Data Acquisition and Processing I: Big Data @ Principle Component Analysis (PCA)

Rensheng Wang,
rwang1@stevens.edu
Dept. of Electrical and Computer Engineering
Stevens Institute of Technology

Data Transformations

- Transformations of predictor variables may be needed for several reasons. Some modeling techniques may have strict requirements, such as the predictors having a common scale. In other cases, creating a good model may be difficult due to specific characteristics of the data (e.g., outliers).
- The most straightforward and common data transformation is to center scale the predictor variables.
- To center a predictor variable, the average predictor value is subtracted from all the values. As a result of centering, the predictor has a zero mean.
- Similarly, to scale the data, each value of the predictor variable is divided by its standard deviation. Scaling the data coerce the values to have a common standard deviation of one.
- These manipulations are generally used to improve the numerical stability of some calculations.
- The only real downside to these transformations is a loss of interpretability of the individual values since the data are no longer in the original units.

Transformations to Resolve Outliers

- We will generally define outliers as samples that are exceptionally far from the mainstream of the data. Under certain assumptions, there are formal statistical definitions of an outlier.
- Even with a thorough understanding of the data, outliers can be hard to define. However, we can often identify an unusual value by looking at a figure. When one or more samples are suspected to be outliers, the first step is to make sure that the values are scientifically valid (e.g., positive blood pressure) and that no data recording errors have occurred.
- Great care should be taken not to hastily remove or change values, especially if the sample size is small. With small sample sizes, apparent outliers might be a result of a skewed distribution where there are not yet enough data to see the skewness.
- Also, the outlying data may be an indication of a special part of the population under study that is just starting to be sampled. Depending on how the data were collected, a "cluster" of valid points that reside outside the mainstream of the data might belong to a different population than the other samples.

Transformations to Resolve Outliers

- There are several predictive models that are resistant to outliers. Tree-based classification models create splits of the training data and the prediction equation is a set of logical statements such as "if predictor A is greater than X, predict the class to be Y ," so the outlier does not usually have an exceptional influence on the model.
- Also, support vector machines for classification generally disregard a portion of the training set samples when creating a prediction equation. The excluded samples may be far away from the decision boundary and outside of the data mainstream.
- If a model is considered to be sensitive to outliers, one data transformation that can minimize the problem is the spatial sign.

Transformations to Resolve Outliers

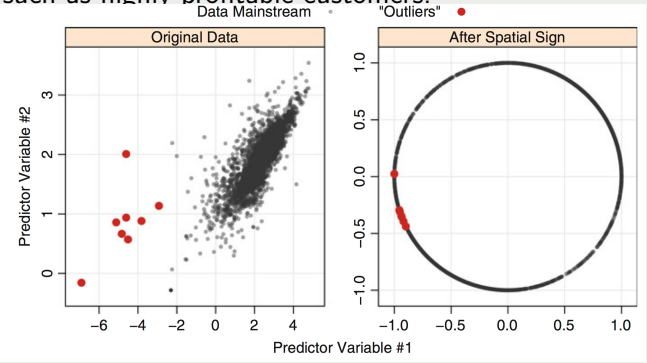
- This procedure projects the predictor values onto a multidimensional sphere. This has the effect of making all the samples the same distance from the center of the sphere. Mathematically, each sample is divided by its squared norm:

$$x^*_{ij} = \frac{x_{ij}}{\sum_{j=1}^{P} x^2_{ij}}$$

- Since the denominator is intended to measure the squared distance to the center of the predictor's distribution, it is important to center and scale the predictor data prior to using this transformation.

Transformations to Resolve Outliers

- Note that, unlike centering or scaling, this manipulation of the predictors transforms them as a group. Removing predictor variables after applying the spatial sign transformation may be problematic.
- Below plot shows another data set with two correlated predictors. In these data, at least eight samples cluster away from the majority of other data. These data points are likely a valid, but poorly sampled subpopulation of the data.
- The modeler would investigate why these points are different; perhaps they represent a group of interest, such as highly profitable customers.



Data Reduction and Feature Extraction

- Data reduction techniques are another class of predictor transformations. These methods reduce the data by generating a smaller set of predictors that seek to capture a majority of the information in the original variables.
- In this way, fewer variables can be used that provide reasonable fidelity to the original data. For most data reduction techniques, the new predictors are functions of the original predictors; therefore, all the original predictors are still needed to create the surrogate variables. This class of methods is often called signal extraction or feature extraction techniques.
- PCA (principal components analysis) is a commonly used data reduction technique. This method seeks to find linear combinations of the predictors, known as principal components (PCs), which capture the most possible variance.
- The first PC is defined as the linear combination of the predictors that captures the most variability of all possible linear combinations.
- Then, subsequent PCs are derived such that these linear combinations capture the most remaining variability while also being uncorrelated with all previous PCs.

Data Reduction and Feature Extraction (PCA)

- Mathematically, the jth PC can be written as:

$$\text{PC}_j = (a_{j1} \times \text{Predictor } 1) + (a_{j2} \times \text{Predictor } 2) + \cdots + (a_{jP} \times \text{Predictor } P).$$

  where P is the number of predictors.

- The coefficients a_{j1}, a_{j2}, $\cdots, a_{jP}$ are called component weights and help us understand which predictors are most important to each PC.

- The primary advantage of PCA, and the reason that it has retained its popularity as a data reduction method, is that it creates components that are uncorrelated, i.e., PC_1 is not correlated to $\text{PC}_2, \cdots, \text{PC}_P$.

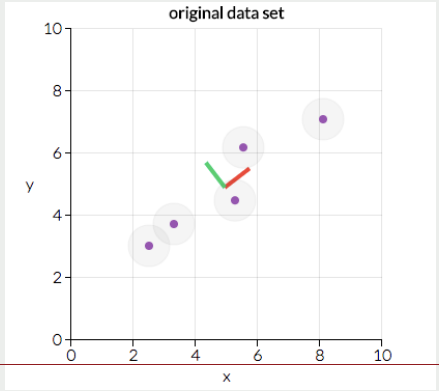
- Some predictive models prefer predictors to be uncorrelated (or at least low correlation) in order to find solutions and to improve the modelâĂŹs numerical stability. PCA pre-processing creates new predictors with desirable characteristics for these kinds of models.

Data Reduction and Feature Extraction (PCA)

- While PCA delivers new predictors with desirable characteristics, it must be used with understanding and care.
- Notably, practitioners must understand that PCA seeks predictor-set variation without regard to any further understanding of the predictors (i.e., measurement scales or distributions) or to knowledge of the modeling objectives (i.e., response variable).
- Hence, without proper guidance, PCA can generate components that summarize characteristics of the data that are irrelevant to the underlying structure of the data and also to the ultimate modeling objective.

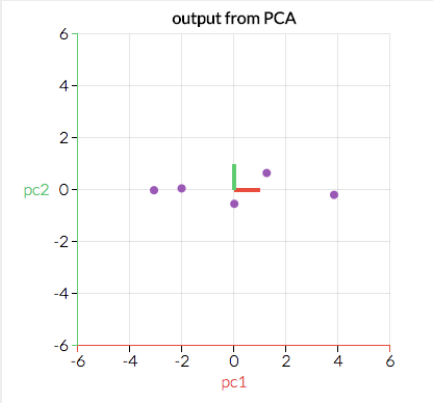
How does PCA work?

- We first calculate a covariance matrix that summarizes how our variables all relate to one another.
- We then break this matrix down into two separate components: direction and magnitude. We can then understand the "directions" of our data and its "magnitude" (or how "important" each direction is).
- The screenshot below, from the setosa.io applet, displays the two main directions in this data: the "red direction" and the "green direction." In this case, the "red direction" is the more important one.



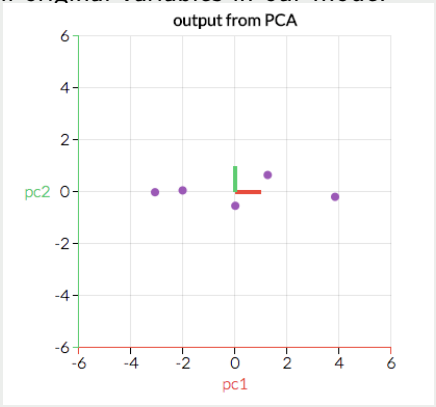
How does PCA work?

- We will transform our original data to align with these important directions (which are combinations of our original variables).
- The screenshot below is the same exact data as above, but transformed so that the x- and y-axes are now the "red direction" and "green direction."



How does PCA work?

- While the visual example here is two-dimensional (and thus we have two "directions"), think about a case where our data has more dimensions. By identifying which "directions" are most "important," we can compress or project our data into a smaller space by dropping the "directions" that are the "least important."
- By projecting our data into a smaller space, we're reducing the dimensionality of our feature space ⋯ but because we've transformed our data in these different "directions," we've made sure to keep all original variables in our model!



Why does PCA work?

- First, the covariance matrix is a matrix that contains estimates of how every variable relates to every other variable. Understanding how one variable is associated with another is quite powerful.
- Second, eigenvalues and eigenvectors are important. Eigenvectors represent directions. Think of plotting your data on a multidimensional scatterplot. Then one can think of an individual eigenvector as a particular "direction" in your scatterplot of data. Eigenvalues represent magnitude, or importance. Bigger eigenvalues correlate with more important directions.
- Finally, we make an assumption that more variability in a particular direction correlates with explaining the behavior of the dependent variable. Lots of variability usually indicates signal, whereas little variability usually indicates noise. Thus, the more variability there is in a particular direction is, theoretically, indicative of something important we want to detect.

PCA example

- One PCA example in imaging processing: a) 25 randomly chosen 64 × 64 pixel images from the face database. (b) The mean and the first three principal component basis vectors (or eigenfaces, i.e., linear combinations of the 25 faces).

PCA in Stock Market

- An important topic in multivariate time series analysis is the study of the covariance (or correlation) structure of the series.
- For example, the covariance structure of a vector return series plays an important role in portfolio selection. In what follows, we discuss some statistical methods useful in studying the covariance structure of a vector time series.
- Given a k-dimensional random variable $\mathbf{r} = (r_1, \cdots, r_k)$
- If $\mathbf{r}$ denotes the monthly log returns of k assets, then PCA can be used to study the main source of variations of these k asset returns. Here the keyword is *few* so that simplification can be achieved in multivariate analysis.
- Let $\mathbf{w}_i = (w_{i1}, \cdots, w_{ik})'$ be a k-dimensional real-valued vector, where $i = 1, \cdots, k$. Then

$$y_i = \mathbf{w}_i' \mathbf{r} = \sum_{j=1}^{k} w_{ij} r_j$$

is a linear combination of the random vector r. If $\mathbf{r}$ consists of the simple returns of k stocks, then y_i is the return of a portfolio that assigns weight w_{ij} to the jth stock. Since multiplying a constant to w_i does not affect the proportion of allocation assigned to the jth stock, we standardize the vector w_i so that $\mathbf{w}_i' \mathbf{w}_i = 1 \sum_{j=1}^{k} w_{ij}^2 = 1$.

PCA in Stock Market

- Using properties of a linear combination of random variables, we have

$$\begin{aligned} \text{Var}(y_i) &= \mathbf{w}_i' \mathbf{\Sigma}_r \mathbf{w}_i, \qquad i = 1, \cdots, k \\ \text{Cov}(y_i, y_j) &= \mathbf{w}_i' \mathbf{\Sigma}_r \mathbf{w}_j, \qquad i,\ j = 1, \cdots, k \end{aligned}$$

- The idea of PCA is to find linear combinations w_i such that y_i and y_j are uncorrelated for $i \neq j$ and the variances of y_i are as large as possible. More specifically:

  1. The first principal component of $\mathbf{r}$ is the linear combination $y_1 = \mathbf{w}_1' \mathbf{r}$ that maximizes $\text{Var}(y_1)$ subject to the constraint $\mathbf{w}_1' \mathbf{w}_1 = 1$.
  2. The second principal component of $\mathbf{r}$ is the linear combination $y_2 = \mathbf{w}_2' \mathbf{r}$ that maximizes $\text{Var}(y_2)$ subject to the constraints $\mathbf{w}_2' \mathbf{w}_2 = 1$ and $\text{Cov}(y_2, y_1) = 0$.
  3. The ith principal component of $\mathbf{r}$ is the linear combination $y_i = \mathbf{w}_i' \mathbf{r}$ that maximizes $\text{Var}(y_i)$ subject to the constraints $\mathbf{w}_i' \mathbf{w}_i = 1$ and $\text{Cov}(y_i, y_j) = 0$ for $j = 1, ..., i$âĹŠ1.

PCA in Stock Market

- Since the covariance matrix $\mathbf{\Sigma}_r$ is positive definite, it has the positive eigen-decompositions.
- Let $(\lambda_1, \mathbf{e}_1), \cdots, (\lambda_k, \mathbf{e}_k)$ be the eigenvalue-eigenvector pairs of $\mathbf{\Sigma}_r$, where $\lambda_1 > \lambda_2 > \cdots > \lambda_k > 0$ and eigenvectors $\mathbf{e}_i$ are also properly normalized.
- The ith principal component of $\mathbf{r}$ is $y_i = \mathbf{e}_i' \mathbf{r} = \sum_{j=1}^{k} e_{ij} r_j$ for $i = 1, \cdots, k$.

$$\begin{aligned} \mathrm{Var}(y_i) &= \mathbf{e}_i' \mathbf{\Sigma}_r \mathbf{e}_i = \lambda_i, \quad i = 1, \cdots, k \\ \mathrm{Cov}(y_i, y_j) &= \mathbf{e}_i \mathbf{\Sigma}_r \mathbf{e}_j = 0, \quad i \neq j \end{aligned} \tag{1}$$

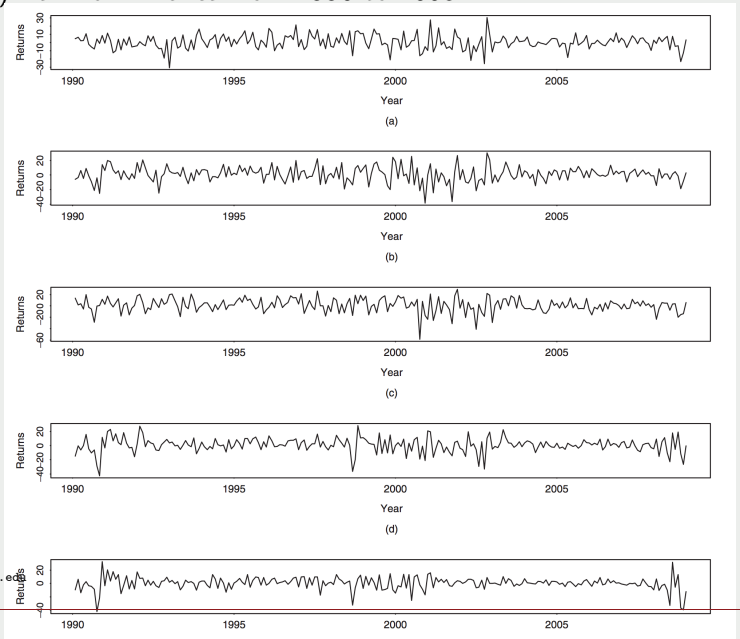
- The variance of new principal components over original data variance

$$\frac{\mathrm{Var}(y_i)}{\sum_{i=1}^{k} \mathrm{Var}(r_i)} = \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_k}$$

- The proportion of total variance in $\mathbf{r}$ explained by the ith principal component is simply the ratio between the ith eigenvalue and the sum of all eigenvalues with $\mathbf{\Sigma}_r$.

PCA in Stock Market

- Time plots of monthly stock returns in % for (a) IBM, (b) HP, (c) Intel, (d) J.P.Morgan Chase, and (e) Bank of America from 1990 to 2008.



PCA in Stock Market

- Denote the return vector $\mathbf{r}' = [\text{IBM, HP, Intel, JPM, BOA}]$. The sample mean vector and covariance matrix

$$\hat{\mathbf{\Sigma}}_r = \frac{1}{T-1}\sum_{t=1}^{T}(\mathbf{r}_t - \bar{\mathbf{r}})(\mathbf{r}_t - \bar{\mathbf{r}})', \quad \hat{\mathbf{r}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_t$$

- Based on the sample covariance matrix $\hat{\mathbf{\Sigma}}_r$

Using Sample Covariance Matrix

Eigenvalue	284.17	112.93	57.43	46.81	29.87
Proportion	0.535	0.213	0.108	0.088	0.056
Cumulative	0.535	0.748	0.856	0.944	1.000
Eigenvector	0.330	0.139	−0.264	0.895	−0.014
	0.483	0.279	−0.701	−0.430	−0.116
	0.581	0.478	0.652	−0.096	−0.016
	0.448	−0.550	0.013	−0.064	0.702
	0.347	−0.610	0.119	−0.009	−0.702

PCA in Stock Market

- From the last Table, we have the first 2 principal components contribute 75% total variance of 5 time series.

$$\hat{\lambda}_1 = 284.17, \quad \hat{\mathbf{e}}_1 = [0.33,\ 0.48,\ 0.58,\ \ 0.45,\ \ 0.34]'$$
$$\hat{\lambda}_2 = 112.93, \quad \hat{\mathbf{e}}_2 = [0.14,\ 0.28,\ 0.48,\ -0.55,\ -0.61]'$$

- The 1st component is roughly equally weighted linear combinations of the stock returns. This component might represent the general movement of the stock market and hence is a market component.
- The 2nd component represents the difference between the two industrial sectors – namely, technologies versus financial services. It might be an industrial component.
- In other words, the market movements explain the majority of the stock returns movement and plusing industrial component explains the 3/4 variances of all the stock returns.

Data Reduction and Feature Extraction

- Because PCA seeks linear combinations of predictors that maximize variability, it will naturally first be drawn to summarizing predictors that have more variation.
- If the original predictors are on measurement scales that differ in orders of magnitude [consider demographic predictors such as income level (in $) and height (in feet)], then the first few components will focus on summarizing the higher magnitude predictors (e.g., income), while latter components will summarize lower variance predictors (e.g., height).
- This means that the PC weights will be larger for the higher variability predictors on the first few components.
- In addition, it means that PCA will be focusing its efforts on identifying the data structure based on measurement scales rather than based on the important relationships within the data for the current problem.
- For most data sets, predictors are on different scales. Hence, to help PCA avoid summarizing distributional differences and predictor scale information, it is best to first center and scale the predictors prior to performing PCA.
- Centering and scaling enables PCA to find the underlying relationships in the data without being influenced by the original measurement scales.

Data Reduction and Feature Extraction

- The second caveat of PCA is that it does not consider the modeling objective or response variable when summarizing variability. Because PCA is blind to the response, it is an unsupervised technique.
- If the predictive relationship between the predictors and response is not connected to the predictorsâĂŹ variability, then the derived PCs will not provide a suitable relationship with the response. In this case, a supervised technique, like PLS (partial least squares), will derive components while simultaneously considering the corresponding response.
- For data sets with many predictor variables, we must decide how many components to retain. A heuristic approach for determining the number of components to retain is to create a scree plot, which contains the ordered component number (x-axis) and the amount of summarized variability (y-axis). For most data sets, the first few PCs will summarize a majority of the variability, and the plot will show a steep descent; variation will then taper off for the remaining components.

Data Reduction and Feature Extraction

- Generally, the component number prior to the tapering off of variation is the maximal component that is retained.
- In below plot, the variation tapers off at component 5.