PRINCIPAL COMPONENT ANALYSIS

MIKE MEI

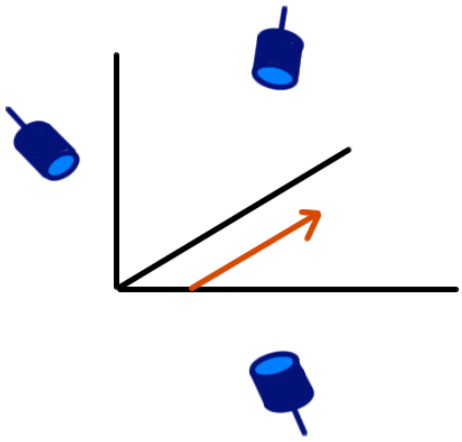
Abstract. The following paper will explore the concepts of linear algebra that are relevant to the statistical method of principal component analysis (PCA). We will prove the spectral theorem for real inner product spaces and explain how spectral decomposition is essential for finding principal components of random vectors. The reader is assumed to have knowledge of basic concepts in linear algebra and be familiar with statistical terms and their fundamental properties.

Contents

1. Introduction

In many theoretical and real-life situations that involve the collection of vector data for statistical purposes, the vectors are determined by as many random elements as dimensions measured. However, the actual variation in the vector data may occur in only a few dimensions, a number which is less than the number of vector components. In other words, the data can be repetitive, confusing, and unnecessarily complicated because the method of collection looked at more dimensions than necessary.

Date: AUGUST 21, 2009.



For example, suppose we have a vector that varies mostly in one direction on a line and we record two-dimensional projections of this vector with three different cameras. If these cameras are in different positions and are at angles that are not necessary perpendicular to each other, we may record substantial variation in all the data from each camera. How we discover the simple underlying behavior of a one-dimensional vector from a collection of seemingly complicated six-dimensional vector data (two dimensions from each camera) is of interest to us.

We are, in short, interested in representing most of the variation in the data by transforming the original random vector into variables called *principal components*. These components are all orthogonal and are ordered so that the first few explain most of the variation of the random vector. Therefore, the goal is to find an orthogonal basis that aligns itself with the data and thus explains a substantial amount of variation in just a few dimensions. The methods that allow us to find such bases and principal components, of course, come from interesting ideas rooted in linear algebra.

2. The Spectral Theorem

Definition 2.1. An *inner product space* is a vector space over $\mathbb{R}$ or $\mathbb{C}$ that has an inner product operation satisfying the following properties:

1) $\langle x, y\rangle = \overline{\langle y, x\rangle}$
2) $\alpha\langle x, y\rangle = \langle \alpha x, y\rangle$
3) $\langle x + z, y\rangle = \langle x, y\rangle + \langle z, y\rangle$
4) $\langle x, x\rangle > 0 \ \forall \mathrm{x} \neq 0$

Definition 2.2. An *adjoint* of a linear map A: V $\rightarrow$ V is another linear map A*: V $\rightarrow$ V that satisfies

$$\langle Ax, y\rangle = \langle x, A^*y\rangle \ \forall x, y \in V$$

Proposition 2.3. *If A is a linear map A: $V \rightarrow V$, then the adjoint A^* exists.*

Proof. Let ϕ be any linear map ϕ: V $\to \mathbb{R}$ or $\mathbb{C}$ for some inner product space V. Let $\{e_1, e_2, ..., e_n\}$ be an orthonormal basis in V. Note that

$$\begin{aligned}\phi(u) &= \phi(\langle u, e_1\rangle e_1 + \langle u, e_2\rangle e_2 + ... + \langle u, e_n\rangle e_n) \\ &= \phi(e_1)\langle u, e_1\rangle + \phi(e_2)\langle u, e_2\rangle + ... + \phi(e_n)\langle u, e_n\rangle \\ &= \langle \phi(e_1)u, e_1\rangle + \langle \phi(e_2)u, e_2\rangle + ... + \langle \phi(e_n)u, e_n\rangle \\ &= \langle u, \overline{\phi(e_1)}e_1\rangle + \langle u, \overline{\phi(e_2)}e_2\rangle + ... + \langle u, \overline{\phi(e_n)}e_n\rangle \\ &= \langle u, \overline{\phi(e_1)}e_1 + \overline{\phi(e_2)}e_2 + ... + \overline{\phi(e_n)}e_n\rangle\end{aligned}$$

Let v ∈ V be given. Note that the inner product $\langle Au, v\rangle$ is, like ϕ, a linear map that takes a vector u ∈ V to an element in $\mathbb{R}$ or $\mathbb{C}$. Let ϕ^*(u) = $\langle Au, v\rangle$. We know from above that

$$\langle Au, v\rangle = \langle u, \overline{\phi^*(e_1)}e_1 + \overline{\phi^*(e_2)}e_2 + ... + \overline{\phi^*(e_n)}e_n\rangle$$

We can set

$$A^*v = \overline{\phi^*(e_1)}e_1 + \overline{\phi^*(e_2)}e_2 + ... + \overline{\phi^*(e_n)}e_n$$

This shows that

$$\langle Au, v\rangle = \langle u, A^*v\rangle$$

So the adjoint A^* exists.

□

Definition 2.4. The linear map A on a real inner product space is $self-adjoint$ if $A = A^*$.

Proposition 2.5. *Suppose the linear map A has a matrix T with respect to an orthonormal basis in a real inner product space V. Then T^*, the matrix of A^* with respect to the same basis, is the transpose of T.*

Proof. Let $\{e_1, e_2, ..., e_n\}$ be the orthonormal basis of V. Because A is a linear map, matrix T is composed of n columns, the i^{th} one being the coordinates of Ae_i with respect to the basis. Specifically, there are scalars a_1, a_2,...,a_n such that

$$Ae_i = a_1e_1 + a_2e_2 + ... + a_ne_n$$

Because the basis is orthonormal, it follows that

$$\begin{aligned}\langle Ae_i, e_j\rangle &= \langle a_1e_1 + a_2e_2 + ... + a_ne_n, e_j\rangle \\ &= \langle a_1e_1, e_j\rangle + \langle a_2e_2, e_j\rangle + ... + \langle a_ne_n, e_j\rangle \\ &= a_1\langle e_1, e_j\rangle + a_2\langle e_2, e_j\rangle + ... + a_n\langle e_n, e_j\rangle \\ &= a_j\end{aligned}$$

Therefore, $Ae_i = \langle Ae_i, e_1\rangle e_1 + \langle Ae_i, e_2\rangle e_2 + ... + \langle Ae_i, e_n\rangle e_n$. In other words, T_{ji}, the i^{th} column, j^{th} row of T is $\langle Ae_i, e_j\rangle$. By repeating this derivation for the matrix entries of A^*, we get that T^*_{ji} is likewise

$$\begin{aligned}\langle A^*e_i, e_j\rangle &= \langle e_i, Ae_j\rangle \\ &= \overline{\langle Ae_j, e_i\rangle} \\ &= \langle Ae_j, e_i\rangle\end{aligned}$$

The last line follows from the fact that we are in a real inner product space. The result shows that $T^*_{ji} = T_{ij}$. Thus T is the transpose of T^*. □

This proposition shows that a self-adjoint operator has a matrix that is equal to its transpose. In other words, the matrix is *symmetric* across the diagonal. Symmetric matrices which represent linear maps have important properties that are very useful to PCA, as we will see later with covariance matrices.

Now we will attempt to prove a very important theorem related to symmetric matrices and their self-adjoint operators. Given a symmetric matrix or self-adjoint operator, the Spectral Theorem will allow us to find orthonormal eigenvectors, which is exactly what we want for PCA.

Lemma 2.6. *If A: $V \to V$ is a self-adjoint linear map, then A has a real eigenvalue.*

Proof. A has a matrix T with respect to some basis. Consider the matrix (T - Iz), where I is the identity matrix. Then $\det(T - Iz)$ is a polynomial function of z. By the fundamental theorem of algebra, this polynomial vanishes at a complex value λ. Therefore, there is a v ∈ V such that

$$\begin{aligned}(T - \lambda I)v &= 0\\ Tv &= \lambda Iv\\ Tv &= \lambda v\\ Av &= \lambda v\end{aligned}$$

So λ is an eigenvalue. In addition we also know that λ is real because

$$\begin{aligned}\lambda|v|^2 &= \langle Tv, v\rangle\\ &= \langle v, Tv\rangle\\ &= \overline{\langle Tv, v\rangle}\\ &= \overline{\lambda|v|^2}\\ &= \overline{\lambda}|v|^2\end{aligned}$$

This gives $\overline{\lambda} = \lambda$, which must be real.

□

Theorem 2.7. (The Real Spectral Theorem) *Let V be a real inner product space, and let A: $V \to V$ be a linear map. Then there is an orthonormal basis in V consisting of eigenvectors of A if and only if A is self-adjoint.*

Proof. Let $\{e_1, e_2, ..., e_n\}$ be an orthonormal eigenbasis in V. Then, with respect to that basis,

$$\begin{aligned}Ae_1 &= \lambda_1 e_1\\ Ae_2 &= \lambda_2 e_2\\ &\vdots\\ Ae_n &= \lambda_n e_n\end{aligned}$$

It is clear that A has a diagonal matrix

$$\begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \lambda_n \end{pmatrix}$$

This matrix is equal to its transpose, and therefore A is self-adjoint.

The converse will be proven by induction on the dimension of V. Our induction hypothesis will be that for dimension n greater than 1, the theorem holds for vector spaces of dimension less than n. However, we will first need to construct our proof by introducing a special subspace U that has the desired properties.

If A is self-adjoint, we know from Lemma 2.6 that the matrix of A has a real eigenvalue. Let u ∈ V be the corresponding eigenvector. We can scale u so that the norm is 1.

Let U be the set of all scalar multiples of u. U is a subspace of V with dimension 1. Let $\mathrm{U}^{\perp} = \{v \in V | \langle u, v\rangle = 0\}$. Given $v_0 \in U^{\perp}$, note that

$$\begin{aligned} \langle u, Av_0\rangle &= \langle Au, v_0\rangle \\ &= \langle \lambda u, v_0\rangle \\ &= \lambda\langle u, v_0\rangle \\ &= 0 \end{aligned}$$

Therefore, $v_0 \in U^{\perp}$ implies $Av_0 \in U^{\perp}$. In other words, $U^{\perp}$ is invariant under A. Let S be the linear map on V defined by $\mathrm{S} = \mathrm{A}|_{U^{\perp}}$ (i.e., A restricted to the domain of $U^{\perp}$). Let x, y ∈ $U^{\perp}$.

$$\langle Sx, y\rangle = \langle Ax, y\rangle = \langle x, Ay\rangle = \langle x, Sy\rangle$$

This shows that S is self-adjoint. Now suppose that the theorem holds for the n-dimensional subspace $U^{\perp}$. Let $\{u_2, u_2, ..., u_n\}$ be the orthonormal eigenbasis for $U^{\perp}$. Then $\{u, u_2, u_2, ..., u_n\}$ is an orthonormal eigenbasis for V, which is n+1 dimensional.

The trivial step is to show that the theorem holds when the dimension = 1; the theorem holds because all transformations by linear maps on a real one-dimensional inner product space are simply scalar multiplications by real numbers.

□

Corollary 2.8. *Let T be the self-adjoint matrix of a linear map of a real inner product space V. Then T can be decomposed into* $T = E\Lambda E^T = \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + ... + \lambda_n e_n e_n^T$, *where E is a diagonalizing matrix with orthonormal columns.*

Proof. Since the linear map is self-adjoint, we know there is an orthonormal eigenbasis in V. Let this basis be $\{e_1, e_2, ..., e_n\}$. Let E be the matrix with the basis vectors on the columns,

$$\mathrm{E} = \begin{pmatrix} e_1 & e_2 & \dots & e_n \end{pmatrix}$$

and let

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \lambda_n \end{pmatrix}$$

It is clear that TE = EΛ since $Te_i = \lambda_i e_i$. This implies that T = EΛE^{-1}.

Note that $E^{-1} = E^T$ because EE^T = I (since $\langle e_i, e_i \rangle = 1$ and $\langle e_i, e_j \rangle = 0$ for i ≠ j). Thus we have a diagonalizing matrix E (with orthonormal columns) such that T = EΛE^T.

$$\begin{aligned} \mathrm{E\Lambda E^T} &= \begin{pmatrix} e_1 & e_2 & \dots & e_n \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \lambda_n \end{pmatrix} \begin{pmatrix} e_1^T \\ e_2^T \\ \vdots \\ e_n^T \end{pmatrix} \\ &= \begin{pmatrix} e_1 & e_2 & \dots & e_n \end{pmatrix} \begin{pmatrix} \lambda_1 e_1^T \\ \lambda_2 e_2^T \\ \vdots \\ \lambda_n e_n^T \end{pmatrix} \\ &= \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + ... + \lambda_n e_n e_n^T \end{aligned} \tag{2.9}$$

□

3. Principal Component Analysis

We are interested in finding vectors in a real inner product space that are statistically uncorrelated. In the language of linear algebra, this means that they are orthogonal (and thus linearly independent). The Real Spectral Theorem helps us use the covariance matrix, which is a symmetric matrix (and therefore is the matrix of a self-adjoint linear map). The Real Spectral Theorem guarantees that we will find an orthonormal basis of eigenvectors. As we will see, these eigenvectors will have corresponding eigenvalues of great significance.

Definition 3.1. Let E(u) be the expected value, or the mean, of a random variable u. The *covariance* of two random variables x, y is $cov[x, y]$, where

$$cov[x, y] = E(xy) - E(x)E(y)$$

Definition 3.2. The *variance* of a random variable x is $var[x]$, where

$$var[x] = cov[x, x]$$

Definition 3.3. With regards to a basis, x is a vector in an n-dimensional real inner product space (equipped with a dot product) determined by n random scalars.

Definition 3.4. For k = 1, 2,...,n, α_k is the vector satisfying the following properties:

1) α_kx = $a_{k1}x_1 + a_{k2}x_2 + ... + a_{kn}x_n$, where $\alpha_k = (a_{k1}\ a_{k2}\ ...\ a_{kn})$ and x = $(x_1\ x_2\ ...\ x_n)$
2) The variance of α_kx is maximized under the constraint that $\langle \alpha_k, \alpha_k \rangle = 1$
3) α_{k+1} is calculated after α_k and is uncorrelated to α_k

Definition 3.5. Let x = $(x_1\, x_2 \ldots x_n)$. The *covariance matrix* Σ of x is the matrix with entries $\Sigma_{ij} = \text{cov}[x_i,\, x_j]$.

Proposition 3.6. *Let Σ be the covariance matrix for the elements of x. Then var[$\alpha_k x$] = $\alpha_k \Sigma \alpha_k^T$.*

Proof. Let $\alpha_k = (a_{k1}\, a_{k2} \ldots a_{kn})$ and x = $(x_1\, x_2 \ldots x_n)$. Note that cov[α_kx,x] is the 1 x n matrix M where $M_{ij} = \text{cov}[\alpha_k x, x_j]$. M equals

$$\begin{pmatrix} cov[(\alpha_{k1}x_1 + \alpha_{k2}x_2 + \ldots + \alpha_{kn}x_n), x_1] & \ldots & cov[(\alpha_{k1}x_1 + \alpha_{k2}x_2 + \ldots + \alpha_{kn}x_n), x_n] \end{pmatrix}$$

Because α_k is constant, we can factor it out of the covariance bracket.

$$\begin{aligned} M &= \begin{pmatrix} \sum_{i=1}^n \alpha_{ki} cov[x_i, x_1] & \sum_{i=1}^n \alpha_{ki} cov[x_i, x_2] & \ldots & \sum_{i=1}^n \alpha_{ki} cov[x_i, x_n] \end{pmatrix} \\ &= \begin{pmatrix} \alpha_{k1} & \alpha_{k2} & \ldots & \alpha_{kn} \end{pmatrix} \begin{pmatrix} cov[x_1,x_1] & cov[x_1,x_2] & \ldots & cov[x_1,x_n] \\ cov[x_1,x_2] & cov[x_2,x_2] & \ldots & cov[x_2,x_n] \\ \vdots & \vdots & \ddots & \vdots \\ cov[x_1,x_n] & cov[x_2,x_n] & \ldots & cov[x_n,x_n] \end{pmatrix} \\ &= \alpha_k \Sigma \end{aligned}$$

Since cov[α_kx, x] = $\alpha_k \Sigma$, by a similar argument, cov[x, α_kx] = $\Sigma \alpha_k^T$. Thus, var[α_kx] = cov[α_kx, α_kx] = $\alpha_k \Sigma a_k^T$.

□

To actually get interesting results from the properties set in Definition 3.4, Proposition 3.6 suggests that it will be necessary to maximize $\alpha_k \Sigma \alpha_k^T$. However, there is one assumption in PCA not fully disclosed yet, and it is that the covariance matrix Σ has distinct eigenvalues. Though having eigenvalues of multiplicities greater than one is theoretically possible and does not pose problems for the purposes of this paper, it does pose problems in statistics because principal components are not uniquely defined (i.e., they are not necessarily ordered). Because such occurrences are very uncommon, the assumption usually holds (Jolliffe, 27).

Theorem 3.7. (Principal Components) *Let Σ be the covariance matrix for x. If Σ has distinct eigenvalues, then for k = 1, 2,...,n, α_k^T is an eigenvector corresponding to the kth largest eigenvalue of Σ.*

Proof. By the third property of Definition 3.4, α_1 is derived first. The maximization of var[$\alpha_k x$] = $\alpha_k \Sigma \alpha_k^T$ is set up as a Lagrange maximization problem. Let β be the Lagrange multiplier.

$$L = \alpha_1 \Sigma \alpha_1^T - \beta(\alpha_k \alpha_k^T - 1)$$

Differentiating both sides with respect to α_1 yields:

$$\begin{aligned} \Sigma \alpha_1^T - \beta \alpha_1^T &= 0 \\ \Sigma \alpha_1^T &= \beta \alpha_1^T \end{aligned}$$

So β is an eigenvalue of Σ. The fact that β is the largest eigenvalue follows from the observation that the variance that is being maximized is $\alpha_1 \Sigma \alpha_1^T = \alpha_1 \beta \alpha_1^T = \beta \alpha_1 \alpha_1^T = \beta$, so β must not only be maximized as an eigenvalue, it must also equal the largest variance.

We will now look at α_2. To satisfy the third property of Definition 3.4, we need to maximize $\alpha_2 \Sigma \alpha_2^T$ with the additional restriction that $\alpha_2 x$ is uncorrelated with $\alpha_1 x$, i.e., $\text{cov}[\alpha_1 x, \alpha_2 x] = 0$. Note that

$$cov[\alpha_1 x, \alpha_2 x] = \alpha_1 \Sigma \alpha_2^T = \alpha_2 \Sigma \alpha_1^T = \alpha_1 \beta \alpha_2^T = \beta \alpha_1 \alpha_2^T = \beta \alpha_2 \alpha_1^T = 0$$

So $\alpha_2 \alpha_1^T = 0$. We can therefore set up the Lagrange maximization problem to be the following, where γ and δ are the Lagrange multipliers.

$$L = \alpha_2 \Sigma \alpha_2^T - \delta(\alpha_k \alpha_k^T - 1) - \gamma(\alpha_2 \alpha_1^T)$$

Differentiating both sides with respect to α_2 yields:

$$\begin{aligned} \Sigma \alpha_2^T - \delta \alpha_2^T - \gamma \alpha_1^T &= 0 \\ \alpha_1 \Sigma \alpha_2^T - \delta \alpha_1 \alpha_2^T - \gamma \alpha_1 \alpha_1^T &= 0 \\ 0 - 0 - \gamma &= 0 \\ \gamma &= 0 \end{aligned}$$

This means $\Sigma \alpha_2^T = \delta \alpha_2^T$.

Thus, δ is an eigenvalue and α_2^T is the corresponding eigenvector, and it is likewise maximized, but not greater than or equal to β because β was maximized first and Σ has distinct eigenvalues. By the repetition of this process for $\alpha_k x$ for k≥3, we have eigenvectors α_k^T corresponding to the kth largest eigenvalue.

□

Definition 3.8. *Principal components* are the variables $\alpha_k x$, where α_k are the transposes of the eigenvectors α_k^T from Theorem 3.7.

Corollary 3.9. *Let A be the matrix that has α_k, the kth eigenvector of Σ, as the kth column. Then $\Sigma = \lambda_1 \alpha_1^T \alpha_1 + \lambda_2 \alpha_2^T \alpha_2 + ... + \lambda_n \alpha_n^T \alpha_n$.*

Proof. Let Λ be the diagonal matrix with eigenvalues from the maximization process in Theorem 3.7. Because Σ has eigenvalues and eigenvectors,

$$\begin{aligned} \Sigma A &= A\Lambda \\ \Rightarrow \Sigma &= A\Lambda A^T \end{aligned}$$

Therefore, $\Sigma = \lambda_1 \alpha_1^T \alpha_1 + \lambda_2 \alpha_2^T \alpha_2 + ... + \lambda_n \alpha_n^T \alpha_n$ by equation 2.9.

□

Theorem 3.7 shows that the random vector x can be transformed into numerous ordered principal components and that the eigenvalue corresponding to a given eigenvector is actually the variance of a_kx. Corollary 3.9 expands on this and decomposes the covariance matrix into parts $\lambda_i \alpha_i^T \alpha_i$. This gives more detail into exactly how the variation is spread out among individual principal components.

Going back to our example of the one-dimensional vector recorded in six-dimensional inner product space, the first principal component $\alpha_1 x$ will probably explain an

overwhelming majority of the variation in the data simply because the vector varies in one direction. Of course, the reduction of dimension via PCA can be unclear in circumstances where there is no clear cutoff for a principal component. Often, dimensions are reduced until a set percentage of variation, e.g., 80 percent, is accounted for.

Theorem 3.10. *Let T be an $n \times n$ symmetric matrix that has distinct eigenvalues. Then T has exactly n eigenvalues.*

Proof. The Spectral Theorem already guarantees at least n eigenvectors, which gives n eigenvalues. For any eigenvector v $\in$ V, we know that

$$Tv = \lambda v$$
$$(\lambda I - T)v = 0$$

Note that since v $\neq$ 0, (λI - T) is a singular matrix that has a determinant of zero. As seen in Lemma 2.6, the determinant is a polynomial function of λ, which vanishes at at most n points. T has exactly n eigenvalues.

□

The significance of the deceptively simple theorem above is that by simply finding the eigenvalues and corresponding eigenvectors of the covariance matrix Σ and ranking the eigenvalues by size, we are effectively performing PCA. There is no "maximization" procedure needed because the eigenvalues of Σ, as shown in Theorem 3.7, are already the largest variances of some α_k. Of course, the actual process of finding eigenvalues and eigenvectors in a space with many dimensions is much more complicated than it seems, and further investigation into this topic may include statistical inference on principal components and procedures when the covariance matrix is not fully known. It may also be important to understand the limitations of PCA in explaining vector behavior. In any case, we hope you enjoyed this brief overview of some ideas in linear algebra.

Acknowledgments. I would like to thanks my mentors Blair Davey and Shawn Drenning who have both been very supportive and helpful.

References

[1] Gilbert Strang. Linear Algebra and Its Applications. Academic Press. 1976.
[2] Sheldon Axler. Linear Algebra Done Right. Springer. 1997.
[3] I. T. Jolliffe. Principal Component Analysis. Springer. 2002.
[4] Jon Schlens. A Tutorial on Principal Component Analysis. New York University. 2005.