# Principal Components Theory Notes

Charles J. Geyer

November 6, 2006

## 1 Introduction

These are class notes for Stat 5601 (nonparametrics) taught at the University of Minnesota, Spring 2006. This not a theory course, so the bit of theory we do here is very simple, but very important in multivariate analysis, which is not really the subject of this course. It is necessary to fully understand the principle components example (our first really messy bootstrap example).

## 2 Random Vectors

### 2.1 Definitions

A random vector is just a vector whose components are random variables. Its mean is the vector whose components are the means of the components. If $\mathbf{X} = (X_1, \ldots, X_n)$ is a random vector, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ is its mean vector, then

$$\mu_i = E(X_i), \qquad i = 1, \ldots, n,$$

but we usually write this as a vector equation $\boldsymbol{\mu} = E(\mathbf{X})$.

The analogy for variances is not obvious. We define $\operatorname{var}(\mathbf{X})$ to be a matrix whose $i$, $j$ element is $\operatorname{cov}(X_i, X_j)$. Different people have different names for this matrix. Some call it the variance matrix, some call it the covariance matrix, some call it the variance-covariance matrix — because the diagonal elements are variances, $\operatorname{cov}(X_i, X_i) = \operatorname{var}(X_i)$ — and some call it the dispersion matrix. Whatever you call it, there isn't any other matrix that plays any analogous role in the theory (there's just one theoretically useful matrix, but people can't agree what to call it).

A variance matrix is always symmetric, because the covariance operator is a symmetric function of its arguments: $\operatorname{cov}(X_i, X_j) = \operatorname{cov}(X_j, X_i)$.

## 2.2 Linear Transformations

If $X$ and $Y$ are random variables and $Y$ is a linear function of $X$, that is, $Y = aX + b$ for some constants $a$ and $b$, then

$$E(Y) = aE(X) + b \tag{1a}$$

$$\operatorname{var}(Y) = a^2 \operatorname{var}(X) \tag{1b}$$

$$\operatorname{sd}(Y) = |a| \operatorname{sd}(X) \tag{1c}$$

If $\mathbf{X}$ and $\mathbf{Y}$ are random vectors, then we say $\mathbf{Y}$ is a linear function of $\mathbf{X}$ when $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ for some constant matrix $\mathbf{A}$ and some constant vector $\mathbf{b}$ (in real math, say in a linear algebra course, this would be called an "affine" function, which would be called "linear" only if $\mathbf{b} = \mathbf{0}$, but most people call these functions "linear"), then

$$E(\mathbf{Y}) = \mathbf{A}E(\mathbf{X}) + \mathbf{b} \tag{2a}$$

$$\operatorname{var}(\mathbf{Y}) = \mathbf{A} \operatorname{var}(\mathbf{X}) \mathbf{A}^T \tag{2b}$$

are formulas for vector-to-vector linear transformations analogous to (1a) and (1b).

## 2.3 Positive Semi-Definiteness

Consider the case where $Y = \mathbf{A}\mathbf{X}$ is scalar, so $\mathbf{A}$ is a matrix with just one row (a row vector). Since it is a convention to consider vectors as matrices with just one column (column vectors), we write $\mathbf{A} = \mathbf{a}^T$ where $\mathbf{a}$ is a column vector. Then (2b) becomes

$$0 \le \operatorname{var}(Y) = \mathbf{a}^T \operatorname{var}(\mathbf{X}) \mathbf{a}$$

the inequality coming from the fact that the variance of a random variable is always nonnegative.

This property has a name. An arbitrary symmetric matrix $\mathbf{V}$ (not necessarily a variance matrix) is said to be *positive semi-definite* if

$$\mathbf{a}^T \mathbf{V} \mathbf{a} \ge 0, \qquad \text{for all vectors } \mathbf{a}.$$

Thus we now have two properties that all variance matrices must satisfy: they are symmetric and positive semi-definite.

## 2.4 Spectral Decomposition

Any symmetric matrix $\mathbf{A}$ can has a *spectral decomposition*

$$\mathbf{A} = \mathbf{ODO}^T \tag{3}$$

where $\mathbf{D}$ is diagonal and $\mathbf{O}$ is orthogonal, which means $\mathbf{O}^{-1} = \mathbf{O}^T$.

The reason an orthogonal matrix is called orthogonal is because its columns are orthogonal vectors (vectors whose scalar product is zero). Let $\mathbf{w}_i$ denote the $i$-th column of $\mathbf{O}$. Then $\mathbf{w}_i^T \mathbf{w}_j$ is the $i, j$-th element of $\mathbf{O}^T\mathbf{O} = \mathbf{O}^{-1}\mathbf{O} = \mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix. This says each column of $\mathbf{O}$ has length one and is perpendicular (orthogonal) to every other column.

Thinking of $\mathbf{O}$ as a change of coordinate systems, we see that it corresponds to a rigid rotation from one frame of reference with perpendicular coordinate axes to another frame of reference with perpendicular coordinate axes.

The spectral decomposition can be used to determine whether a matrix is positive semi-definite. A diagonal matrix is positive semidefinite if and only if all its elements are nonnegative, because

$$\mathbf{a}^T \mathbf{Da} = \sum_i a_i^2 d_{ii}.$$

A general symmetric matrix is positive semi-definite if and only if the diagonal matrix in its spectral decomposition is positive semi-definite, because

$$\mathbf{a}^T \mathbf{Aa} = \mathbf{a}^T \mathbf{ODO}^T \mathbf{a} = \mathbf{b}^T \mathbf{Db}$$

where $\mathbf{b} = \mathbf{O}^T \mathbf{a}$ and $\mathbf{a} = \mathbf{Ob}$.

## 2.5 Eigenvalues and Eigenvectors

Multiplying (3) on the right by $\mathbf{O}$ gives

$$\mathbf{AO} = \mathbf{ODO}^T \mathbf{O} = \mathbf{OD}$$

If we interpret this by looking at what it does to the columns $\mathbf{w}_i$ of $\mathbf{O}$, we get

$$\mathbf{Aw}_i = \lambda_i \mathbf{w}_i,$$

where $\lambda_i$ is the $i, i$-th element of $\mathbf{D}$.

This property also has a name. If $\mathbf{A}$ is any matrix and if

$$\mathbf{A}\mathbf{w} = \lambda\mathbf{w}$$

holds for some nonzero vector $\mathbf{w}$ and scalar $\lambda$, then we say $\mathbf{w}$ is an *eigenvector* of $\mathbf{A}$ corresponding to the *eigenvalue* $\lambda$. Thus in the spectral decomposition, the columns of $\mathbf{O}$ are eigenvectors of $\mathbf{A}$ and the corresponding elements of the diagonal of $\mathbf{D}$ are the corresponding eigenvalues.

Eigenvectors are not uniquely determined by eigenvalues. The system of linear equations $(\mathbf{A} - \lambda\mathbf{I})\mathbf{w}$ always has multiple solutions. Any nonzero scalar multiple of an eigenvector is an eigenvector. Any linear combination of eigenvectors corresponding to the same eigenvalue is an eigenvector.

Since the eigenvectors produced by a spectral decomposition are orthogonal, they are linearly independent and form a basis for $n$-dimensional space. Any other eigenvectors are linear combinations of those given by the spectral decomposition.

### 2.6 Matrix Square Roots

A symmetric positive semi-definite matrix $\mathbf{A}$ has a natural matrix square root calculated using the spectral decomposition (3)

$$\mathbf{A}^{1/2} = \mathbf{O}\mathbf{D}^{1/2}\mathbf{O}^T,$$

where $\mathbf{D}^{1/2}$ is the diagonal matrix whose diagonal elements are the square roots of the diagonal elements of $\mathbf{D}$ (which are nonnegative because $\mathbf{D}$ is positive semi-definite).

It is easily seen that $\mathbf{D}^{1/2}\mathbf{D}^{1/2} = \mathbf{D}$, and this implies $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$, because of $\mathbf{O}^T = \mathbf{O}^{-1}$.

We can use this notion to define "standard deviations" of random vectors, but they are nowhere near as useful as standard deviations of scalar random variables. The analog of (1c) would be

$$\operatorname{var}(\mathbf{Y})^{1/2} = \bigl(\mathbf{A}\operatorname{var}(\mathbf{X})\mathbf{A}^T\bigr)^{1/2}. \tag{4}$$

## 3 Principal Components

### 3.1 Definition

If $X$ is any random vector having finite variance, let

$$\operatorname{var}(\mathbf{X}) = \mathbf{O}\mathbf{D}\mathbf{O}^T \tag{5}$$

be the spectral decomposition of its variance matrix.

Consider the random vector $\mathbf{Y} = \mathbf{O}^T\mathbf{X}$. (We are using a linear transformation that is derived from the spectral decomposition, $\mathbf{O}$ being the same matrix in both places.) Then

$$\begin{aligned} \operatorname{var}(\mathbf{Y}) &= \mathbf{O}^T \operatorname{var}(\mathbf{X})\mathbf{O} \\ &= \mathbf{O}^T\mathbf{O}\mathbf{D}\mathbf{O}^T\mathbf{O} \\ &= \mathbf{D} \end{aligned}$$

where the first equality is (2b), the second is (5), and the third is the defining property of orthogonal matrices ($\mathbf{O}^T = \mathbf{O}^{-1}$).

Thus $\mathbf{Y}$ has a diagonal variance matrix. Hence, since the off-diagonal elements of the variance matrix (a. k. a., covariance, variance-covariance, or dispersion matrix) are covariances, the components of $\mathbf{Y}$ are uncorrelated. And, since the diagonal elements of the variance matrix are variances and the diagonal elements of $\mathbf{D}$ are the eigenvalues of $\operatorname{var}(\mathbf{X})$, the variances of the components of $\mathbf{Y}$ are the eigenvalues of the variance matrix of $X$.

The components of $\mathbf{Y}$ are called the *principal components* of $\mathbf{X}$. Since an orthogonal matrix is invertible, we also have $\mathbf{X} = \mathbf{O}\mathbf{Y}$. This expresses an arbitrary random vector $\mathbf{X}$ as a linear combination of uncorrelated random variables (its principal components).

The process of doing the spectral decomposition (a. k. a., finding eigenvalues and eigenvectors) of the variance matrix of $\mathbf{X}$ is called principal components analysis (PCA).

### 3.2 Dimension Reduction

PCA is often used as a method of dimension reduction. If we only keep a few of the principal components, then we get a "simple" explanation of the structure of $\mathbf{X}$ involving a few random variables. Order the components of $\mathbf{Y}$ putting the components with larger variance (larger eigenvalues) first. As in Sections 2.4 and 2.5, let $\mathbf{w}_i$ denote the columns of $\mathbf{O}$, which are eigenvectors of $\operatorname{var}(\mathbf{X})$, and let $\lambda_i$ denote the diagonal elements of $\mathbf{D}$, which are the eigenvalues of $\operatorname{var}(\mathbf{X})$. Then

$$\mathbf{X} = \sum_{i=1}^{n} Y_i \mathbf{w}_i,$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)$. A sum of fewer terms

$$\widetilde{\mathbf{X}} = \sum_{i=1}^{k} Y_i \mathbf{w}_i$$

may explain most of the variance in the following sense.

### 3.3 Fraction of Variance Explained

Let $\| \cdot \|$ denote the Euclidean norm of a vector, so

$$\|\mathbf{Y}\|^2 = \mathbf{Y}^T\mathbf{Y} = \sum_i Y_i^2.$$

Then because the components of $\mathbf{Y}$ are independent we have

$$E\big\{\|\mathbf{Y} - \boldsymbol{\nu}\|^2\big\} = \sum_{i=1}^{n} \lambda_i,$$

where $\boldsymbol{\nu} = E(\mathbf{Y})$. This is the most natural scalar measure of the variablility of $\mathbf{Y}$ (of course, the complete measure is its entire variance matrix $\mathbf{D}$).

Because an orthogonal transformation is a rotation, it does not affect lengths. So if $\boldsymbol{\mu} = E(\mathbf{X})$, we have

$$\begin{aligned} E\big\{\|\mathbf{X} - \boldsymbol{\mu}\|^2\big\} &= E\big\{(\mathbf{X} - \boldsymbol{\mu})^T(\mathbf{X} - \boldsymbol{\mu})\big\} \\ &= E\big\{(\mathbf{Y} - \boldsymbol{\nu})^T\mathbf{O}^T\mathbf{O}(\mathbf{Y} - \boldsymbol{\nu})\big\} \\ &= E\big\{(\mathbf{Y} - \boldsymbol{\nu})^T(\mathbf{Y} - \boldsymbol{\nu})\big\} \\ &= E\big\{\|\mathbf{Y} - \boldsymbol{\nu}\|^2\big\}. \end{aligned}$$

Because $\mathbf{Y} = \mathbf{O}^T\mathbf{X}$ implies $\mathbf{X} = \mathbf{O}\mathbf{Y}$ and $\boldsymbol{\mu} = \mathbf{O}\nu$, because $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$ for any matrices $\mathbf{A}$ and $\mathbf{B}$, and because $\mathbf{O}^T\mathbf{O} = \mathbf{I}$.

Similarly, if $\tilde{\boldsymbol{\mu}} = E(\widetilde{\mathbf{X}})$, then

$$E\big\{\|\widetilde{\mathbf{X}} - \tilde{\boldsymbol{\mu}}\|^2\big\} = \sum_{i=1}^{k} \lambda_i.$$

Thus the fraction of the variance of $\mathbf{X}$ explained by the first $k$ principal components is $\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{n} \lambda_i$.