Statistics for Applications

Chapter 9: Principal Component Analysis (PCA)

Multivariate statistics and review of linear algebra (1)

- Let $\mathbf{X}$ be a d-dimensional random vector and $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be n independent copies of $\mathbf{X}$.
- Write $\mathbf{X}_i = (X_i^1, \ldots, X_i^d)^\top, \quad i = 1, \ldots, n.$
- Denote by $\mathbb{X}$ the random $n \times d$ matrix

$$\mathbb{X} = \begin{pmatrix} \cdots & \mathbf{X}_1^\top & \cdots \\ & \vdots & \\ \cdots & \mathbf{X}_n^\top & \cdots \end{pmatrix}.$$

Multivariate statistics and review of linear algebra (2)

Assume that $\mathbb{E}[\|\mathbf{X}\|_2^2] < \infty$.

Mean of $\mathbf{X}$:

$$\mathbb{E}[\mathbf{X}] = \Big(\mathbb{E}[X^1], \dots, \mathbb{E}[X^d]\Big)^\top.$$

Covariance matrix of $\mathbf{X}$: the matrix $\Sigma = (\sigma_{j,k})_{j,k=1,\dots,d}$, where

$$\sigma_{j,k} = cov(\mathbf{X}^j, \mathbf{X}^k).$$

It is easy to see that

$$\Sigma = \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^\top = \mathbb{E}\Big[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top\Big].$$

Multivariate statistics and review of linear algebra (3)

Empirical mean of $\mathbf{X}_1, \ldots, \mathbf{X}_n$:

$$\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i = \left(\bar{X}^1, \ldots, \bar{X}^d \right)^\top .$$

Empirical covariance of $\mathbf{X}_1, \ldots, \mathbf{X}_n$: the matrix $S = (s_{j,k})_{j,k=1,\ldots,d}$ where $s_{j,k}$ is the empirical covariance of the X_i^j, X_i^k, $i = 1 \ldots, n$.

It is easy to see that

$$S = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top - \bar{\mathbf{X}} \bar{\mathbf{X}}^\top = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{X}_i - \bar{\mathbf{X}} \right) \left(\mathbf{X}_i - \bar{\mathbf{X}} \right)^\top .$$

Multivariate statistics and review of linear algebra (4)

Note that $\bar{\mathbf{X}} = \dfrac{1}{n}\mathbb{X}^\top \mathbb{1}$, where $\mathbb{1} = (1, \ldots, 1)^\top \in \mathbb{R}^d$.

Note also that

$$S = \frac{1}{n}\mathbb{X}^\top \mathbb{X} - \frac{1}{n^2}\mathbb{X}\mathbb{1}\mathbb{1}^\top \mathbb{X} = \frac{1}{n}\mathbb{X}^\top H \mathbb{X},$$

where $H = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}^\top$.

H is an orthogonal projector: $H^2 = H, H^\top = H$. *(on what subspace ?)*

If $\mathbf{u} \in \mathbb{R}^d$,

- $\mathbf{u}^\top \Sigma \mathbf{u}$ is the variance of $\mathbf{u}^\top \mathbf{X}$;
- $\mathbf{u}^\top S \mathbf{u}$ is the sample variance of $\mathbf{u}^\top \mathbf{X}_1, \ldots, \mathbf{u}^\top \mathbf{X}_n$.

Multivariate statistics and review of linear algebra (5)

In particular, $\mathbf{u}^\top S\mathbf{u}$ measures how spread (i.e., diverse) the points are in direction $\mathbf{u}$.

If $\mathbf{u}^\top S\mathbf{u} = 0$, then all $\mathbf{X}_i$' s are in an affine subspace orthogonal to $\mathbf{u}$.

If $\mathbf{u}^\top \Sigma\mathbf{u} = 0$, then $\mathbf{X}$ is almost surely in an affine subspace orthogonal to $\mathbf{u}$.

If $\mathbf{u}^\top S\mathbf{u}$ is large with $\|\mathbf{u}\|_2 = 1$, then the direction of $\mathbf{u}$ explains well the spread (i.e., diversity) of the sample.

Multivariate statistics and review of linear algebra (6)

In particular, Σ and S are symmetric, positive semi-definite.

Any real symmetric matrix $A \in \mathbb{R}^{d \times d}$ has the decomposition

$$A = PDP^\top,$$

where:

- P is a $d \times d$ orthogonal matrix, i.e., $PP^\top = P^\top P = I_d$;
- D is diagonal.

The diagonal elements of D are the *eigenvalues* of A and the columns of P are the corresponding *eigenvectors* of A.

A is semi-definite positive iff all its eigenvalues are nonnegative.

Principal Component Analysis: Heuristics (1)

The sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ makes a cloud of points in $\mathbb{R}^d$.

In practice, d is large. If $d > 3$, it becomes impossible to represent the cloud on a picture.

Question: Is it possible to project the cloud onto a linear subspace of dimension $d' < d$ by keeping as much information as possible ?

Answer: PCA does this by keeping as much covariance structure as possible by keeping orthogonal directions that discriminate well the points of the cloud.

Principal Component Analysis: Heuristics (2)

Idea: Write $S = PDP^\top$, where

$P = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$ is an orthogonal matrix, i.e.,
$\|\mathbf{v}_j\|_2 = 1, \mathbf{v}_j^\top \mathbf{v}_k = 0, \forall j \neq k$.

$$D = \begin{pmatrix} \lambda_1 & & & & \\ & \lambda_2 & & \mathbf{0} & \\ & & \ddots & & \\ & \mathbf{0} & & \ddots & \\ & & & & \lambda_d \end{pmatrix}, \text{ with } \lambda_1 \geq \ldots \geq \lambda_d \geq 0.$$

Note that D is the empirical covariance matrix of the $P^\top \mathbf{X}_i$'s, $i = 1, \ldots, n$.

In particular, λ_1 is the empirical variance of the $\mathbf{v}_1^\top \mathbf{X}_i$'s; λ_2 is the empirical variance of the $\mathbf{v}_2^\top \mathbf{X}_i$'s, etc...

Principal Component Analysis: Heuristics (3)

So, each λ_j measures the spread of the cloud in the direction $\mathbf{v}_j$.

In particular, $\mathbf{v}_1$ is the direction of maximal spread.

Indeed, $\mathbf{v}_1$ maximizes the empirical covariance of $\mathbf{a}^\top \mathbf{X}_1, \ldots, \mathbf{a}^\top \mathbf{X}_n$ over $\mathbf{a} \in \mathbb{R}^d$ such that $\|\mathbf{a}\|_2 = 1$.

Proof: For any unit vector $\mathbf{a}$, show that

$$\mathbf{a}^\top \Sigma \mathbf{a} = \quad P^\top \mathbf{a} \quad^\top D \quad P^\top \mathbf{a} \quad \le \lambda_1,$$

with equality if $\mathbf{a} = \mathbf{v}_1$.

Principal Component Analysis: Main principle

Idea of the PCA: Find the collection of orthogonal directions in which the cloud is much spread out.

Theorem

$$\mathbf{v}_1 \in \underset{\|\mathbf{u}\|=1}{\operatorname{argmax}} \; \mathbf{u}^\top S \mathbf{u},$$

$$\mathbf{v}_2 \in \underset{\|\mathbf{u}\|=1, \mathbf{u}\perp\mathbf{v}_1}{\operatorname{argmax}} \; \mathbf{u}^\top S \mathbf{u},$$

$$\dots$$

$$\mathbf{v}_d \in \underset{\|\mathbf{u}\|=1, \mathbf{u}\perp\mathbf{v}_j, j=1,\dots,d-1}{\operatorname{argmax}} \; \mathbf{u}^\top S \mathbf{u}.$$

Hence, the k orthogonal directions in which the cloud is the most spread out correspond exactly to the eigenvectors associated with the k largest values of S.

Principal Component Analysis: Algorithm (1)

1. Input: $\mathbf{X}_1, \ldots, \mathbf{X}_n$: cloud of n points in dimension d.

2. Step 1: Compute the empirical covariance matrix.

3. Step 2: Compute the decomposition $S = PDP^\top$, where $D = Diag(\lambda_1, \ldots, \lambda_d)$, with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ and $P = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$ is an orthogonal matrix.

4. Step 3: Choose $k < d$ and set $P_k = (\mathbf{v}_1, \ldots, \mathbf{v}_k) \in \mathbb{R}^{d \times k}$.

5. Output: $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, where

$$\mathbf{Y}_i = P_k^\top \mathbf{X}_i \in \mathbb{R}^k, \quad i = 1, \ldots, n.$$

Question: How to choose k ?

Principal Component Analysis: Algorithm (2)

Question: How to choose k ?

Experimental rule: Take k where there is an inflection point in the sequence $\lambda_1, \ldots, \lambda_d$ (scree plot).

Define a criterion: Take k such that

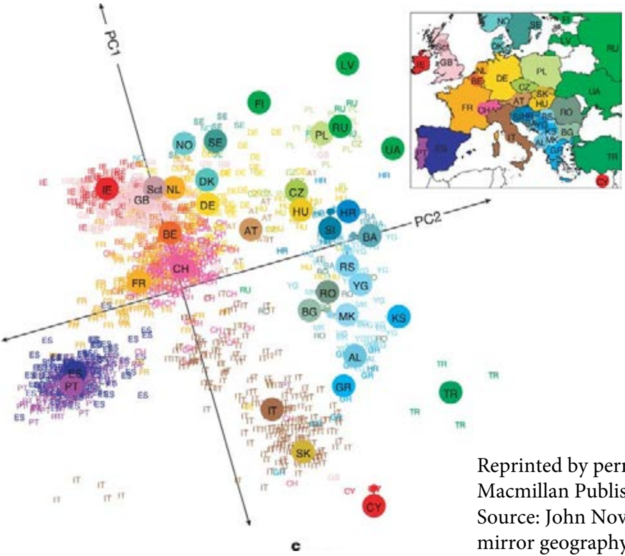
$$\frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_d} \geq 1 - \alpha,$$

for some $\alpha \in (0, 1)$ that determines the approximation error that the practitioner wants to achieve.

Remark: $\lambda_1 + \ldots + \lambda_k$ is called *the variance explained by the PCA* and $\lambda_1 + \ldots + \lambda_d = Tr(S)$ is *the total variance*.

Data visualization: Take $k = 2$ or 3.

Example: Expression of 500,000 genes among 1400 Europeans



Reprinted by permission from Macmillan Publishers Ltd: Nature. Source: John Novembre, et al. "Genes mirror geography within Europe." Nature 456 (2008): 98-101. © 2008.

Principal Component Analysis - Beyond practice (1)

PCA is an algorithm that reduces the dimension of a cloud of points and keeps its covariance structure as much as possible.

In practice this algorithm is used for clouds of points that are not necessarily random.

In statistics, PCA can be used for estimation.

If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. random vectors in $\mathbb{R}^d$, how to estimate their population covariance matrix Σ ?

If $n \gg d$, then the empirical covariance matrix S is a consistent estimator.

In many applications, $n \ll d$ (e.g., gene expression). Solution: sparse PCA

Principal Component Analysis - Beyond practice (2)

It may be known beforehand that Σ has (almost) low rank.

Then, run PCA on S: Write $S \approx S'$, where

$$S' = P \begin{pmatrix} \lambda_1 & & & & & & \\ & \lambda_2 & & & \mathbf{0} & & \\ & & \ddots & & & & \\ & & & \lambda_k & & & \\ & & & & 0 & & \\ & \mathbf{0} & & & & \ddots & \\ & & & & & & 0 \end{pmatrix} P^\top.$$

S' will be a better estimator of S under the low-rank assumption.

A theoretical analysis would lead to an optimal choice of the tuning parameter k.