COMPSCI 589
Lecture 20: Linear Dimensionality Reduction, SVD and PCA

Benjamin M. Marlin

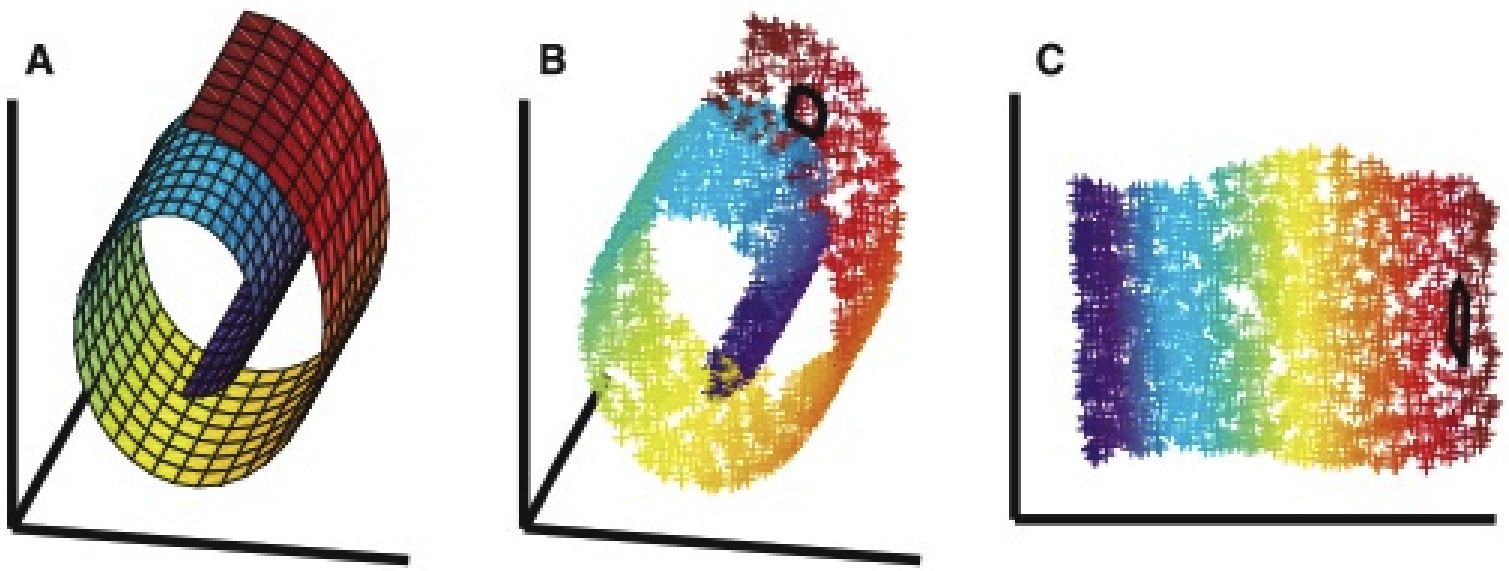
College of Information and Computer Sciences
University of Massachusetts Amherst

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

The Dimensionality Reduction Task

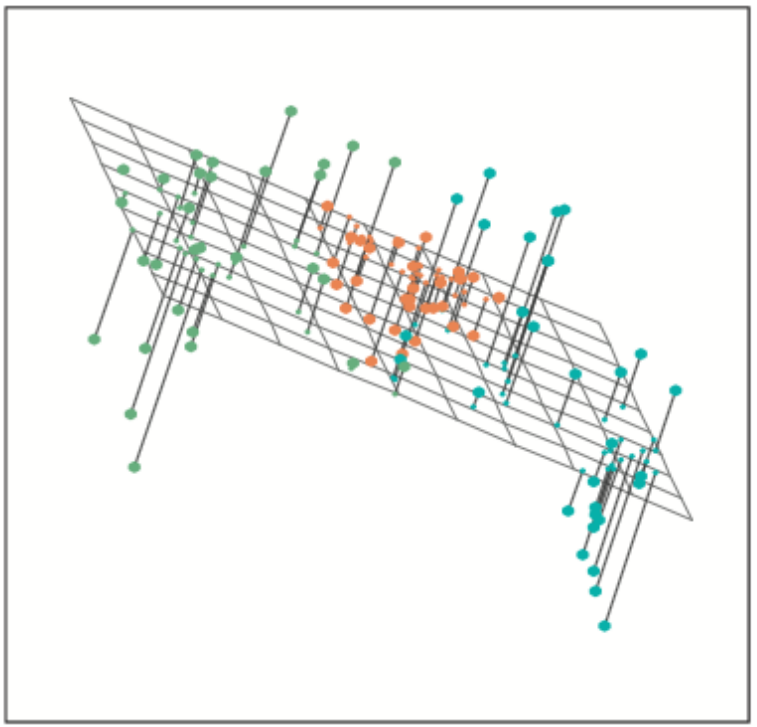
Definition: The Dimensionality Reduction Task

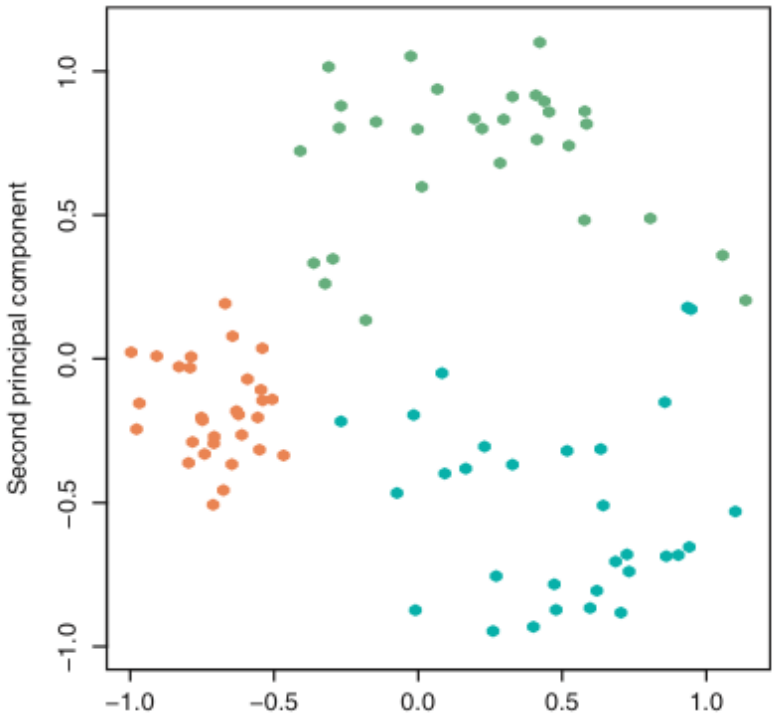
Given a collection of feature vectors $\mathbf{x}_i \in \mathbb{R}^D$, map the feature vectors into a lower dimensional space $\mathbf{z}_i \in \mathbb{R}^K$ where $K < D$ while preserving certain properties of the data.



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 $\mathbb{R}^D$.





Linear Dimensionality Reduction

- Mathematically, the linear sub-space assumption can be written as follows:

$$\mathbf{x}_i = \mathbf{z}_i \mathbf{B}$$

- $\mathbf{x}_i \in \mathbb{R}^D$ is a data case in the high-dimensional data space.
- $\mathbf{z}_i \in \mathbb{R}^K$ is the representation of $\mathbf{x}_i$ in the lower-dimensional data space, also called the embedding space.
- $\mathbf{B}$ is a $K \times D$ matrix of basis vectors parameterizing a K-dimensional linear sub-space of $\mathcal{R}^D$.

Matrix Form

- If we let $\mathbf{X}$ be a data matrix where the i^{th} row is $\mathbf{x}_i$ and $\mathbf{Z}$ be a matrix of embeddings where the i^{th} row is $\mathbf{z}_i$, we can define $\mathbf{X}$ under the linear sub-space assumption as follows:

$$\mathbf{X} = \mathbf{ZB}$$

- Most real world data will be subject to noise. If we assume that $\epsilon \in \mathbb{R}^{N \times D}$ is a matrix of noise values from some probability distribution, we have:

$$\mathbf{X} = \mathbf{ZB} + \epsilon$$

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:

$$\hat{\mathbf{Z}}, \hat{\mathbf{B}} = \underset{\mathbf{Z},\mathbf{B}}{\arg\min} \, ||\mathbf{X} - \mathbf{ZB}||_F$$

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

Learning: SGD

- The obvious learning algorithm to apply is a version of stochastic gradient descent (e.g., Adam).
- We select a subset of data cases for each batch and compute a stochastic gradient based on the reconstruction error for that batch.
- With modern tools like PyTorch, we can specify the model and objective function and use automatic differentiation to obtain gradients.
- Each gradient step will update all of the parameters in the $\mathbf{B}$ matrix, and the rows of $\mathbf{Z}$ for data cases that are included in the batch.

Learning: ALS

- An alternative approach to learning is obtained by leveraging the OLS solution to linear regression. The algorithm is often referred to as Alternating Least Squares or ALS.
- Starting from a random initialization, ALS iterates between assuming $\mathbf{Z}$ are known features and optimizing $\mathbf{B}$ as the unknown weights, and assuming that $\mathbf{B}$ are the known features and optimizing $\mathbf{Z}$ as the unknown weights:

$$\mathbf{B} \leftarrow (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X}$$

$$\mathbf{Z}^\top \leftarrow (\mathbf{B}\mathbf{B}^\top)^{-1} \mathbf{B}\mathbf{X}^\top$$

Non-Uniqueness of Solution

- Suppose we learn the model to convergence (using any method) and obtain estimates for $\hat{\mathbf{Z}}$ and $\hat{\mathbf{B}}$.
- Now let $\mathbf{R}$ be any invertible matrix and define $\tilde{\mathbf{Z}} = \hat{\mathbf{Z}}\mathbf{R}$ and $\tilde{\mathbf{B}} = \mathbf{R}^{-1}\hat{\mathbf{B}}$.
- We can easily see that if $\hat{\mathbf{Z}}$ and $\hat{\mathbf{B}}$ are solutions to the learning problem, so are $\tilde{\mathbf{Z}}$ and $\tilde{\mathbf{B}}$ since:

$$\tilde{\mathbf{Z}}\tilde{\mathbf{B}} = (\hat{\mathbf{Z}}\mathbf{R})(\mathbf{R}^{-1}\hat{\mathbf{B}}) = \hat{\mathbf{Z}}(\mathbf{R}\mathbf{R}^{-1})\hat{\mathbf{B}} = \hat{\mathbf{Z}}\hat{\mathbf{B}}$$

- Interestingly, this optimization problem has a continuous subspace of solutions that all obtain the same global minimum value of the objective function.
- Each optimal solution is simply a representation the same linear subspace using different basis vectors.

Inference

- So far, we have only seen how to learn the low-dimensional embeddings $\mathbf{z}_i$ along with the basis matrix $\mathbf{B}$ during the model training phase.
- What happens if we have a new data case $\mathbf{x}_*$ and needs to compute its embedding?
- Given the value of $\hat{\mathbf{B}}$, the embedding $\hat{\mathbf{z}}_*$ for $\mathbf{x}_*$ can be obtaind by solving the optimization problem:

$$\hat{\mathbf{z}}_* = \arg\min_{\mathbf{z}} ||\mathbf{x}_* - \mathbf{z}\hat{\mathbf{B}}||_F$$

- The solution is again available in closed form:

$$\hat{\mathbf{z}}_* = \left((\hat{\mathbf{B}}\hat{\mathbf{B}}^\top)^{-1}\hat{\mathbf{B}}\mathbf{x}_*^\top\right)^\top$$

Singular Value Decomposition

- Classical Rank-K Singular Value Decomposition (K-SVD) is another equivalent approach to dimensionality reduction using the following decomposition:

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

- $\mathbf{S}$ is a $K \times K$ diagonal matrix with positive elements listed in decreasing order.
- $\mathbf{U}$ is an $N \times K$ orthonormal matrix such that $\mathbf{U}^\top\mathbf{U} = I$
- $\mathbf{V}$ is a $D \times K$ orthonormal matrix such that $\mathbf{V}^\top\mathbf{V} = I$.
- Under these constraints, the parameters are unique so long as all of the diagonal elements $\mathbf{S}_{ii}$ (the singnular values) are unique.

Singular Value Decomposition

- To convert back to our original two-factor representation, we use the following mapping:

$$\hat{\mathbf{Z}} = \mathbf{US}$$

$$\hat{\mathbf{B}} = \mathbf{V}^\top$$

- This choice ensures that $\hat{\mathbf{B}}$ is an orthonormal representation of the K-dimensional linear subspace of $\mathbb{R}^D$.
- Specifically, since $\mathbf{V}^\top\mathbf{V} = I$, we have $\hat{\mathbf{B}}\hat{\mathbf{B}}^\top = I$.

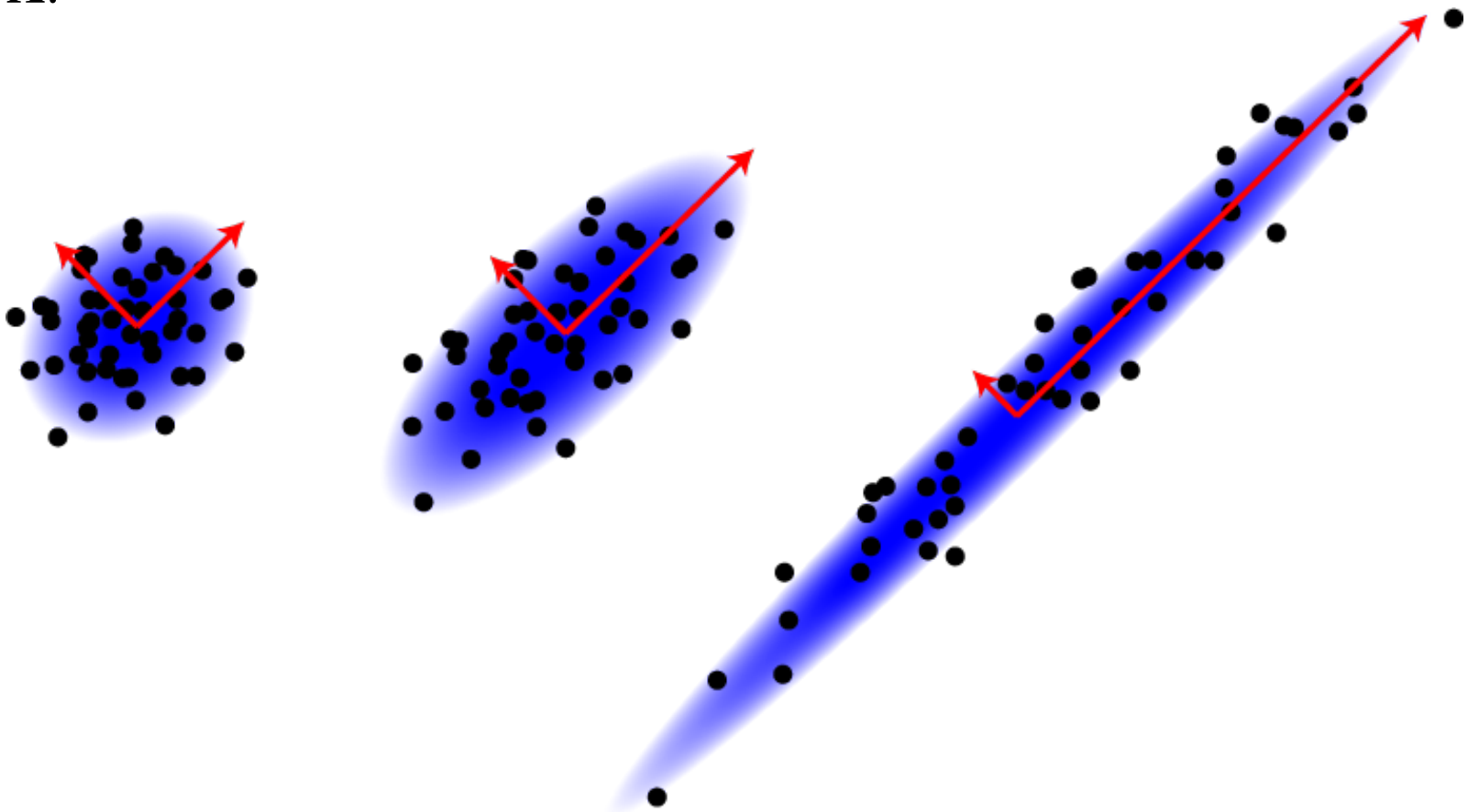
Inference

- In the special case where we fit the model using SVD, the orthonormal property $\hat{\mathbf{B}}\hat{\mathbf{B}}^\top = I$ leads to a simplified equation for dimensionality reduction of new data:

$$
\begin{aligned}
\hat{\mathbf{z}}_* &= \left((\hat{\mathbf{B}}\hat{\mathbf{B}}^\top)^{-1} \hat{\mathbf{B}}\mathbf{x}_*^\top \right)^\top \\
&= \left((I)^{-1} \hat{\mathbf{B}}\mathbf{x}_*^\top \right)^\top \\
&= \left(\hat{\mathbf{B}}\mathbf{x}_*^\top \right)^\top \\
&= \mathbf{x}_* \mathbf{B}^T
\end{aligned}
$$

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 an orthonormal basis $\mathbf{B} \in \mathbb{R}^{K \times D}$ corresponding to the the K directions of maximum variation in $\mathbf{X}$.



Principal Component Analysis

- Once the maximum variation basis $\mathbf{B}$ has been obtained, the data set can then be represented in terms of a matrix K-dimensional coordinates $\mathbf{Z}$ in this new space, accomplishing dimensionality reduction.
- The solution to the PCA learning problem is to select as the basis $\mathbf{B}$ the K eigenvectors of the $D \times D$ empirical covariance matrix $\mathbf{C} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$ that have the largest eigenvalues.
- The PCA projection into the lower dimensional space is defined to be $\mathbf{Z} = \mathbf{X}\mathbf{B}^T$.
- While this sounds very different compared to error-minimizing reconstruction of $\mathbf{X}$, it turns out to identify exactly the same orthonormal basis as SVD and result in identical low-dimensional representations.

Principal Component Analysis

- Let Σ be a diagonal matrix containing the eigenvalues of $\mathbf{C} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$ in decreasing order and $\mathbf{W}$ be the corresponding basis matrix, the full rank PCA satisfies the identity:

$$\mathbf{C} = \mathbf{B}^\top\Sigma\mathbf{B}$$

- We can equivalently represent $\mathbf{X}$ using the full rank SVD as $X = \mathbf{USV}^\top$. This gives us:

$$\mathbf{C} = \frac{1}{N}\mathbf{VSU}^T\mathbf{USV}^T$$

- Due to orthonormality of $\mathbf{U}$, we get the further simplification:

$$\mathbf{C} = \mathbf{V}\left(\frac{1}{N}\mathbf{S}^2\right)\mathbf{V}^T$$

- This shows that the PCA basis matrix $\mathbf{B} = \mathbf{V}^T$

Summary

- Basic linear dimensionality reduction will converge to an arbitrary representation of the reconstruction error-minimizing K-dimensional linear subspace of $\mathbb{R}^D$ given a data set $\mathbf{X} \in \mathbb{R}^{N \times D}$.
- The reconstruction error-minimizing K-dimensional linear subspace does have a unique representation that can be found using the rank-K singular value decomposition of $\mathbf{X}$.
- The connection to PCA shows that the reconstruction error-minimizing K-dimensional linear subspace and the K-dimensional variance maximizing linear subspace are identical.
- By transitivity, this means Basic linear dimensionality reduction also identifies the K-dimensional variance maximizing linear subspace.

Complexity

- The basic ALS algorithm scales as $O(K^3 + D^3)$ per pair of iterations.
- The full SVD algorithm scales as $O(\min(DN^2, ND^2))$.
- The PCA approach has complexity $O(D^3)$.
- In practice, there are randomized algorithms for computing the rank K SVD and these are used for both PCA and SVD-based dimensionality reduction.
- However, SGD on the Frobenius norm objective is much, much more scalable for large data.
- As with clustering, the rank K of the latent sub-space is a free parameter that can be set using validation set methods, at additional cost.

Limitations

- A significant limitation of linear dimensionality reduction is that it can fail to achieve a useful compression of the data if the underlying manifold is not actually linear.