PCA Analysis

Eric Walton

1 Problem Statement

1.1 Input

n vectors x_i in $\mathbf{R}^d$ where each vector is a training example, with d features. The matrix X, composed of the columns x_i. X is $d \times n$.

1.2 Goal

Choose:

- n vectors θ_i.
- a normalizing vector μ,
- a linear mapping $x_i \to \theta_i$, represented by a matrix A.

Choose these so that:

$$x_i - \mu \approx A\theta_i \, for \, i = (1, 2, \ldots, n)$$

That is, minimize:

$$\sum_{i=1}^{n} \|x_i - \mu - A\theta_i\|^2$$

Lastly, we choose A such that it is orthogonal, that is:

$$A^T A = I$$

2 Solution

2.1 Find θ

Use least squares.

$$A^T A\theta_i = A^T(x_i - \mu)$$
$$\theta_i = (A^T A)^{-1} A^T(x_i - \mu)$$

Since A is orthogonal:

$$= A^T(x_i - \mu)$$

2.2 Find μ

Minimize:

$$\sum_{i=1}^{n} \|x_i - \mu - A\theta_i\|^2$$

$$= \sum_{i=1}^{n} \|x_i - \mu - AA^T(x_i - \mu)\|^2$$

$$= \sum_{i=1}^{n} \|x_i - \mu - AA^T x_i - AA^T \mu)\|^2$$

$$= \sum_{i=1}^{n} \|Ix_i - AA^T x_i - I\mu - AA^T \mu)\|^2$$

$$= \sum_{i=1}^{n} \|(I - AA^T)x_i - (I - AA^T)\mu)\|^2$$

$$= \sum_{i=1}^{n} \|(I - AA^T)(x_i - \mu)\|^2$$

$$= \sum_{i=1}^{n} \big((I - AA^T)(x_i - \mu)\big)^T \big((I - AA^T)(x_i - \mu)\big)$$

$$= \sum_{i=1}^{n} (x_i - \mu)^T (I - AA^T)^T (I - AA^T)(x_i - \mu)$$

Isolate the middle:

$$(I - AA^T)^T (I - AA^T)$$

$$= (I - AA^T)(I - AA^T)$$

$$= I - AA^T - AA^T + AA^T AA^T$$

$$= I - AA^T$$

Continue:

$$= \sum_{i=1}^{n} (x_i - \mu)^T (I - AA^T)(x_i - \mu)$$

Take the derivative with respect to μ.
The derivative of a quadratic form $v^T M v$ is simply $2Mv$.

$$\frac{\delta}{\delta \mu} = -2(I - AA^T) \sum_{i=1}^{n} (x_i - \mu)$$

To find a local minimum (or maximum), set the derivative to 0.

$$0 = -2(I - AA^T) \sum_{i=1}^{n} x_i - \mu$$

Set $\sum_{i=1}^{n} x_i - \mu = 0$

$$\sum_{i=1}^{n} x_i = n\mu$$

$$1/n \sum_{i=1}^{n} x_i = \mu$$

μ is the sample mean.

2.3 Find A

Let $\hat{x}_i = x_i - \mu$.

Cost function:

$$\sum_{i=1}^{n} \hat{x}_i^T (I - AA^T) \hat{x}_i$$

$$= \sum_{i=1}^{n} \hat{x}_i^T \hat{x}_i - \hat{x}_i^T AA^T \hat{x}_i$$

The first term is independent of A. So maximize the second.

$$max \sum_{i=1}^{n} \hat{x}_i^T AA^T \hat{x}_i$$

$$= \sum_{i=1}^{n} (A^T \hat{x}_i)^T (A^T \hat{x}_i)$$

$$= \sum_{i=1}^{n} \|A^T \hat{x}_i\|^2$$

Define new variables.

$$X = [\hat{x}_1, \hat{x}_2, ..., \hat{x}_n]$$

$$S = \frac{1}{n} X X^T$$

X is $d \times n$. S is $d \times d$. Since S is symmetric, there is an orthogonal eigen decomposition:

$$S = U \Lambda U^T$$

where U and Λ are d $\times$ d. Arrange U and Λ with the eigenvalues in descending order.
Let $B = U^T A$. B is d $\times$ k.
Let $Y = U^T X$. Y is d $\times$ n.

Meanwhile, we're maximizing:

$$\sum_{i=1}^{n} \|A^T \hat{x}_i\|^2$$

$$= \|A^T X\|_F^2$$

$$= \|A^T U U^T X\|_F^2$$

$$= \|B^T Y\|_F^2$$

$$= trace[(B^T Y)^T (B^T Y)]$$

$$= trace(Y^T B B^T Y)$$

$$= trace(B^T Y Y^T B)$$

Note that $Y Y^T = U^T X (U^T X)^T = U^T X X^T U = n U^T S U = n \Lambda$. So:

$$= trace(B^T n\Lambda B)$$

$$= n * trace(\Lambda B B^T)$$

$$= n \sum_{i=1}^{d} \lambda_i \, \|\text{row i of } B\|^2$$

2.3.1 Sum of $\|$row i of B$\|^2$

$$\sum_{i=1}^{d} \|\text{row i of } B\|^2$$

$$= \|B\|_F^2$$

$$= trace(B^T B)$$

Note that $B^T B = A^T U U^T A = A^T A = I$ where I is k × k.

$$= trace(I_k) = k$$

2.3.2 Bound on$\|$row i of B$\|^2$

As shown above, B is orthogonal, with dimensions d × k. Therefore each column of B has length 1. Note that each row of B, being in $\mathbf{R}^k$, has fewer elements than the columns of B, which are in $\mathbf{R}^d$ and $k < d$. If B were square, than B^T would also be orthogonal and each row of B would have length 1. But in fact the rows of B have fewer elements than the columns. Thus the length of each row is less than 1. And since length is always positive:

$$0 \leq \|\text{row i of } B\|^2 \leq 1$$

.

3 Conclusion

Back to maximizing:

$$n \sum_{i=1}^{d} \lambda_i \, \|\text{row i of } B\|^2$$

Since all $\|\text{row i of B}\|^2 \leq 1$, and they must add to k, our best choice is:

$$\|\text{row i of B}\|^2 = \begin{cases} 1 & \text{if } i \leq k \\ 0 & \text{otherwise} \end{cases}$$

(Remember that our eigenvalues are sorted in descending order).
Let $B =$ rows 1..k from the identity matrix, with zero rows below.
Since $B = U^T A$ we need to set $A = [u_1, u_2, \ldots, u_k]$.

4 Error

Calculate the error, given those choices for A, μ, and θ_i.
Error is given by our cost function:

$$\sum_{i=1}^{n} \|\hat{x}_i - AA^T \hat{x}_i\|^2$$

$$= \|X - AA^T X\|_F^2$$

$$= trace[(X - AA^T X)^T (X - AA^T X)]$$

$$= trace(X^T X - X^T AA^T X)$$

$$= trace(X^T X) - trace(A^T X X^T A)$$

$$= trace(nS) - trace(A^T nSA)$$

$$= n(trace(S) - trace(A^T SA))$$

The first term is easy:

$$trace(S) = \sum_{i=1}^{d} \lambda_i$$

Isolate the matrix in the second trace:

$$A^T SA = A^T U \Lambda U^T A$$

Note that A consists of the first k columns of U. So:

$$A^T U = \begin{bmatrix} I_k & 0_{k \times d-k} \end{bmatrix}$$

And:

$$U^T A = \begin{bmatrix} I_k \\ 0_{d-k \times k} \end{bmatrix}$$

Let Λ_k = the slice of Λ with eigenvalues 1..k.
Let Λ_d = the slice of Λ with eigenvalues k+1..d. Therefore:

$$A^T S A = (A^T U)\Lambda(U^T A) = \begin{bmatrix} I_k & 0_{k \times d-k} \end{bmatrix} \begin{bmatrix} \Lambda_k & 0_k \\ 0_{d-k} & \Lambda_d \end{bmatrix} \begin{bmatrix} I_k \\ 0_{d-k \times k} \end{bmatrix} = \Lambda_k$$

$$trace(A^T S A) = trace(\Lambda_k) = \sum_{i=1}^{k} \lambda_i$$

So put it all together:

$$Error = n \sum_{i=1}^{d} \lambda_i - n \sum_{i=1}^{k} \lambda_i = n \sum_{i=k+1}^{d} \lambda_i$$

5 Uniqueness

Question: Does $(I - AA^T)$ have a nullspace other than 0? Yes, there are many choices for μ. Simply adjust θ to compensate. For some vector v:

$$\sum_{i=1}^{n} \|x_i - (\mu + v) - AA^T(x_i - (\mu + v))\|^2 = \sum_{i=1}^{n} \|x_i - \mu - v + AA^T v - AA^T(x_i - \mu)\|^2$$

If we choose v to be in the column space of A, then its projection $AA^T v = v$. (Because there exists a w such that $v = Aw$ and $AA^T v = AA^T(Aw) = Aw = v$).

$$= \sum_{i=1}^{n} \|x_i - \mu - AA^T(x_i - \mu)\|^2$$

In other words, varying our choice of μ by any vector in the column space of A yields the same error value.

6 Acknowledgements

This was originally an assignment from Gopal Nataraj, and portions of this were adapted from his lectures. I freely consulted, and used the wikipedia article on the topic:

`http://en.wikipedia.org/wiki/Principal_component_analysis`