Analytical derivation of principal component analysis

N. Fordui, O. Galganov

Problem description

Let $\{x_1, ..., x_m\}$ be a set of m points in $\mathbb{R}^n$. Task is to find for given $k < n$ a k-dimensional hyperplane, which is the closest to these points in the sense Euclidean distance: in other words, difference between points and their projections onto a plane must be the smallest possible.
This problem is called **principal component analysis** and is widely used in statistics and machine learning.

Problem statement

Let $\{x_1, ..., x_m\}$ be m vectors in $\mathbb{R}^n$. It is known that k-dimensional hyperplane H_k in $\mathbb{R}^n$ can be described by k orthogonal unit vectors that can be completed to the orthonormal basis of $\mathbb{R}^n$ (ONB), and bias vector b $H_k = \{x = b + c_1 u_1 + ... + c_k u_k : c_1, ..., c_k \in \mathbb{R}\}$.

Let $\{u_i\}_{i=1}^n$ be some ONB in $\mathbb{R}^n$. Then $\forall\, i = 1, ..., n : x_i = b + \sum\limits_{j=1}^{n} c_{i,j} u_j$, where $c_{i,j} = (x_i, u_j)$ (it is Fourier expansion in $\mathbb{R}^n$).

Firstly, we will show that we can set $b = 0$ after applying some transformation to given vectors.

Auxiliary problem

For given m points $\{x_1, ..., x_m\}$ in $\mathbb{R}^n$ find the closest point to all of them in the sense of Euclidean distance.

Auxiliary problem solution

This problem can be formalized like $F(x) = \sum\limits_{k=1}^{m} \|x - x_k\|^2 \to \min,\ x \in \mathbb{R}^n$.

$F'_x(x^*) = 2\sum\limits_{k=1}^{m}(x^* - x_k) = 0 \Rightarrow x^* = \frac{1}{m}\sum\limits_{k=1}^{m} x_k$ — is stationary point.

As $F(x)$ is convex function, $x^* = \frac{1}{m}\sum\limits_{k=1}^{m} x_k$ is solution of given minimization problem.

Problem statement

So, after replacing x_i with $y_i = x_i - \frac{1}{m}\sum_{i=1}^{m} x_i$, we can assume $b = 0$, because closest point in $\mathbb{R}^n$ to all of them is 0, hence desired hyperplane must contain 0.

Thus, $\forall\ i = 1, ..., n : y_i = \sum_{j=1}^{n} c_{i,j} u_j$, where $c_{i,j} = (y_i, u_j)$. Hyperplane projections y_i are in form $\hat{y}_i = \sum_{j=1}^{k} c_{i,j} u_j$, $k < n$. Let's define error vectors $\varepsilon_i = y_i - \hat{y}_i = \sum_{j=k+1}^{n} c_{i,j} u_j$ and put them in one matrix:

$$E = \left(\varepsilon_1, \varepsilon_2, ..., \varepsilon_m\right) = \left(u_{k+1}, u_{k+2}, ..., u_n\right) \cdot \begin{pmatrix} c_{1,k+1} & c_{2,k+1} & \dots & c_{m,k+1} \\ c_{1,k+2} & c_{2,k+2} & \dots & c_{m,k+2} \\ \dots & \dots & \dots & \dots \\ c_{1,n} & c_{2,n} & \dots & c_{m,n} \end{pmatrix}$$

We will denote it $E = UC$. Note that $U^T U = I$ because of orthogonality and unit norm of u_i vectors, and $C = Y^T U$.

Problem solution

Problem

Find orthonormal vectors $\{u_i\}_{i=1}^{n}$ such that $\|E\|^2 \to min$, where $\|E\| = \sqrt{\sum_{i,j=1}^{n} e_{ij}^2}$ denotes Frobenius norm of error matrix E.

Let's denote $Y = (y_1, ..., y_m)$, $F = YY^T$.
$\|E\|^2 = \mathrm{Tr}\left(E^T E\right) = \mathrm{Tr}\left(C^T U^T U C\right) = \mathrm{Tr}\left(C^T C\right) = \mathrm{Tr}\left(U^T Y Y^T U\right) =$
$= \mathrm{Tr}\left(U^T F U\right)$. As $U = \sum\limits_{j=k+1}^{n} \left(0, ...0, u_j, 0..., 0\right) = \sum\limits_{j=k+1}^{n} U_j$, by linearity of Tr (trace of matrix, sum of diagonal elements) we have $\mathrm{Tr}\left(U^T F U\right) = \sum\limits_{j=k+1}^{n} \mathrm{Tr}\left(U_j^T F U_j\right)$. In each U_j matrix only one columns is non-zero, so $\sum\limits_{j=k+1}^{n} \mathrm{Tr}\left(U_j^T F U_j\right) = \sum\limits_{j=k+1}^{n} (F u_j, u_j)$.
So, we arrive to constrained optimization problem.

Problem solution

$$\begin{cases} F(u_{k+1}, ..., u_n) = \sum\limits_{j=k+1}^{n} (Fu_j, u_j) \to \min \\ \|u_j\|^2 = 1, \ j = k+1, ..., n \\ \{u_{k+1}, ..., u_n\} \text{ — linearly independent} \end{cases}$$

This problem is regular because gradients of constrains are linearly independent. Let's write Lagrange function and it's derivatives:

$$\mathcal{L}(u_{k+1}, ..., u_n, \lambda_{k+1}, ..., \lambda_n) = \sum_{j=k+1}^{n} \left((Fu_j, u_j) + \lambda_j \left(\|u_j\|^2 - 1 \right) \right)$$

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial u_j} = 2Fu_j + 2\lambda_j u_j = 0 \\ (j = k+1, ..., n) \\ \|u_j\|^2 = 1, \ j = k+1, ..., n \\ \{u_{k+1}, ..., u_n\} \text{ — linearly independent} \end{cases}$$

As $2Fu_j + 2\lambda_j u_j = 0 \Leftrightarrow Fu_j = -\lambda_j u_j$, solutions of this systems are u_j — eigenvectors of F with unit norm. Because we deal with minimization problem, $\{u_{k+1}, ..., u_n\}$ must correspond to smallest eigenvalues $\mu_j = -\lambda_j$.

Problem solution

$F^T = (YY^T)^T = YY^T$, $F \geq 0$, because $\forall\ x \in \mathbb{R}^n : (Fx, x) = (YY^T x, x) = (Y^T x, Y^T x) \geq 0$. So, all $\mu_j = -\lambda_j \geq 0$.
Target function is bounded from below and increasing, so u_j are solutions to minimization problem. Moreover, vectors of ONB $\{u_i\}_{i=1}^n$ are eigenvectors of F, ordered by decreasing of their eigenvalues, and $u_1, ..., u_k$ correspond to k largest eigenvalues (recall that $\hat{y}_i = \sum\limits_{j=1}^{k} c_{i,j} u_j$).

Also, recall that $y_i = x_i - \frac{1}{m} \sum\limits_{i=1}^{m} x_i$. Let's denote $X = (x_1, ..., x_m)$, then $Y = X - \left(\frac{1}{m} \sum\limits_{i=1}^{m} x_i \right) \cdot \underbrace{(1, 1, ..., 1)}_{m}$. Now we will compute YY^T.

Problem solution

let $\overline{x} = \frac{1}{m}\sum\limits_{i=1}^{m} x_i$.

$$YY^T = (X - \overline{x}\cdot(1,1,...,1))\cdot\left(X^T - (1,1,...,1)^T\cdot\overline{x}^T\right) =$$
$$= XX^T - \overline{x}\cdot(1,1,...,1)\cdot X^T - X\cdot(1,1,...,1)^T\cdot\overline{x}^T + \overline{x}\cdot(1,1,...,1)\cdot(1,1,...,1)^T\cdot\overline{x}^T = XX^T - m\cdot\overline{x}\cdot\overline{x}^T - m\cdot\overline{x}\cdot\overline{x}^T + m\cdot\overline{x}\cdot\overline{x}^T = XX^T - m\cdot\overline{x}\cdot\overline{x}^T =$$
$$= XX^T - \frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)\cdot\left(\sum_{i=1}^{m} x_i^T\right)$$

So, u_j are eigenvectors of $XX^T - m\cdot\overline{x}\cdot\overline{x}^T$.

Solution

Desired k-dimensional hyperplane is $\overline{x} + L(u_1, ..., u_k)$.

Here $\overline{x} = \frac{1}{m} \sum\limits_{i=1}^{m} x_i$ and $L(u_1, ..., u_k)$ is linear span of eigenvectors of $XX^T - m \cdot \overline{x} \cdot \overline{x}^T$ ($X = (x_1, ..., x_m)$), which correspond to k largest eigenvalues. Note that projection error (sum of squares of Euclidean norms of ε_i) is equal $\sum\limits_{j=k+1}^{n} \mu_j$ — it is sum of $n-k$ smallest eigenvalues of this matrix.

It's useful to note that projections $y_i = x_i - \overline{x}$ onto $L(u_1, ..., u_k)$ can be calculated using this formula: $\mathrm{pr}(y_i) = (u_1, ..., u_k)^T \cdot y_i$, or in matrix form: $\mathrm{pr}(Y) = (u_1, ..., u_k)^T \cdot Y$

Appendix: derivative of quadratic form

We used derivative of quadratic form and squared norm in solution. Vector derivative of vector-argument and scalar-valued function is a vector of partial derivatives with respect to each vector coordinated. We provide a proof of used derivative formulas.

$F(x) = (Ax, x)$, $x \in \mathbb{R}^n$, A — real symmetric $n \times n$ matrix.

$F(x+h) - F(x) = (Ax + Ah, x + h) - (Ax, x) = (Ax, x) + (Ax, h) + (Ah, x) + (Ah, h) - (Ax, x) = \left[A = A^T\right] = (2Ax, h) + (Ah, h)$

Linear part of increment is $2Ax$, so $F'(x) = 2Ax$. In particular, derivative of square of norm $\|x\|^2$ is simply $2x$.