Derivation of Principal Components Analysis

Given a set of data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, we want to find the linear subspace (plus an affine offset) that is the best fit in the least-squares sense. Mathematically, we want to solve

$$\underset{\boldsymbol{\mu}, \boldsymbol{A}, \{\boldsymbol{\theta}_i\}}{\text{minimize}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\theta}_i\|_2^2, \quad \text{subject to } \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \mathbf{I},$$

where $\boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\theta}_i \in \mathbb{R}^k$, and $\boldsymbol{A}$ is a $n \times k$ matrix; the constraint $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \mathbf{I}$ means that we will consider $\boldsymbol{A}$ with orthonormal columns.

Minimizing the expression above over the $\{\boldsymbol{\theta}_i\}$ and $\boldsymbol{\mu}$ is straightforward. For the $\{\boldsymbol{\theta}_i\}$, suppose that $\boldsymbol{A}$ and $\boldsymbol{\mu}$ are fixed. Then we have a series of n decoupled least-squares problems: for $i = 1, \ldots, n$, we solve

$$\underset{\boldsymbol{\theta}_i}{\text{minimize}} \; \|\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\theta}_i\|_2^2$$

This is a standard unconstrained least-squares problem that has solution

$$\widehat{\boldsymbol{\theta}}_i = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{\mu}) = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{\mu}),$$

where the second equality follows from $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \mathbf{I}$. With $\boldsymbol{A}$ still fixed, we solve for $\boldsymbol{\mu}$ by plugging in our expression for the $\boldsymbol{\theta}_i$:

$$\begin{aligned}
\underset{\boldsymbol{\mu}}{\text{minimize}} \; & \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{\mu})\|_2^2, \\
= \; & \sum_{i=1}^{n} \|(\mathbf{I} - \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})(\boldsymbol{x}_i - \boldsymbol{\mu})\|_2^2, \\
= \; & \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}}(\mathbf{I} - \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})(\boldsymbol{x}_i - \boldsymbol{\mu}),
\end{aligned}$$

Georgia Tech ECE 6254 Spring 2017; Notes by M. A. Davenport and J. Romberg. Last updated 12:53, March 1, 2017

where the last step comes from expanding out the norm squared as an inner product, and using the fact that $(\mathbf{I} - \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})$ is a projector; it is symmetric, and $(\mathbf{I} - \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^2 = (\mathbf{I} - \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})$. Taking the gradient of the expression above and setting it to zero means that $\widehat{\boldsymbol{\mu}}$ will obey

$$\begin{aligned}\mathbf{0} &= -2\sum_{i=1}^{n}(\mathbf{I} - \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}) \\ &= -2(\mathbf{I} - \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})\left(\sum_{i=1}^{n}\boldsymbol{x}_i - n\widehat{\boldsymbol{\mu}}\right).\end{aligned}$$

This can be satisfied by taking

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i.$$

Note that this is not the only choice for $\boldsymbol{\mu}$ — any choice that puts $\sum_i \boldsymbol{x}_i - n\boldsymbol{\mu}$ into the column space of $\boldsymbol{A}$ will work. But the choice above is intuitive, so we will go with it.

With $\{\widehat{\boldsymbol{\theta}}_i\}$ and $\widehat{\boldsymbol{\mu}}$ solved for, we now optimize over $\boldsymbol{A}$. We want to solve

$$\underset{\boldsymbol{A}\in\mathbb{R}^{n\times k}}{\text{minimize}} \ \sum_{i=1}^{n}\|\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}} - \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})\|_2^2 \quad \text{subject to} \quad \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \mathbf{I}.$$

We will assume, without loss of generality, that $\widehat{\boldsymbol{\mu}} = \mathbf{0}$, as we could simply use the variable substitution $\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}$ above. The program becomes

$$\underset{\boldsymbol{A}\in\mathbb{R}^{n\times k}}{\text{minimize}} \ \sum_{i=1}^{n}\|(\mathbf{I} - \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})\boldsymbol{x}_i\|_2^2 \quad \text{subject to} \quad \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \mathbf{I}.$$

Georgia Tech ECE 6254 Spring 2017; Notes by M. A. Davenport and J. Romberg. Last updated 12:53, March 1, 2017

Expanding the functional, and again using the fact that $(\mathbf{I} - \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})$ is a projector,

$$\begin{aligned}\sum_{i=1}^{n} \|(\mathbf{I} - \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})\boldsymbol{x}_i\|_2^2 &= \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathrm{T}}(\mathbf{I} - \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})\boldsymbol{x}_i \\ &= \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i.\end{aligned}$$

The first term does not depend on $\boldsymbol{A}$, and the second term is always negative, so our problem is equivalent to

$$\underset{\boldsymbol{A}\in\mathbb{R}^{n\times k}}{\text{maximize}} \ \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i \quad \text{subject to} \quad \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \mathbf{I}.$$

For any vector $\boldsymbol{v}$, it is easy to see that $\|\boldsymbol{v}\|_2^2 = \operatorname{trace}(\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}})$. Thus, the objective function above can also be written as

$$\begin{aligned}\sum_{i=1}^{n} \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i &= \sum_{i=1}^{n} \|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i\|_2^2 \\ &= \sum_{i=1}^{n} \operatorname{trace}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{A}) \\ &= \operatorname{trace}\left(\boldsymbol{A}^{\mathrm{T}}\left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right)\boldsymbol{A}\right) \\ &= \operatorname{trace}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{A}),\end{aligned}$$

where $\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$ is a scaled version of the sample covariance matrix.

By construction, $\boldsymbol{S}$ is symmetric positive semi-definite, so it has eigenvalue decomposition

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}},$$

where $\boldsymbol{U}$ is a $d \times d$ orthonormal matrix, $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} = \mathbf{I}$, and $\boldsymbol{\Lambda} = \mathrm{diag}(\{\lambda_i\})$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. Then

$$\mathrm{trace}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{A}) = \mathrm{trace}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}) = \mathrm{trace}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{W}),$$

where $\boldsymbol{W} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}$. Notice that $\boldsymbol{W}$ also has orthonormal columns, as $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \mathbf{I}$. So we can solve the program

$$\underset{\boldsymbol{W}\in\mathbb{R}^{n\times k}}{\mathrm{maximize}} \;\; \mathrm{trace}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{W}) \quad \text{subject to} \quad \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \mathbf{I},$$

and then take $\widehat{\boldsymbol{A}} = \boldsymbol{U}^{\mathrm{T}}\widehat{\boldsymbol{W}}$.

We can show that the last maximization program above is equivalent to a simple linear program that we can solve by inspection. Let $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ be the columns of $\boldsymbol{W}$. Then

$$\begin{aligned}
\mathrm{trace}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{W}) &= \sum_{i=1}^{k} \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{w}_i \\
&= \sum_{i=1}^{k}\sum_{j=1}^{d} w_i(j)^2\lambda_j \\
&= \sum_{j=1}^{d} h_j\lambda_j, \quad h_j = \sum_{i=1}^{k} w_i(j)^2 = \sum_{i=1}^{k} W(i,j)^2.
\end{aligned}$$

The $h_j,\ j = 1, \ldots, d$ above are the sums of the squares of the *rows* of $\boldsymbol{W}$. It is clear that $h_j \geq 0$. It is also true that

$$\sum_{j=1}^{d} h_j = k,$$

as the fact that the norm of each columns of $\boldsymbol{W}$ is 1 means that

$$\sum_{j=1}^{d}\sum_{i=1}^{k} W(i,j)^2 = \sum_{i=1}^{k}\left(\sum_{j=1}^{d} W(i,j)^2\right) = \sum_{i=1}^{k} 1 = k.$$

Finally, it is also true that $h_j \leq 1$. Here is why: since the columns of $\boldsymbol{W}$ are orthonormal, they can be considered as part of an orthonormal basis for all of $\mathbb{R}^d$. That is, there is a (and actually there are many) $d \times (d-k)$ matrix $\boldsymbol{W}_0$ such that the columns of

$$\boldsymbol{W}' = \begin{bmatrix}\boldsymbol{W} & \boldsymbol{W}_0\end{bmatrix}$$

form an orthonormal basis for $\mathbb{R}^d$. Since $\boldsymbol{W}'$ is square, $\boldsymbol{W}'\boldsymbol{W}^{'\mathrm{T}} = \mathbf{I}$, meaning the sum of the squares of each row are equal to 1. Thus

$$h_j = \sum_{i=1}^{k} W(i,j)^2 \leq \sum_{i=1}^{d} W'(i,j)^2 = 1.$$

With these constraints on the h_j, let's see how large we can make the quantity of interest:

$$\underset{\boldsymbol{h}\in\mathbb{R}^d}{\text{maximize}} \ \sum_{j=1}^{d} h_j\lambda_j \quad \text{subject to} \quad \sum_{j=1}^{d} h_j = k, \quad 0 \leq h_j \leq 1.$$

This is a linear program, but we can intuit the answer. Since all of the λ_j are positive, we want to have their weights (i.e., the h_j) as large as possible for the largest entries. Since the weights are constrained to be less than 1, and their sum is k, this simply means we assign a weight of 1 to the k largest terms, and 0 to the others:

$$\widehat{h}_j = \begin{cases} 1, & j = 1, \ldots, k, \\ 0, & \text{otherwise.} \end{cases}$$

This means that the sum of the squares of the entries in the rows of the corresponding $\widehat{\boldsymbol{W}}$ are 1 for the first k, and zero below — there are many matrices with orthonormal columns which fit the bill, but a specific one which does is

$$\widehat{\boldsymbol{W}} = \begin{bmatrix} \mathbf{I}_{k\times k} \\ \mathbf{0}_{(d-k)\times k} \end{bmatrix}. \tag{1}$$

Taking $\widehat{\boldsymbol{A}} = \boldsymbol{U}^{\mathrm{T}}\widehat{\boldsymbol{W}}$, this results in

$$\widehat{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_k \end{bmatrix},$$

where the $\boldsymbol{u}_i$ above are the first k columns of $\boldsymbol{U}$.

PCA Theorem

$$\underset{\boldsymbol{\mu},\boldsymbol{A},\{\boldsymbol{\theta}_i\}}{\text{minimize}} \quad \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\theta}_i\|_2^2, \quad \text{subject to } \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \mathbf{I},$$

has solution

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i, \quad \widehat{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_k \end{bmatrix}, \quad \widehat{\boldsymbol{\theta}}_i = \widehat{\boldsymbol{A}}^{\mathrm{T}}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}),$$

where $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ are the eigenvectors corresponding to the k largest eigenvalues of

$$\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}.$$

Note that our analysis above shows that the choice of $\boldsymbol{A}$ is not unique — we are really choosing the subspace spanned by the columns of $\boldsymbol{A}$, and do not care which orthobasis we use to span it. In the end, taking $\widehat{\boldsymbol{A}}' = \widehat{\boldsymbol{A}}\boldsymbol{Q}$, for any $k \times k$ orthonormal matrix $\boldsymbol{Q}$ would also work, as

$$\widehat{\boldsymbol{A}}'\widehat{\boldsymbol{A}}'^{\mathrm{T}} = \widehat{\boldsymbol{A}}\boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}}\widehat{\boldsymbol{A}}^{\mathrm{T}} = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{A}}^{\mathrm{T}}.$$

In our choice for $\widehat{\boldsymbol{W}}$ in (1) above, we would take

$$\widehat{\boldsymbol{W}} = \begin{bmatrix} \boldsymbol{Q} \\ \mathbf{0}_{(d-k)\times k} \end{bmatrix},$$

which also meets the constraints dictated by the $\widehat{h}_j$ — the sum of the squares of the entries in the rows if 1 for the first k, zero for the last $d-k$, and the columns are orthonormal.