}^{D} \sigma_d (\mathbf{w}^T\mathbf{V}_d)^2 \text{ ... st } ||\mathbf{w}||_2 = 1$$

The Direction of Maximum Variance

- $\mathbf{w}$ can also be expressed in the orthonormal basis $\mathbf{V}_1, ..., \mathbf{V}_D$ by letting $\mathbf{w} = \sum_{d=1}^{D} \omega_d \mathbf{V}_d$.
- The constraint that $||\mathbf{w}||_2 = 1$ becomes $\sqrt{\sum_{d=1}^{D} \omega_d^2} = 1$.
- This means $\sum_{d=1}^{D} \omega_d^2 = 1$ and $\omega_d^2 > 0$, so the ω_d^2 values act like a discrete probability distribution.

The Direction of Maximum Variance

- Plugging this back into the objective function, we have:

$$\max_{\mathbf{w}} \sum_{d=1}^{D} \sigma_d (\mathbf{w}^T \mathbf{V}_d)^2 \ \ldots \text{ st } ||\mathbf{w}||_2 = 1$$

$$\max_{\omega} \sum_{d=1}^{D} \sigma_d \left(\sum_{d'=1}^{D} \omega_{d'} \mathbf{V}_{d'}^T \mathbf{V}_d \right)^2 \ \ldots \text{ st } \sum_{d=1}^{D} \omega_d^2 = 1$$

$$\max_{\omega} \sum_{d=1}^{D} \sigma_d \omega_d^2 \ \ldots \text{ st } \sum_{d=1}^{D} \omega_d^2 = 1$$

The Direction of Maximum Variance

- At this point, the solution is clear.
- To maximize the variance, we need to set $\omega_1 = 1$ and set $\omega_d = 0$ otherwise. This put's all the weight on the maximum eigenvalue of Σ, which is σ_1 by assumption.
- Working our way back to $\mathbf{w}_1$, we put all our weight on the maximum eigenvalue, so $\mathbf{w} = \sum_{d=1}^{D} \omega_d \mathbf{V}_d = \mathbf{V}_1$.
- **This shows that the maximum variance direction given a data matrix X is the eigenvector of $\mathbf{X}^T\mathbf{X}$ with the largest eigenvalue.**

K Largest Directions of Variance

- Suppose instead of just the direction of maximum variance, we want the K largest directions of variance that are all mutually orthogonal.
- Finding the second-largest direction of variance corresponds to solving the problem:

$$\mathbf{w}_2 = \max_{\mathbf{w}} \sum_{d=1}^{D} \sigma_d (\mathbf{w}^T \mathbf{V}_d)^2 \; ... \text{ st } ||\mathbf{w}||_2 = 1 \text{ and } \mathbf{w}^T \mathbf{w}_1 = 0$$

- It's easy to see that this is going to be the eigenvector corresponding to the second largest eigenvalue.
- **In general, the top K directions of variance $\mathbf{w}_1, ..., \mathbf{w}_K$ are given by the K eigenvectors corresponding to the K largest eigenvalues of $\mathbf{X}^T\mathbf{X}$.**

Dimensionality Reduction with PCA

1. Given centered data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$, compute unscaled sample covariance matrix $\Sigma = \mathbf{X}^T\mathbf{X}$.
2. Compute the K leading eigenvectors $w_1, \ldots, w_K$ of Σ where $\mathbf{w}_k \in \mathbb{R}^D$.
3. Stack the eigenvectors together into a $D \times K$ matrix $\mathbf{W}$ where each column k of $\mathbf{W}$ corresponds to $\mathbf{w}_k$.
4. Project the matrix $\mathbf{X}$ into the rank-K sub-space of maximum variance by computing the matrix product $\mathbf{Z} = \mathbf{X}\mathbf{W}$.
5. To reconstruct $\mathbf{X}$ given $\mathbf{Z}$ and $\mathbf{W}$, we use $\hat{\mathbf{X}} = \mathbf{Z}\mathbf{W}^T$.

Connection to SVD

- Last class we saw that the minimum Frobenius norm linear dimensionality reduction problem could be solved using the the rank-K SVD of $\mathbf{X}$:

  $$\arg\min_{\mathbf{U},\mathbf{S},\mathbf{V}} ||\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}^T||_F$$

  where the matrix product $\mathbf{Z} = \mathbf{U}\mathbf{S}$ gives the optimal rank-K representation of $\mathbf{X}$ with respect to Frobenius norm minimization.

Connection to SVD

- If we let $K = D$ then $\mathbf{X} = \mathbf{USV}^T$ and $\mathbf{X}^T\mathbf{X} = \mathbf{VSU}^T\mathbf{USV}^T$.
- Due to orthogonality of U this gives: $\mathbf{X}^T\mathbf{X} = \mathbf{VS}^2\mathbf{V}^T$.
- This means that the right singular vectors of $\mathbf{X}$ are exactly the eigenvectors of $\mathbf{X}^T\mathbf{X}$, so SVD's $\mathbf{V}$ and PCA's $\mathbf{W}$ are identical (assuming $\mathbf{X}$ is centered).
- We can also see that the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are the squares of the diagonal elements of $\mathbf{S}$.
- This means that the K largest singular values and K largest eigenvalues correspond to the same K basis vectors.

Connection to SVD

- According to PCA, the projection operation is $\mathbf{Z} = \mathbf{X}\mathbf{W}$.
- Using $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ and $\mathbf{V} = \mathbf{W}$ we have:

$$\mathbf{Z} = \mathbf{X}\mathbf{W} = (\mathbf{U}\mathbf{S}\mathbf{V}^T)(\mathbf{V}) = \mathbf{U}\mathbf{S}$$

- Finally, note that if the decompositions are based only on the K leading basis vectors, which are identical under both PCA and SVD, the projections $\mathbf{Z} = \mathbf{X}\mathbf{W}$ and $\mathbf{Z} = \mathbf{U}\mathbf{S}$ will still be identical.

Connection to SVD

- These manipulations show that PCA on $\mathbf{X}^T\mathbf{X}$ and SVD on $\mathbf{X}$ identify exactly the same sub-space and result in exactly the same projection of the data into that sub-space.
- As a result, generic linear dimensionality reduction simultaneously minimizes the Frobenius norm of the reconstruction error of $\mathbf{X}$ and maximizes the retained variance in the learned sub-space.
- Both SVD and PCA provide the same refinement of generic linear dimensionality reduction: an orthogonal basis for exactly the same optimal linear subspace.

Issues

- The computational complexity of PCA is $O(D^2N + D^3)$ if the full eigendecomposition is obtained and then truncated, compared to $O(min(DN^2, ND^2))$ for SVD.
- If $K << D$, then PCA can also be computed iteratively, as can SVD.
- The basic SVD and PCA algorithms are not suitable for large-scale data. Instead, randomized algorithms are often used.
- The value of K can sometimes be chosen based on looking for eigenvalue gaps in the eigenspectrum of the covariance matrix. Otherwise, a supervised end/side-task is needed or a criteria like AIC/BIC must be applied.