Ordinary Principal Components

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

SVD for $\boldsymbol{X}$ with centered columns

When all the columns of $\mathbf{X}$ have been centered (each $\mathbf{1}'\mathbf{x}_j = 0$ for $\mathbf{x}_j$ the jth column of $\mathbf{X}$), there is additional terminology and insight associated with SVD as describing the structure of $\mathbf{X}$. Note that centering is often sensible in unsupervised learning contexts because the object is to understand the internal structure of the data cases $\mathbf{x}_i \in \Re^p$, not the location of the data cloud (that is easily represented by the sample mean vector). So accordingly, we first translate the data cloud to the origin.

Principal components ideas are then based on the singular value decomposition of $\mathbf{X}$

$$\underset{N\times p}{\mathbf{X}} = \underset{N\times r}{\mathbf{U}}\ \underset{r\times r}{\mathbf{D}}\ \underset{r\times p}{\mathbf{V}'}$$

(and related spectral/eigen decompositions of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$).

PC directions, scores, and loadings

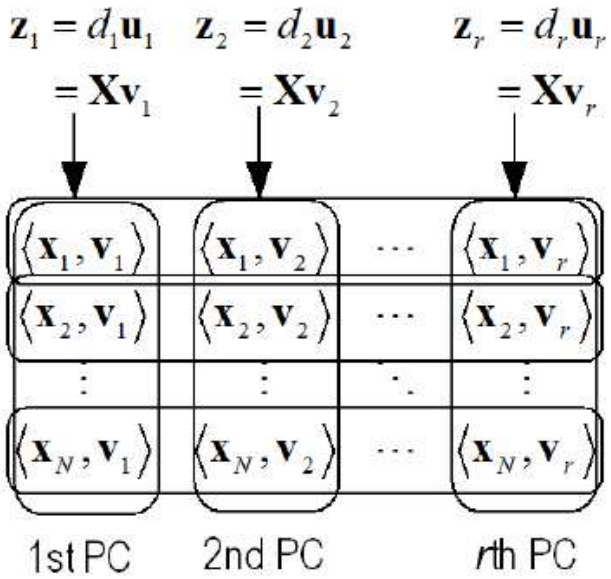
Columns of $\mathbf{V}$ (namely $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$) are called the **principal component directions in** $\Re^p$ of the $\mathbf{x}_i$, and the elements of the vectors

$$\mathbf{z}_j = \mathbf{X}\mathbf{v}_j = \begin{pmatrix} \langle \mathbf{x}_1, \mathbf{v}_j \rangle \\ \vdots \\ \langle \mathbf{x}_N, \mathbf{v}_j \rangle \end{pmatrix} = d_j \mathbf{u}_j$$

namely the $\langle \mathbf{x}_i, \mathbf{v}_j \rangle$, are the **principal components** of the $\mathbf{x}_i$. (The ith element of $\mathbf{z}_j$, $\langle \mathbf{x}_i, \mathbf{v}_j \rangle$, is the value of the jth principal component for case i, or the corresponding **principal component score**. The entries of the $p \times 1$ vector $\mathbf{v}_j$ are the **component weights** or **loadings** for the jth component. A 0 loading means that the corresponding column of $\mathbf{X}$ is ignored in the creation of $\mathbf{z}_j$.) Notice that $\langle \mathbf{x}_i, \mathbf{v}_j \rangle \mathbf{v}_j$ is the projection of $\mathbf{x}_i$ onto the 1-dimensional space spanned by $\mathbf{v}_j$.

Summary of terminology

Below is a summary of the language just introduced. (The $N \times p$ matrix of inner products $\langle \mathbf{x}_i, \mathbf{v}_j \rangle$ is **UD**.)



PC scores for case 1

PC scores for case 2

PC scores for case N

$$\mathbf{v}_j = \begin{pmatrix} v_{j1} \\ v_{j2} \\ \vdots \\ v_{jp} \end{pmatrix}$$

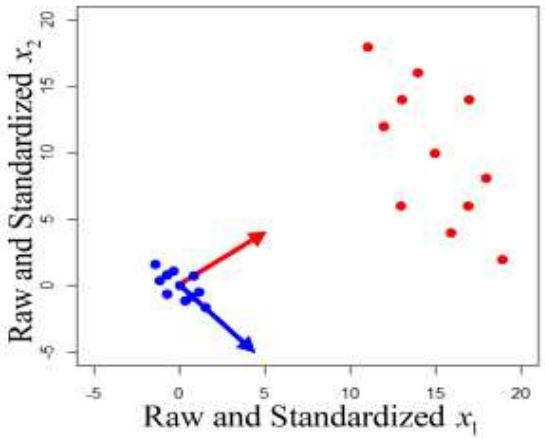
PC j loading on variable 1

PC j loading on variable 2

PC j loading on variable p

Toy *p=2* example

The figure below shows scatterplots of a raw (red dots) and corresponding standardized (blue dots) $p = 2$ dataset. The red arrow points in the direction of the *raw data* first right singular vector (i.e. points "at" the raw data). The blue arrow is in the **first principal component direction** of the *standardized data* (pointing in the direction of their greatest variation).



More Interpretation for SVDs

It is worth thinking a bit more about the form of the product

$$\mathbf{X}^{*l} = \mathbf{U}_l \, \mathbf{diag}\left(d_1, d_2, \ldots, d_l\right) \mathbf{V}_l'$$

that we've already said is the best rank l approximation to $\mathbf{X}$. In fact it is

$$\mathbf{X}^{*l} = \sum_{j=1}^{l} d_j \mathbf{u}_j \mathbf{v}_j' = \sum_{j=1}^{l} \mathbf{z}_j \mathbf{v}_j' = \left(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_l\right) \begin{pmatrix} \mathbf{v}_1' \\ \mathbf{v}_2' \\ \vdots \\ \mathbf{v}_l' \end{pmatrix}$$

and its ith row is $\sum_{j=1}^{l} \left\langle \mathbf{x}_i, \mathbf{v}_j \right\rangle \mathbf{v}_j'$, which (since the $\mathbf{v}_j$ are orthonormal) is the transpose of the projection of $\mathbf{x}_i$ onto $C\left(\mathbf{V}_l\right)$.

More Interpretation

That is,

$$\begin{aligned}\mathbf{X}^{*l} &= \begin{pmatrix} \langle \mathbf{x}_1, \mathbf{v}_1 \rangle \\ \vdots \\ \langle \mathbf{x}_N, \mathbf{v}_1 \rangle \end{pmatrix} \mathbf{v}_1' + \begin{pmatrix} \langle \mathbf{x}_1, \mathbf{v}_2 \rangle \\ \vdots \\ \langle \mathbf{x}_N, \mathbf{v}_2 \rangle \end{pmatrix} \mathbf{v}_2' + \cdots + \begin{pmatrix} \langle \mathbf{x}_1, \mathbf{v}_l \rangle \\ \vdots \\ \langle \mathbf{x}_N, \mathbf{v}_l \rangle \end{pmatrix} \mathbf{v}_l' \\ &= \mathbf{z}_1 \mathbf{v}_1' + \mathbf{z}_2 \mathbf{v}_2' + \cdots + \mathbf{z}_l \mathbf{v}_l' \\ &= \mathbf{X}\mathbf{v}_1\mathbf{v}_1' + \mathbf{X}\mathbf{v}_2\mathbf{v}_2' + \cdots + \mathbf{X}\mathbf{v}_l\mathbf{v}_l'\end{aligned}$$

a sum of rank 1 summands, producing for $\mathbf{X}^{*l}$ a matrix with each $\mathbf{x}_i$ in $\mathbf{X}$ replaced by the transpose of its projection onto $C(\mathbf{V}_l)$.

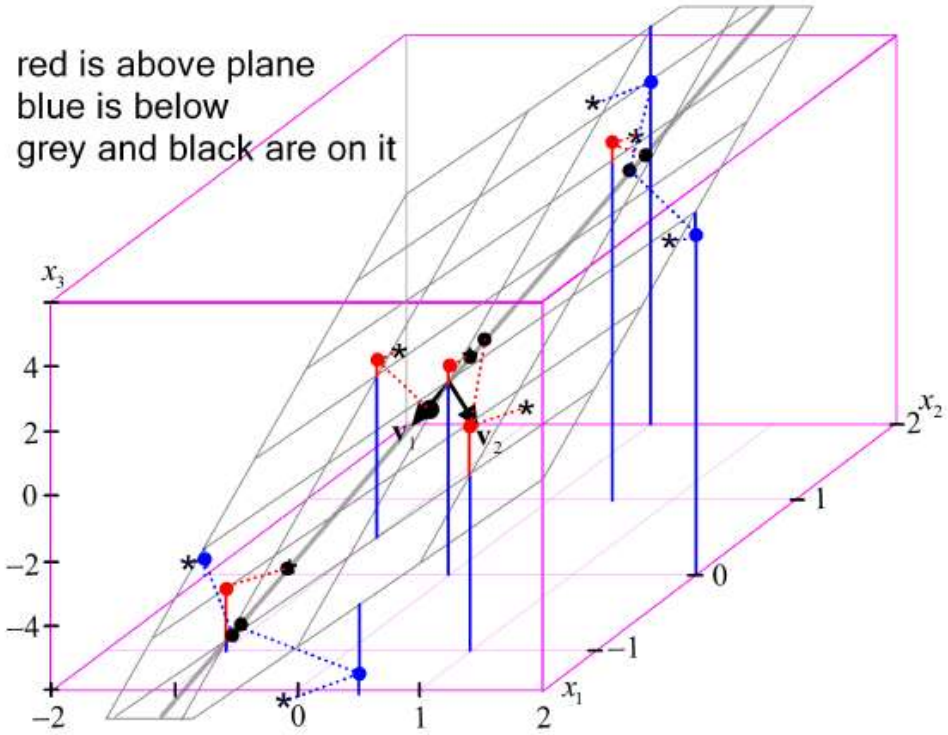
More interpretation

Since $\mathbf{z}_j = d_j \mathbf{u}_j$, $\mathbf{z}_j \mathbf{v}_j' = d_j \mathbf{u}_j \mathbf{v}_j'$. Then since the $\mathbf{u}_j$s and $\mathbf{v}_j$s are unit vectors, the sum of squared entries of both $\mathbf{z}_j$ and $\mathbf{z}_j \mathbf{v}_j'$ is d_j^2. These are non-increasing in j. So the $\mathbf{z}_j$ and $\mathbf{z}_j \mathbf{v}_j'$ decrease in "size" with j, and directions $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ are successively "less important" in describing variation in the $\mathbf{x}_i$ and in reconstructing $\mathbf{X}$.

This agrees with common interpretation of cases where a few singular values are much bigger than the others. There "simple structure" in the data is that observations can be more or less reconstructed as linear combinations of a few orthonormal vectors.

Low rank approximations to *N=9* data points

Below is a portrayal of a toy $p = 3$ dataset. Shown are $N = 9$ data points, the *rank* $= 1$ approximation (black balls on the line defined by the first PC direction) and the *rank* $= 2$ approximation (black stars on the plane).



Interpretation of PCs

To summarize interpretation of principal components of a centered dataset, one can say the following:

> *Principal components analysis amounts to the development of an alternative coordinate system in which to represent a p-dimensional dataset. One effectively finds a rotation of the original coordinate system to a new one where axes are defined by the p-vectors* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ *in which variation of the data in the directions* $\mathbf{v}_j$ *decreases with increasing* j *(as much as possible with each increment of* j*). The N-vectors* $\mathbf{u}_j$ *are unit vectors and their multiples* $\mathbf{z}_j = d_j\mathbf{u}_j$ *are the vectors of coordinates of the N data vectors in the new/rotated coordinate system. (And the* d_j *are the magnitudes of these vectors of new coordinates in* $\Re^N$*.)*

A few small singular values

Izenman, in his discussion of "polynomial principal components" points out that in some circumstances the existence of a few very *small* singular values can *also* identify important simple structure in a dataset. Suppose, for example, that all singular values except $d_p \approx 0$ are of appreciable size. One simple feature of the dataset is then that all $\langle \mathbf{x}_i, \mathbf{v}_p \rangle \approx 0$, i.e. there is one linear combination of the p coordinates x_j that is essentially constant (namely $\langle \mathbf{x}, \mathbf{v}_p \rangle$). The data fall nearly on a $(p-1)$-dimensional hyperplane in $\Re^p$. In cases where the p coordinates x_j are not functionally independent (for example consisting of centered versions of 1) all values, 2) all squares of values, and 3) all cross products of values of a smaller number of functionally independent variables), a single "nearly 0" singular value identifies a quadratic function of the functionally independent variables that must be essentially constant, a potentially useful insight about the dataset.

SVD of $\boldsymbol{X}$ and eigen analysis of of $\boldsymbol{X'X}$

The singular value decomposition of $\mathbf{X}$ means that both $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$ have useful representations in terms of singular vectors and singular values. Consider first $\mathbf{X}'\mathbf{X}$ (that is most of the sample covariance matrix). The SVD of $\mathbf{X}$ means that

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}'$$

and it's then clear that the columns of $\mathbf{V}$ are eigenvectors of $\mathbf{X}'\mathbf{X}$ and the squares of the diagonal elements of $\mathbf{D}$ are the corresponding eigenvalues. An eigen analysis of $\mathbf{X}'\mathbf{X}$ then directly yields the principal component directions of the data, and through the further computation of the inner products $\langle \mathbf{x}_i, \mathbf{v}_j \rangle$, the principal components $\mathbf{z}_j$ (and hence the singular vectors $\mathbf{u}_j$) are available.

Sample covariance matrix for N data p-vectors

Note that

$$\frac{1}{N}\mathbf{X}'\mathbf{X}$$

is the (N-divisor) sample covariance matrix for the p input variables $x_1, x_2, \ldots, x_p$. (**When X has standardized columns**–i.e. each column of $\mathbf{X}$, $\mathbf{x}_j$, has $\langle \mathbf{1}, \mathbf{x}_j \rangle = 0$ and $\langle \mathbf{x}_j, \mathbf{x}_j \rangle = N$– the matrix $\frac{1}{N}\mathbf{X}'\mathbf{X}$ is the sample correlation matrix for the p input variables $x_1, x_2, \ldots, x_p$.) The principal component directions of $\mathbf{X}$ in $\Re^p$, namely $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$, are also unit eigenvectors of the sample covariance matrix. The squared lengths of the principal components $\mathbf{z}_j$ in $\Re^N$ divided by N are the (N-divisor) sample variances of entries of the $\mathbf{z}_j$, and their values are

$$\frac{1}{N}\mathbf{z}_j'\mathbf{z}_j = \frac{1}{N}d_j\mathbf{u}_j'\mathbf{u}_j d_j = \frac{d_j^2}{N}$$

SVD of $\boldsymbol{X}$ and eigen analysis of $\boldsymbol{X}\,\boldsymbol{X}'$

The SVD of $\mathbf{X}$ also implies that

$$\mathbf{X}\mathbf{X}' = \mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{V}\mathbf{D}\mathbf{U}' = \mathbf{U}\mathbf{D}^2\mathbf{U}'$$

and it's then clear that the columns of $\mathbf{U}$ are eigenvectors of $\mathbf{X}\mathbf{X}'$ and the squares of the diagonal elements of $\mathbf{D}^2$ are the corresponding eigenvalues. $\mathbf{U}\mathbf{D}$ then produces the $N \times r$ matrix of principal components of the data. It does not appear that the principal component directions are available even indirectly based only on this second eigen analysis.