Principal Components Analysis
Part 2

The objective of Principal Components Analysis (PCA) is to uncover structure in a dataset by replacing the old variables with new variables that have more desirable properties. The new variables are linear combinations of the old variables.

What are these more desirable properties?

- Information……..……... The new variables capture most of the information (variability, explanatory power) of the old variables.
- Dimensionality reduction… There are fewer new variables than old variables - parsimony.
- Uncorrelated…….. The new variables are uncorrelated with each other.
- Insight…… The new variables may reveal hidden structure and unsuspected interpretations.

A potential killer disadvantage: It may be hard to understand what the new variables mean.

In most statistical methodology (e.g., regression), it is better to have a small set of uncorrelated variables than a large set of correlated variables, *ceteris paribus*. PCA enables the analyst to replace a dataset that has a big-*p* problem (large number of predictors) with a dataset having fewer predictors but with almost the same information content and with nicer properties.

So PCA has two major applications:

1) To massage the data before applying another statistical technique.
   For example, suppose the predictor variables in a regression exhibit multicollinearity.[1] PCA processes the original predictors into new predictors (principal components) that have *no* correlation at all with each other. Furthermore, a subset of the principal components often captures most of the variability of the original predictors. Thus, you can replace the original predictors by this subset of principal components in the regression. The benefits: No multicollinearity problems, fewer predictors (potentially more parsimonious final model). The disadvantages: May be hard to understand the meaning of the new predictors, some loss (usually modest) in explanatory power. Another example is the frequent use of PCA in lossy compression algorithms for images.
2) As a tool for discovering structure in data.
   In this respect, PCA can be thought of as a special case of factor analysis. It is often very illuminating to interpret the meaning of those principal components that explain most of the variance in a dataset to see what underlying theoretical constructs are "driving" the data. A related use is to calculate the values that individual observations have on the most important principal component dimensions and plot these component scores against each other. Such a plot reveals where individual observations lie on the dimensions of the theoretical constructs that the main principal components represent.

[1] Multicollinearity is a problem in regression in which the predictor variables are highly correlated with each other. The presence of multicollinearity in regression can distort the apparent relationships of the predictors with the response and can make predictors appear less important than they really are.

Abbreviated Discussion of Aspects of How PCA Works

Start with a set of *p* old variables $X_1, X_2, ..., X_p$. These may be predictor variables, or response variables, or just a collection of miscellaneous variables. PCA produces *p* new variables, called principal components, denoted $\xi_1, \xi_2, ..., \xi_p$ (the Greek letter ξ is pronounced "ksee"). SAS denotes principal components as PRIN1, PRIN2, …,PRIN*p*.

(1) Each ξ_i is a linear combination of all of the old variables:

$$\left\{\begin{array}{lcll} \xi_1 & = & \text{PRIN1} & = \beta_{11}X_1 + \beta_{12}X_2 + \cdots + \beta_{1p}X_p \\ \xi_2 & = & \text{PRIN2} & = \beta_{21}X_1 + \beta_{22}X_2 + \cdots + \beta_{2p}X_p \\ \xi_3 & = & \text{PRIN3} & = \beta_{31}X_1 + \beta_{32}X_2 + \cdots + \beta_{3p}X_p \\ \vdots & & \vdots & \vdots \\ \xi_p & = & \text{PRINp} & = \beta_{p1}X_1 + \beta_{p2}X_2 + \cdots + \beta_{pp}X_p \end{array}\right\} \text{(Eqn 4)}^2$$

(2) It is customary practice to standardize the X's before running PCA on them. That is, subtract the mean and divide by the stdev of each X: $X_i \rightarrow \dfrac{(X_i - \overline{X}_i)}{s_{X_i}}$. [*Note: To standardize or not is a nontrivial decision, since standardization gives each of the X_1, X_2, ..., X_p the same equal weight. See paragraph (10) below.]*

(3) The main competitor to standardizing is centering: just subtract the mean without dividing by the stdev. I.e., $X_i \rightarrow X_i - \overline{X}_i$. Rarely is PCA applied to the raw data without standardizing or centering.

It will be assumed throughout these notes that each X_i has been standardized prior to running the PCA, unless the notes state specifically to the contrary.

(4) There are potentially as many PRINs as there are old variables, but it is desirable to use only a few PRINs in most applications.

(5) If you replace a set of original $X_1, X_2, ..., X_p$, by the complete set of *p* principal components $\xi_1, \xi_2, ..., \xi_p$, no information that was in the original variables is lost.[3]

(6) If you replace a set of original $X_1, X_2, ..., X_p$, by a subset of the *p* principal components, some information that was in the original $X_1, X_2, ..., X_p$ will be lost. In this case, it is important that the subset of principal components capture the "essence" of the old variables.

---

[2] The equation numbering continues from Part 1 of these PCA notes.

[3] For example, if you replace the predictors in a regression by the complete set of their principal components, the new regression will have the same R-square and Root MSE as the old regression. Of course, the coefficients and their standard errors will change, as will other features. But the information is just redistributed, not lost.

(7) That "essence" is variability.[4] The total variability of $X_1, X_2, ..., X_p$ is the sum of their variances,[5] called the **total variance**. The sum of the variances of PRIN1, …, PRINp also equals the total variance.

Total variance = Var(X_1) + … + Var (X_p) = Var(ξ_1) + … + Var(ξ_p) (Eqn 5)

If you standardize the X's, as is customary, then Var(X_i) = 1 for all X's; so all X's have the same importance in terms of total variance. Thus the total variance = *p* if the variables have been standardized. If the X's are not standardized, then they have unequal importance in terms of total variance.

(8) When you are deciding to replace $X_1, X_2, ..., X_p$ by a subset of PRIN1, …, PRINp, there are two criteria to weigh:

- First, you want the sum of the variances of the replacing principal components to be as close to the total variance as possible (**explanatory power**).[6]
- Second, you want to minimize the number of principal components used to replace $X_1, X_2, ..., X_p$ (**parsimony**).

You can get 100% of the total variance by using all *p* principal components. You can minimize the number of principal components by using only one. Sometimes one principal component is all that you need. But usually there must be a trade-off between explanatory power and parsimony.[7]

(9) What percentage of the total variance of $X_1, X_2, ..., X_p$ should be captured by the subset of principal components used to replace them? And how few principal components should be in the subset? In general, one should not be dogmatic in answering these questions. Practical experience suggests that it is desirable to capture about 80% or more of the variance with 6 or fewer principal components, but there is no theoretical justification for such a rule. *[See paragraph (21).]*

---

[4] It may seem strange to call variability the "essence" of a variable. (Then again, it may not -- after all a *variable* is variable!) The effects of a variable are felt through the changes it induces in other variables. Unless the variable changes, it cannot produce any effects. And the more it changes, the larger the changes it can produce. We gain information by studying the effects that variability in one phenomenon induces in other phenomena.

[5] Heads up! If the $X_1, X_2, ..., X_p$ have only been centered, how is the sum of their variances affected? (Answer: not at all – the variance remains the same after centering as before). But if the $X_1, X_2, ..., X_p$ have been standardized, how is the sum of their variances affected? (Answer: sum of variances after standardizing = *p* since each standardized variable has unit variance). Thus, to standardize or not is a nontrivial decision, since standardization gives each of the $X_1, X_2, ..., X_p$ the same equal weight.

[6] In some applications, like regression, you may prefer to retain certain low-variance PCs in preference to higher-variance PCs if the low-variance PCs are more strongly related to the Y variable. But you will still seek to discard PCs with low relevance to Y.

[7] The selection of principal components to replace $X_1, X_2, ..., X_p$ recapitulates the issue of model selection in regression – wherein the goals of *explanatory power* and *parsimony* must be balanced against each other. One furthers the goal of explanatory power by including more predictors; one furthers parsimony by including fewer predictors.

(10) The principal components are chosen to "front-load" their variability. That is, PRIN1 will have the largest variance among all *p* principal components, and Var(PRIN1) will exceed the variance of all other possible linear combinations of $X_1, X_2, ..., X_p$ [8]. PRIN2 will have the second highest variance among the *p* principal components, and Var(PRIN2) will exceed the variance of all other possible linear combinations of $X_1, X_2, ..., X_p$ that are uncorrelated with PRIN1.[9,10] Let λ_i = Var(PRINi). Then $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_p$, and $\lambda_1 + \lambda_2 + \lambda_3 + ... + \lambda_p$ = Var(X_1) + Var(X_2) + ... + Var(X_p), which = *p* if the X's have been standardized.

Further note on standardizing the X's (or not): Since PRIN1 tries to load up on variance, PRIN1 will give extra weight to an X with large variance in order to grab as much of the variance of the X as possible. That is, the principal components will cater toward X's with large variance and shun X's with small variance. This property seems undesirable, since a change in the units of measurement of an X from feet to inches (say) would multiply Var(X) by 144, making X seem 144 times as important – when nothing intrinsically has changed. If the X's are all standardized, then their variances are unaffected by the scale of measurement: The variances are all 1. Standardization neutralizes catering on the basis of size (mean) and scale (variance). That is the primary reason that the X's are usually standardized. But if there is nothing to choose among the X's on the basis of mean or variance, then how does PCA choose coefficients? Answer: Solely on the basis of correlations among the X's (for standardized X's).

(11) A mathematical note:

a) PRINi is the eigenvector of the i^{th} largest eigenvalue of the correlation matrix of $X_1, X_2, ..., X_p$ (or, of the covariance matrix of $X_1, X_2, ..., X_p$ if the data are centered instead of standardized).
b) Var(PRINi) = i^{th} largest eigenvalue of the correlation matrix of $X_1, X_2, ..., X_p$ (or, of the covariance matrix of $X_1, X_2, ..., X_p$ if the data are centered instead of standardized).

[8] A problem: The variance of a linear combination can be made arbitrarily large by multiplying the linear combination by a large number: Var(k*PRINi) = k^2 Var(PRINi). But multiplying by a constant does not change the *relative* weights given to the X-variables involved in the linear combination. In attempting to maximize the variance of the principal component, you could fall into an unending fruitless search by just multiplying the chosen component by ever larger numbers. To avoid this, it is necessary to impose a constraint on the coefficients of the principal component: the sum of squared coefficients = 1. This constraint makes the choice of coefficients unique for the principal components.

[9] Subject to the sum of squared coefficients = 1.

[10] In this respect, PCA parallels forward stepwise regression, in that each new PRIN takes as big a bite as possible out of the remaining variability of $X_1, X_2, ..., X_p$. PCA is a "greedy" optimizer.

(12) Some properties of PRIN1 = $\beta_{11}X_1 + \beta_{12}X_2 + \cdots + \beta_{1p}X_p$:

a) PRIN1 is the unique line which is closest to the X-observations in the sense of minimizing the sum of squared *perpendicular* distances from the observations to the line. This is in contrast to regression, which minimizes the sum of squared distances in the direction of the dependent variable.

b) The coefficients of PRIN1 are chosen to make PRIN1 have larger variance than any other linear combination of $X_1, X_2, ..., X_p$, subject to $\beta_{11}^{\ 2} + \beta_{12}^{\ 2} + \cdots + \beta_{1p}^{\ 2} = 1$. This restriction says that the length of the coefficient vector ($\beta_{11}, \beta_{12}, \cdots, \beta_{1p}$) is 1. Because doubling the coefficients quadruples the variance, there is no maximum variance unless we impose some restriction on the coefficients. This restriction also permits the principal component transformation to be orthonormal because it ensures that the unit basis vectors will retain unit length under the transformation.

c) The preceding property can be demonstrated mathematically: Let **X** be a *n*x*p* matrix of standardized variables. The rows are the *n* observations and the columns are the *p* variables. Let $\boldsymbol{\beta} = (\beta_{11}, \beta_{12}, \cdots, \beta_{1p})^T$ be a column vector of variance-maximizing coefficients yet to be determined. The problem in PCA is to choose $\boldsymbol{\beta}$ to maximize the variance of the *n*x1 column vector $\mathbf{X}\boldsymbol{\beta} = \left(\sum_{j=1}^{p} \beta_{1j} \frac{x_{1j} - \bar{x}_{.j}}{s_j}, \sum_{j=1}^{p} \beta_{1j} \frac{x_{2j} - \bar{x}_{.j}}{s_j}, \cdots, \sum_{j=1}^{p} \beta_{1j} \frac{x_{nj} - \bar{x}_{.j}}{s_j} \right)^T$ subject to $\boldsymbol{\beta}^T\boldsymbol{\beta} = 1$. Observe that the variance of $\mathbf{X}\boldsymbol{\beta}$ is the scalar $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \sum_{m=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{n} \beta_{1m}\beta_{1j} \frac{x_{km} - \bar{x}_{.m}}{s_m} \frac{x_{kj} - \bar{x}_{.j}}{s_j}$. Also note that $\mathbf{X}^T\mathbf{X}$ is the *p*x*p* correlation matrix of the data. To maximize $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$ subject to $\boldsymbol{\beta}^T\boldsymbol{\beta} = 1$, introduce the Lagrange multiplier λ and maximize $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \lambda(1 - \boldsymbol{\beta}^T\boldsymbol{\beta})$. Differentiate partially with respect to each of $\beta_{11}, \beta_{12}, \cdots, \beta_{1p}$ and λ, and set the resulting *p* + 1 partial derivatives equal to zero: $2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\lambda\boldsymbol{\beta} = \mathbf{0}$ and $1 - \boldsymbol{\beta}^T\boldsymbol{\beta} = 0$. Hence, critical points satisfy $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \lambda\boldsymbol{\beta}$ and the side constraint $\boldsymbol{\beta}^T\boldsymbol{\beta} = 1$. But $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \lambda\boldsymbol{\beta}$ is the defining condition for an eigenvalue and associated eigenvector. Thus, the variance-maximizing $\boldsymbol{\beta}$ is an eigenvector of the correlation matrix $\mathbf{X}^T\mathbf{X}$ of the data, and the Lagrange multiplier λ is the associated eigenvalue. Moreover, by pre-multiplying $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \lambda\boldsymbol{\beta}$ by $\boldsymbol{\beta}^T$, we have $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}^T\lambda\boldsymbol{\beta} = \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} = \lambda$, which says that the maximum variance $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$ is the eigenvalue λ.

d) The formula for $\beta_{1j} = corr(\text{PRIN1}, X_j) / \sqrt{\lambda_1}$,[11] where λ_1 = Var(PRIN1). This formula provides a key to interpreting the meaning of PRIN1 *[see (15) below.]* The formula says that the coefficients of PRIN1 are all proportional to the correlations between PRIN1 and the X's.

[11] This formula applies for standardized data. If the data are only centered, replace λ_1 in this formula by $\lambda_1 / Var(X_j)$.

e) Therefore, $corr(\text{PRIN1}, X_j) = \beta_{1j}\sqrt{\lambda_1}$. Corr(PRIN1, X_j) is called a **loading** and plays a big role in interpreting the meaning of PRIN1 *[see (15) below.]*

f) Therefore, if you do a simple linear regression of PRIN1 as the response variable on X_j as the predictor variable, the R-square of that regression = $\beta_{1j}^2\lambda_1$. But also if you do a simple linear regression of X_j on PRIN1, the R-square of that regression is the same = $\beta_{1j}^2\lambda_1$. Thus, this R-square value can be interpreted:

- as the proportion of variance of PRIN1 explained by X_j (hence the importance of X_j to PRIN1), or
- as the proportion of variance of X_j explained by PRIN1 (hence the importance of PRIN1 to X_j).

(13) Some properties of PRIN2 = $\beta_{21}X_1 + \beta_{22}X_2 + \cdots + \beta_{2p}X_p$:

a) PRIN2 is the unique line which is closest to the perpendicular residuals from fitting PRIN1 to the observations *[see (12)]*, in the sense of minimizing the sum of squared perpendicular distances from the residuals to the line.

b) The coefficients of PRIN2 are chosen to make PRIN2 have larger variance than any other linear combination of $X_1, X_2, ..., X_p$, *subject to PRIN2 being uncorrelated with PRIN1* and $\beta_{21}^2 + \beta_{22}^2 + \cdots + \beta_{2p}^2 = 1$. The latter restriction says that the length of the coefficient vector $(\beta_{21}, \beta_{22}, \cdots, \beta_{2p})$ is 1.

c) The preceding property can be demonstrated mathematically in a manner similar to the corresponding demonstration for PRIN1 in (12).

d) The formula for $\beta_{2j} = corr(\text{PRIN2}, X_j)/\sqrt{\lambda_2}$,[12] where λ_2 = Var(PRIN2). This formula provides a key to interpreting the meaning of PRIN2 *[see (15) below.]* The formula says that the coefficients of PRIN2 are all proportional to the correlations between PRIN2 and the X's.

e) Therefore, $corr(\text{PRIN2}, X_j) = \beta_{2j}\sqrt{\lambda_2}$. Corr(PRIN2, X_j) is called a **loading** and plays a big role in interpreting the meaning of PRIN2 *[see (15) below.]*

f) Therefore, if you do a simple linear regression of PRIN2 on X_j, the R-square of that regression = $\beta_{2j}^2\lambda_2$. But also if you do a simple linear regression of X_i on PRIN2, the R-square of that regression is the same = $\beta_{2j}^2\lambda_2$. Thus, this value can be interpreted:

- as the proportion of variance of PRIN2 explained by X_j (hence the importance of X_j to PRIN2), or
- as the proportion of variance of X_j explained by PRIN2 (hence the importance of PRIN2 to X_j).

[12] This formula applies for standardized data. If the data are only centered, replace λ_2 in this formula by λ_2 / Var(X_i).

(14) In general, the pattern established in the above discussion of PRIN1 and PRIN2 continues with subsequent PRIN's. PRINj is uncorrelated with all j-1 preceding PRIN's, provides the best perpendicular fit to the residuals from the fit of PRIN(j-1), and maximizes the variance among all uncorrelated linear combinations with coefficient vector length = 1, etc.

(15) **INTERPRETATION OF THE MEANING OF PRINCIPAL COMPONENTS**. Either the loadings (correlations) or the β-coefficients may be used to divine a meaning for PRINi. Many analysts prefer to use the loadings for this purpose because the loadings are really correlations, although the β_{ij}'s may also be used because, for a given PRINi, the β_{ij}'s are equal to the loadings times a constant of proportionality *[see (12d,e) and (13d,e) above]*. Analysts look at the pattern of loadings on PRINi to interpret the meaning of PRINi. For example, suppose X_1 and X_2 have high positive loadings on PRINi, and X_4 has a high negative loading on PRINi, and no other X has a high loading. Then these three X's are highly correlated with PRINi and largely determine the way PRINi behaves: When X_1 and/or X_2 go up, so does PRINi; when X_4 goes up, PRINi goes down. Thus, the meaning of PRINi derives from the common meaning of X_1 and X_2 in contrast with the meaning of X_4. In general, look to the variables that load high on PRINi to determine a meaning for PRINi.[13] Often, the analyst's imagination takes flight, with many creative or imaginative renderings offered for the meaning. The interpretation of PRIN's is an art.

(16) It follows immediately from the above relationships in (12) or (13) that $\lambda_i = \lambda_i \cdot 1 = \lambda_i(\beta_{i1}^2 + \beta_{i2}^2 + \cdots + \beta_{ip}^2) = \lambda_i\beta_{i1}^2 + \lambda_i\beta_{i2}^2 + \cdots + \lambda_i\beta_{ip}^2$ = R-square(PRINi, X_1) + R-square(PRINi, X_2) + ... + R-square(PRINi, X_p).[14] This can be interpreted as saying that the explanation for the variability of the i^{th} principal component can be found by adding up the explanations provided by each of the X's.

(17) Just as (16) says that PRINi can be explained by the X's collectively, so each X can be explained by the PRIN's collectively: It is a mathematical fact that the coefficients of the predictors in multiple regression are the same as in simple regression if the predictors are uncorrelated with each other. Now, the PRIN's are all uncorrelated with each other by the way that they are constructed. Therefore, if you regress X_j on PRIN1, PRIN2, …, PRINp, then the coefficients of the multiple regression are the same as in simple linear regressions of X_j on each PRIN. But we know what those simple linear regression coefficients are, because in any simple linear regression of a generic response Y on a generic predictor X, the slope coefficient has the general formula $\text{cov}(Y, X)\dfrac{stdev(Y)}{stdev(X)}$.

[13] It is unwise to be dogmatic about how high a loading must be to be considered high enough to be included in the meaning of a PRIN. But note that if the loading (correlation) of X_j on PRINi is below 0.50, then X_j explains less than 25% of the variability of PRINi. Some analysts use 0.50 as a guideline, but the whole pattern of loadings should be examined.

[14] This formula applies for standardized data.

Thus, the coefficient of PRINi is $corr(\text{PRINi}, X_j)\frac{StDev(X_j)}{StDev(\xi_i)} = corr(\text{PRINi}, X_j)\frac{1}{\sqrt{\lambda_i}} = \beta_{ij}$.

Thus the multiple regression equation is $\hat{X}_j = \beta_{1j}\xi_1 + \beta_{2j}\xi_2 + \beta_{3j}\xi_3 + \cdots + \beta_{pj}\xi_p$. Because no information is lost by using the PRINs (ξ's) instead of the X's, the residuals are zero, and so

$$X_j = \beta_{1j}\xi_1 + \beta_{2j}\xi_2 + \beta_{3j}\xi_3 + \cdots + \beta_{pj}\xi_p \text{ } \textit{exactly}. \text{ (}\mathbf{Eqn\ 6}\text{)}$$

The R-square of this multiple regression is 1 = R-square(PRIN1, X_j) + ... + R-square(PRINp, X_j) = corr^2(PRIN1, X_j) + ... + corr^2(PRINp, X_j) = $\lambda_1\beta_{1j}{}^2 + \lambda_2\beta_{2j}{}^2 + \cdots + \lambda_p\beta_{pj}{}^2$.[15]

(18) Eqn 6 provides the inverse equations for changing the PRINs back into X's:

$$\left\{\begin{array}{lcl} X_1 &=& \beta_{11}\xi_1 + \beta_{21}\xi_2 + \beta_{31}\xi_3 + \cdots + \beta_{p1}\xi_p \\ X_2 &=& \beta_{12}\xi_1 + \beta_{22}\xi_2 + \beta_{32}\xi_3 + \cdots + \beta_{p2}\xi_p \\ X_3 &=& \beta_{13}\xi_1 + \beta_{23}\xi_2 + \beta_{33}\xi_3 + \cdots + \beta_{p3}\xi_p \\ \vdots & \vdots & \vdots \\ X_p &=& \beta_{1p}\xi_1 + \beta_{2p}\xi_2 + \beta_{3p}\xi_3 + \cdots + \beta_{pp}\xi_p \end{array}\right\} \text{(Eqn 7)}$$

Eqn 4 and Eqn 7 therefore provide the principal components coordinate rotations that change our perspective on the data from the original X's (uninteresting) perspective to the new PRINs (interesting) perspective – and back. Expressed in matrix notation, these rotations are:

$$\begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_p \end{bmatrix} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} \text{ (Eqn 8 – equiv Eqn 4)}$$

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \beta_{11} & \beta_{21} & \cdots & \beta_{p1} \\ \beta_{12} & \beta_{22} & \cdots & \beta_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{1p} & \beta_{2p} & \cdots & \beta_{pp} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_p \end{bmatrix} \text{ (Eqn 9 – equiv Eqn 7)}$$

Both rotations are orthonormal transformations. Therefore, the matrices of the transformations are inverses of each other, in addition to being transposes of each other:

$$\begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pp} \end{bmatrix} \cdot \begin{bmatrix} \beta_{11} & \beta_{21} & \cdots & \beta_{p1} \\ \beta_{12} & \beta_{22} & \cdots & \beta_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{1p} & \beta_{2p} & \cdots & \beta_{pp} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \text{ (Eqn 10)}$$

[15] This formula applies for standardized data.

(19) Recall that for standardized X's, $p = \lambda_1 + ... + \lambda_p = Var(PRIN1) + ... + Var(PRINp)$. If the X's are all independent, then $\lambda_1 = ... = \lambda_p = 1$. Otherwise, some λ's > 1 and others < 1. Note that this can be interpreted as guaranteeing the success of PCA, in the sense that PCA identifies some variables with lower variance (information) that may be candidates for dimensionality reduction – although the actual benefit could be minimal.

(20) Suppose that you are deciding how many PRINs to use *[also see (7) and (8)]*. If Var(PRINi) < 1, then PRINi explains less variance than any standardized X explains, because each standardized X has variance = 1. Therefore, any PRIN with a $\lambda < 1$ is a candidate for being discarded on the grounds that you could do better with an original X: A PRIN with a $\lambda < 1$ could be replaced by any standardized variable and the total variance would go up. However, there may be other reasons to keep PRIN's with low eigenvalues, e.g., some theoretically important variables may not be adequately represented by PRIN's with big λ's.

What if $\lambda = 0$ for a PRIN? Then the variance of that PRIN is zero. A variable that has zero variance is a constant! Therefore, that PRIN has no explanatory power and definitely should be discarded. Geometrically, the meaning of $\lambda = 0$ is that all of the data are contained in a flat hyperplane of zero thickness. Thus, one of the dimensions is definitely unnecessary. And if λ is close to zero, then the variance of its PRIN is nearly zero, so the data are concentrated in a flat hyperplane of *nearly* zero thickness. Thus, one of the dimensions is very close to being unnecessary, and plays little role in the explanatory power of the data.

(21) Rules that have been proposed for discarding PRINs:

a) Kaiser: Discard all PRINs with $\lambda < 1$.
b) Jolliffe: Discard all PRINs with $\lambda < 0.7$.
c) Cattell: Draw "scree" plot (a plot of λ vs. eigenvalue number); this will be monotone decreasing; discard all PRINs after the point where the graph appears to level out.
d) Use enough PRINs to explain 75-80% of the total variance.
e) Use all interpretable PRINs with sizable λ's.
f) Horn's "parallel procedure" (formalized scree plot).

(22) Only the correlation matrix of the X's is necessary to calculate the PRINs *[see (11)]*. The original values of the X's are not necessary unless you want to calculate scores of the X's on each PRIN (often interesting). SAS has an option for inputting the correlation matrix only. Analysis of the standardized X's is equivalent to analyzing the correlation matrix.

(23) PCA on the centered X's is equivalent to analyzing the variance-covariance matrix of the X's (the **COV** option in SAS). SAS has an option for inputting the covariance matrix only. The PRINs that result from PCA on centered X's are not the same as on standardized X's, nor are they simple algebraic functions of the PRINs for standardized X's.

(24) Criteria for choosing standardized vs centered PCA *[also see (10)]*:

- With standardized X's, all variables are treated as equally important because the variance of any standardized variable is 1. So unimportant variables influence PRINs as much as important variables.
- With centered X's, the variances of the X's are not equal, so the variance-maximizing feature of PCA results in PRINs that emphasize X's with large variances, even if the only reason for those large variances is a larger scale of measurement.

Doing Principal Components Analysis with SAS

The SAS procedure for PCA is **PROC PRINCOMP**. SAS will automatically standardize your data when you run PROC PRINCOMP, unless you use the **COV** option, in which case SAS will only center your data:

SAS set-up (SAS automatically standardizes):	Example:
PROC PRINCOMP DATA=<sas ds>; **VAR <X1 X2 ...Xp>;**	**PROC PRINCOMP DATA=STOCKS;** **VAR PROFIT GROWTH;**

SAS set-up (SAS centers data):	Example:
PROC PRINCOMP DATA=<sas ds> COV; **VAR <X1 X2 ...Xp>;**	**PROC PRINCOMP DATA=STOCKS COV;** **VAR PROFIT GROWTH;**

SAS output from this set-up comes in four parts:

1) Simple summary statistics about $X_1, X_2, ..., X_p$
2) Correlation matrix of $X_1, X_2, ..., X_p$ (covariance matrix if **COV** option used).
3) Table of eigenvalues of the correlation matrix (of the covariance matrix if COV option is used). The eigenvalue for each PRIN is the variance of the PRIN. So the sum of the eigenvalues is the total variance. The table also shows the proportion of the total variance accounted for by each PRIN and the proportion of variance accounted for by each PRIN and all lower-numbered PRINs. This is useful in deciding how many PRINs to retain.
4) Table of eigenvectors. These are the coefficients of the principal components. The column of values under PRINi in this table shows the β_{ij}'s for PRINi. This is useful in divining a meaning for PRINi.