SVD and PCA

Mohammad Emtiyaz Khan
EPFL

Nov 19, 2015

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Motivation

Principal component analysis (PCA) is a popular method for dimensionality reduction. It is also the simplest example of a latent factor model. It is very similar to matrix factorization and can be obtained using singular value decomposition (SVD).

PCA can be seen as a method that minimizes the reconstruction error or maximizes the variance of the projection, as well as a method to decorrelate the data.

Matrix factorization and PCA

In matrix factorization, we compute an approximation $\mathbf{X} \approx \widetilde{\mathbf{X}} = \mathbf{W}\mathbf{Z}^T$. If we restrict columns of $\mathbf{W}$ to be orthogonal, then the factorization is equivalent to PCA. This is also a regularizer similar to an L_2 regularizer used in the alternating least-squares algorithm.

SVD

Such orthogonal factorization can be obtained using SVD:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices of size $D \times D$ and $N \times N$ respectively, and $\mathbf{S}$ is a diagonal matrix of size $D \times N$ with non-negative entries which are called singular values. Columns of $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors, respectively.

The singular values appear in a descending order in $\mathbf{S}$, i.e. we have $s_1 \geq s_2 \geq s_3 \ldots$, where s_i is the i'th singular value.

We let $\mathbf{W} = \mathbf{U}\mathbf{S}^{1/2}$ and $\mathbf{Z} = \mathbf{V}\mathbf{S}^{1/2}$ to obtain the low rank approximation. This minimizes the reconstruction error (a result known as the Eckart-Young theorem).

Spectral view of SVD

Assuming $D < N$, we can express SVD as follows

$$\mathbf{X} = \sum_{j=1}^{D} s_j \mathbf{u}_j \mathbf{v}_j^T$$

Easy to see that, for all j, $\mathbf{X}\mathbf{v}_j = s_j \mathbf{u}_j$. Note the similarity to the eigenvalue decomposition. Zero singular values correspond to the basis vector in the null space.

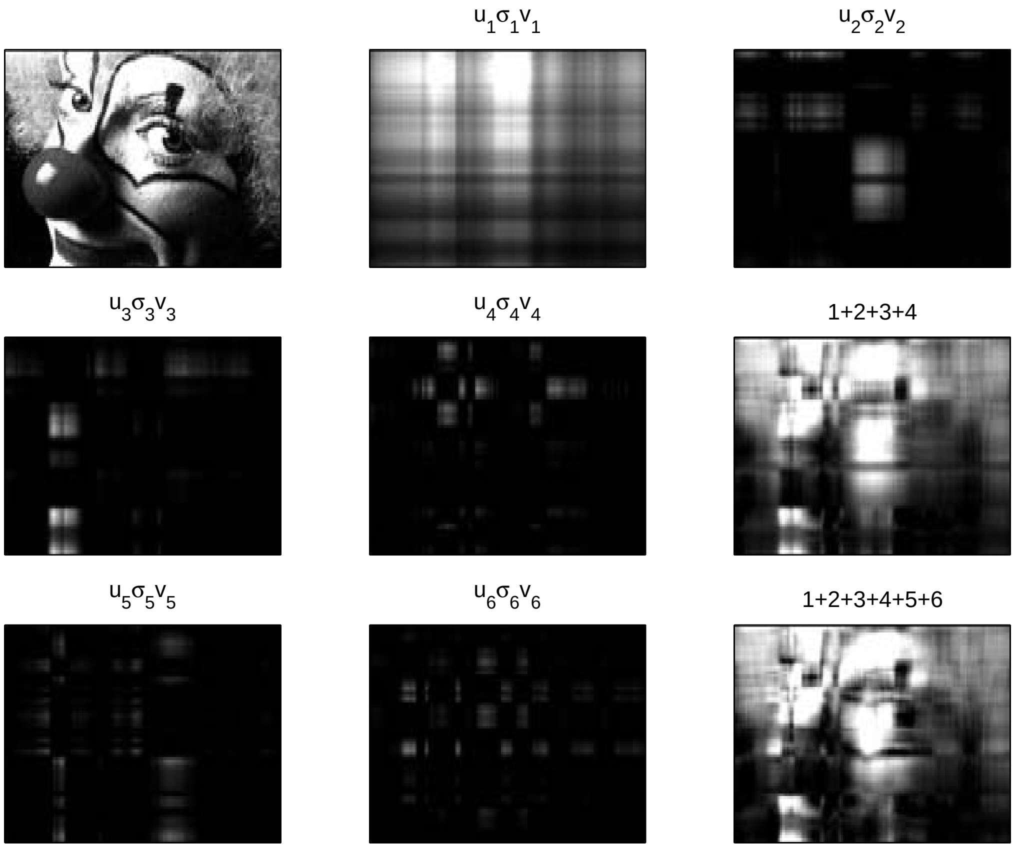
Since s_j are ordered, this tell you about the *spectrum* of $\mathbf{X}$, where higher singular vectors contain the *low-frequency information* and lower singular values contain the *high-frequency information.*

Therefore, if you have a reason to believe that the low-frequency content contains more useful *information* than the high-frequency content, then a low-rank approximation is justified.

An example

The following example is taken from lecture notes of Nando De Frietas's.

```
[U,S,V] = svd(X);
imshow(U(:,1:M)*S(1:M,1:M)*V(:,1:M)')
```



PCA and decorrelation

Define the sample *mean* and sample *covariance matrix* of the data vector $\mathbf{x}_n$ as follows:

$$\bar{\mathbf{x}} := \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \quad , \quad \mathbf{S} := \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$$

If $\mathbf{x}_n$ are i.i.d. samples drawn from some $p(\mathbf{x})$, then the sample mean and covariance will indeed converge to the true mean and covariance of $p(\mathbf{x})$ as $N \to \infty$.

Suppose that $\bar{\mathbf{x}} = 0$, i.e. the data is zero mean (or centered). Then, $\mathbf{S} = \frac{1}{N}\mathbf{X}\mathbf{X}^T$. Using SVD, we can write the following:

$$\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{S}^2\mathbf{U}^T$$

Multiplying the left by $\mathbf{U}^T$ and the right by $\mathbf{U}$, we get the following:

$$\mathbf{U}^T\mathbf{X}\mathbf{X}^T\mathbf{U} = \mathbf{S}^2$$

The columns of matrix $\mathbf{U}$ are called the principal components and they *decorrelate* the covariance matrix. The matrix $\mathbf{U}$ can also be used to visualize the factors.

To do

1. Read Section 12.1.1 and 12.1.2 of Bishop. Understand the two viewpoints: maximizing variance and minimizing reconstruction error.

2. Read Section 12.1.4 of Bishop to learn about the computational complexity of PCA.