Dimensionality Reduction Techniques

Adrian M. Peter

May 12, 2014

Caveat

These are my personal notes and as such I am a bit lackadaisical when it comes to referencing everything. Hence, I want to *clearly state* that some parts of these notes are *verbatim copies* of material from the various sources I drew upon. Sometimes, I made very little attempt to change the wording, equations, etc. from the original source. At other moments, I have put in my own additional insight, which may or may not serve to elucidate the discussion at hand. Here's an attempt to list the major sources: [1]

1 Principal Component Analysis

Principal component analysis, or PCA, is a technique that is widely used for applications such as dimensionality reduction, lossy data compression, feature extraction, and data visualization. It is also known as the *Karhunen-Loéve transform*. There are two commonly used definitions of PCA that give rise to the same algorithm. PCA can be defined as *the orthogonal projection of the data onto a lower dimensional linear space*, known as the principal subspace, such that the variance of the projected data is maximized (Hotelling, 1933)

$$\max_U Var(\hat{X}) = \frac{1}{N}\sum_{i=1}^{N} \left\| U^T(x_i - \bar{x}) \right\|^2 .$$

Equivalently, it can be defined as *the linear projection that minimizes the average projection cost*, defined as the mean squared distance between the data points and their projections (Pearson, 1901)

$$\min_U err(\hat{X}) = \frac{1}{N}\sum_{i=1}^{N} \left\| (x_i - \bar{x}) - UU^T(x_i - \bar{x}) \right\| .$$

[1] We consider each of these definitions in turn.

[1] Confusion sometimes arises when you see the projection written as $U^T x$ vs. $UU^T x$. $U^T x$ gives you the just the coordinates in the projected basis. These coordinates may be less than the original dimension of the data if projecting on to a subspace. $UU^T x$ on the other hand uses the coordinates from $U^T x$ to expand (represent) the projectED vector in the subspace basis U; this will have the same dimensions as the original data.

1.1 Maximum Variance Formulation

Consider a data set of observations $\{x_i\}$ where $i = 1 \ldots N$, and x_i is a Euclidean variable with dimensionality D, i.e. $x_i \in \mathbb{R}^D$. Our goal is to project the data onto a space having dimensionality $M < D$ while maximizing the variance of the projected data. For the moment, we shall assume that the value of M is given.

To begin with, consider the projection onto a one-dimensional space ($M = 1$). We can define the direction of this space using a D-dimensional vector u_1, which for convenience (and without loss of generality) we shall choose to be a unit vector so that $u_1^T u_1 = 1$ (note that we are only interested in the direction defined by u_1, not in the magnitude of u_1 itself). Each data point x_i is then projected onto a scalar value $u_1^T x_i$. The mean of the projected data is $u_1^T \bar{x}$ where $\bar{x}$ is the sample set mean given by

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

and the variance of the projected data is given by

$$\frac{1}{N} \sum_{i=1}^{N} \left(u_1^T x_i - u_1^T \bar{x} \right)^2 . \tag{1}$$

We can write this in a more compact matrix representation. To do this, first notice

$$\begin{aligned}
\left(u_1^T x_i - u_1^T \bar{x} \right)^2 &= \left(u_1^T x_i - u_1^T \bar{x} \right) \left(u_1^T x_i - u_1^T \bar{x} \right) \\
&= u_1^T x_i u_1^T x_i - u_1^T x_i u_1^T \bar{x} - u_1^T \bar{x} u_1^T x_i + u_1^T \bar{x} u_1^T \bar{x} \\
&= u_1^T x_i x_i^T u_1 - u_1^T x_i \bar{x}^T u_1 - u_1^T \bar{x} x_i^T u_1 + u_1^T \bar{x} \bar{x}^T u_1 \\
&= u_1^T \left(x_i x_i^T - x_i \bar{x}^T - \bar{x} x_i^T + \bar{x} \bar{x}^T \right) u_1 \\
&= u_1^T \left[\left(x_i - \bar{x} \right) \left(x_i^T - \bar{x}^T \right) \right] u_1 \\
&= u_1^T \left[\left(x_i - \bar{x} \right) \left(x_i - \bar{x} \right)^T \right] u_1,
\end{aligned}$$

and putting this back into eq. (1) yields

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^{N} \left(u_1^T x_i - u_1^T \bar{x} \right)^2 &= \frac{1}{N} \sum_{i=1}^{N} u_1^T \left[\left(x_i - \bar{x} \right) \left(x_i - \bar{x} \right)^T \right] u_1 \\
&= u_1^T \left[\frac{1}{N} \sum_{i=1}^{N} \left(x_i - \bar{x} \right) \left(x_i - \bar{x} \right)^T \right] u_1 \\
&= u_1^T S u_1,
\end{aligned}$$

where S is the data covariance matrix

$$S = \frac{1}{N} \sum_{i=1}^{N} \left(x_i - \bar{x} \right) \left(x_i - \bar{x} \right)^T .$$

We now maximize the projected variance $u_1^T S u_1$ with respect to u_1. This has to be a constrained maximization to prevent $\|u_1\| \to \infty$. The appropriate constraint comes from the normalization condition $\|u_1\| = u_1^T u_1 = 1$. To enforce this constraint, we introduce a Lagrange multiplier that we shall denote by λ_1, and then make an unconstrained maximization of

$$u_1^T S u_1 + \lambda_1(1 - u_1^T u_1).$$

By setting the derivative with respect to u_1 equal to zero, we see that this quantity will have a stationary point when[2]

$$S u_1 = \lambda_1 u_1$$

which says that u_1 must be an eigenvector of S. If we left-multiply by u_1^T and make use of $u_1^T u_1 = 1$, we see that the variance is given by

$$u_1^T S u_1 = \lambda_1$$

and so the variance will be a maximum when we set u_1 equal to the eigenvector having the largest eigenvalue λ_1. This eigenvector is known as the first *principal component*.

We can define additional principal components in an incremental fashion by choosing each new direction to be that which maximizes the projected variance amongst all possible directions orthogonal to those already considered. If we consider the general case of an M-dimensional projection space, the optimal linear projection for which the variance of the projected data is maximized is now defined by the M eigenvectors $u_1, \ldots, u_M$ of the data covariance matrix S corresponding to the M largest eigenvalues $\lambda_1, \ldots, \lambda_M$. This is can be shown using proof by induction.

To summarize, *principal component analysis* involves evaluating the mean $\bar{x}$ and the covariance matrix S of the data set and then finding the M eigenvectors of S corresponding to the M largest eigenvalues. Algorithms for finding eigenvectors and eigenvalues, as well as additional theorems related to eigenvector decomposition, can be found in Golub and Van Loan (1996). Note that the computational cost of computing the full eigenvector decomposition for a matrix of size $D \times D$ is $O(D^3)$. If we plan to project our data onto the first M principal components, then we only need to find the first M eigenvalues and eigenvectors. This can be done with more efficient techniques, such as the power method (Golub and Van Loan, 1996), that scale like $O(MD^2)$, or alternatively we can make use of the Expectation-Maximization algorithm.

[2] Derivation:

$$\max_{u_1} u_1^T S u_1 + \lambda_1(1 - u_1^T u_1)$$

$$\begin{aligned} \frac{\partial}{\partial u_1} u_1^T S u_1 + \lambda_1(1 - u_1^T u_1) &= 2 S u_1 - 2\lambda_1 u_1 = 0 \\ \implies S u_1 &= \lambda u_1. \end{aligned}$$

1.2 Minimum Error Formulation

We now discuss an alternative formulation of PCA based on projection error minimization. To do this, we introduce a complete orthonormal set of D-dimensional basis vectors $\{u_j\}$ where $j = 1, \ldots, D$ that satisfy

$$u_j^T u_j = \delta_{ij}.$$

Because this basis is complete, each data point can be represented exactly by a linear combination of the basis vectors

$$x_i = \sum_{j=1}^{D} \alpha_{ij} u_j$$

where the coefficients α_{ij} will be different for different data points. This simply corresponds to a rotation of the coordinate system to a new system defined by the $\{u_j\}$, and the original D components $\{x_{i1}, \ldots, x_{iD}\}$ are replaced by an equivalent set $\{\alpha_{i1}, \ldots, \alpha_{iD}\}$. Taking the inner product with u_j, and making use of the orthonormality property, we obtain $\alpha_{ij} = x_i^T u_j$, so without loss of generality we can write

$$x_i = \sum_{j=1}^{D} \left(x_i^T u_j\right) u_j. \tag{2}$$

Our goal, however, is to approximate this data point using a representation involving a restricted number $M < D$ of variables corresponding to a projection onto a lower-dimensional subspace. The M-dimensional linear subspace can be represented, without loss of generality, by the first M of the basis vectors, and so we approximate each data point x_i by

$$\hat{x}_i = \sum_{j=1}^{M} z_{ij} u_j + \sum_{j=M+1}^{D} b_j u_j \tag{3}$$

where the $\{z_{ij}\}$ depend on the particular data point, whereas the $\{b_j\}$ are constants that are the same for all data points. We are free to choose the $\{u_j\}$, the $\{z_{ij}\}$, and the $\{b_j\}$ so as to minimize the distortion introduced by the reduction in dimensionality. As our distortion measure, we shall use the squared distance between the original data point x_i and its approximation $\hat{x}_i$, averaged over the data set, so that our goal is to minimize

$$J = \frac{1}{N} \sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2 .$$

Consider first of all the minimization with respect to the quantities $\{z_{ij}\}$. Substituting for $\hat{x}_i$, setting the derivative with respect to z_{ij} to zero, and making use of the orthonormality conditions, we obtain

$$z_{ij} = x_i^T u_j \tag{4}$$

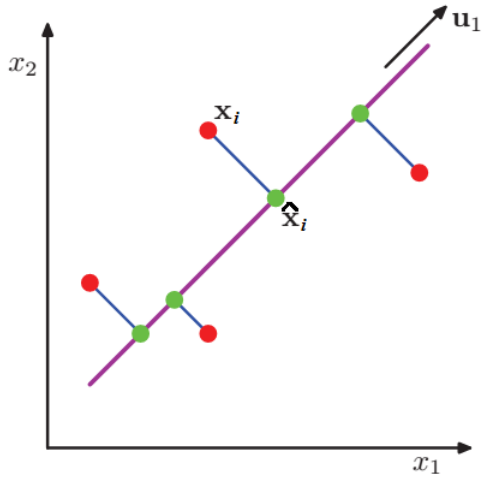

Figure 1: PCA projection. Principal component analysis seeks a space of lower dimensionality, known as the principal subspace and denoted by the magenta line, such that the orthogonal projection of the data points (red dots) onto this subspace maximizes the variance of the projected points (green dots). An alternative definition of PCA is based on minimizing the sum-of-squares of the projection errors, indicated by the blue lines.

where $j = 1, \ldots, M$. Similarly, setting the derivative of J with respect to b_j to zero, and again making use of the orthonormality relations, gives

$$b_j = \bar{x}^T u_j \tag{5}$$

where $j = M+1, \ldots, D$. If we substitute for z_{ij} and b_j, and make use of the general expansion in eq. (2), we obtain

$$x_i - \hat{x}_i = \sum_{j=M+1}^{D} \left[(x_i - \bar{x})^T u_j \right] u_j$$

from which we see that the displacement vector from x_i to $\hat{x}_i$ lies in the space orthogonal to the principal subspace, because it is a linear combination of $\{u_j\}$ for $j = M+1, \ldots, D$ (see Fig. 1). This is to be expected because the projected points $\hat{x}_i$ must lie within the principal subspace, but we can move them freely within that subspace, and so the minimum error is given by the orthogonal projection.

We therefore obtain an expression for the distortion measure J as a function

purely of the $\{u_j\}$ in the form

$$\begin{aligned} J &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=M+1}^{D}\left(x_i^T u_j - \bar{x}^T u_j\right) \\ &= \sum_{j=M+1}^{D} u_j^T S u_j. \end{aligned}$$

There remains the task of minimizing J with respect to the $\{u_j\}$, which must be a constrained minimization otherwise we will obtain the vacuous result $u_j = 0$. The constraints arise from the orthonormality conditions and, as we shall see, the solution will be expressed in terms of the eigenvector expansion of the covariance matrix. Before considering a formal solution, let us try to obtain some intuition about the result by considering the case of a two-dimensional data space $D = 2$ and a one-dimensional principal subspace $M = 1$. We have to choose a direction u_2 so as to minimize $J = u_2^T S u_2$, subject to the normalization constraint $u_2^T u_2 = 1$. Using a Lagrange multiplier λ_2 to enforce the constraint, we consider the minimization of

$$\hat{J} = u_2^T S u_2 + \lambda_2\left(1 - u_2^T u_2\right).$$

Setting the derivative with respect to u_2 to zero, we obtain $Su_2 = \lambda_2 u_2$ so that u_2 is an eigenvector of S with eigenvalue λ_2. Thus any eigenvector will define a stationary point of the distortion measure. To find the value of J at the minimum, we back-substitute the solution for u_2 into the distortion measure to give $J = {\lambda_2}^3$. We therefore obtain the minimum value of J by choosing u_2 to be the eigenvector corresponding to the smaller of the two eigenvalues. Thus we should choose the principal subspace to be aligned with the eigenvector having the larger eigenvalue. This result accords with our intuition that, in order to minimize the average squared projection distance, we should choose the principal component subspace to pass through the mean of the data points and to be aligned with the directions of maximum variance. For the case when the eigenvalues are equal, any choice of principal direction will give rise to the same value of J.

The general solution to the minimization of J for arbitrary D and arbitrary $M < D$ is obtained by choosing $\{u_j\}$ to be eigenvectors of the covariance matrix given by

$$Su_i = \lambda_i u_i$$

---

[3]

$$\begin{aligned} \hat{J} &= u_2^T S u_2 + \lambda_2\left(1 - u_2^T u_2\right) \\ &= u_2^T \lambda_2 u_2 + \lambda_2\left(1 - u_2^T u_2\right) \\ &= \lambda_2 u_2^T u_2 \\ &= \lambda_2. \end{aligned}$$

where $i = 1, \ldots, D$, and as usual the eigenvectors $\{u_j\}$ are chosen to be orthonormal. The corresponding value of the distortion measure is then given by

$$J = \sum_{i=M+1}^{D} \lambda_i$$

which is simply the sum of the eigenvalues of those eigenvectors that are orthogonal to the principal subspace. We therefore obtain the minimum value of J by selecting these eigenvectors to be those having the $D-M$ smallest eigenvalues, and hence the eigenvectors defining the principal subspace are those corresponding to the M largest eigenvalues.

Although we have considered $M < D$, the PCA analysis still holds if $M = D$, in which case there is no dimensionality reduction but simply a rotation of the coordinate axes to align with principal components. This represents a compression of the data set, because for each data point we have replaced the D-dimensional vector x_i with an M-dimensional vector having components $\left(x_i^T u_j - \bar{x}^T u_j\right)$. The smaller the value of M, the greater the degree of compression.

1.3 PCA Reconstruction

If we substitute eq. (4) and (5) int eq. (3), we can write the PCA approximation to a data vector x_i in the form

$$\begin{aligned} \hat{x}_i &= \sum_{j=1}^{M} \left(x_i^T u_j\right) u_j + \sum_{j=M+1}^{D} \left(\bar{x}^T u_j\right) u_j \\ &= \bar{x} + \sum_{j=1}^{M} \left(x_i^T u_j - \bar{x}^T u_j\right) u_j \end{aligned}$$

where we have used the fact

$$\bar{x} = \sum_{j=1}^{D} \left(\bar{x}^T u_j\right) u_j$$

which follows from the completeness of the basis $\{u_j\}$.

1.4 Whitening

Another application of principal component analysis is to data pre-processing. In this case, the goal is not dimensionality reduction but rather the transformation of a data set in order to standardize certain of its properties. This can be important in allowing subsequent pattern recognition algorithms to be applied successfully to the data set. Typically, it is done when the original variables are measured in various different units or have significantly different variability. One type of normalization to combat this measurement discrepancy

is to perform separate linear re-scaling of the individual variables such that each variable had zero mean and unit variance (independently normalize each dimension). This is known as *standardizing* the data, and the covariance matrix for the standardized data has components

$$\rho_{lk} = \frac{1}{N} \sum_{i=1}^{N} \frac{(x_{il} - \bar{x}_l)}{\sigma_l} \frac{(x_{ik} - \bar{x}_k)}{\sigma_k}$$

where σ_l is the variance of x_l. This is known as the *correlation matrix* of the original data and has the property that if two components x_i and x_j of the data are perfectly correlated, then $\rho_{lk} = 1$, and if they are uncorrelated, then $\rho_{lk} = 0$.

However, using PCA we can make a more substantial normalization of the data to give it zero mean and unit covariance, so that different variables become decorrelated. To do this, we first write the eigenvector equation in the form

$$SU = U\Lambda$$

where Λ is a $D \times D$ diagonal matrix with elements λ_i, and U is a $D \times D$ orthogonal matrix with columns given by u_i. Then we define, for each data point x_i, a transformed value given by

$$y_i = \Lambda^{-\frac{1}{2}} U^T (x_i - \bar{x})$$

where $\bar{x}$ is the sample mean. The set of new transformed data $\{y_i\}$ has zero mean and its covariance is given by the identity matrix

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^{N} y_i y_i^T &= \frac{1}{N} \sum_{i=1}^{N} \Lambda^{-\frac{1}{2}} U^T (x_i - \bar{x})(x_i - \bar{x})^T U \Lambda^{-\frac{1}{2}} \\ &= \Lambda^{-\frac{1}{2}} U^T S U \Lambda^{-\frac{1}{2}} \\ &= \Lambda^{-\frac{1}{2}} \Lambda \Lambda^{-\frac{1}{2}} \\ &= I. \end{aligned}$$

This operation is known as *whitening* or *sphereing* the data.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.