



به نام خدا



کارگاه علم داده با پایتون پیشرفته

جلسه هفتم: تقلیل ابعاد (dimensionality reduction)

مدرس :

مهرناز جلیلی

دانشجو کارشناسی ارشد علم داده ها

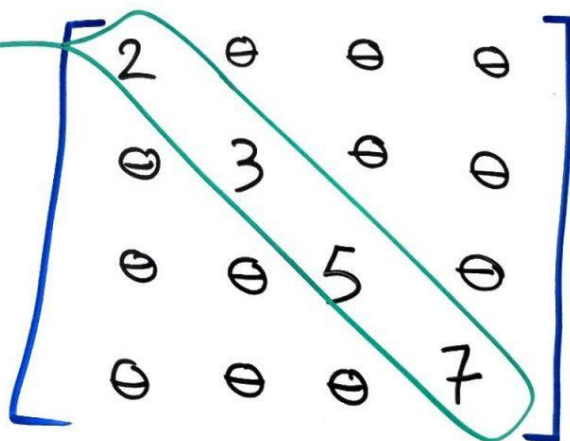
دانشگاه شهید بهشتی



مقدمه

$$AA^T = A^T A = I$$

ماتریس متعامد


$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 7 \end{bmatrix}$$

ماتریس قطری

← قطر اصلی

دترمینان ماتریس

$$\det A = |A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

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

بردار ویژه و مقدار ویژه یک ماتریس

بردار v ، بردار ویژه ماتریس مربعی A است، اگر ثابت λ وجود داشته باشد بطوری که:

$$A \cdot v = \lambda \cdot v$$

محاسبه مقدار ویژه

$$\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$$

مقادیر ویژه: ریشه‌های $|\mathbf{A} - \lambda \cdot \mathbf{I}|$

مثال

$$A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$$

$$\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$$

$$\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$$

ماتریس کو اریانس

X_1	X_2
1	2
3	5
5	8

ماتریس کواریانس (روش دوم)

X_1	X_2
1	2
3	5
5	8

تقلیل ابعاد

الگوریتم تقلیل ابعاد، یکی از الگوریتم های یادگیری بدون ناظر است.

کاربردهای تقلیل ابعاد:

- 1- به تصویر کشیدن داده ها (Data Visualization)
هدف از تجسم داده ها، تقلیل داده ها به $k = 2$ یا $k = 3$ است طوری که داده ها را بتوان بصورت نمودار ترسیم کرد و تجسم آن ها ساده شود.
- 2- فشرده کردن داده ها (Data Compression)
فشرده سازی به منظور کاهش حجم حافظه کامپیوتر و تسریع الگوریتم های یادگیری انجام می شود.
هدف فشرده سازی، تقلیل ابعاد داده ها از n به k است طوری که درصد بالایی از واریانس داده ها حفظ شود.

استخراج ویژگی

استخراج ویژگی (Feature Extraction):
تبدیل خطی یا غیر خطی بر روی فضای ویژگی اصلی
ایجاد ویژگی های جدید از ویژگی های اصلی

انتخاب ویژگی (Feature selection):
انتخاب یک زیرمجموعه از مجموعه
ویژگی های داده شده

All Features



Feature Selection



Final Features



الگوریتم های خطی Feature extraction

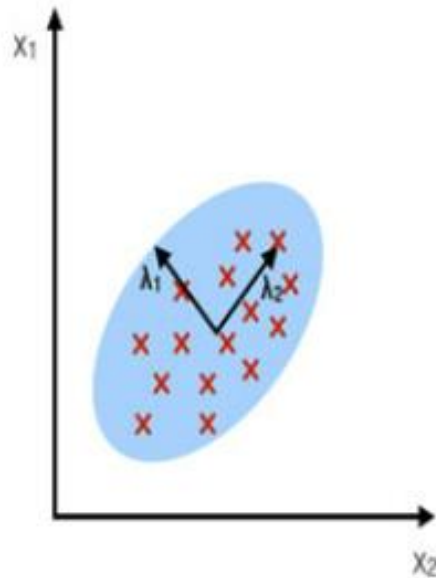
بدون نظارت : PCA (Principal Component Analysis)

با نظارت : LDA (Linear Discriminant Analysis)

در LDA از برچسب ها استفاده می کند و راستایی را پیدا می کند که اگر داده ها بر روی آن پروجکت شوند، داده های دو کلاس از همدیگر فاصله دارند و داده های یک کلاس به هم نزدیک تر باشند.

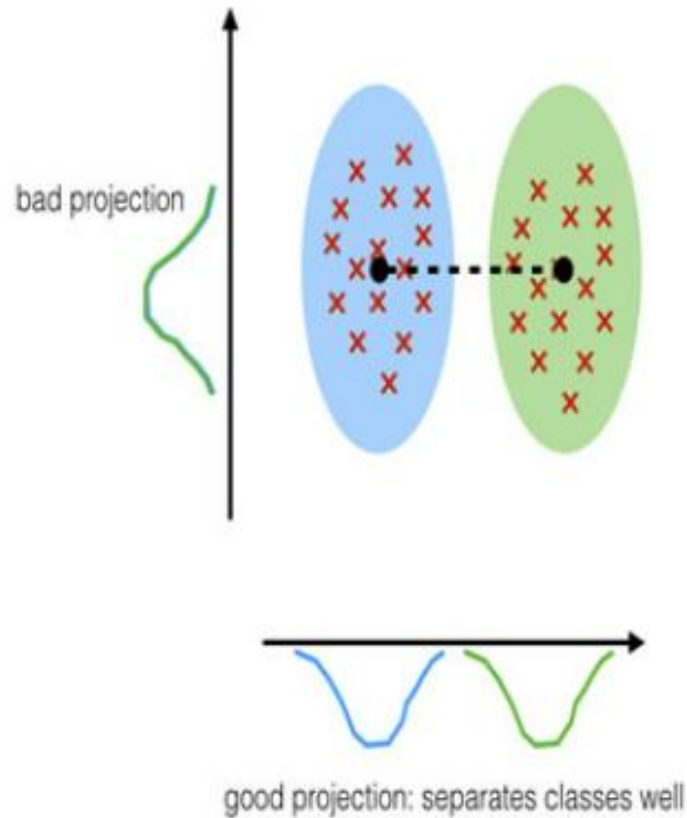
PCA:

component axes that
maximize the variance



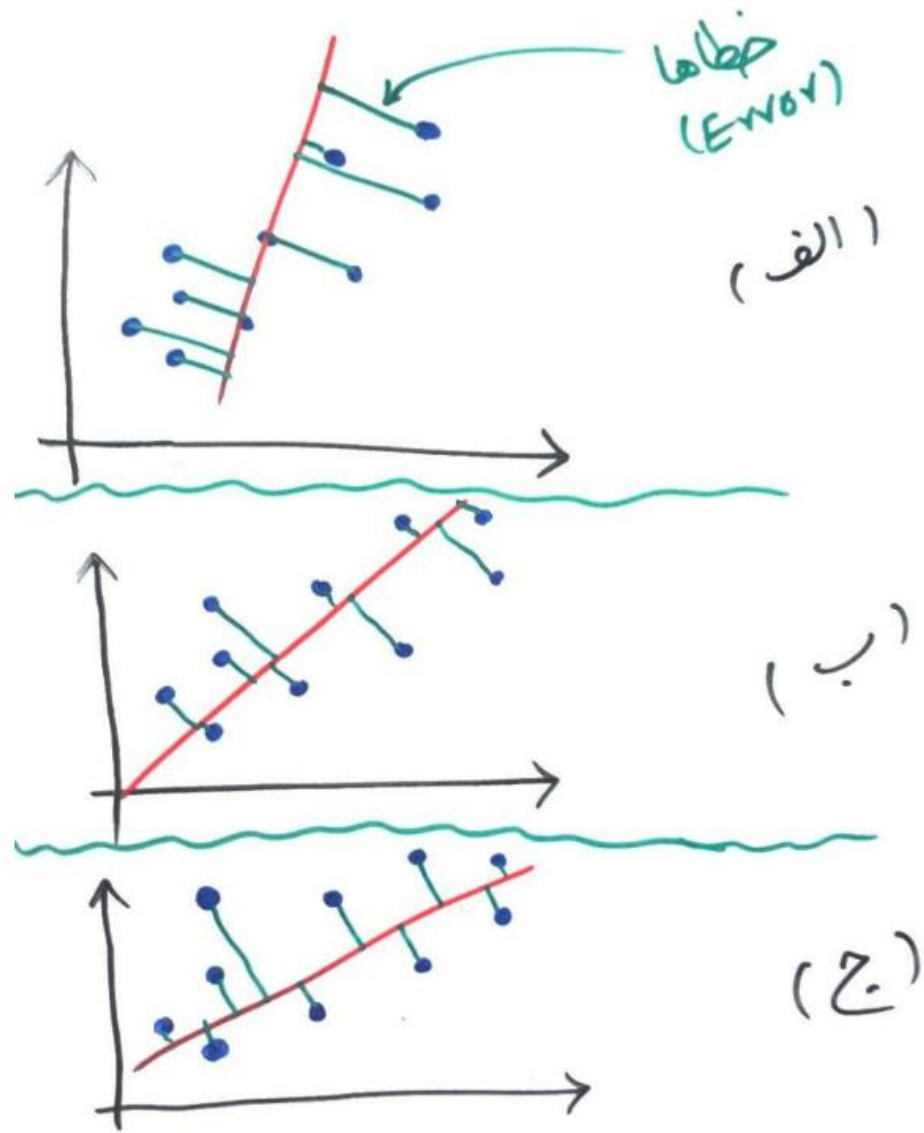
LDA:

maximizing the component
axes for class-separation

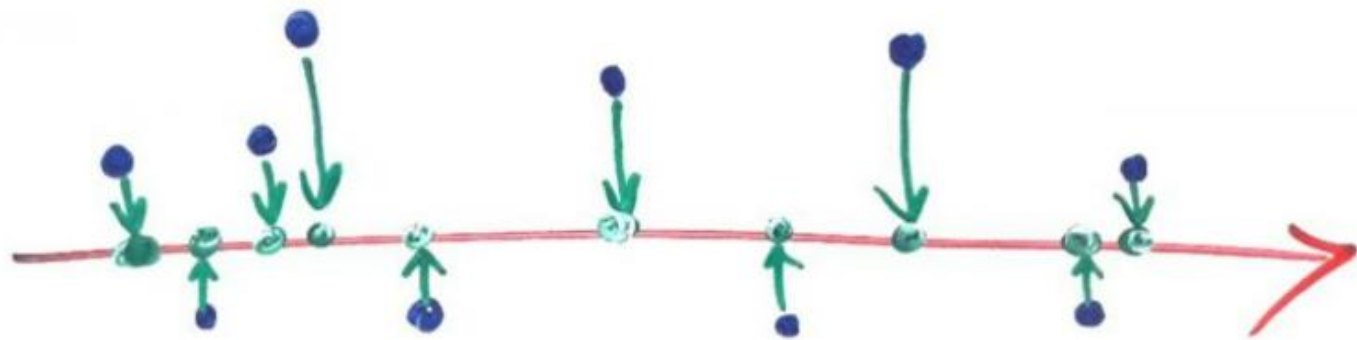


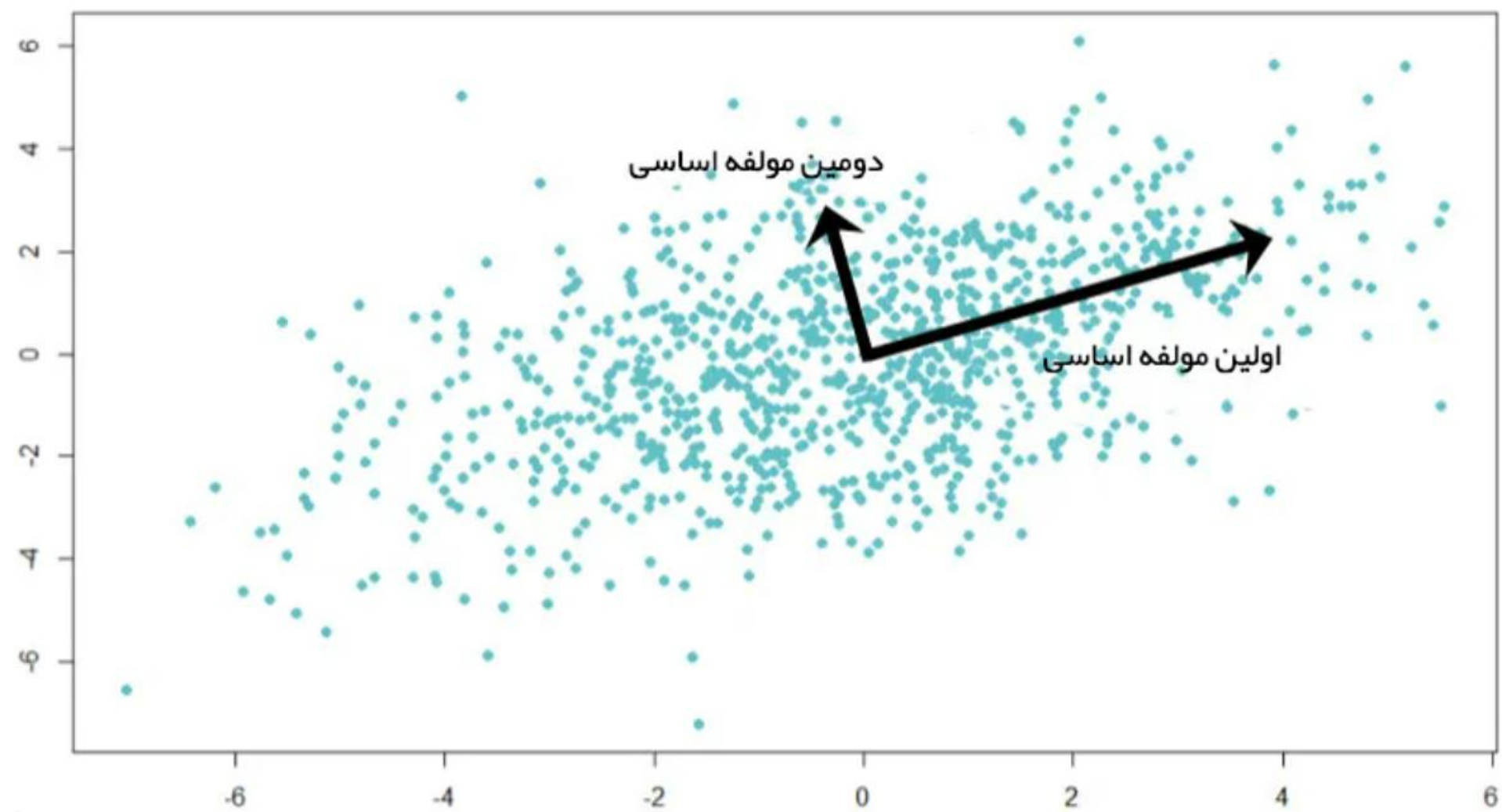
PCA

یافتن خطی که اگر از نقاط بر آن خط عمود کنیم، مجموع مربعات فاصله های نقاط تا تصاویر آن ها بر آن خط کمینه شود.



کمی دقت کنید. فاصله نقاط نسبت به خط قرمز با خط با رنگ سبز مشخص شده‌اند. اگر جمع این فاصله (خطا) را برای هر نمودار برابر خطای کلی داده‌ها نسبت به خط قرمز در نظر بگیریم، تصویر الف بیشترین میزان خطا را دارد و بعد از آن تصویر ب و در نهایت تصویر ج کمترین خطا را دارد. PCA به دنبال ساختن خطی مانند خط ج است (که در واقع همان بردار ویژه ماست) که کمترین خطا را داشته باشد. با این کار هر کدام از نقاط بر روی خط قرمز نگاشت می‌شوند و در تصویر بالا که بُعد ۱ است می‌توان این بُعد ۲ را به بُعد ۱ (که همان خط قرمز رنگ است) نگاشت کرد.

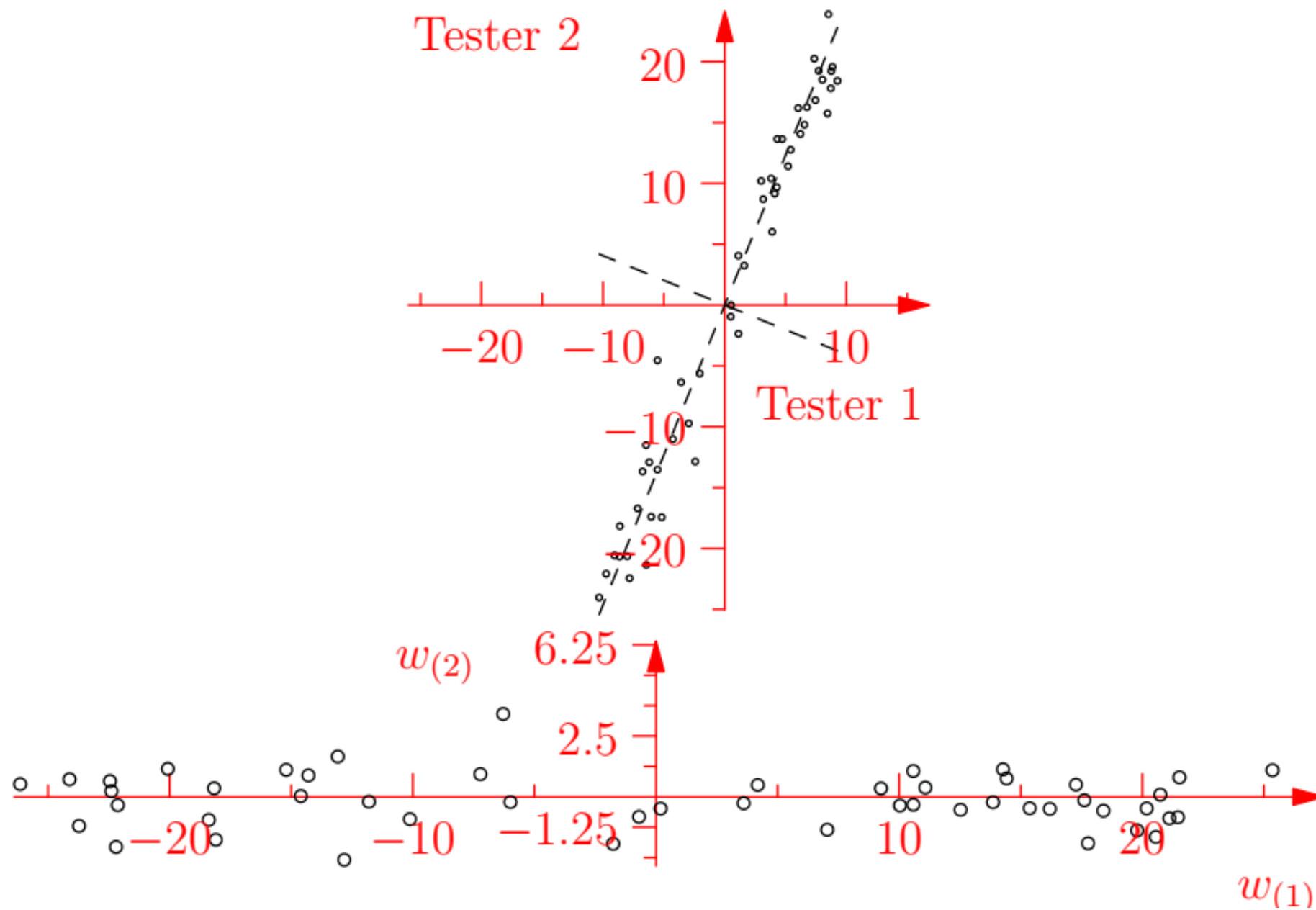




اگر بخواهیم یک بردار روی فضا پیدا کنیم که داده ها روی آن پروجکت شوند و دارای بیشترین واریانس باشند، آن بردار، بردار ویژه متناظر با بزرگترین مقدار ویژه ماتریس کواریانس است.

بردارهای متناظر با مقادیر ویژه کوچک را می توان دور ریخت تا ابعاد کاهش پیدا کنند تا در فضای کوچکتری عملیات را انجام داد. بردارهای متناظر با مقادیر ویژه کوچک، در بازسازی تاثیر زیادی ندارند.

یا عبارتی بردارهای ویژه Eigenvectors مولفه های اساسی برای مجموعه داده محاسبه و همه آنها در یک «ماتریس تصویر» (Projection Matrix) گردآوری می شوند. به هر یک از این بردارهای ویژه یک مقدار ویژه تخصیص داده می شود که می تواند به عنوان طول یا بزرگنمایی بردار ویژه متناظر در نظر گرفته شود. اگر برخی از مقدارهای ویژه دارای بزرگنمایی به طور قابل توجهی بزرگتر از دیگر موارد بوده اند، کاهش مجموعه داده با تحلیل مولفه اساسی (PCA) به یک زیرفضای ابعاد کوچکتر با حذف جفت ویژه هایی با «اطلاعات کمتر» معقول است.

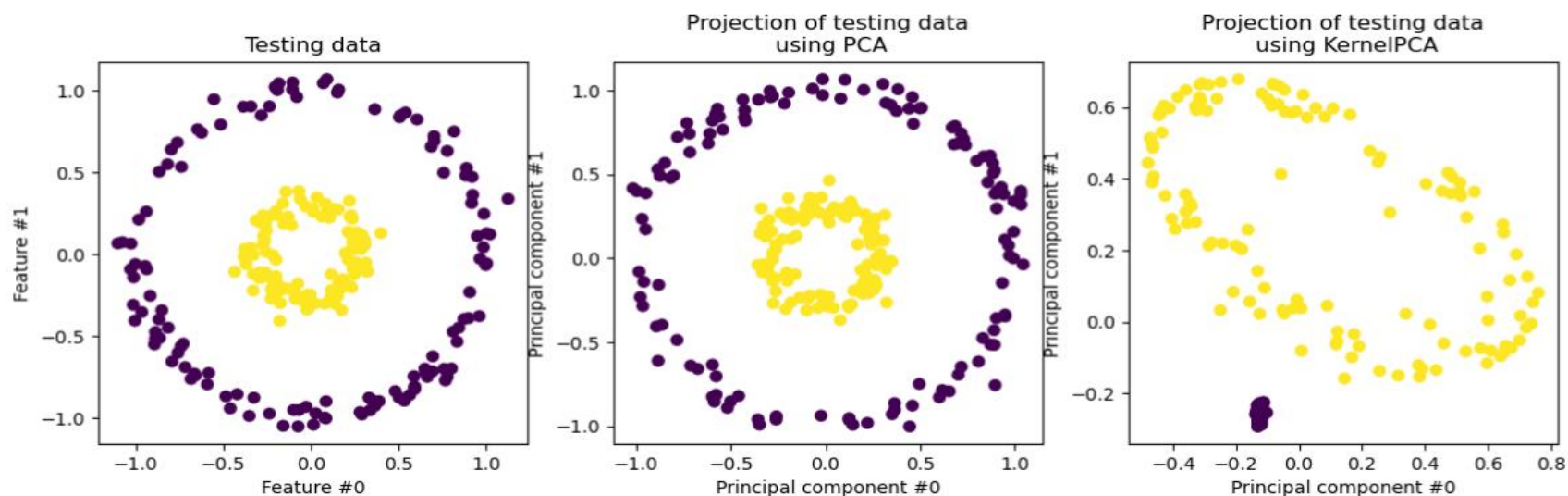


خلاصه‌ای از رویکرد PCA

- استانداردسازی داده‌ها
- به دست آوردن بردارهای ویژه و مقدارهای ویژه از «ماتریس کواریانس» (Covariance matrix) یا «ماتریس همبستگی» (Correlation Matrix)، یا انجام «تجزیه مقدارهای منفرد» (Singular Vector Decomposition)
- مرتب‌سازی مقدارهای ویژه به ترتیب نزولی و انتخاب k بردار ویژه‌ای که متناظر با K بزرگ‌ترین مقدار ویژه هستند. K تعداد ابعاد زیرفضای ویژگی جدید است ($k \leq d$).
- ساخت ماتریس تصویر W از K بردار ویژه انتخاب شده
- تبدیل مجموعه داده اصلی X به وسیله W ، برای به دست آوردن زیرفضای K بُعدی Y

Kernel PCA

تجزیه و تحلیل مولفه اصلی می تواند با استفاده از ترند هسته در یک روش غیر خطی استفاده شود.
تکنیک حاصل قادر به ساخت نقشه های غیر خطی است که واریانس را در داده ها به حداکثر می رساند.



t-SNE

یک نگاشت غیر خطی است که فاصله ی هر دو نقطه را حفظ می کند یعنی فاصله ی نقاط را در فضای دومی همانند فضای اولی حفظ می کند (distance-based)

