PCA & Min-Max Thms

Recap

Given $x_1, \ldots, x_m \in \mathbb{R}^n$ i.i.d random vectors with

mean vector $\qquad\qquad$ covariance matrix

$$\mu = \mathbb{E}[x] \quad \text{and} \quad C = \mathbb{E}[(x-\mu)(x-\mu)^T]$$

Cov. matrix C is symmetric positive semidefinite.

eigenvalue decomposition

$$C = \overset{U^TU = UU^T = I}{U \underset{\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n \geq 0}{\Lambda} U^T}$$

Principle component analysis (PCA) analyzes $x_1, \ldots, x_m$ in the eigenbasis

$$y_j = U^T(x_j - \mu)$$

New random vectors are mean zero and have diagonal covariance (uncorrelated entries)

$$\underbrace{\mathbb{E}[y] = 0}_{\text{mean}} \quad \text{and} \quad \underbrace{\mathbb{E}[yy^T] = \Lambda}_{\text{covariance matrix}}$$

Principle Components

The k^{th} principle component of x is

$$y^{(k)} = u_k^T x$$

$\Rightarrow$ Coordinate of x in k^{th} direction of eigenbasis.

Remarkably, the leading principle components capture the "most" variance in the distribution of random vector X in the following sense:

$$\underset{\|v\|=1}{\operatorname{argmax}} \; \underbrace{\mathbb{E}\left[|v^T(x-\mu)|^2\right]}_{\text{Variance in } y = v^T(x-\mu)} = \underset{\|v\|=1}{\operatorname{argmax}} \; \mathbb{E}\left[v^T(x-\mu)(x-\mu)^T v\right]$$

$$= \underset{\|v\|=1}{\operatorname{argmax}} \; v^T \mathbb{E}\left[(x-\mu)(x-\mu)^T\right] v$$

$$= \underset{\|v\|=1}{\operatorname{argmax}} \; v^T C v \; = u_1$$

The Rayleigh Quotient $v^T C v$ is maximized precisely when v is the leading eigenvector of C.

pf sketch

$$\Rightarrow \quad v^T C v = v^T\left(\sum_{j=1}^{n} \lambda_j u_j u_j^T\right) v^T \qquad \|v\|=1$$

$$= \sum_{j=1}^{n} \lambda_j |v^T u_j|^2 \leq \lambda_1 \sum_{j=1}^{n} |v^T u_j|^2$$

$$= \lambda_1 \|v\|^2 = \lambda_1 \qquad \text{upper bound}$$

$$\Rightarrow \quad u_1^T C u_1 = u_1^T\left(\sum_{j=1}^{n} \lambda_j u_j u_j^T\right) u_1 = \sum_{j=1}^{n} \lambda_j |u_1^T u_j|^2$$

$$= \lambda_1 |u_1^T u_1|^2 = \lambda_1$$

upper bound achieved when $v = u_1$

So u_1 maximizes the variance of $y^{(1)}$, the first component of new random vector.
$\Rightarrow$ And λ_1 (eigenvalue) is max value of RQ.

We can use a similar principle to "find" the remaining rows of the transformation $(x-\mu) \to y$.
If we restrict to $V_1^\perp = (\text{span}\{u_1\})^\perp$, then

$$u_2 = \underset{\substack{\|v\|=1 \\ v \in V_1^\perp}}{\operatorname{argmax}} \mathbb{E}\left[|v^T(x-\mu)|^2\right] = \underset{\substack{\|v\|=1 \\ v \in V_1}}{\operatorname{argmax}} v^T C v$$

pf sketch

$$v^T C v = \sum_{j=1}^{n} \lambda_j |v^T u_j|^2 = \sum_{j=2}^{n} \lambda_j |v^T u_j|^2 \le \lambda_2 \overbrace{\sum_{j=2}^{n} |v^T u_j|^2}^{\le \|v\|^2 = 1}$$

$\to v^T u_1 = 0$

$$u_2^T C u_2 = \sum_{j=1}^{n} \lambda_j |u_2^T u_j|^2 = \lambda_2 \underbrace{|u_2^T u_2|^2}_{u_2^T u_2 = 1} = \lambda_2$$

$\to u_2^T u_j = 0$ for $j \neq 2$

So λ_2 is the maximum of $v^T C v$ when v is restricted to $\|v\| = 1$ and $v^T u_1 = 0$, and u_2 is the vector achieving the max.

$\Rightarrow$ u_2 maximizes the variance of the second principle component subject to constraint that $u_2 \perp u_1$ (b/c we want ONB).

$\Rightarrow$ Equivalently, Rayleigh Quotient is maximized over $v \in V_1$ by $v = u_1$.

In general, if $V_k = \text{span}\{u_1, \ldots, u_k\}$, then

$$u_{k+1} = \underset{v \in V_k^{\perp}}{\arg\max}\ \mathbb{E}\left[|v^T(x-\mu)|^2\right] = \underset{v \in V_k^{\perp}}{\arg\max}\ v^T C v$$

Courant-Fisher-Weyl Min-Max Principle

We can formulate the eigenvalues themselves as extrema of the Rayleigh Quotient using the same idea:

$$\lambda_k = \max_{\substack{\dim(M)\\=k}} \min_{\substack{v\in M\\ \|v\|=1}} v^T C v$$

selects $\mathrm{span}\{u_1, \dots, u_k\}$ — selects $\underline{\lambda_k} \leq \dots \leq \lambda_1$

"Variational characterization of eigenvalues"

Courant Fischer Weyl

$$\lambda_k = \min_{\substack{\dim(M)\\=n-k+1}} \max_{\substack{v\in M\\ \|v\|=1}} v^T C v$$

selects $\mathrm{span}\{u_k, \dots, u_n\}$ — selects $\underline{\lambda_k} \geq \dots \geq \lambda_n$

This characterization replaces the orthogonality constraints with an outer optimization over an appropriately sized subspace.

The min-max (max-min) characterizations hold for any self-adjoint matrix (real-symm or Hermitian) and can be extended to self-adjoint compact ops.

Computing Principle Components

In practice, the mean and covariance of the random variable $x \in \mathbb{R}^n$ are not usually known. Instead, they can be estimated directly from the data.

$$\tilde{u} = \frac{1}{m} \sum_{j=1}^{m} x_j$$ (Sample mean)

$$\tilde{C} = \frac{1}{m-1} \sum_{j=1}^{m} (x_j - \tilde{u})(x_j - \tilde{u})^T$$ (Sample covariance)

$$= \frac{1}{m-1} BB^T \quad \text{where} \quad B = X - \tilde{u}\mathbf{1}^T$$

subtract mean from each column

Note that $\tilde{C}$ remains symmetric PSD.

$\Rightarrow$ We can calculate the eigenvalue decomp. of the sample covariance of the data:

$$\tilde{C} = U \Lambda U^T$$

$\Rightarrow$ Maximizes the sample variance along each successive eigendirection s.t. orthogonal constraints