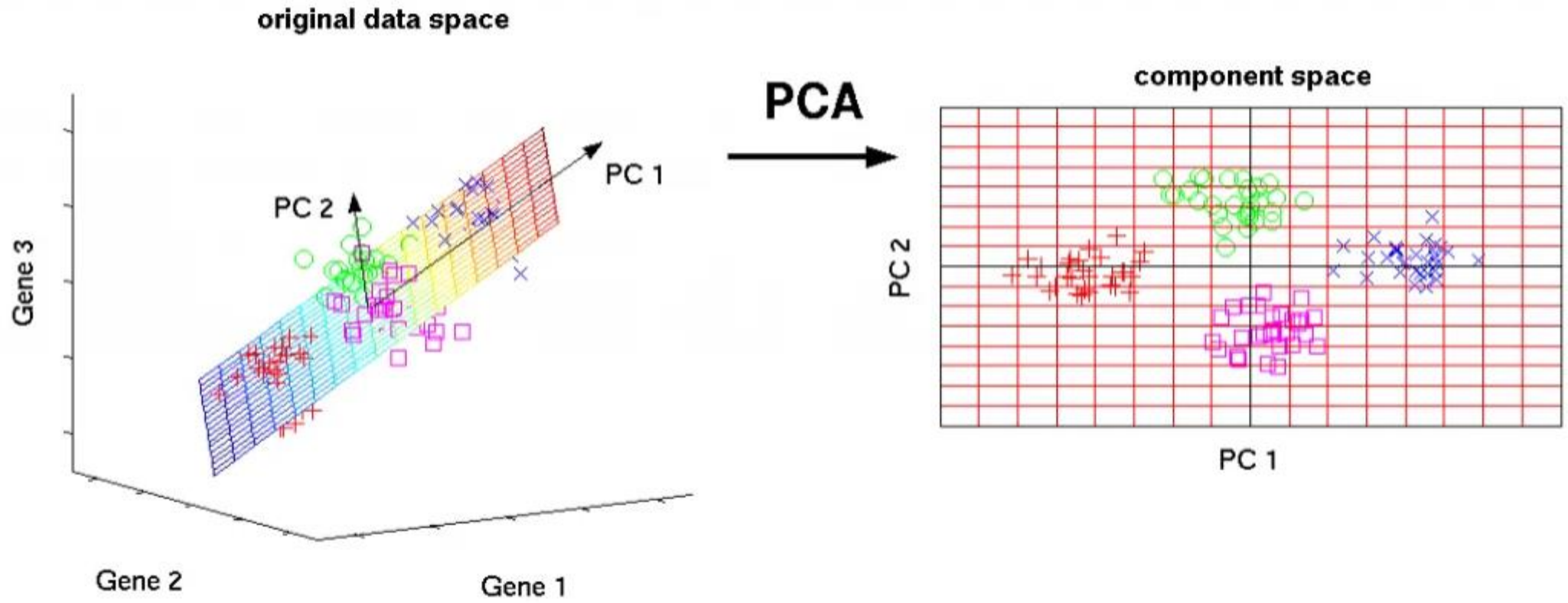
ضروریات آمار برای علم داده

تحلیل مولفه اصلی

تحلیل عاملی


original data space
PC 1
PC 2
Gene 3
Gene 2
Gene 1
PCA
component space
PC 2
PC 1

تحلیل مؤلفه اصلی

یکی از روش‌های تحلیل چند متغیره است

هدف اصلی آن تقلیل بعد مسئله مورد مطالعه است

از کاربردهای مهم تحلیل مؤلفه اصلی در رگرسیون است

با استفاده از این روش تعداد زیادی متغیر توضیحی هم‌بسته با تعداد محدودی متغیر توضیحی جدید که مولفه اصلی نامیده می‌شوند و ناهمبسته هستند جایگزین می‌شود؛ به این ترتیب نه تنها بعد مسئله کاهش می‌یابد بلکه مسئله هم‌خطی نیز حل می‌شود.

تحلیل مؤلفه اصلی

- انجام بهترین بصری‌سازی از داده‌های دارای ابعاد بالا
- برای غلبه بر افزونگی متغیرهای موجود
- روی مجموعه داده‌های دارای ویژگی‌های عددی قابل اعمال است
  مولفه‌های این متغیرها حاصل ترکیب خطی نرمال شده متغیرهای پیش‌بین اصلی هستند.
- هدف این مولفه‌ها حفظ بیشترین اطلاعات ممکن با واریانس‌های بالا است
- اولین مولفه اصلی بالاترین واریانس را داراست و پس از آن مولفه اصلی دوم دارای بیشترین مقدار واریانس است و این موضوع برای مولفه‌های اصلی سوم و دیگر مولفه‌های اصلی نیز صادق است.
- مولفه‌ها باید ناهمبسته باشند (جهت‌های آن‌ها متعامد است).
- نرمال‌سازی داده‌ها هنگامی‌که متغیرها دارای واحدهای (یکاهای) گوناگونی هستند، فوق‌العاده مهم است

Principal component Analysis

$$z = \frac{value - mean}{standard\ deviation}$$

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Compute the eigenvectors and eigenvalues of the covariance matrix

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

مفاهیم جبری

$$M \text{ positive semi-definite} \iff z^{\mathsf{T}} M z \geq 0 \text{ for all } z \in \mathbb{R}^n \setminus \mathbf{0}$$

$$det(A - \lambda I) = 0$$

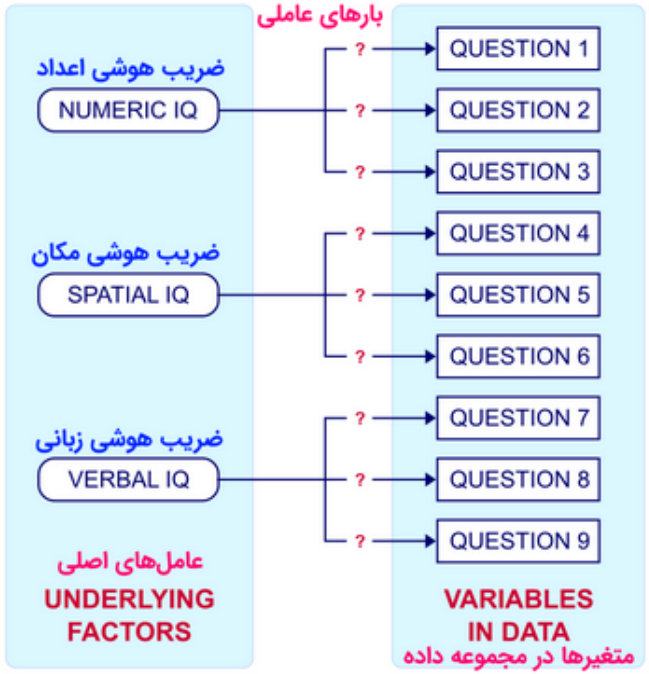
$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \qquad |A| = a(ei - fh) - b(di - fg) + c(dh - eg)$$

$$(A - \lambda I)\,\mathbf{v} = \mathbf{0}$$

تحلیل عاملی

هدف اصلی تحلیل عاملی در صورت امکان بیان روابط کوواریانس میان بسیاری از متغیرها براساس چند کمیت تصادفی غیرقابل مشاهده است که عامل‌ها نامیده می‌شوند.

فرض کنید متغیرها را بتوان براساس همبستگی‌یشان دسته‌بندی کرد یعنی تمام متغیرهای یک گروه خاص در میان خودشان همبستگی بالایی دارند ولی با متغیرهای یک گروه دیگر همبستگی نسبتا کمی دارند؛ می‌توان گفت هرگروه از متغیرها یک ترکیب یا عامل مورد نظر را نشان می‌دهند که نشان‌دهنده همبستگی‌های مشاهده شده است.



تحلیل عاملی

$$x_{i,m} - \mu_i = l_{i,1} f_{1,m} + \cdots + l_{i,k} f_{k,m} + \varepsilon_{i,m}$$

whereby

- $x_{i,m}$ is the value of the ith observation of the mth individual,
- μ_i is the observation mean for the ith observation,
- $l_{i,j}$ is the loading for the ith observation of the jth factor,
- $f_{j,m}$ is the value of the jth factor of the mth individual, and
- $\varepsilon_{i,m}$ is the (i, m)th *unobserved stochastic error term* with mean zero and finite variance.