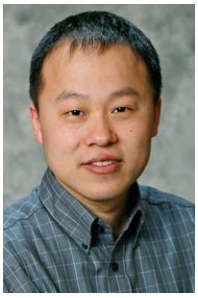
ECONOMICS COMMENTATOR

South Dakota State University No. 510 September 21, 2009



Principal Component Analysis

by

Jing Li[1]
Assistant Professor

Principal component analysis (PCA) is a useful statistical tool. PCA is widely used because it is a simple, non-parametric method of extracting relevant information from confusing data sets. Consider this example.

Suppose an advisor has 20 advisees. At the end of one semester, for each advisee the advisor collects the score data of three courses: English, History and Calculus. Now the department head hopes that the advisor can nominate one advisee for a scholarship. The nomination should be solely based on the students' performance in these three courses. What could the advisor do?

Different advisors may use different methods. Advisors may prefer the simple method of ranking average scores. That is, the advisor first calculates the average score of the three courses for each advisee, and then nominates the student with the highest average score. The issue is, is this method optimal, or fair? Essentially the approach treats the three courses equally. The average puts equal weight (1/3) on each course. People may frown upon this equal treatment. For example, students may perceive Calculus as a harder course than English and History. Statistically speaking, the mean of the Calculus score may be the lowest among the three courses. Therefore it is sensible to argue that an A in Calculus should count more than an A in the two other courses.
(continued on page 2)

[1] Contact the author at jing.li@sdstate.edu or 605-688-4141.

Fall Crop Markets

by

Alan May[2]
Extension Grain Marketing Specialist

The monthly World Agricultural Supply and Demand Estimates report (WASDE) issued by USDA in mid-September continued to confirm fundamental expectations in the grain trade. This year the United States will produce its largest soybean crop in history and the second largest corn crop in history. Although the production projections for corn and soybeans were bearish in this report, the usage projections have a more positive tone. All usage categories for corn and soybeans are expected to be slightly higher than a year ago. However, carryover supply for soybeans is still expected to grow to a projected 220 million bushels. Corn supplies will decline somewhat from a year ago, but the 1.63 billion bushel carryover supply is considered manageable as long as there are no late season problems that might impact production.

Wheat production in the United States this year is not a record-setter but the projected carryover supplies of wheat are expected to be the largest since 2001. Recall that wheat prices were hovering around the $3.00 per bushel mark in 2001. Even though wheat production in the U.S. is lower this year, prices are lower because of the bearish nature of wheat demand. Wheat exports are expected to fall by 65 million bushels compared to a year ago. With the large carry-in supply from 2008, projected carryover supplies for the 2009 crop will grow by nearly 70 million bushels.
(continued on page 3)

[2] Contact the author at Alan.May@sdstate.edu or 605-688-4862.

Principal Component Analysis…..(Cont'd from p. 1)
The advisor should also take into account the variation of scores. If the standard deviation of the History score is greater than the English score, then History may differentiate students better than English. Therefore, the History score should receive a higher weight than the English score.

Now the issue seems to get more and more complicated. How do we find the proper weight so as to account for these concerns? We can use an *ad hoc* weight like (0.25, 0.25, 0.5), but this choice is hardly optimal. Fortunately we can get the optimal weight via PCA. To illustrate how it works, let us abstract the problem a little bit.

Basically we want to reduce a complex data set into a lower dimension to reveal the underlying, hidden, simplified structure. In this case, we want to transform a 20×3 data set (there are 20 advisees, or observations, for which we observe 3 courses, or variables) into a 20×1 vector (for example, we may obtain 20 average scores, one for each student). The dimension of data is reduced from 3 to 1. The advantage of dimensional reduction is evident. For this problem, it is much easier (and less controversial) to rank a 20×1 vector than a 20×3 matrix.

Our goal is to uncover the underlying structure. In this case we are seeking a summary of hidden factors including the students' academic capabilities, willingness to learn, etc. These factors are unobservable, or latent. Nevertheless we can still extract those factors from observable data via PCA. Similar attempts are tried in other disciplines such as physics, where people hope to extract a "signal" from "noise".

PCA method
Mathematically, let $\boldsymbol{X}$ be the raw data, a 20 x 3 matrix. Then PCA provides an optimal weighting vector, $\boldsymbol{c}$, so that the weighted data, $\boldsymbol{Xc}$, can serve as an estimate for the underlying structure. This is done by maximizing the variance of weighted data subject to normalization, i.e., we want to solve the following constrained optimization problem:

$$\max \boldsymbol{Var(Xc)},$$

under certain constraints.

Intuitively, variance measures the amount of information contained in data. A pattern can be seen only when data vary (Can you determine a line with just one point? Absolutely not. You need at least two distinct points.) The bigger variance is, the more information is available. We want to reduce the dimension of raw data (because the original data is clouded, confusing or redundant), but meanwhile we hate to lose useful information and so we want to maximize the variance of transformed data.

It follows from basic statistics theory that $\boldsymbol{Var(Xc) = c'Var(X)c}$. This is a quadratic form that is maximized by a three-step procedure. First we obtain the variance-covariance matrix of $\boldsymbol{X}$, $\boldsymbol{Var(X)}$. Next we compute the eigenvalues and eigenvectors of $\boldsymbol{Var(X)}$. Finally, we choose the eigenvector for the biggest eigenvalue as $\boldsymbol{c}$.

Go back to our example. We need to go through the following nomination procedure. First, save the score data as a 20×3 table called $\boldsymbol{X}$, and compute the variance-covariance matrix of $\boldsymbol{X}$. Next use the eigenvector for the biggest eigenvalue of $\boldsymbol{Var(X)}$ as the optimal weight vector. Notice that only extremely rarely will this eigenvector equal (1/3, 1/3, 1/3), the vector used for the average. After post-multiplying $\boldsymbol{X}$ with $\boldsymbol{c}$, we obtain a 20×1 vector. This vector quantifies or summarizes the underlying factors for the test scores. The i^{th} component in this vector is the scalar summary of academic performance for the i^{th} advisee. Finally we rank this 20×1 vector and nominate the student with highest value.

If we believe (or assume) one vector is sufficient to summarize the raw data, then the vector $\boldsymbol{Xc}$ is called the principal component. The procedure to obtain the principal component is called PCA. The 20×1 vector we get for the nomination problem is the principal component for the score data. Of course there may be more than one component underlying the data. In that case, the second component is obtained by multiplying the data with the eigenvector corresponding to the second largest eigenvalue, and so on.

It is not accidental to focus on the variance-covariance matrix. The diagonal terms of the variance-covariance matrix are the variances, and off-diagonal terms are covariances. Variance measures the pattern of a variable, whereas covariance measures the degree of linear association. If variables are highly correlated,

we say data are redundant in the sense that some variables are (approximately) linear combinations of others.

At this point, you may realize that if we collect all the eigenvectors in ***C, C´Var(X)C*** will produce a diagonal matrix of the eigenvalues of ***Var(X)***. The diagonal of the resulting matrix implies that the transformed data, ***XC,*** is not redundant (because the off-diagonal term, covariance, is zero). The principal component is the first column of ***XC***. Moreover, the principle component is unique in the sense that it is uncorrelated with the second column of ***XC,*** the second component.

A regression application

One important application of PCA is to remedy multicollinearity in regression. The solution is called principal component regression (PCR). Let's look at another example.

One topic in Finance is to explain the stock price of a firm. People may run a regression and use various financial ratios as regressors. A problem with the regression analysis is caused by the correlation among ratios. For instance, a firm with a high return on equity (ROE) is likely to have a high return on assets (ROA). After all, ratios may overlap and, at least partially, measure the same thing.

Correlated regressors give rise to multicollinearity, which makes OLS estimates imprecise by inflating the standard errors. Intuitively, if a regression includes two correlated ratios, the OLS cannot tell the effect of one ratio from the other. Consequently, none of the coefficients for the correlated ratios can be accurately estimated.

How about just using the principal components for ratios? Yes, if you are thinking in this way, you are talking about PCR. Instead of using all available financial ratios, one may just use one, two or three components of ratios. We run PCA first, and then use the transformed data, the components, as regressors.

Fall Crop Markets ….. (Cont'd from p.1)

In order to gauge what the future holds for grain and oilseed prices in the next year, one should look back several years for perspective on the impact of supply and demand. Prior to the fall of 2006, the grain market was often viewed as a supply dominated market. U.S. farmers tended to produce sufficient quantities of grains and oilseeds that outpaced usage resulting in large carryover supplies. However, beginning in the fall of 2006 corn prices were driven to unprecedented levels by December of that year due in part to the expansion of the ethanol industry. This shifted the market to being more demand driven. As a result of the growing demand for corn, U.S. farmers planted 15 million more acres of corn in the spring of 2007 than were planted in 2006. This huge increase in corn acres led to a decline of 11 million acres planted to soybeans in 2007 compared to 2006. Wheat production in the U.S. and worldwide had already shrunk by 2006, leaving wheat prices susceptible to strong demand at a time of very tight domestic and world wheat supplies.

The growth in demand for grains and oilseeds along with the dramatic shift of acres between corn and soybeans in 2007 and 2008 led to even more significant price increases for corn and soybeans through the summer of 2008. Extraordinarily tight supplies of wheat led to significant price increases for wheat as well; particularly from the fall of 2007 through the spring of 2008. While these price increases were influenced by grain supply and demand fundamentals, prices were also very sensitive to other factors such as energy markets, the stock market and the value of the dollar.

By late summer of 2008 the price euphoria that existed in grain commodity markets ebbed considerably. Prices for corn, soybeans and wheat fell considerably through the end of 2008 and the trend has continued through the fall of 2009. The pressure of the strongest recession in years combined with the huge downturn in the stock market and other outside markets led to a grain market that is currently more influenced by supply than it is with demand. Yet, while the pendulum may be poised to swing back to a supply driven market, demand still matters.

The challenge will be to determine if demand can strengthen enough in the next year to reduce the current projections for carryover supply. If corn and soybean production is finalized at the current projections, the chance for growth in ending supplies becomes greater if current demand projections weaken. The length of the recovery from the recession will influence demand in the export market as well as in the domestic market.

Even though wheat harvest is completed and the corn and soybean harvest is just starting, this year's production will likely influence relative profitability through 2010. Although the mix of crop acres and total production in 2010 are unknown at this time, there is a risk that supplies could continue to grow after the 2010 harvest unless demand can strengthen beyond current expectations. This translates into the risk of longer term price weakness and uncertainty. For the short-term, however, it appears that the current projections of record or near record setting production and growing or steady carryover supplies will mean a bearish outlook for grain prices into next year.

****

ECONOMICS COMMENTATOR

****

Department of Economics
South Dakota State University Phone: 605-688-4141
Box 504 Scobey Hall Fax: 605-688-6386
Brookings, SD 57007-0895 E-Mail: Penny_Stover@ sdstate.edu

120 copies of this newsletter were produced at a cost of less than $100

SOUTH DAKOTA STATE UNIVERSITY
Department of Economics
Box 504
Brookings SD 57007-0895

Change Service Requested