Some Intuitions behind PCA

William W. Cohen

October 30, 2013

1 The covariance matrix C_X

Consider a data matrix X where the t-th row corresponds to an instance $\mathbf{x}^t$, with n instances each with m features.

$$X = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_m^1 \\ \vdots & \ddots & & \vdots \\ x_1^n & \dots & & x_m^n \end{bmatrix} = \begin{bmatrix} \vdots \\ - \ \mathbf{x}^t \ - \\ \vdots \end{bmatrix}$$

Sometimes, to remind myself I'm talking about feature values, I will use f_j^t to denote the j-th feature of $\mathbf{x}^t$ (aka, $X(t, j)$) Likewise I will use $\mathbf{f}_i$ to denote the i-th column vector of X: the vector of all values taken on by the i-th feature of the examples.

$$X = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_m^1 \\ \vdots & \ddots & & \vdots \\ x_1^n & \dots & & x_m^n \end{bmatrix} = \begin{bmatrix} & | & \\ \dots & \mathbf{f}_i & \dots \\ & | & \end{bmatrix}$$

If you like, $\mathbf{f}_i$ is a "signature" of the i-th feature on the dataset X. Sometimes I will use F_i for the corresponding random variable.

A first observation: consider the matrix $C_X = X^T X$, whose entries are the pairwise inner products, not of the instances $\mathbf{x}^t$, but of the column vectors $\mathbf{f}_i$ of the matrix X. So the entries of C_X measure the similarity of two features:

$$C_X(i, j) = \sum_t f_i^t f_j^t$$

The notation C_X is chosen because if X has a zero mean, then $\frac{1}{n} C_X(i, j)$ is the *sample covariance* of features i and j on the sample X.

if you're happier thinking about discrete features, then another way of thinking about this is the following: if the features f_i are always $+1$ or -1, then define

$$\text{AGREE}(i,j) = \text{number of examples } \mathbf{x}^t \in X \text{ where } f_i^t \neq f_j^t$$

and define DISAGREE(i,j) analogously. It is also true that

$$C_X(i,j) = \text{AGREE}(i,j) - \text{DISAGREE}(i,j)$$

and that

$$\frac{1}{n} C_X(i,j) = P(f_i^t = f_j^t) - P(f_i^t \neq f_j^t)$$

where the probabilities are the empirical probabilities taken over the sample X. To summarize: $C_X(i,j)$ is (some sort of) measure of agreement between the features $\mathbf{f}_i$ ande $\mathbf{f}_j$, and if for all the features $f_i^t \in \{+1,-1\}$, then $C_X(i,j) \in [-1,+1]$.

2 Eigenvectors of C_X are "consistent predictors"

Suppose I wanted to predict the likely value of F_i from another feature F_j. This is easiest if $C_X(i,j)$ is close to an extreme; then the obvious cases are

- $C_X(i,j) \approx +1$ and $F_j = +1$: predict $F_i = +1$.
- $C_X(i,j) \approx -1$ and $F_j = +1$: predict $F_i = -1$.
- $C_X(i,j) \approx +1$ and $F_j = -1$: predict $F_i = -1$.
- $C_X(i,j) \approx -1$ and $F_j = -1$: predict $F_i = +1$.

On the other hand, if $C_X(i,j) \approx 0$, then it seems that no prediction for F_i can be made confidently. So a simple formula for predicting F_i from F_j using a single real number in $[-1,+1]$, where small predicted values indicate low confidence, might be

$$f_i \text{ is predicted as } C_X(i,j) \cdot f_j$$

Of course, f_i could be predicted just as easily from $f_{j'}$ for $j' \neq j$. If I wanted to combine all of these predictions I might weight them all equally to get

$$f_i \text{ is predicted as } \frac{1}{n} \sum_{j \neq i} C_X(i,j) \cdot f_j \tag{1}$$

Now, suppose I want to predict an entire instance $\mathbf{e}$ that is "likely" according to the sample X. A natural goal is a set of feature values $\mathbf{e} = \langle e_1, \ldots, e_m \rangle$ that are internally consistent with respect to the (confidence-weighted) prediction scheme of Equation 1, i.e., a potential instance $\mathbf{e}$ where

$$\forall i, e_i = \frac{1}{n} \sum_j C_X(i,j) e_j$$

A slightly weaker condition would be that there's some constant λ so that

$$\forall i, \lambda e_i = \frac{1}{n} \sum_j C_X(i,j) e_j$$

or in other words

$$\exists \lambda : \lambda \mathbf{e} = C_X \mathbf{e}$$

or in still other words, $\mathbf{e}$ *is an eigenvector of* C_X.

To summarize: eigenvectors $\mathbf{e}^i$ of C_X are the same length as instances $\mathbf{x}$, and they have the nice property that their feature values are internally consistent with respect to the (admittedly simple-minded) prediction scheme of Equation 1.

The eigenvectors are thus broadly similar to clusters in k-means, or mixture components in a generative model, in that some subsets of the features can be used to predict the other features' values.

3 Using the consistent predictors

Let $\mathbf{e}^1, \ldots, \mathbf{e}^m$ be the eigenvectors of C_X, in decreasing order by their eigenvalues, and let Λ be the diagonal matrix of eigenvalues. Let E be a matrix where the row vectors are the $\mathbf{e}^i$'s, and let E_k be a matrix with just the first k eigenvectors:

$$E = \begin{bmatrix} e_1^1 & e_2^1 & \ldots & e_m^1 \\ \vdots & \ddots & & \vdots \\ e_1^k & \ldots & & e_m^k \end{bmatrix} = \begin{bmatrix} \vdots \\ - \mathbf{e}^i - \\ \vdots \end{bmatrix}$$

Let's consider the matrix product $Z = XE^T$, or more interestingly perhaps, the matrix $Z_k = XE_k^T$:

$$Z_k = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_m^1 \\ \vdots & \ddots & & \vdots \\ x_1^n & \dots & & x_m^n \end{bmatrix} \times \begin{bmatrix} e_1^1 & e_1^2 & \dots & e_k^1 \\ e_2^1 & e_2^2 & \dots & e_k^2 \\ \vdots & \ddots & & \vdots \\ e_m^1 & \dots & & e_k^m \end{bmatrix} = \begin{bmatrix} \mathbf{x}^1 \cdot \mathbf{e}^1 & \mathbf{x}^2 \cdot \mathbf{e}^2 & \dots & \mathbf{x}^1 \cdot \mathbf{e}^k \\ \vdots & \ddots & & \vdots \\ \mathbf{x}^n \cdot \mathbf{e}^1 & \mathbf{x}^n \cdot \mathbf{e}^2 & \dots & \mathbf{x}^n \cdot \mathbf{e}^k \end{bmatrix}$$

This matrix has one row $\mathbf{z}^t$ for each instance $\mathbf{x}^t$, but the columns (features) are different. Instead of the original feature space, we now have the values

$$\mathbf{z}^t = \langle z_1^t, \dots, z_k^t \rangle = \langle \mathbf{x}^t \cdot \mathbf{e}^1, \dots, \mathbf{x}^t \cdot \mathbf{e}^k \rangle$$

If you think of dot-product as kind of similarity score (as I do) then *the i-th feature of $\mathbf{z}^t$ is the similarity of $\mathbf{x}^t$ to the i-th eigenvector* $\mathbf{e}^i$. In otherwords, the instances $\mathbf{x}^t$ have been mapped to a new space where each dimension indicates how similar/different the instance $\mathbf{x}^t$ to some consistently-predictable potential instance.

If you again think of the eigenvectors as similar to clusters, or mixture components, the new space that $\mathbf{x}^t$ has been mapped into is like a space of posteriors for the components.

4 From PCA to SVD

We started out looking at the correlations between the variables of X, by computing the dot-products of the "feature signatures" $\mathbf{f}_i$, via computing $C_X = X^T X$. What if we do the same trick to Z? It turns out the answer is "not much": in particular, the corresponding signatures in Z are not correlated.

Let's use $\mathbf{g}_i$ for the i-th column of Z:

$$Z = XE^T = \begin{bmatrix} \vdots \\ - \ \mathbf{z}^t \ - \\ \vdots \end{bmatrix} = \begin{bmatrix} & | & \\ \dots & \mathbf{g}_i & \dots \\ & | & \end{bmatrix}$$

In $C_Z = Z^T Z$, what do the entries look like? Well,

$$C_Z(i,j) = \mathbf{g}_i \cdot \mathbf{g}_j$$

and, treating $\mathbf{e}^j$ as a column vector, $\mathbf{g}_i = X\mathbf{e}_i$. So

$$\begin{aligned}
C_Z(i,j) &= \mathbf{g}_i \cdot \mathbf{g}_j \\
&= \mathbf{g}_i^T \mathbf{g}_j \\
&= (X\mathbf{e}_i)^T (X\mathbf{e}_j) \\
&= \mathbf{e}_i^T X^T X \mathbf{e}_j \\
&= \mathbf{e}_i^T (X^T X) \mathbf{e}_j \\
&= \mathbf{e}_i^T \lambda_j \mathbf{e}_j \\
&= \lambda_j \mathbf{e}_i^T \mathbf{e}_j
\end{aligned}$$

the last step holding since $\mathbf{e}_j$ is an eigenvector of $X^T X$. If we assume we have scaled the $\mathbf{e}_i$'s to unit L2 norm, and recall that the eigenvectors are all orthogonal, then we finally get that

$$C_Z(i,j) = \begin{cases} \lambda_i & \text{if } i = j \\ 0 & else \end{cases}$$

So Z is a somewhat special transformation of X: in this transformed space, the correlation coefficient between any pair of distinct variables is zero, and the variance of each individual variable is λ_i.

If we like, we can also rescale Z so that it has unit variances as well. Let Σ be a diagonal matrix with $\Sigma(i,i) = \sqrt{\lambda_i}$, and consider $U = Z\Sigma^{-1}$. It's pretty simple to show that $U^T U = I$.

So this suggests some alternative ways to represent the original matrix X. Since $Z = XE^T$, and $Z = U\Sigma$, we have

$$\begin{aligned}
XE^T &= Z \\
X &= ZE \\
X &= U\Sigma E
\end{aligned}$$

So now X is decomposed into a product of three factors,

- U, a unit-variance matrix with uncorrelated features, formed by projecting X using PCA and scaling;
- Σ, a diagonal matrix; and
- E, the matrix of eigenvectors of C_X, aka "to self-consistent predictions".

This is called SVD, the *singular valued decomposition* for X: the "decomposition" is factoring X, and the "singular values" are the diagonal elements of Σ. The more usual notation is

$$X = U\Sigma V$$

An important point is that we can replace Z with Z_k and Σ with Σ_k (the first k rows and columns of Σ). The full decomposition then becomes

$$X \approx Z_k E_k = U_k \Sigma_k E_k = U_k \Sigma_k V_k$$

Note here are Z_k is a tall narrow matrix (n rows and k columns), and hence so is U_k, Σ_k is a square matrix, and E_k (aka V_k) is a long wide matrix k rows and n columns). Again, the rows of Z_k (and hence U_k) corresponding to instances $\mathbf{x}^t$, represented by their similarity to the first few eigenvectors $\mathbf{e}_1,\ldots,\mathbf{e}_k$. The rows of E_k (aka V_k) correspond to hypothetical instances that are "self-consistent" according to C_X.