COMPSCI 589
Lecture 20: Sparse Coding, NMF and ICA

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.

Linear Dimensionality Reduction

- 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).

PCA and SVD

- PCA on $\mathbf{X}^T\mathbf{X}$ and SVD on $\mathbf{X}$ identify exactly the same linear sub-space and result in exactly the same projection of the data into that linear 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.
- SVD and PCA provide the same refinement of generic linear dimensionality reduction: an orthogonal basis for exactly the same optimal linear subspace.
- To extend PCA and SVD to the non-linear case, we can use basis expansions or kernels (next class).

Limitations

- PCA and SVD constrain the basis elements to be orthonormal.
- In some cases we may want to extract representations where the basis elements and factor loadings are non-negative, representations where the factor loadings are maximally independent, or representations where the factor loadings are sparse.
- The reason is that these constraints may better model the process that generates the data. These constraints may also help with recognition tasks.

Sparse Coding

- Sparse coding is an extension of linear dimensionality reduction where the factor loadings are constrained to be sparse.
- This model is closely related to the Lasso (ℓ_1 regularized linear regression).
- This gives rise to the following optimization problem:

$$\min_{\mathbf{Z},\mathbf{B}} ||\mathbf{X} - \mathbf{ZB}||_F - \lambda||\mathbf{Z}||_1$$

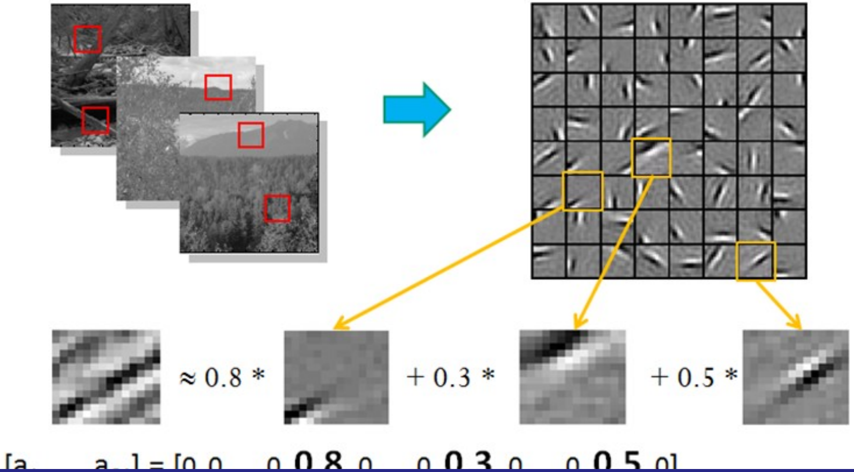
$$\text{such that } ||B_k||_2 = 1 \text{ for all } k$$

where $||\mathbf{A}||_1$ is the sum of the absolute values of the elements in $\mathbf{A}$ and $||\mathbf{A}||_F$ is the sum of the squares of the elements in $\mathbf{A}$.

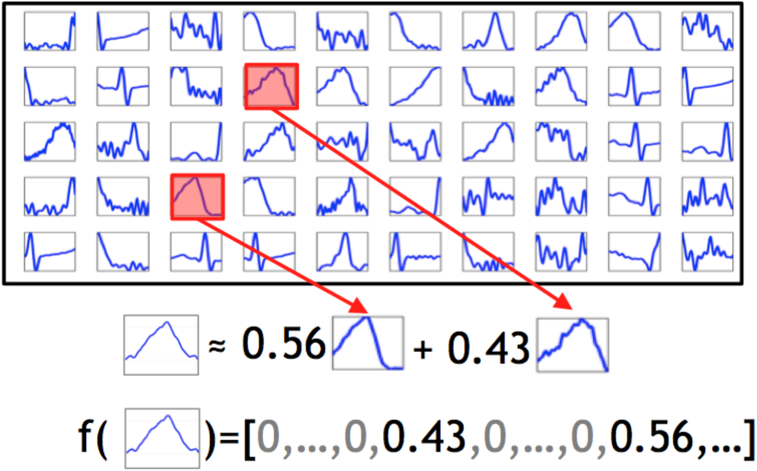
Motivation

- By the early 2000's several theoretical, computational, and experimental studies suggested that neurons encode sensory information using a small number of active neurons at any given point in time, a strategy that was named sparse coding in the computational neuroscience literature.
- Olshausen and Field (2004) argued that sparse coding makes the structure in natural signals more apparent, represents complex data in a way that is easier to read out in later stages of processing, and saves energy.
- As λ increases, the representation becomes sparser, typically using a small number of the K available basis vectors to encode each signal. By comparison, the PCA representation of a natural signal normally puts non-zero weight on all basis elements.

Example: Image Patches



Example: Time Series



Non-Negative Matrix Factorization

- NMF is an extension of linear dimensionality reduction where the factor loadings and the basis elements are constrained to be positive.
- This gives rise to the following optimization problem:

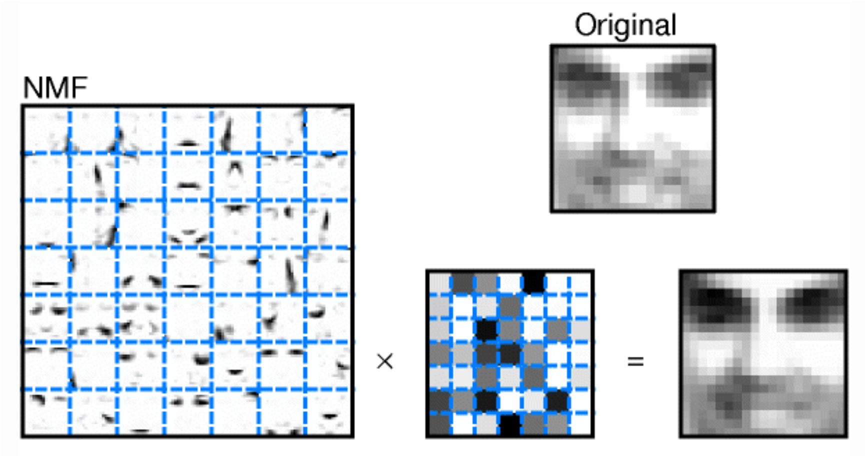
$$\min_{\mathbf{Z},\mathbf{B}} ||\mathbf{X} - \mathbf{Z}\mathbf{B}||_F$$

$$\text{such that } \mathbf{B} \geq 0, \mathbf{Z} \geq 0$$

Motivation

- Data including natural images, gene expressions, and word count representations of text are naturally non-negative.
- In many cases, complex non-negative data arise from a non-negative composition of simpler non-negative parts.
- This is exactly the intuition that non-negative matrix factorization is designed to capture.

Example: Learning Face Parts



Independent Components Analysis

- ICA is an extension of linear dimensionality reduction where the random variables that represent the factor loadings are constrained to be problematically independent of each other.
- This gives rise to the following optimization problem:

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

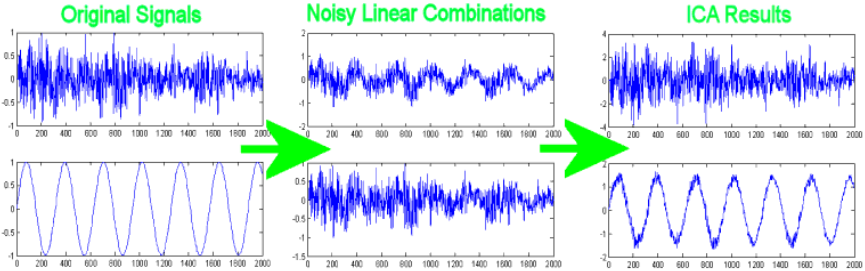
$$\text{such that } Z_i \perp Z_j \text{ for all } 1 < i < j < k$$

- In practice, a surrogate criterion must be used in place of independence and a number of different functions have been explored in the literature.

Motivation

- Linear mixing of independent sources is exactly what occurs when you listen to multiple audio sources at the same time.
- Humans are somehow able to automatically de-mix multiple sources of audio (multiple people speaking) into distinct source channels very accurately.
- ICA was designed to solve exactly this problem (called blind source separation) and can do so very reliably when the number of observed linearly mixed channels is equal to the number of sources.
- The method has also been applied to images and many other types of data.

Example: Blind Source Separation



Example: Independent Components of Natural Images