36-350: Data Mining

Handout 9
September 24, 2003

Principal Component Analysis (PCA)

What is the best direction to look at a cloud of points? PCA provides one answer. It can be understood in three equivalent ways: maximizing variance, maximizing information, and minimizing information loss.

Maximizing variance

Let $\mathbf{X}$ be the data matrix, whose rows are cases and whose columns have been standardized to have zero mean. PCA chooses combination weights $\mathbf{w}$ so that the projected data

$$\mathbf{h} = \mathbf{X}\mathbf{w} \tag{1}$$

has maximal variance. The sample variance of $\mathbf{h}$ can be written in terms of the sample covariance matrix of $\mathbf{X}$:

$$\mathrm{var}(\mathbf{h}) = \mathbf{h}^{\mathrm{T}}\mathbf{h} = \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} = \mathbf{w}^{\mathrm{T}}\mathrm{cov}(\mathbf{X})\mathbf{w} \tag{2}$$

Thus to maximize variance we want $\mathbf{w}$ to maximize $\mathbf{w}^{\mathrm{T}}\mathrm{cov}(\mathbf{X})\mathbf{w}$. The vector which does this is the top **eigenvector** of $\mathrm{cov}(\mathbf{X})$. $\mathbf{h}$ is called the *first principal component* of $\mathbf{X}$. It is one-dimensional. To get two dimensions, we find another vector $\mathbf{w}_2$, which maximizes the variance of $\mathbf{h}_2 = \mathbf{X}\mathbf{w}_2$, under the condition that it is orthogonal to $\mathbf{w}_1$. This ends up being the second eigenvector of $\mathrm{cov}(\mathbf{X})$, and $\mathbf{h}_2$ is called the *second principal component* of $\mathbf{X}$. You can repeat this process to get any number of principal components.

Maximizing information

Another interpretation of the first principal component is that it provides the most *information* about the attributes of a car than any other projection. Consider the correlation coefficient between $\mathbf{h}$ and any column of $\mathbf{X}$. If $\mathbf{X}$ is standardized, these correlations are given by

$$\rho = \frac{\mathbf{h}^{\mathrm{T}}\mathbf{X}}{\sqrt{\mathbf{h}^{\mathrm{T}}\mathbf{h}}} \tag{3}$$

The sum of squared correlations is

$$\rho\rho^{\mathrm{T}} = \frac{\mathbf{h}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{h}}{\mathbf{h}^{\mathrm{T}}\mathbf{h}} = \frac{\mathbf{w}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{w}}{\mathbf{w}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{w}} \tag{4}$$

The maximum correlation is achieved by $\mathbf{w}$ equal to the top eigenvector of $\mathrm{cov}(\mathbf{X})$, in other words, PCA. Thus the first principal component is *maximally correlated* with the original attributes.

The second principal component is the next most correlated projection, and so on. An interesting fact about the principal components is that they are uncorrelated with each other, that is:

$$\text{cov}(\mathbf{h}_1, \mathbf{h}_2) = \mathbf{h}_1^{\mathrm{T}} \mathbf{h}_2 = 0 \tag{5}$$

This means that $\mathbf{h}_1$ and $\mathbf{h}_2$ carry independent pieces of information.

Is PCA optimal?

Note that PCA interprets "information" as correlation, which is more restricted than our earlier definition of information as a reduction in entropy. It is possible to define a different sort of projection, which seeks to maximize information in the general sense. This projection wants the data to have maximum *entropy*, not just maximum variance. In other words, it wants the data to be spread out uniformly, which is generally desirable for visualization. PCA allows some of the data to be 'bunched up' in the projection.

Minimizing information loss

When a point is projected, it has an **image** located at $\mathbf{h}\mathbf{w}^{\mathrm{T}}$. The difference between the original point and its image is the information lost due to projection (also known as a **residual**):

$$\mathbf{X} - \mathbf{h}\mathbf{w}^{\mathrm{T}} \tag{6}$$

The projection which minimizes the sum of squared residuals is PCA. The R^2 **value** is a convenient summary of the amount of information captured in a projection:

$$R^2 = \frac{\text{variance of projected data}}{\text{variance of original data}} \tag{7}$$

$$1 - R^2 = \text{percent of information lost} \tag{8}$$

(For multivariate data, the total variance is given by the determinant of the covariance matrix.)

- PCA is the projection with maximal R^2.
- $R^2 \approx 1 \rightarrow$ data is flat (highly correlated attributes).
- $R^2 \approx 0 \rightarrow$ data is ball-shaped (uncorrelated attributes).

Interpreting the arrows

The arrows on a projection plot are the projections of unit vectors pointing along each of the original attributes. Changing that attribute will move in the given direction in the projected space. Thus it is easy to determine where a point will project, given its attributes. In reverse, from a projected point you can estimate its attribute values by looking at its position along the arrows. This is the **image** of the point, as defined above. The estimates will be good if the R^2 value is large.

The angle between two arrows gives an estimate of the correlation between attributes:

Angle	Correlation cos(angle)
0° (aligned)	1
90° (orthogonal)	0
180° (opposite)	-1

The quality of the estimate depends on the R^2 value—it is exact when $R^2 = 1$.

In PCA plots, most arrows point horizontally (along h_1) rather than vertically (along h_2). Can you explain why?

Interpreting a PCA plot

Coordinates Using the arrows, summarize what each coordinate (h_1 and h_2) is measuring. For the cars data, h_1 measures "size" and h_2 measures "sporty".

Correlations For many datasets, the arrows cluster into groups of highly correlated attributes. Describe the clusters. Also determine the overall level of correlation (given by the R^2 value).

Clusters Clusters indicate a preference for particular combinations of attribute values. Summarize each cluster by its prototypical member. For the cars data, the vans form a cluster (this is mostly clearly seen in 3D).

Funnels Funnels are wide at one end and narrow at the other. They happen when one dimension affects the variance of a perpendicular dimension. Thus, even though the dimensions are uncorrelated (because they are perpendicular) they still affect each other. The cars data has a funnel, showing that small cars are similar in sportiness, while large cars are more varied.

Voids Voids are areas inside the range of the data which are unusually unpopulated. A **permutation plot** is a good way to spot voids. (Randomly permute the data in each column, and see if any new areas become occupied.) For the cars data, there is a void of sporty cars which are very small or very large. This suggests that such cars are undesirable or difficult to make.

References

[1] A. O'Hagan. "Motivating Principal Components, And A Stronger Optimality Result." *The Statistician* 33(3): 313-315, 1984.