CPH 636 — Spring 2009 — Dr. Charnigo

Lecture 11

Principal components analysis

Introduction. The goal of principal components analysis is to convert a (possibly large) collection of highly correlated continuous variables $X_1, \ldots, X_k$ into a (smaller) collection of uncorrelated continuous variables $P_1, \ldots, P_j$ that, in a sense to be made precise later, capture as much "energy" as possible from $X_1, \ldots, X_k$.

Principal components analysis, while itself an unsupervised learning technique, is a useful adjunct to supervised learning techniques.

Suppose that we wish to incorporate $X_1, \ldots, X_k$ into a linear regression model but that we encounter collinearity. [[For background on collinearity, see {web.as.uky.edu/statistics/users/rjchar2/CPH930F06/L3930F06.pdf}.]] One approach to addressing collinearity is to convert $X_1, \ldots, X_k$ into $P_1, \ldots, P_j$ through principal components analysis. Then we can use $P_1, \ldots, P_j$ in the linear regression model instead of $X_1, \ldots, X_k$. [[For another way to address collinearity, see {www.richardcharnigo.net/CPH931F08/L1931F08.pdf}.]]

Even if we are not concerned about collinearity, principal components analysis can help us to overcome the curse of dimensionality. Thus, principal components analysis may also be useful as a data preprocessing device before we conduct a nearest neighbors analysis.

Mathematical formulation. In what follows, we assume that $X_1, \ldots, X_k$ have been standardized (i.e., have mean 0 and standard deviation 1). Fernandez's FACTOR macro takes care of this automatically before performing principal components analysis.

We define

$$P_1 := a_{11}X_1 + a_{12}X_2 + \cdots + a_{1k}X_k,$$

where $a_{11}, a_{12}, \ldots, a_{1k}$ are chosen to maximize $Var(P_1)$ subject to the constraint that

$$a_{11}^2 + a_{12}^2 + \cdots + a_{1k}^2 = 1.$$

This maximization is done by computer. We refer to P_1 as the first principal component.

We then define

$$P_2 := a_{21}X_1 + a_{22}X_2 + \cdots + a_{2k}X_k,$$

where $a_{21}, a_{22}, \ldots, a_{2k}$ are chosen to maximize $Var(P_2)$ subject to the constraints that

$$a_{21}^2 + a_{22}^2 + \cdots + a_{2k}^2 = 1$$

and $Corr(P_1, P_2) = 0$.

We define $P_3, \ldots, P_k$ similarly. Note that $P_{j+1}, \ldots, P_k$ are defined mathematically even though we have no intention of using them practically.

Also, even though we have defined $P_1, \ldots, P_k$ in terms of $X_1, \ldots, X_k$, we can express $X_1, \ldots, X_k$ in terms of $P_1, \ldots, P_k$. In particular, there exist constants $b_{11}, b_{12}, \ldots, b_{1k}$ such that

$$X_1 = b_{11}P_1 + b_{12}P_2 + \cdots + b_{1k}P_k,$$

there exist constants $b_{21}, b_{22}, \ldots, b_{2k}$ such that

$$X_2 = b_{21}P_1 + b_{22}P_2 + \cdots + b_{2k}P_k, \quad \text{and so forth.}$$

[[Let $\mathbf{X}$ be the $k \times 1$ vector whose i^{th} element is X_i, let $\mathbf{P}$ be the $k \times 1$ vector whose i^{th} element is P_i, and let $\mathbf{A}$ be the $k \times k$ matrix whose $(i,j)^{th}$ element is a_{ij}. Then the representation $\mathbf{P} = \mathbf{AX}$ yields $\mathbf{X} = \mathbf{BP}$, where $\mathbf{B} = \mathbf{A}^{-1}$ is the $k \times k$ matrix whose $(i,j)^{th}$ element is b_{ij}.]]

Illustrative example. I will discuss some practical aspects of principal components analysis in the course of presenting an illustrative example. In what follows, I refer to output {RCLib.Chol81.doc} and {RCLib.Chol82.doc} available from my web page.

The data set {chol.sas7bdat} contains eleven variables. A natural response variable in this data set would be TC (total serum cholesterol). Natural explanatory variables would be EN (dietary energy), TFAT (total fat intake), SFAT (saturated fat intake), PFAT (polyunsaturated fat intake), VFAT (vegetable fat intake), AFAT (animal fat intake), CHOL (dietary cholesterol), FIBER (dietary fiber), AL (alcohol intake), and BMI (body mass index). Unfortunately, the natural explanatory variables have a severe collinearity problem. Thus, I would like to apply principal components analysis to acquire a (smaller) collection of uncorrelated continuous variables, which could then take the place of the natural explanatory variables in a linear regression model.

For convenience, I let $X_1, \ldots, X_{10}$ denote the standardized versions of EN, TFAT, SFAT, PFAT, VFAT, AFAT, CHOL, FIBER, AL, and BMI. Although you do not have a written assignment on principal components analysis, instructions for using Fernandez's FACTOR macro are available in {SASMacroInstr7.txt}.

Exploratory output. The content of {RCLib.Chol81.doc} is purely exploratory and does not present the results of principal components analysis. The main reason for showing you {RCLib.Chol81.doc} is to demonstrate that the FACTOR macro can be a useful adjunct to the UNIVAR macro. While the UNIVAR macro examines the distributions of individual continuous variables, the FACTOR macro can be employed to examine the distributions of pairs of continuous variables.

Assessing multivariate normality. Page 4 of {RCLib.Chol82.doc} displays a quantile-quantile plot. If $X_1, \ldots, X_{10}$ arose from a multivariate normal distribution, then the configuration of points in this plot should be roughly linear. However, we see a clear departure from linearity, along with very small p-values for tests of multivariate skewness and multivariate kurtosis. Hence, we do not believe that $X_1, \ldots, X_{10}$ arise from a multivariate normal distribution. Fortunately, there is no requirement for multivariate normality in principal components analysis.

Eigenvalues. Page 7 provides "eigenvalues". [[Let $\mathbf{V}$ be the $k \times k$ matrix whose $(i,j)^{th}$ element is $Corr(X_i, X_j)$. A nonzero number λ is called an eigenvalue of $\mathbf{V}$ if there exists a nonzero $k \times 1$ vector $\mathbf{U}$ such that $\mathbf{VU} = \lambda\mathbf{U}$. The nonzero $k \times 1$ vector $\mathbf{U}$ is called an eigenvector.]] These eigenvalues are precisely $Var(P_1)$ through $Var(P_{10})$.

Using linear algebra, one can prove that

$$Var(P_1) + \cdots + Var(P_{10}) = Var(X_1) + \cdots + Var(X_{10}) = 10.$$

If we regard $Var(X_1) + \cdots + Var(X_{10})$ as the "energy" in $X_1, \ldots, X_{10}$, then $Var(P_1)$ is the amount of energy captured by P_1 and $Var(P_1)/10$ is the fraction of the energy captured by P_1. Likewise, $Var(P_2)$ is the amount of energy captured by P_2 and $Var(P_2)/10$ is the fraction of the energy captured by P_2.

In the present example, P_1 captures 52.49% of the energy in $X_1, \ldots, X_{10}$, while P_1 and P_2 together capture 64.43% of the energy.

Eigenvectors. Page 7 also provides "eigenvectors". These eigenvectors define the principal components. Explicitly,

$$P_1 = 0.38332X_1 + 0.42317X_2 + \cdots + 0.03169X_{10},$$

$$P_2 = 0.27230X_1 - 0.09239X_2 + \cdots + 0.39668X_{10},$$

and

$$P_3 = -0.11471X_1 + 0.08967X_2 + \cdots + 0.65651X_{10}.$$

The remaining seven eigenvectors are suppressed; when I ran the FACTOR macro, I stated that I wanted details only for the first three principal components.

Discussion questions. How can we interpret P_1 scientifically? What about P_2 and P_3?

Correlations with principal components. Page 8 provides Pearson correlations between $X_1, \ldots, X_{10}$ and P_1, P_2, P_3. For instance, the Pearson correlation between X_1 and P_1 is 0.87820.

Scree plot. Page 9 provides a scree plot, which is a graphical representation of the eigenvalues. Some people use a scree plot to choose j, the number of principal components with which they will replace $X_1, \ldots, X_k$. Specifically, these people look for an "elbow" point in the scree plot (i.e., a point at which the eigenvalues plateau).

In the present example, there are elbow points at 2 and at 6, which suggests that 1 and 5 may be reasonable choices for j.

On the other hand, some people choose j to be the number of principal components with greater-than-unit variances (i.e., the number of eigenvalues greater than 1), while others choose j to be the smallest number at which a certain percentage of the energy has been captured.

In the present example, there are 3 principal components with greater-than-unit variances, and 3 is the smallest number of principal components at which 70% of the energy has been captured.

Scatter plots. Page 12 provides a scatter plot of P_2 against P_1. The arrows labeled with variable names are intended to suggest that individuals with large values of $X_1, \ldots, X_8$ have large values of P_1, while individuals with large values of X_9 have large values of P_2.

Pages 13 and 14 provide scatter plots of P_3 against P_1 and P_3 against P_2.

Factor analysis

Introduction. The goal of factor analysis is to model a (possibly large) collection of highly correlated continuous variables $X_1, \ldots, X_k$ in terms of a (smaller) collection of uncorrelated variables $F_1, \ldots, F_j$ called factors.

The preceding description of factor analysis sounds a lot like the description of principal components analysis. However, there is a key conceptual difference. In principal components analysis, we define $P_1, \ldots, P_j$ in terms of $X_1, \ldots, X_k$ because the latter variables are too unwieldy for supervised learning. In factor analysis, we model $X_1, \ldots, X_k$ in terms of $F_1, \ldots, F_j$ because we believe that the former variables are imperfect representations of scientific phenomena described more authentically by the latter variables. Borrowing the language of behavioral scientists, $F_1, \ldots, F_j$ are unobservable constructs while $X_1, \ldots, X_k$ are observable indicators.

Mathematical formulation. In what follows, we assume that $X_1, \ldots, X_k$ have been standardized (i.e., have mean 0 and standard deviation 1). Fernandez's FACTOR macro takes care of this automatically before performing factor analysis.

We postulate that there exists $j \leq k$ such that

$$X_1 = l_{11}F_1 + l_{12}F_2 + \cdots + l_{1j}F_j + \epsilon_1,$$

$$X_2 = l_{21}F_1 + l_{22}F_2 + \cdots + l_{2j}F_j + \epsilon_2, \cdots,$$

$$X_k = l_{k1}F_1 + l_{k2}F_2 + \cdots + l_{kj}F_j + \epsilon_k,$$

where: $F_1, F_2, \ldots, F_j, \epsilon_1, \epsilon_2, \ldots, \epsilon_k$ are mutually independent; $F_1, F_2, \ldots, F_j$ have mean 0 and standard deviation 1; and, $\epsilon_1, \epsilon_2, \ldots, \epsilon_k$ have mean 0. [[**Discussion question**: Why did I not make any assumption about the standard deviations of $\epsilon_1, \epsilon_2, \ldots, \epsilon_k$?]]

We refer to l_{11} as the loading of X_1 on F_1, to l_{12} as the loading of X_1 on F_2, and so forth.

Illustrative example. I will discuss some practical aspects of factor analysis in the course of presenting an illustrative example. In what follows, I refer to output {RCLib.nphpsp85.doc} and {RCLib.nphpsp86.doc} available from my web page.

The data set (not posted on my web page because it is not for public consumption) consists of ten variables called EPHS1, EPHS2, EPHS3, EPHS4, EPHS5, EPHS6, EPHS7, EPHS8, EPHS9, and EPHS10. These ten variables represent the evaluations of local health departments on how well they carried out the ten essential public health services. I am applying factor analysis to see whether the evaluations on the ten essential public health services can be expressed in terms of a few underlying constructs.

For convenience, I let $X_1, \ldots, X_{10}$ denote the standardized versions of EPHS1, EPHS2, EPHS3, EPHS4, EPHS5, EPHS6, EPHS7, EPHS8, EPHS9, and EPHS10. Although you do not have a written assignment on factor analysis, instructions for using Fernandez's FACTOR macro are available in {SASMacroInstr7.txt}.

Exploratory output. The content of {RCLib.nphpsp85.doc} is purely exploratory and does not present the results of factor analysis. However, this output provides us with some preliminary understanding of how EPHS1, EPHS2, EPHS3, EPHS4, EPHS5, EPHS6, EPHS7, EPHS8, EPHS9, and EPHS10 are related.

Assessing multivariate normality. Page 4 of {RCLib.nphpsp86.doc} displays a quantile-quantile plot to assess whether $X_1, \ldots, X_{10}$ have a multivariate normal distribution. [[This is because we want to use maximum likelihood to estimate the loadings, but to use maximum likelihood we must make a distributional assumption.]] While the p-values for multivariate skewness and multivariate kurtosis are small, the overall appearance of the quantile-quantile plot is very good. We would not have serious reservations about assuming multivariate normality in this example.

Hypothesis test. Page 9 provides results for a test of the null hypothesis that 3 is an acceptable choice for j. The p-value is 0.0514, suggesting that 3 is an acceptable choice. This test is predicated on the assumption that $X_1, \ldots, X_{10}$ have a multivariate normal distribution.

Estimates of loadings and varimax rotation. Page 10 provides maximum likelihood estimates of the loadings. For instance, $\hat{l}_{51} = 0.85$. All of $X_1, \ldots, X_{10}$ load heavily on F_1, while none of them load heavily on F_2 or F_3.

However, there are infinitely many combinations of estimated loadings and factors just as compatible with the observed data as the estimated loadings shown on page 10 and the factors to which they correspond. [[Let $\mathbf{F}$ denote the $j \times 1$ vector whose i^{th} element is F_i. Let $\mathbf{M}$ be any $j \times j$ matrix such that $\mathbf{M}^T = \mathbf{M}^{-1}$. Then $\mathbf{G} = \mathbf{MF}$ is a $j \times 1$ vector whose elements could be considered the factors instead of the elements of $\mathbf{F}$. If $\widehat{\mathbf{L}}$ denotes a $k \times j$ matrix of estimated loadings based on $\mathbf{F}$, then $\widehat{\mathbf{L}}\mathbf{M}^T$ is a matrix of estimated loadings based on $\mathbf{G}$. **Discussion question**: Does $\mathbf{G}$ have the same mean vector and variance/covariance matrix as $\mathbf{F}$?]]

Since we are performing factor analysis to identify constructs that underlie the evaluations on the ten essential public health services, we want a combination of estimated loadings and factors that is highly amenable to interpretation. We seek such a combination by performing "varimax rotation". [[This entails finding a matrix $\mathbf{M}$ as above such that a certain function of the estimated loadings is maximized.]]

Page 11 presents the results of varimax rotation. If we parse the ten essential public health services according to the factors on which their evaluations load most heavily, we see that: services 1, 2, 5, 6, and 10 are grouped together; services 3, 7, 8, and 9 are grouped together; and, service 4 is isolated from the others. Thus, after varimax rotation, the first factor can be roughly interpreted as a monitoring/investigating/innovating construct, the second factor can be roughly interpreted as an education/assessment construct, and the third factor can be roughly interpreted as a community partnership construct.

Approximations to the factors. The loadings describe how $X_1, \ldots, X_{10}$ are defined from the factors. We can alter our perspective and ask how the factors could be approximated from $X_1, \ldots, X_{10}$.

Page 12 shows that, after varimax rotation, the first factor could be approximated by

$$0.322X_{10} + 0.287X_2 + \cdots - 0.201X_4.$$

However, a more common practice would be to let the "scale"

$$\mathrm{EPHS1} + \mathrm{EPHS2} + \mathrm{EPHS5} + \mathrm{EPHS6} + \mathrm{EPHS10}$$

serve as a proxy for the first factor.

Scree plot. The scree plot on page 13 is remarkable mainly because it suggests that 1 might have been a reasonable choice for j. In fact, the results prior to varimax rotation were consistent with the idea of a single "overall performance" construct. Thus, this example was unusual in that varimax rotation actually complicated rather than simplified interpretation.