

Principal Components Analysis

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA

1 Population PCA

Let X be an $n \times p$ matrix whose *rows* are iid random vectors X_i with mean $\mu' \in \mathbb{R}^p$ and covariance $\Sigma \in \mathcal{S}_+^p$ —for example, they might be $(X_i)' \stackrel{\text{iid}}{\sim} \text{No}(\mu, \Sigma)$. For many problems (such as multivariate regression of some Y on X) we might wish to *reduce the dimension* p of these rows. For example, if we have a vector of $p = 1000$ possible explanatory variables about each individual, we may hope that a small subset of these or perhaps a few different linear combinations of these will capture most of the information in X .

For any vector $\alpha \in \mathbb{R}^p$ the random vector $z = X\alpha \in \mathbb{R}^n$ will have iid entries z_i whose means and variances are

$$\mathbb{E}[z_i] = \mu' \alpha \quad \mathbb{V}[z_i] = \alpha' \Sigma \alpha$$

If the values $X_{.j}$ don't vary much from μ_j for some fixed $j \in 1 : p$ then the j th column won't be very helpful in regression (or similar) problems; similarly if a linear combination $z = X\alpha$ has small variance then it won't contribute much. Thus we consider finding such linear combinations with the *maximal* variance as candidates for further study.

Of course the variance $\alpha' \Sigma \alpha$ can be made as large as desired by taking arbitrarily large values for α_i ; by restricting to some bounded set of α s we can make the problem of finding the “best” vectors α well-posed. Set:

$$\begin{aligned} \lambda_1 &= \sup \{ \alpha' \Sigma \alpha : \alpha' \alpha = 1 \} \\ \alpha_1 &= \text{The vector } \alpha \in \mathbb{R}^p \text{ where this supremum is attained} \\ z_1 &= X\alpha_1 \end{aligned}$$

and call λ_1 the *first principal value*, α_1 the *first principal direction*, and z_1 the *first principal component* of X .

We can find α_1 and λ_1 by solving the constrained optimization problem above using method of Lagrange Multipliers, seeking a stationary value for the Lagrangian

$$\begin{aligned}\mathcal{L}(\alpha, \lambda) &:= \alpha' \Sigma \alpha + \lambda[1 - \alpha' \alpha] \\ \frac{\partial}{\partial \alpha} \mathcal{L} &= 2 \Sigma \alpha - 2 \lambda \alpha \\ &= 0 \quad \text{implies} \quad \Sigma \alpha = \lambda \alpha; \\ \frac{\partial}{\partial \lambda} \mathcal{L} &= 1 - \alpha' \alpha \\ &= 0 \quad \text{implies} \quad \alpha' \alpha = 1, \text{ whereupon} \\ \alpha' \Sigma \alpha &= \alpha' [\lambda \alpha] = \lambda.\end{aligned}$$

Thus the solution is for λ_1 to be the largest eigenvalue of the positive-definite matrix Σ , and for α_1 to be a corresponding unit eigenvector.

Similarly, for $1 \leq j \leq p$ we can let λ_j be the j th largest (always nonnegative!) eigenvalue, with an orthonormal set $\{\alpha_j\}$ of unit eigenvectors, and call λ_j the j th *principal value*, α_j the j th *principal direction*, and $z_j = X \alpha_j$ the j th *principal component* of X . The eigenvalues are determined uniquely. If they are distinct then the eigenvectors are determined uniquely up to an arbitrary \pm sign; if they are not distinct, the *eigenspaces* are determined uniquely for each eigenvalue but we may have some choice in picking orthonormal bases for them.

The matrix $Z = [z_1 z_2 \dots z_p]$ whose j th column is $X \alpha_j$ can be written $Z = X A$ where $A = [\alpha_1 \alpha_2 \dots \alpha_p]$ is an orthogonal ($p \times p$) matrix that rotates the rows of X in \mathbb{R}^p so that they become uncorrelated with decreasing variances $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$:

$$E[Z] = \mathbf{1} \mu' A \quad \text{Cov}[Z_{ij}, Z_{kl}] = \delta_{ik} \cdot [A' \Sigma A]_{jl} = \delta_{ik} \cdot \Lambda_{jl} = \delta_{ik} \delta_{jl} \lambda_j$$

where $\Lambda := \text{diag}(\lambda_1, \dots, \lambda_p)$ and $\delta_{jk} = 1$ if $j = k$, otherwise zero. Of course since $AA' = A'A = I$ we could replace $X = ZA'$ with $Z = XA$ and have the same possible linear models $Y = XB + U = Z\tilde{B} + U$ with $\tilde{B} = A'B$; the more interesting possibility is to replace X in our modeling or computation with a smaller $n \times r$ matrix $Z_r = XA_r$ where $A_r = [\alpha_1 \alpha_2 \dots \alpha_r]$ consists of the first $r < p$ columns of A . For example, replacing a regression model “ $Y = XB + U$ ” with “ $Y = Z_r B_r + U$ ” reduces the dimension of the coefficient

matrix from $(p \times n)$ for B to $(r \times n)$ for B_r , with the hope of simplifying modeling and analysis without much loss of predictive power.

Note that PCA is NOT invariant under scaling— if a column of X is multiplied by a constant c (for example, if we switch from measuring length in inches to measuring it in millimeters) then its variance will change by a factor of c^2 , completely changing the principal values *and directions*. The customary advice is to measure each quantity in the same units, if possible, or perhaps to standardize each to have mean zero and variance one.

Since determinants and traces are invariant under orthogonal conjugation, $\text{tr } \Sigma = \sum \lambda_j$ and $|\Sigma| = \prod \lambda_j$ are frequently viewed as summaries of the variability of the $\{X_{ij}\}$.

1.1 Example: Two by Two

Consider the standardized 2×2 case with $p = 2$ and constant variance $\mathbb{E}[(X_{ij} - \mu_j)^2] = 1$. If we denote the covariance (also correlation) by $\rho = \mathbb{E}[(X_{i1} - \mu_1)(X_{i2} - \mu_2)]$, then the covariance matrix is

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

If $\rho \geq 0$ then the ordered eigenvalues are $\lambda_1 = (1 + \rho)$ and $\lambda_2 = (1 - \rho)$, with normalized eigenvectors $\alpha_1 = (1, 1)' / \sqrt{2}$ (proportional to the unweighted mean $(X_{\cdot 1} + X_{\cdot 2})/2$) and $\alpha_2 = (1, -1)' / \sqrt{2}$ (proportional to the difference $(X_{\cdot 1} - X_{\cdot 2})/2$). If $\rho < 0$ then the same ev's and EV's appear, but their order is reversed.

2 Sample PCA

In practice of course one rarely knows Σ , and so rarely can compute the necessary ev's $\{\lambda_j\}$ and EV's $\{\alpha_j\}$ as in Section (1). With a sufficiently large sample-size n , however, one can estimate the mean and covariance of X_i by

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i \leq n} X_i \in \mathbb{R}^p \quad \hat{\Sigma} = \frac{1}{n} (X - \mathbf{1}\bar{x}')' (X - \mathbf{1}\bar{x}') \in \mathbb{S}_+^p,$$

and, from these, the ev's and EV's $\{\lambda_j\}$ and $\{\alpha_j\}$ of $\hat{\Sigma}$ by those of $\hat{\Sigma}$:

$$\{\hat{\lambda}_j\} \quad \text{and} \quad \{\hat{\alpha}_j\}$$

to construct the *empirical* or *sample* principal components

$$\{\hat{z}_j = X\hat{\alpha}_j\}.$$

The “fraction of variation explained by the first r PC’s” is

$$F_r = \frac{\sum_{i \leq r} \hat{\lambda}_j}{\sum_{i \leq p} \hat{\lambda}_j},$$

a fraction that ranges from something over $1/p$ up to 1; it’s customary to plot this and choose r to be the smallest value for which F_r exceeds some comforting value like 90% or 99%.

If the eigenvalues $\{\lambda_1 > \lambda_2 > \dots > \lambda_p > 0\}$ of Σ are *distinct* and *strictly positive*, then the estimates are consistent and asymptotically efficient:

Theorem 1. *Under the stated conditions, for each $1 \leq j \leq p$*

$$\begin{aligned} \sqrt{\frac{n-1}{2}} \frac{\hat{\lambda}_j - \lambda_j}{\lambda_j} &\Rightarrow \text{No}(0, 1) \\ \left[\frac{n-1}{2} \right] \frac{(\hat{\lambda}_i - \lambda_i)(\hat{\lambda}_j - \lambda_j)}{\lambda_i \lambda_j} &\rightarrow 0, \quad i \neq j \\ \sqrt{n-1} (\hat{\alpha}_j - \alpha_j) &\Rightarrow \text{No}(0_p, D_j) \end{aligned}$$

as $n \rightarrow \infty$, where

$$D_j = \lambda_j \sum_{i \neq j} \frac{\lambda_i}{(\lambda_j - \lambda_i)^2} \alpha_i \alpha_i'.$$

With this one can construct interval estimates for $\{\lambda_j\}$ or F_r .

2.1 Computation

The MATLAB (or OCTAVE) function `princomp` (in the `Statistics` toolbox) performs sample PCA on an $n \times p$ data matrix X and returns the PCA direction matrix A , after automatically centering the data by subtracting column means. The eigenvalues (called *scores*) are also available, with the syntax `[COEFF SCORE] = princomp(X);`. For $p \geq n$ the option ‘`econ`’ (for econometrics, perhaps) returns the non-zero singular values and is much faster. Online help is available at URL

<http://www.mathworks.com/access/helpdesk/help/toolbox/stats/princomp.html>

The R functions `prcomp()` and `princomp()` offer similar functionality; try “?? `prcomp`” or go to URL <http://stat.ethz.ch/R-manual/R-patched/library/stats/html/prcomp.html> for more info. Thanks to our TA Anirban Bhattacharya for tracking down this information.

2.2 Example using R

Copy the file `dat/TestScores.dat` from our website to your R directory and run the command `scores <- dget("TestScores.dat");` to construct a data-frame consisting of scores (all measured in percent) on five tests (Mechanics, Vectors, Algebra, Analysis, and Statistics) for each of 88 engineering students¹. The first two tests were closed-book; the other three were open-book. The tests’ means and standard deviations are:

mec	vec	alg	ana	sta
38.95455	50.59091	50.60227	46.68182	42.30682
17.48622	13.14695	10.62478	14.84521	17.25559

Now the command `pca <- prcomp(scores);` will generate an object of class “prcomp” that includes the square roots of the eigenvalues $\{\lambda_j\}$ (you can see a plot of the cumulative sum of eigenvalues themselves by executing `plot(0:5, c(0, cumsum(pca$sdev^2)));`). Here is the rotation matrix $A = \{\alpha_j\}$, obtained as `pca$rotation` (the shorter `pca$rot` works too):

	PC1	PC2	PC3	PC4	PC5
mec	-0.5054457	-0.74874751	0.2997888	-0.296184264	-0.07939388
vec	-0.3683486	-0.20740314	-0.4155900	0.782888173	-0.18887639
alg	-0.3456612	0.07590813	-0.1453182	0.003236339	0.92392015
ana	-0.4511226	0.30088849	-0.5966265	-0.518139724	-0.28552169
sta	-0.5346501	0.54778205	0.6002758	0.175732020	-0.15123239

The first principal direction is essentially the (negative) average of the test scores; the second direction distinguishes the closed-book scores from the open-book scores. The third principal direction suggests that the first and last tests were a little different from the others (easier or harder?); these three represent about 90% of the variability of students’ scores.

Many authors recommend standardizing the data by replacing each column $X_{.j}$ (the j th test, in this example) with $(X_{.j} - \bar{X}_j)/s_j$. Here this will reduce the role of the first and fifth tests (which had higher sample variances) and

¹These data first appeared as Table 1.2.1 in *Multivariate Analysis* by Mardia, Kent & Bibby; they’re available in digital format as variable `scor` in the R package `bootstrap`.

increase that of the third. In R this can be accomplished simply by adding the argument “,scale=TRUE” to the `prcomp()` function; the result is:

	PC1	PC2	PC3	PC4	PC5
mec	-0.3996045	-0.6454583	0.62078249	-0.1457865	-0.1306722
vec	-0.4314191	-0.4415053	-0.70500628	0.2981351	-0.1817479
alg	-0.5032816	0.1290675	-0.03704901	-0.1085987	0.8466894
ana	-0.4569938	0.3879057	-0.13618182	-0.6662561	-0.4221885
sta	-0.4382444	0.4704545	0.31253342	0.6589164	-0.2340223

Again the first and second PC are essentially the average (now standardized) test score and the difference between open-book and closed-book scores.

The fractional cumulative total variation of the first r PC's can be found most easily by:

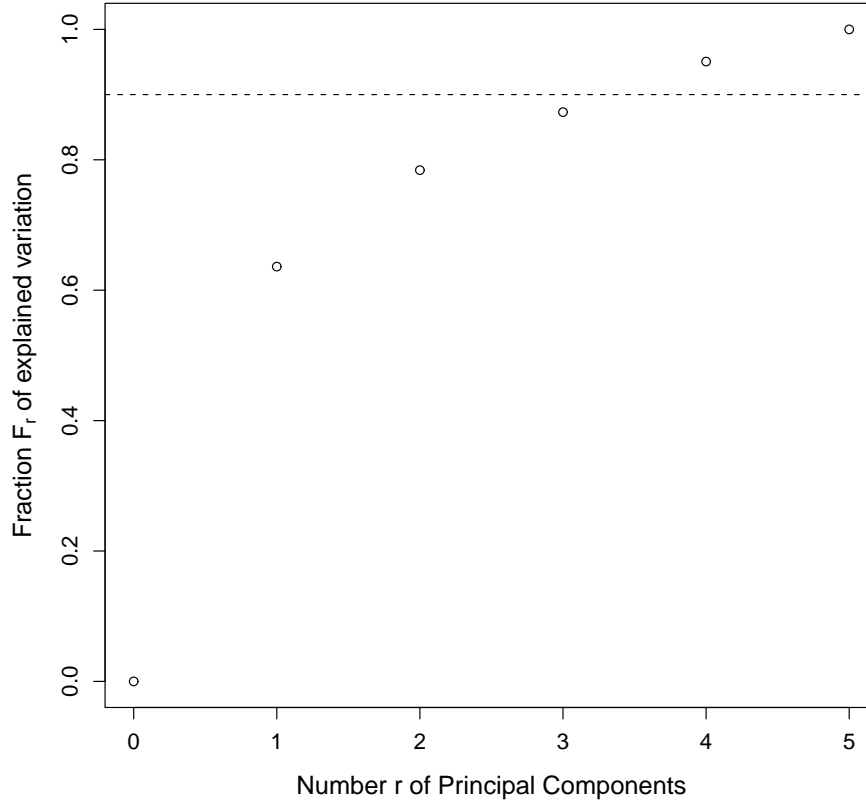
```
> pca <- prcomp(scores, scale=TRUE);
> summary(pca);
Importance of components:
```

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.784	0.860	0.667	0.6228	0.4966
Proportion of Variance	0.636	0.148	0.089	0.0776	0.0493
Cumulative Proportion	0.636	0.784	0.873	0.9507	1.0000

or can be shown as a plot by:

```
> plot(0:5, c(0,cumsum(pca$sdev^2))/sum(pca$sdev^2),
      cex.lab=1.3, cex.axis=1.3,
      xlab="Number r of Principal Components",
      ylab=expression(paste("Fraction ", F[r],
      " of explained variation")));
> abline(h=0.90, lty=2);
```

generating the plot:



Since the first three PC's capture 87% of the variability, and the first four 95%, it is tempting to proceed with a reduced model with only $r = 3$ or 4 variables. How will this affect the analysis?

2.2.1 Correlation and Variable Elimination

First some computation. The covariance between the j th test score and the k th principal component is the jk th element of the covariance matrix (for any fixed $1 \leq i \leq n$)

$$\begin{aligned}
 E(X'_{i\cdot} - \mu)(Z_{i\cdot} - \mu' A) &= E(X'_{i\cdot} - \mu)(X_{i\cdot} - \mu') A \\
 &= \Sigma A = (A \Lambda A') A \\
 &= A \Lambda,
 \end{aligned}$$

where (as before) $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ and $A = [\alpha_1, \dots, \alpha_p]$. The variances of X_{ij} and Z_{ik} are σ_{jj} and λ_k , respectively, so the *correlation* is

$$\rho_{jk} := \text{Cor}(X_{ij}, Z_{ik}) = A_{jk} \lambda_k / \sqrt{\sigma_{jj} \lambda_k} = A_{jk} \sqrt{\lambda_k / \sigma_{jj}}$$

The number ρ_{jk}^2 is often described as the “proportion of the j th variable’s variation *explained by* the k th PC” and, since the $\{z_k\}$ are uncorrelated, the proportion explained by any set K of indices is just the sum

$$\rho_{jK}^2 = \sum_{k \in K} \rho_{jk}^2 = \frac{1}{\sigma_{jj}} \sum_{k \in K} A_{jk}^2 \lambda_k$$

which we must typically estimate from the sample by

$$r_{jK}^2 = \frac{1}{\hat{\sigma}_{jj}} \sum_{k \in K} \hat{A}_{jk}^2 \hat{\lambda}_k. \quad (1)$$

When K includes all indices the sum is the jj th entry of $\hat{\Sigma} = \hat{A}\hat{\Lambda}\hat{A}'$ or $\hat{\sigma}_{jj}$, making r_{jK}^2 one. For the test data, for example, since the data were standardized (so $\hat{\sigma}_{jj} \equiv 1$), r_{jk}^2 is given by

```
> r.jk <- pca$rot %*% diag(pca$sd);
> LastTwo <- r.jk[,4:5]^2;
> cbind(LastTwo, sum=apply(LastTwo,1,sum));
```

	PC4	PC5	sum
mec	0.008244153	0.004210587	0.01245474
vec	0.034477637	0.008145448	0.04262309
alg	0.004574680	0.176776522	0.18135120
ana	0.172184331	0.043953070	0.21613740
sta	0.168411575	0.013504879	0.18191645

which show that removing the last one or two PC’s affects the five tests disproportionately— 18% of the Algebra test’s variability is captured by the last (5th) PC alone, and about a fifth of that of the three open-book tests (Algebra, Analysis and Statistics) are captured by the last two PC’s, so a reduced analysis will rest much more on the (closed-book) Mechanics and Vectors tests than on the other three tests. It’s a matter of judgment (and not of statistical science) whether or not this is acceptable.