Machine Learning

کاهش ابعاد

Dimensionality Reduction

یادگیری ماشین

دانشگاه شهید بهشتی

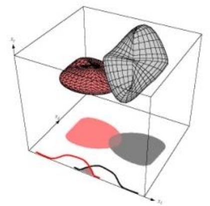
پژوهشکده‌ی فضای مجازی

پاییز ۱۳۹۹

احمد محمودی ازناوه

فهرست مطالب

- مزایای کاهش ابعاد
- انتخاب خصیصه
- استخراج خصیصه
- تحلیل مؤلفه‌ی اصلی
- تحلیل تفکیک خطی
- تحلیل عاملی
- تجزیه به مقادیر تکین
- تغییر مقیاس داده‌های چند بعدی



نکبت ابعاد!

- از لحاظ نظری، افزایش ابعاد منجر به بهبود عملکرد دسته‌بندی می‌شود، اما در عمل همیشه این گونه نیست.

Curse of dimensionality

- انتظار می‌رود در یک فرآیند ایده‌آل دسته‌بندی یا رگرسیون از خصیصه‌های بی‌اهمیت صرفنظر شود و فرآیند «کاهش ابعاد» به صورت جداگانه مورد نیاز نباشد. با این وجود کاهش ابعاد به دلایل زیر مورد توجه قرار می‌گیرد:

مزایای کاهش ابعاد

- «کاهش حجم محاسبات»: حافظه‌ی مصرفی و حجم محاسبات به تعداد (N) و ابعاد (d) داده‌ها بستگی دارد.
  - زمان محاسبات
  - حافظه‌ی مورد نیاز
- «صرفه‌جویی در جمع‌آوری داده»: حذف داده‌های غیرضروری
- «مقاوم‌بودن»(robustness): مدل‌های ساده، هنگامی که داده‌هاي آموزشی کم باشد، «مقاوم‌تر» می‌باشند؛ قدرت پیش‌بینی برای تعداد مشخصی داده، با افزایش ابعاد، کاهش می‌یابد.

مزایای کاهش ابعاد(ادامه...)

- «استخراج دانش»: با تعداد خصیصه‌های کم‌تر، در مورد داده‌ها و فرآیندهای مربوط به آن درک بهتری وجود خواهد داشت. گاهی این خصیصه‌ها را می‌توان به صورت «عوامل پنهان» در نظر گرفت که متغیرهای قابل مشاهده از آن‌ها نشأت می‌گیرند.

Hidden or latent factor

- هنگامی که تعداد خصیصه‌ها(بدون از دست دادن اطلاعات) کم‌تر باشد، «ساختار داده‌ها» بهتر درک می‌شود. داده‌های پرت و غیرمعمول بهتر تشخیص داده می‌شود؛ قابلیت نمایش بهتری دارند.

انتخاب –استخراج خصیصه

Feature Selection vs Extraction

- انتخاب خصیصه:
  - K خصیصه‌ی مهم‌تر $(k<d)$ انتخاب می‌شود.
  - الگوریتم‌های انتخاب زیرمجموعه
- استخراج خصیصه:
  - K خصیصه‌ی جدید، استخراج می‌شود.
  - نگاشت از فضای n-بعدی به فضای k–بعدی
  - روش‌های استخراج خصیصه نیز از دیدگاه‌های مختلفی قابل طبقه‌بندی هستند، روش‌های خطی در برابر روش‌های غیرخطی و یا روش‌های بی‌نظارت در برابر روش‌های بانظارت

انتخاب زیرمجموعه

- در انتخاب زیرمجموعه، هدف انتخاب بهترین زیرمجموعه، زیرمجموعه‌ای با کم‌ترین ابعاد و درست‌ترین نتیجه، می‌باشد.
  - 2^d زیرمجموعه، در یک مجموعه‌ی d-عضوی وجود دارد، بررسی تمام حالات به جز زمانی که d کوچک باشد، امکان‌پذیر نیست.

Forward search

- جستجوی رو به جلو:
  - در گام نخست، مجموعه‌ی خصیصه‌ها، F در حالت اولیه Ø در نظر گرفته می‌شود.
  - در هر گام بهترین خصیصه به مجموعه‌ی خصیصه‌ها افزوده می‌شود. (میزان خطای (E(F)) کم‌تر)
  - برای بررسی خطا باید از داده‌های validation استفاده کرد.

$$j = \arg\min_i E(F \cup x_i)$$

$$\text{Add } x_j \text{ to } F \text{ if } E(F \cup x_j) < E(F)$$

انتخاب زیرمجموعه (ادامه...)

Backward search

- جستجوی رو به عقب:
  - در گام نخست، مجموعه‌ی خصیصه‌ها، F در حالت اولیه تمامی خصیصه‌ها در نظر گرفته می‌شود.
  - در هر گام بدترین خصیصه از مجموعه‌ی خصیصه‌ها حذف می‌شود.
- هنگامی که تعداد خصیصه‌ها زیاد است، روش جستجوی رو به جلو ترجیح داده می‌شود.
- انتخاب زیرمجموعه به صورت بانظارت است.
- در کاربردهایی که یک خصیصه به تنهایی اطلاعات مفیدی ندارد، انتخاب خصیصه مفید نیست.(مانند تشخیص چهره)

Iris data: Single feature

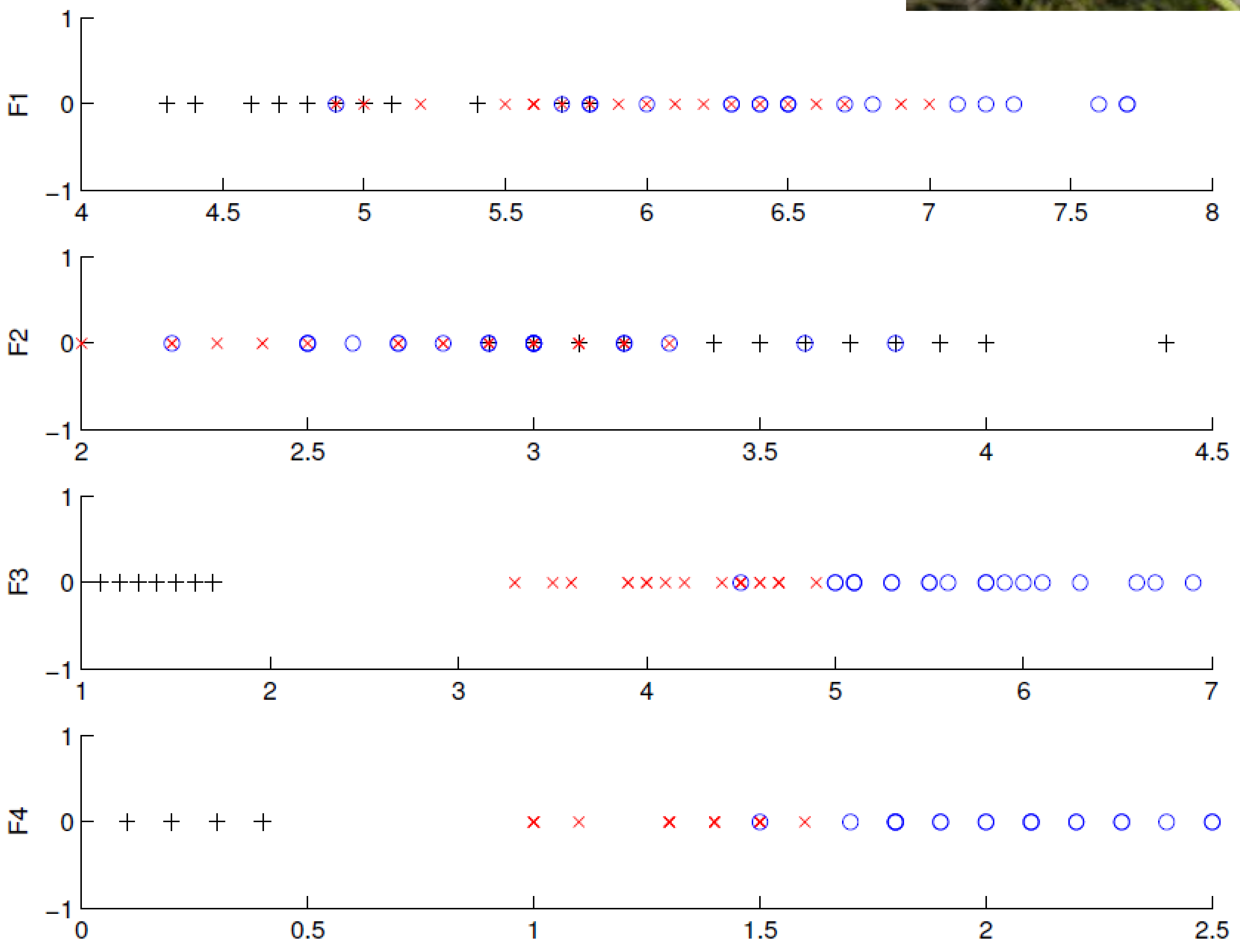
مثال

0.76

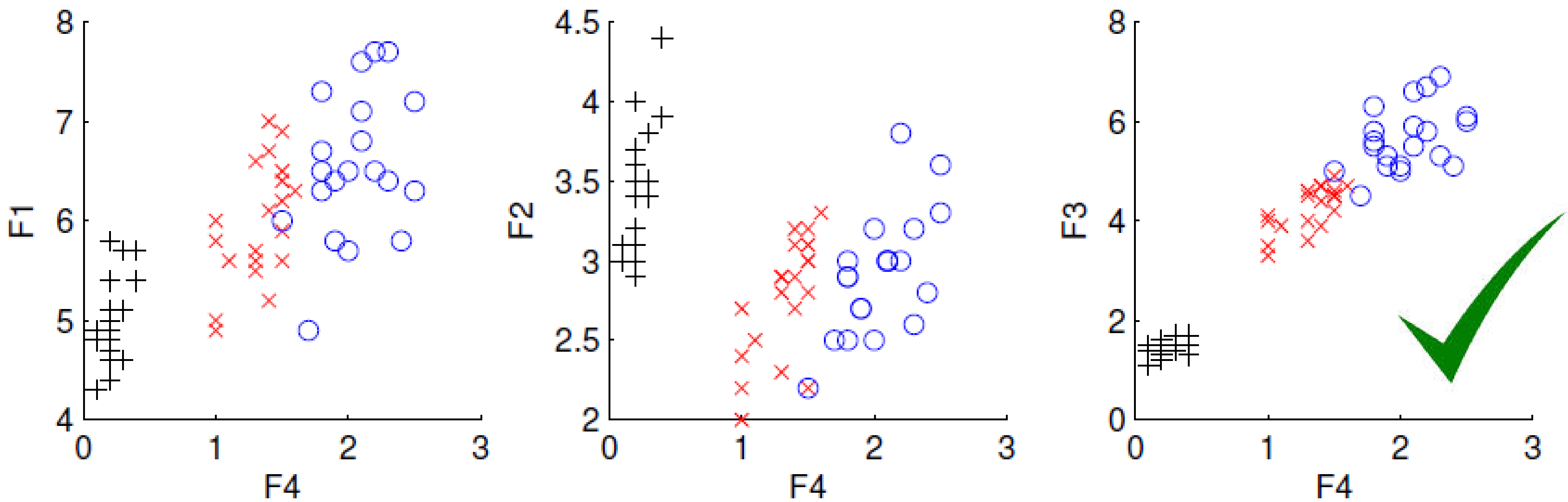
0.57

0.92

0.94



Iris data: Add one more feature to F4



0.87 **0.94** **0.96**

در صورت اضافه کردن خصیصه‌ی بعدی نتایج افت می‌کند!

در بسیاری موارد انتخاب خصیصه‌ها به نوع دسته‌بند بستگی دارد.

در صورت کوچک بودن دسته‌داده، خصیصه‌ی انتخاب شده، می‌تواند به نحوه‌ی تقسیم پایگاه به دو دسته‌ی training و validation مربوط باشد.

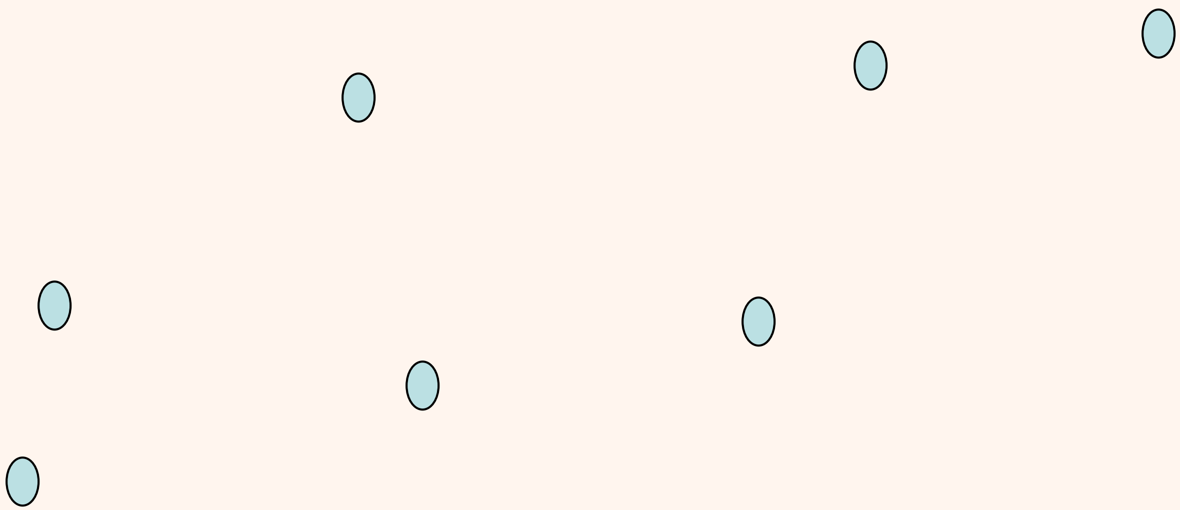
تحلیل مؤلفه‌های اصلی

- هدف نگاشت داده‌ی d-بعدی به فضای k-بعدی است (k<d)، به گونه‌ای که کم‌ترین میزان اتلاف رخ دهد.
  - نگاشت $\boldsymbol{x}$ در راستای $\boldsymbol{w}$:

$$z = \boldsymbol{w}^T\boldsymbol{x}$$

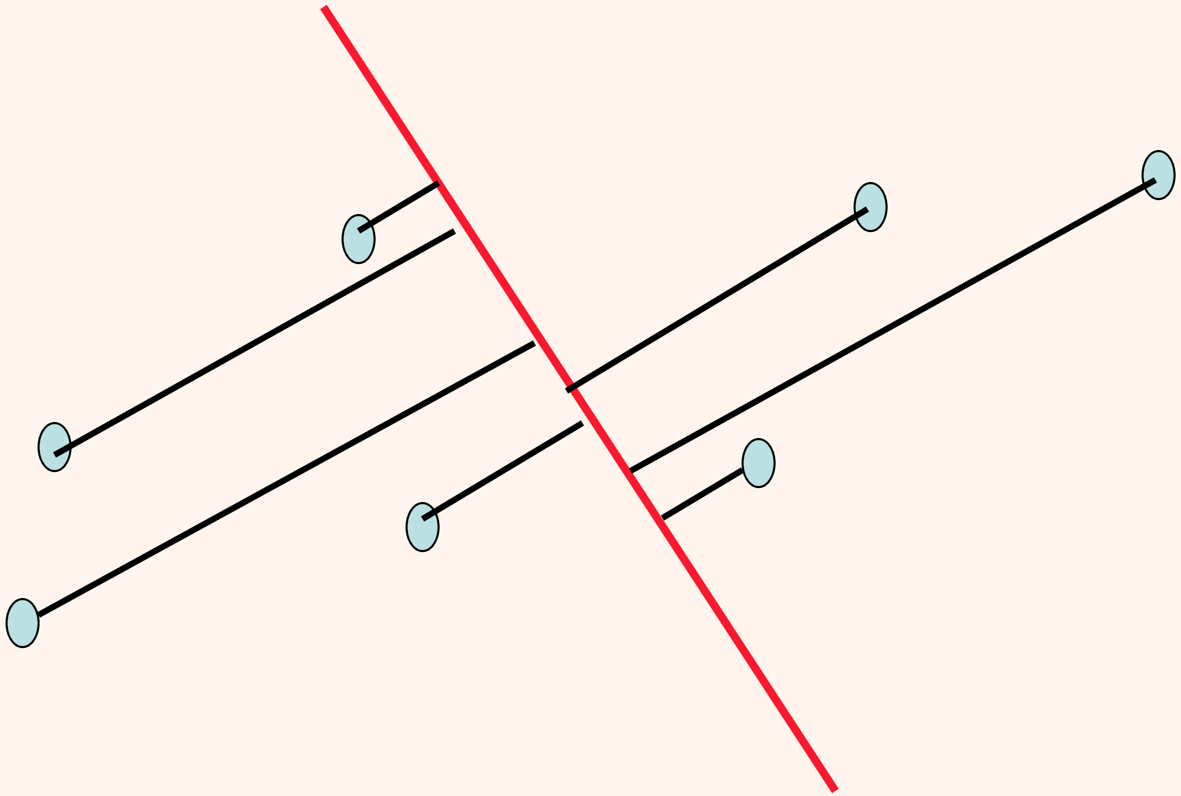
- این راستا به گونه‌ای انتخاب می‌شود که Var(z) ماکزیمم شود، راستایی که داده در امتداد آن بیشترین تغییرات را داشته باشد.
  - این مسأله باعث می‌شود، تفاوت نمونه‌ها آشکارتر شود.
- این شیوه‌ی کاهش بعد به صورت «بی‌نظارت» است.

تحلیل مؤلفه‌های اصلی (ادامه...)



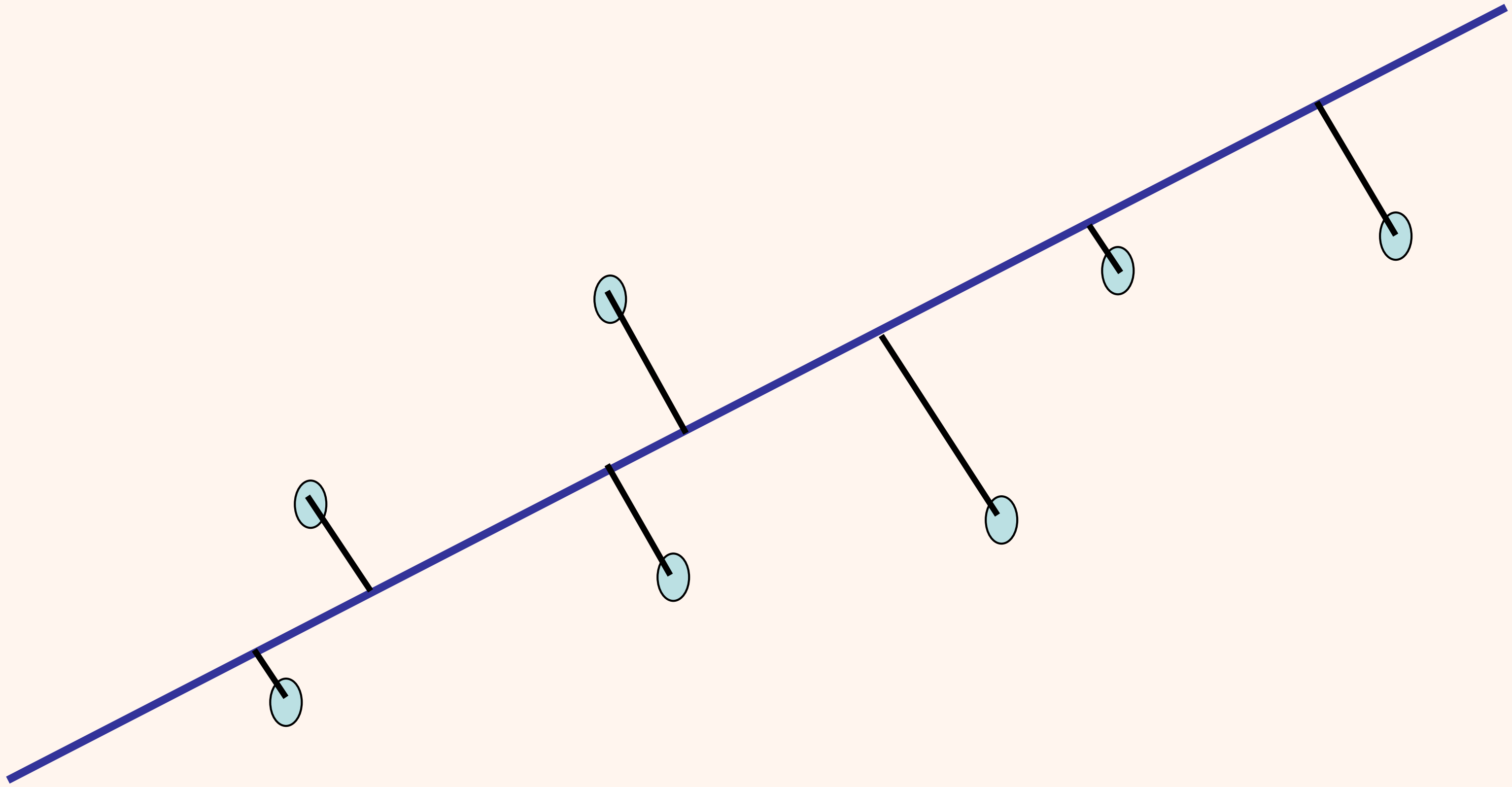
http://www.public.asu.edu/~jye02/

تحلیل مؤلفه‌های اصلی(ادامه...)



http://www.public.asu.edu/~jye02/

تحلیل مؤلفه‌های اصلی(ادامه...)



http://www.public.asu.edu/~jye02/

تحلیل مؤلفه‌های اصلی(ادامه...)

- در راستای $\boldsymbol{w}$، پراکندگی داده ماکزیمم می‌شود:

$$\mathrm{Var}(z) = \mathrm{Var}(\boldsymbol{w}^T\boldsymbol{x}) = \mathrm{E}[(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})^2]$$

$$= \mathrm{E}[(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})]$$

$$= \mathrm{E}[\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{w}]$$

$$= \boldsymbol{w}^T \mathrm{E}[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T]\boldsymbol{w} = \boldsymbol{w}^T \Sigma \boldsymbol{w}$$

where $\mathrm{Cov}(\boldsymbol{x}) = \Sigma$

- در این حالت تنها راستا است که اهمیت دارد، در نتیجه برای یافتن پاسخ یکتا، باید شرط زیر نیز برقرار باشد:

$$\|\boldsymbol{w}\|=1$$

تحلیل مؤلفه‌های اصلی (ادامه...)

- در نتیجه برای اولین مؤلفه‌ی اساسی رابطه‌ی زیر به دست می‌آید:

$$\max_{\mathbf{w}_1} \mathbf{w}_1^T \Sigma \mathbf{w}_1 - \alpha\left(\mathbf{w}_1^T \mathbf{w}_1 - 1\right)$$

- با مشتق گرفتن نسبت به $\mathbf{w}_1$ و برابر صفر قرار دادن آن

$$2\Sigma \mathbf{w}_1 - 2\alpha \mathbf{w}_1 = 0$$

- در نتیجه

$$\Sigma \mathbf{w}_1 = \alpha \mathbf{w}_1$$

در نتیجه w_1 یکی از بردارهای ویژه‌ی ماتریس Σ می‌باشد

- از طرفی $\mathbf{w}_1^T \Sigma \mathbf{w}_1 = \alpha$ ، در واقع واریانس در راستای $\mathbf{w}_1$ برابر مقدار ویژه‌ی متناظر با آن است.

اولین **مؤلفه‌ی اصلی**، برابر بردار ویژه‌ی ماتریس کواریانس با بیشترین مقدار ویژه است.

- برای یافتن دومین مؤلفه‌ی اصلی، علاوه بر شرایط پیش باید بر راستای اولین مؤلفه‌ی اساسی هم عمود باشد، در این حالت داده‌های نگاشت شده «ناهمبسته» (uncorrelated) خواهند بود.

دومین مؤلفه‌ی اصلی، برابر بردار ویژه‌ی ماتریس کواریانس با بیشترین مقدار ویژه در رده‌ی دوم است، به همین ترتیب سایر مقادیر ویژه به دست می‌آیند.

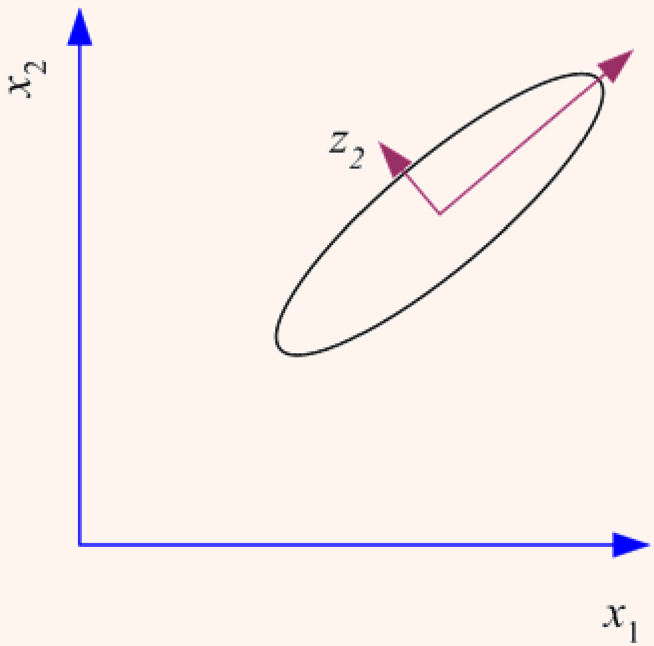
تحلیل مؤلفه‌های اصلی(ادامه...)

- در صورتی که ماتریس متقارن باشد، بردارهای ویژه‌ی آن متعامد هستند.
- در صورتی که ماتریس positive definite باشند، مقادیر ویژه همگی مثبت خواهند بود.
- در صورتی که ماتریس singular باشد، به اندازه‌ی rank ماتریس مقادیر ویژه‌ی غیرصفر خواهیم داشت.

تحلیل مؤلفه‌های اصلی (ادامه...)

$$\boldsymbol{z} = \mathbf{W}^T(\boldsymbol{x} - \boldsymbol{m})$$

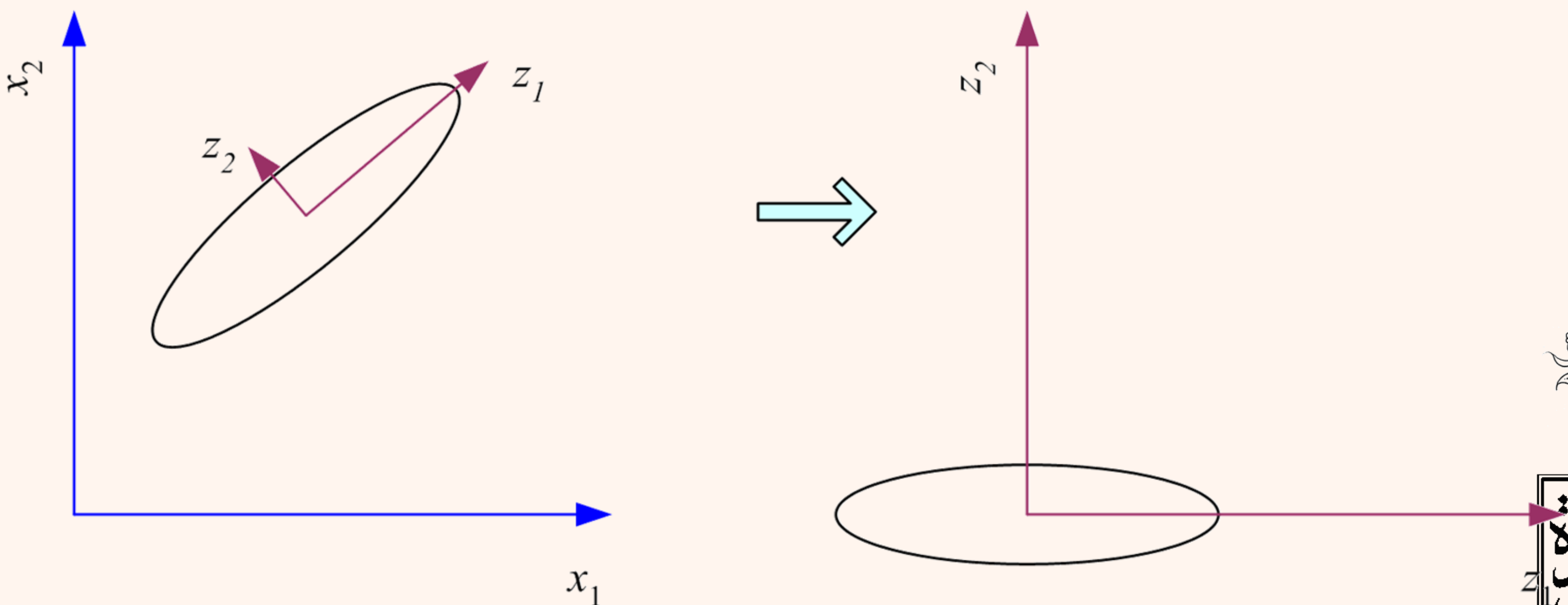
- ستون‌های W، بردارهای ویژه‌ی ماتریس کواریانس هستند.



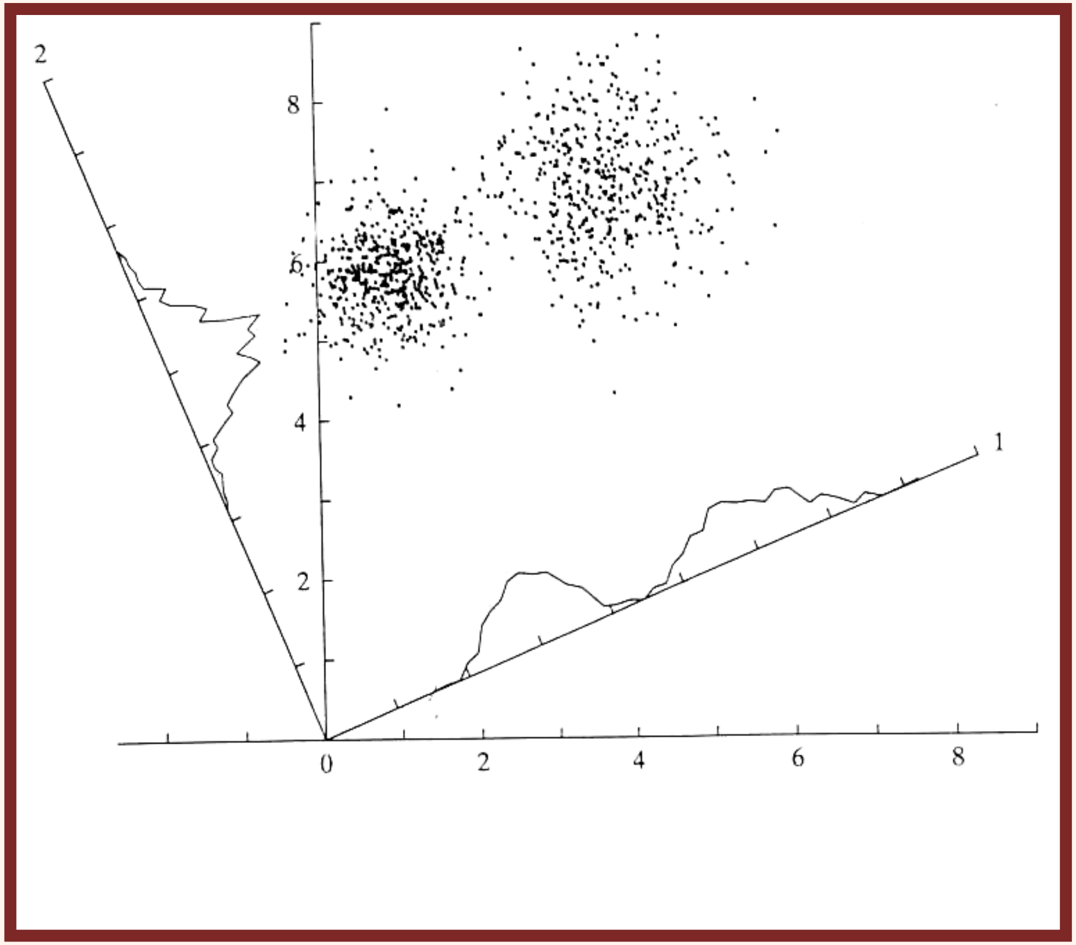
تحلیل مؤلفه‌های اصلی(ادامه...)

$$\boldsymbol{z} = \mathbf{W}^T(\boldsymbol{x} - \boldsymbol{m})$$

- ستون‌های **W**، بردارهای ویژه‌ی ماتریس کواریانس هستند.



تحلیل مؤلفه‌های اصلی(ادامه...)



Haykin, S. <u>Neural Networks: A Comprehensive Foundation,</u>

کاهش بعد

- در صورتی که |S| کوچک باشد، می‌توان نتیجه گرفت برخی مقادیر ویژه، کوچک هستند. در نتیجه داده‌ها در راستای بردار ویژه‌ی متناظر با آن واریانس کمی دارد و قابل صرفنظر کردن است.
  - در این حالت K مؤلفه‌ی پرارزش انتخاب می‌شوند، با فرض آن که مقادیر ویژه به صورت صعودی مرتب شده باشند.

Proportion of Variance (PoV)

$$\frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_k+\cdots+\lambda_d}$$

PoV>0.9

  - در کاربردهای نظیر پردازش تصویر یا صوت، معمولاً کاهش ابعاد قابل توجه است.

کاهش بعد

- در صورتی برخی مقادیر ویژه، کوچک باشند. داده‌ها در راستای بردار ویژه‌ی متناظر با آن واریانس کمی دارد و قابل صرفنظر کردن است.
  - در این حالت K مؤلفه‌ی پرارزش انتخاب می‌شوند، با فرض آن که مقادیر ویژه به صورت صعودی مرتب شده باشند.

Proportion of Variance (PoV)

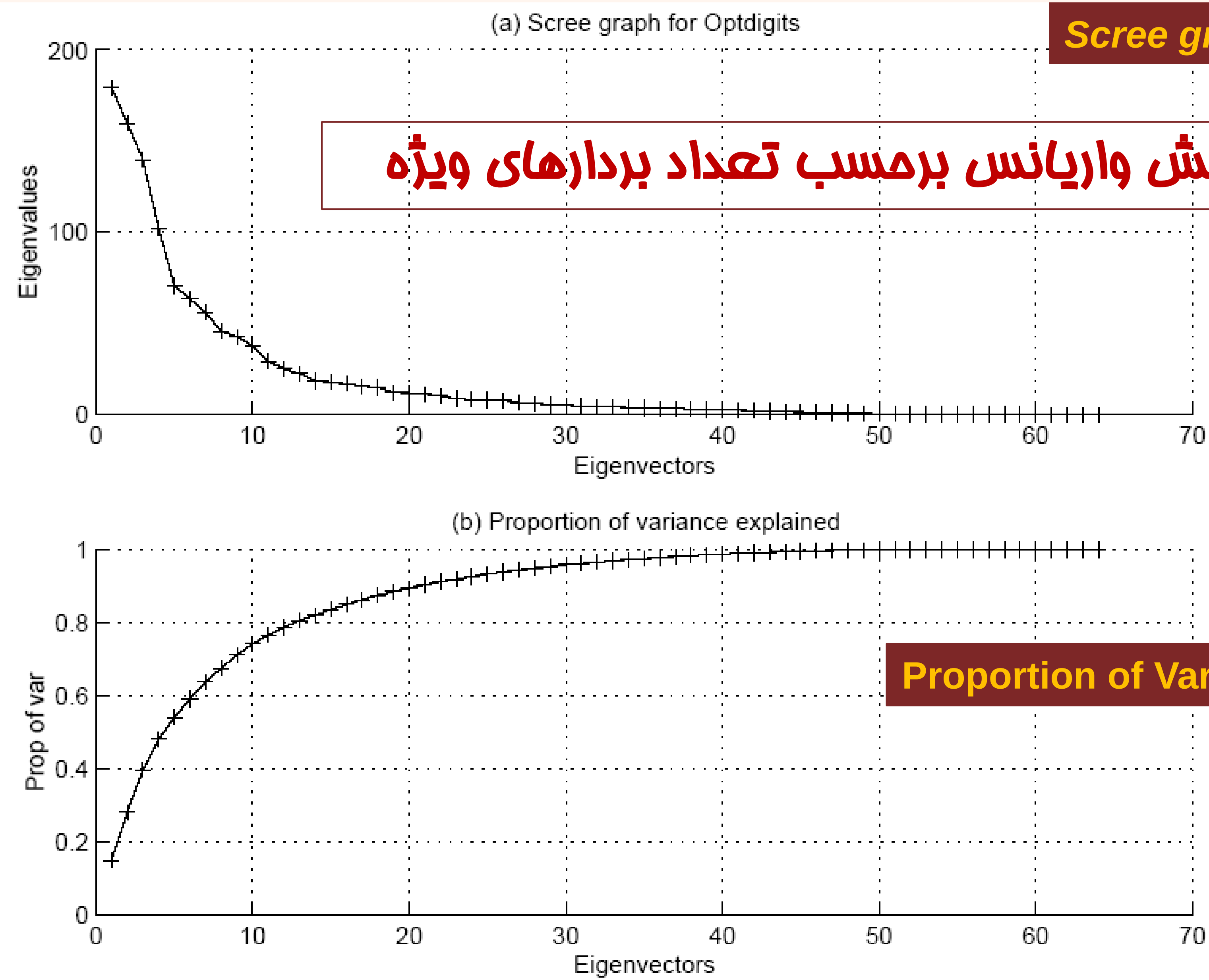
$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_k + \cdots + \lambda_d}$$

PoV>0.9

  - در کاربردهای نظیر پردازش تصویر یا صوت، معمولاً کاهش ابعاد قابل توجه است.

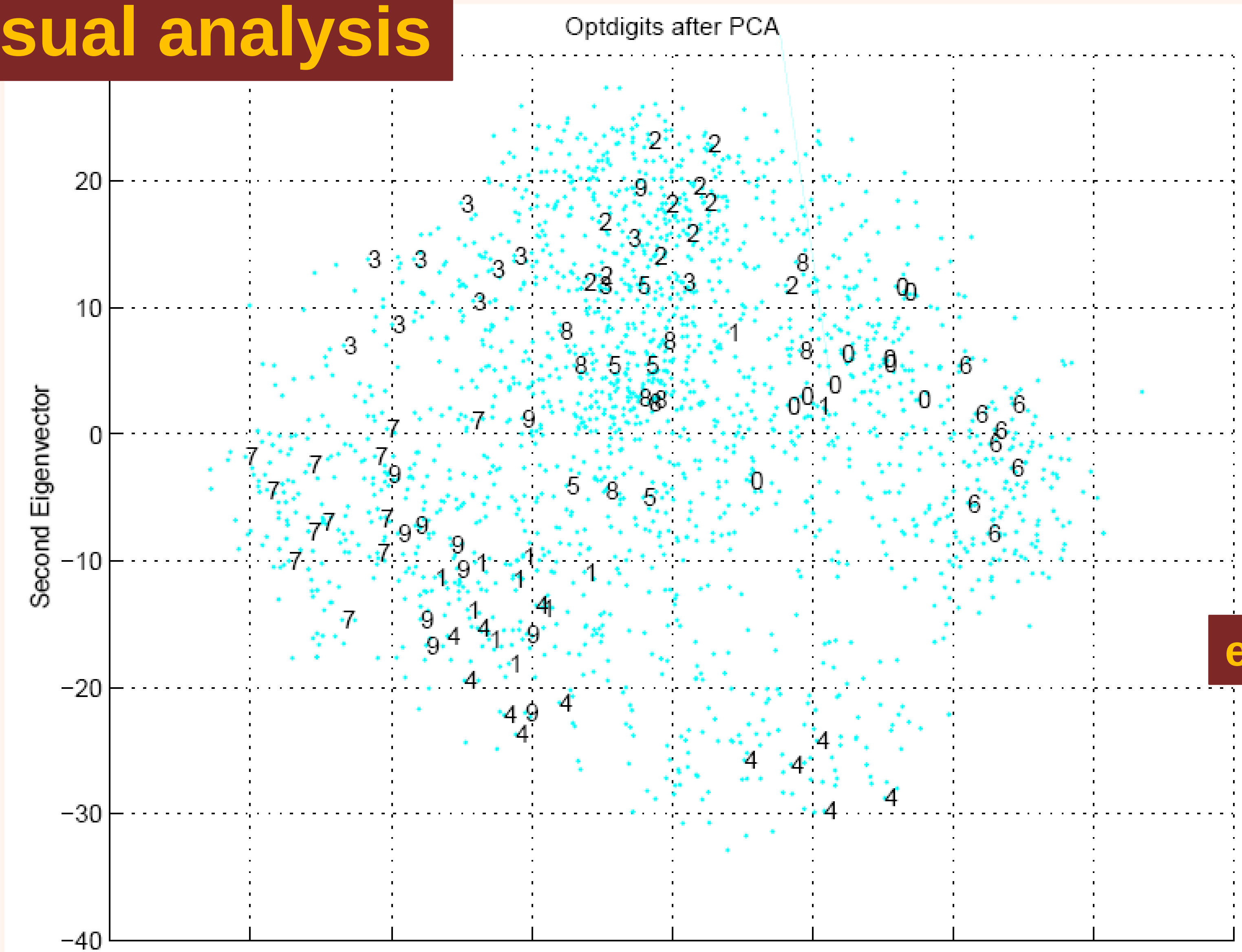
Scree graph

نمایش واریانس برحسب تعداد بردارهای ویژه



Proportion of Variance (PoV)

Visual analysis



eigendigits

در صورتی که سه بعد نخست، حاوی بخش عمده‌ای از واریانس باشند،
می‌توان داده‌ها از آن‌ها برای «بررسی دیداری» بهره برد.

چند نکته

- علاوه بر در نظر گرفتن PoV، می‌توان بردارهای ویژه‌ای که مقدار ویژه‌ی متناظر آن از یک حدآستانه (به عنوان مثال میانگین واریانس) کم‌تر است را حذف نمود.
- در صورتی که واریانس در ابعاد مختلف تغییرات زیادی داشته باشند، بیش از مقدار همبستگی بر روی مؤلفه‌ی اصلی اثرگذار خواهد بود.
  - در این شرایط می‌توان از بردارها و مقادیر ویژه‌ی «ماتریس همبستگی» (R) استفاده کرد یا این که داده‌ها را به گونه‌ای نرمال کرد که همگی واریانس یکسان داشته باشند.

چند نکته

- PCA، نسبت به نویز به شدت حساس است.
  - یک روش ساده حذف داده‌های پرت با استفاده از فاصله‌ی Mahalanobis پیش از محاسبه‌ی ماتریس کوواریانس است.
- در میان تمام بردارهای متعامد، PCA کم‌ترین میزان خطا را دارد.

Reconstruction error

$$\sum_t \left\| \hat{\mathbf{x}}^t - \mathbf{x}^t \right\|$$

- Karhunen-Loève expansion و Hotelling transform نام‌های دیگری هستند برای مفاهیم مشابه به کار می‌روند.

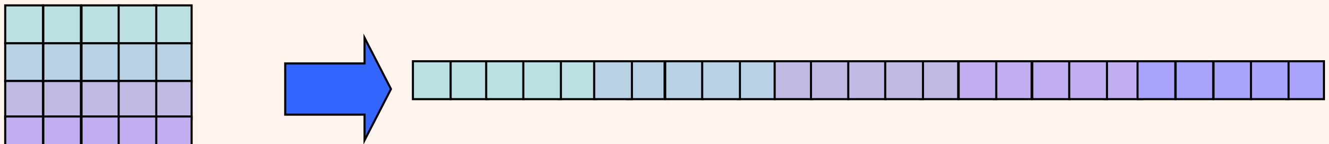
کاربرد PCA در شناسایی چهره

پایگاه داده‌ی ORL

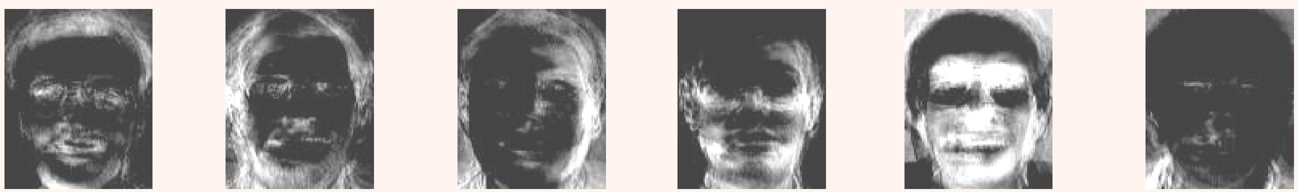
میانگین چهره‌ها

M. Turk, A. Pentland, "Eigenfaces for Recognition", Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71-86, 1991.

کاربرد PCA در شناسایی چهره(ادامه...)



Eigenfaces

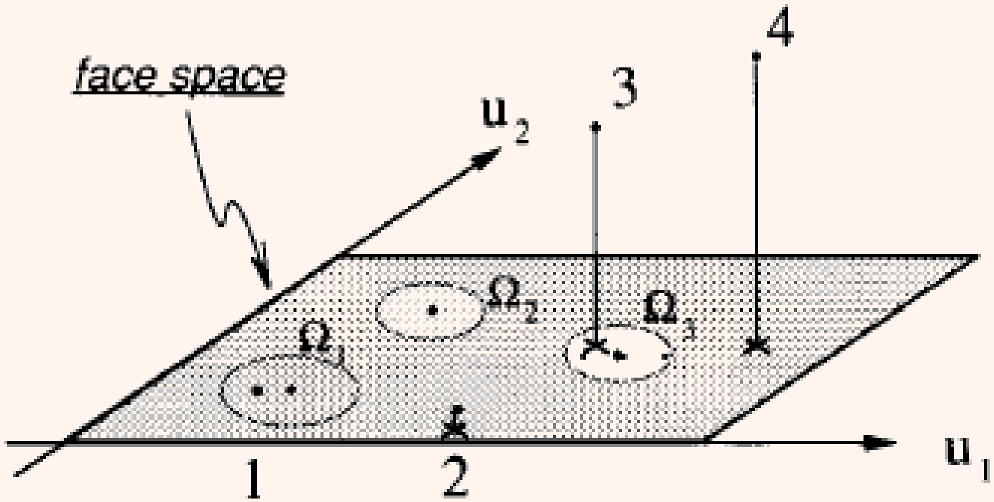


۱ ۲ ۳ ۴ ۵ ۶

۳۵ ۳۶ ۳۷ ۳۸ ۳۹ ۴۰

کاربرد PCA در شناسایی چهره(ادامه...)

تشخیص چهره

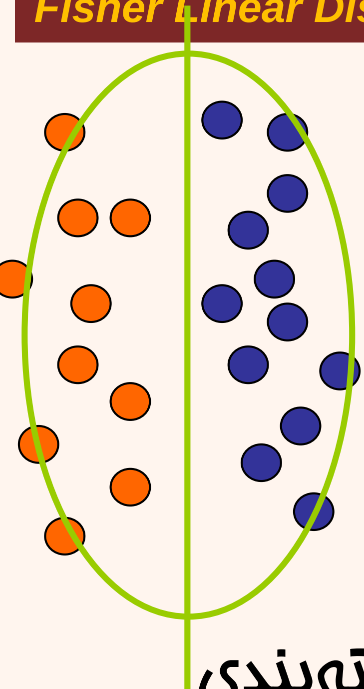


$$\left\| \hat{\mathbf{x}} - \mathbf{x} \right\|$$

تحلیل تفکیک خطی

Fisher Linear Discriminant Analysis

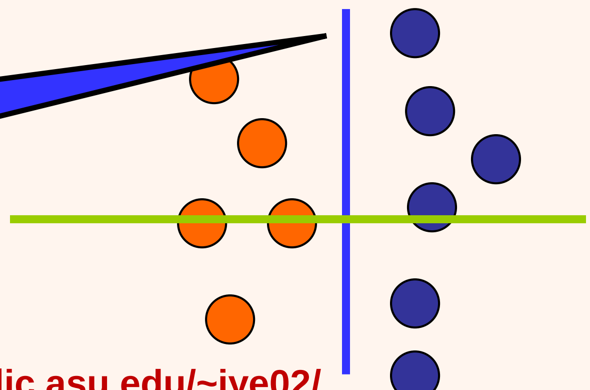
دسته‌بندی دو کلاسه



Ronald Aylmer Fisher

- آیا PCA برای دسته‌بندی مناسب است؟
  - راستای نگاشت بر اساس واریانس، انتخاب می‌شود.
  - در این میان ممکن است اطلاعات دسته‌ها از بین بروند.
- تحلیل تفکیک خطی، «بانظارت» است و برای دسته‌بندی به کار می‌رود.
  - هدف آن کاهش بعد همراه با حفظ اطلاعاتی است که بین دسته‌ها تمایز قائل می‌شود.

در این راستا دو کلاس همپوشانی دارند



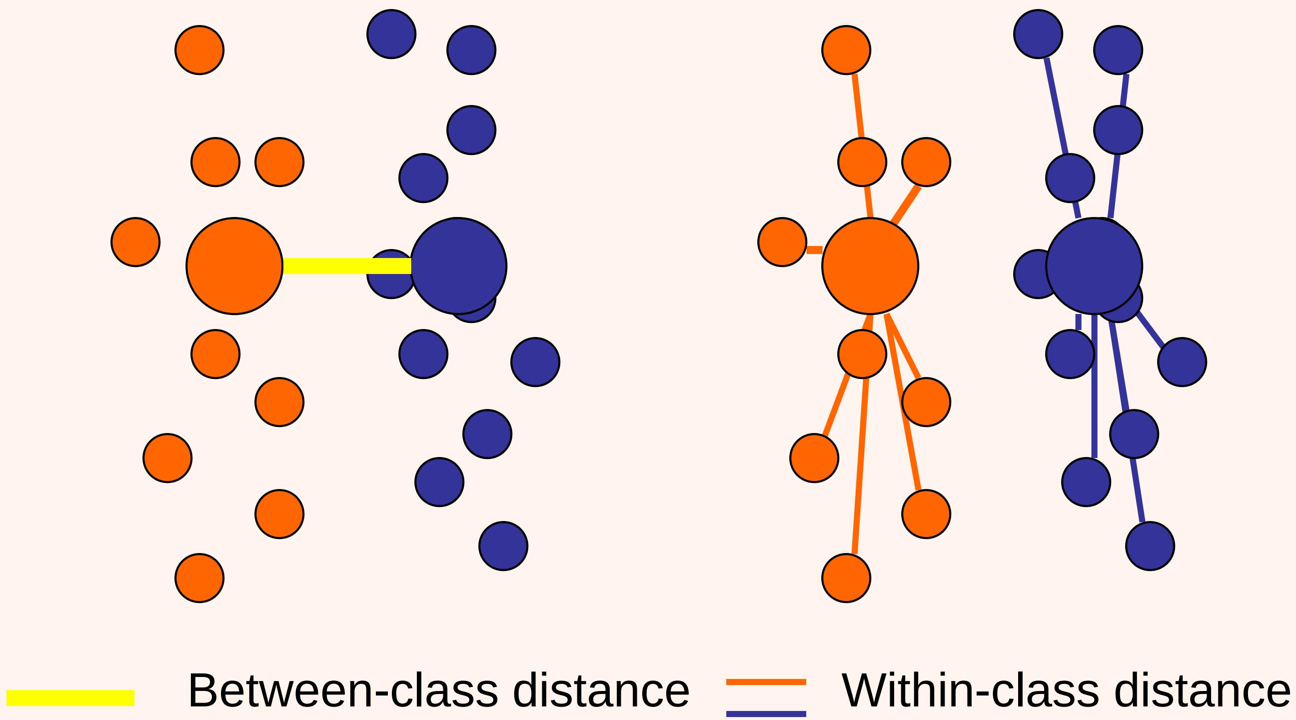
در این راستا دو کلاس بدون خطا طبقه‌بندی می‌شوند

http://www.public.asu.edu/~jye02/

کاهش ابعاد برای دسته‌بندی

دسته‌بندی دو کلاسه

برای انتخاب راستای مناسب برای نگاشت، باید اطلاعات دسته‌ها نیز در نظر گرفته شود.

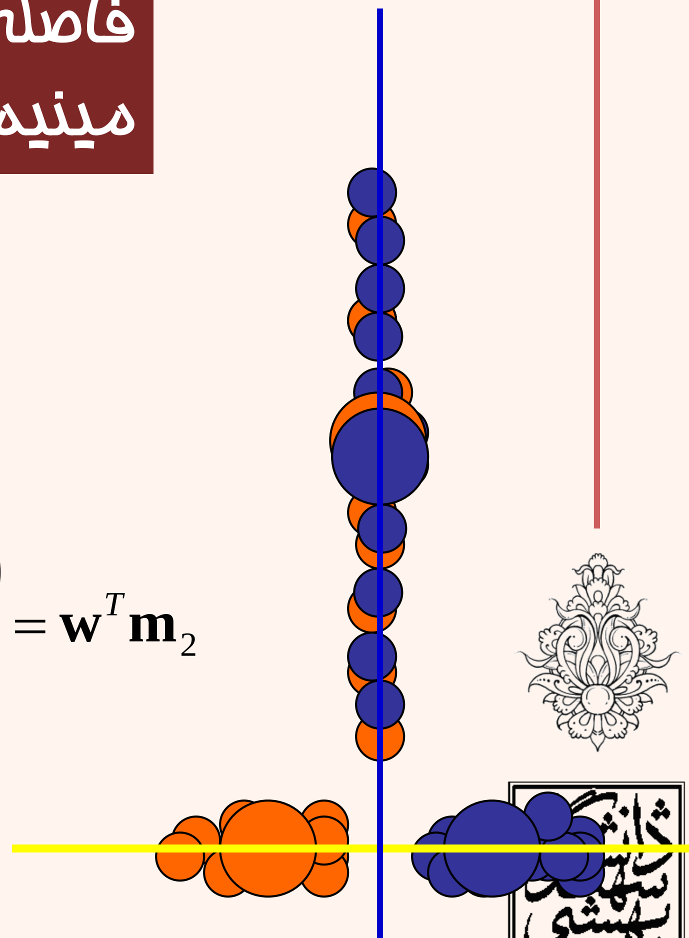


http://www.public.asu.edu/~jye02/

تحلیل تفکیک خطی (ادامه...)

در LDA، نگاشت به گونه‌ای انجام می‌شود که فاصله‌ی بین دو کلاس حداکثر شده و فاصله‌ی نمونه‌های متعلق به یک کلاس مینیمم گردد.

$$J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$$



$$m_1 = \frac{\sum_t \mathbf{w}^T \mathbf{x}^t r^t}{\sum_t r^t} = \mathbf{w}^T \mathbf{m}_1, \quad m_2 = \frac{\sum_t \mathbf{w}^T \mathbf{x}^t (1 - r^t)}{\sum_t (1 - r^t)} = \mathbf{w}^T \mathbf{m}_2$$

$$s_1^2 = \sum_t \left(\mathbf{w}^T \mathbf{x}^t - m_1\right)^2 r^t$$

$$s_2^2 = \sum_t \left(\mathbf{w}^T \mathbf{x}^t - m_2\right)^2 (1 - r^t)$$

$$\boxed{\mathbf{S}_W = \mathrm{S}_1 + S_2}$$

Total within Class scatter

$$\mathbf{w} = c\mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

دسته‌بندی برای بیش از دو کلاس

- زمانی که تعداد کلاس‌ها بیشتر از دو باشد: برای کاهش ابعاد، ماتریس $\mathbf{W}_{d\times k-1}$ برای نگاشت مورد استفاده قرار می‌گیرد:

$$\mathbf{z} = \mathbf{W}^T\mathbf{x}$$

Within-class scatter

$$\mathbf{S}_W = \sum_{i=1}^{K}\mathbf{S}_i \qquad \mathbf{S}_i = \sum_t r_i^t\left(\mathbf{x}^t - \mathbf{m}_i\right)\left(\mathbf{x}^t - \mathbf{m}_i\right)^T$$

Between-class scatter:

$$\mathbf{S}_B = \sum_{i=1}^{K} N_i\left(\mathbf{m}_i - \mathbf{m}\right)\left(\mathbf{m}_i - \mathbf{m}\right)^T \qquad \mathbf{m} = \frac{1}{K}\sum_{i=1}^{K}\mathbf{m}_i \quad N_i = \sum_t r_i^t$$

- پس از نگاشت، $\mathbf{W}^T\mathbf{S}_B\mathbf{W}$ و $\mathbf{W}^T\mathbf{S}_W\mathbf{W}$ ماتریس‌های پراکندگی داده «بین‌دسته‌ها» و «درون‌دسته‌ها» خواهند بود.

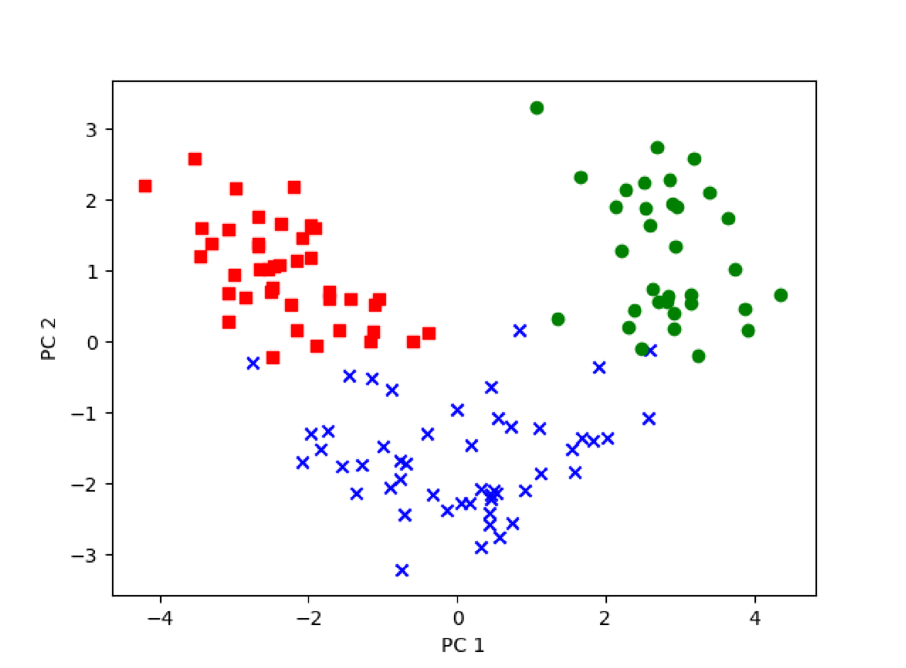
دسته‌بندی برای بیش از دو کلاس

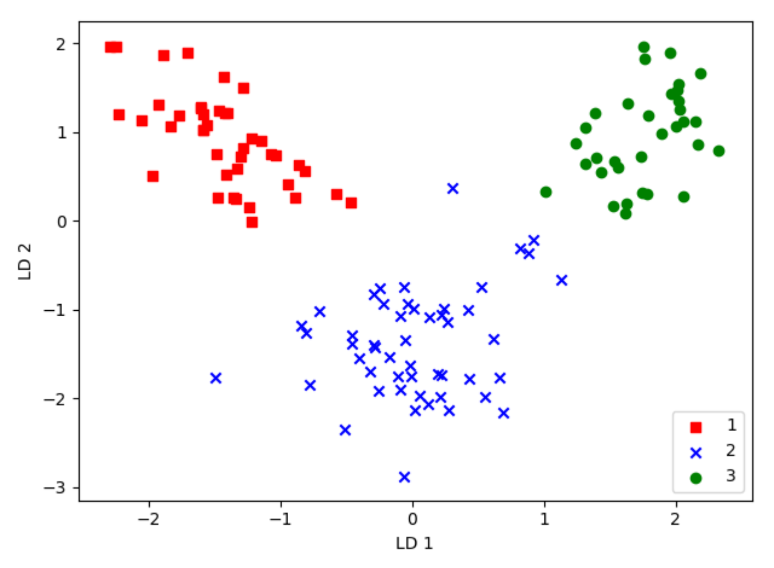
- در نتیجه در صورت بیشینه شدن عبارت زیر، دسته‌بندی به بهترین شکل انجام می‌شود.
  - برای ماتریس کواریانس، دترمینان معیاری است که پراکندگی داده را نشان می‌دهد.

$$J(\mathbf{W}) = \frac{\left|\mathbf{W}^T \mathbf{S}_B \mathbf{W}\right|}{\left|\mathbf{W}^T \mathbf{S}_W \mathbf{W}\right|}$$

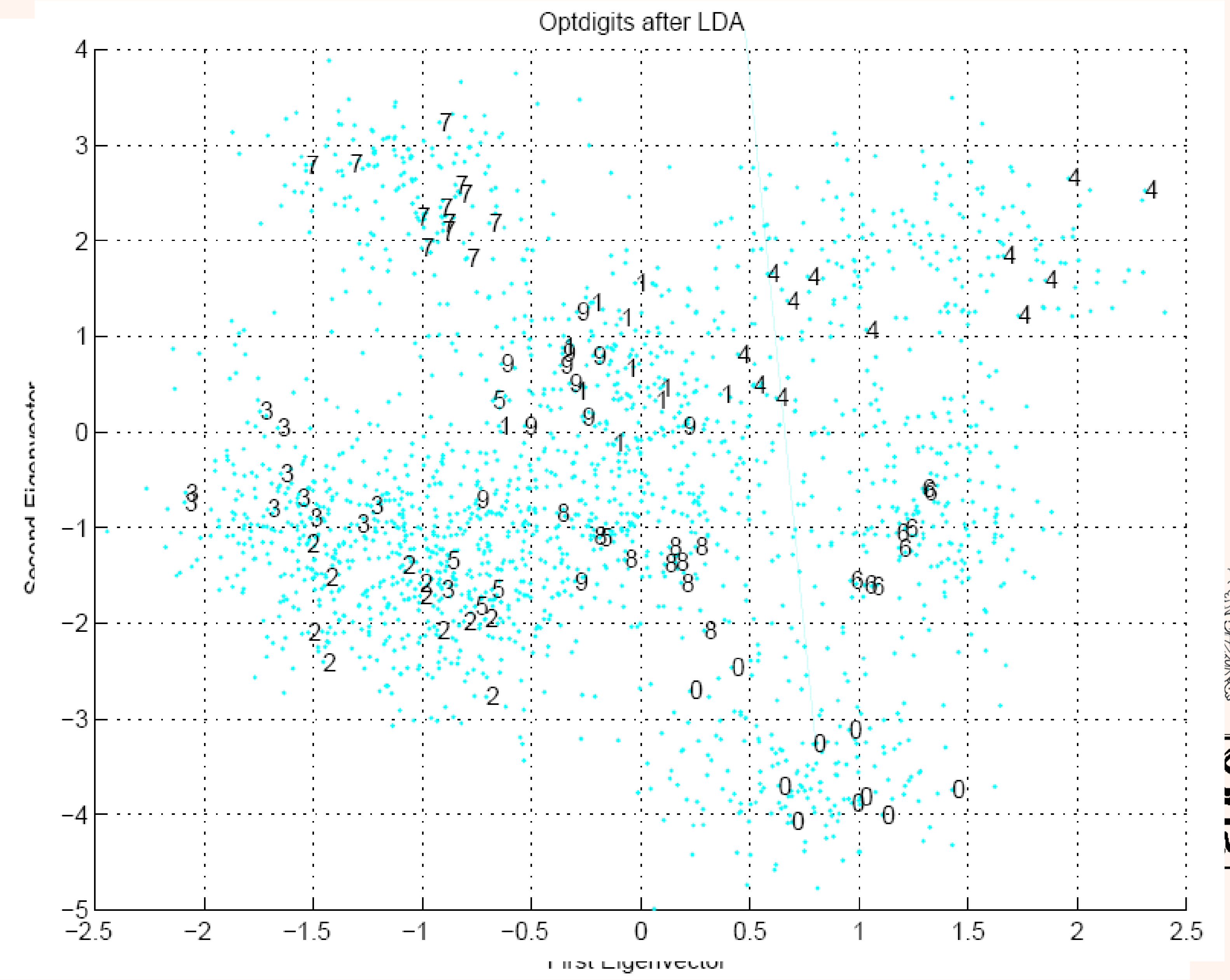
  - در این حالت پاسخ، بردارهای ویژه متناظر با بزرگ‌ترین مقادیر ویژه‌ی ماتریس $\mathbf{S}_W^{-1}\mathbf{S}_B$ خواهد بود.

PCA vs LDA

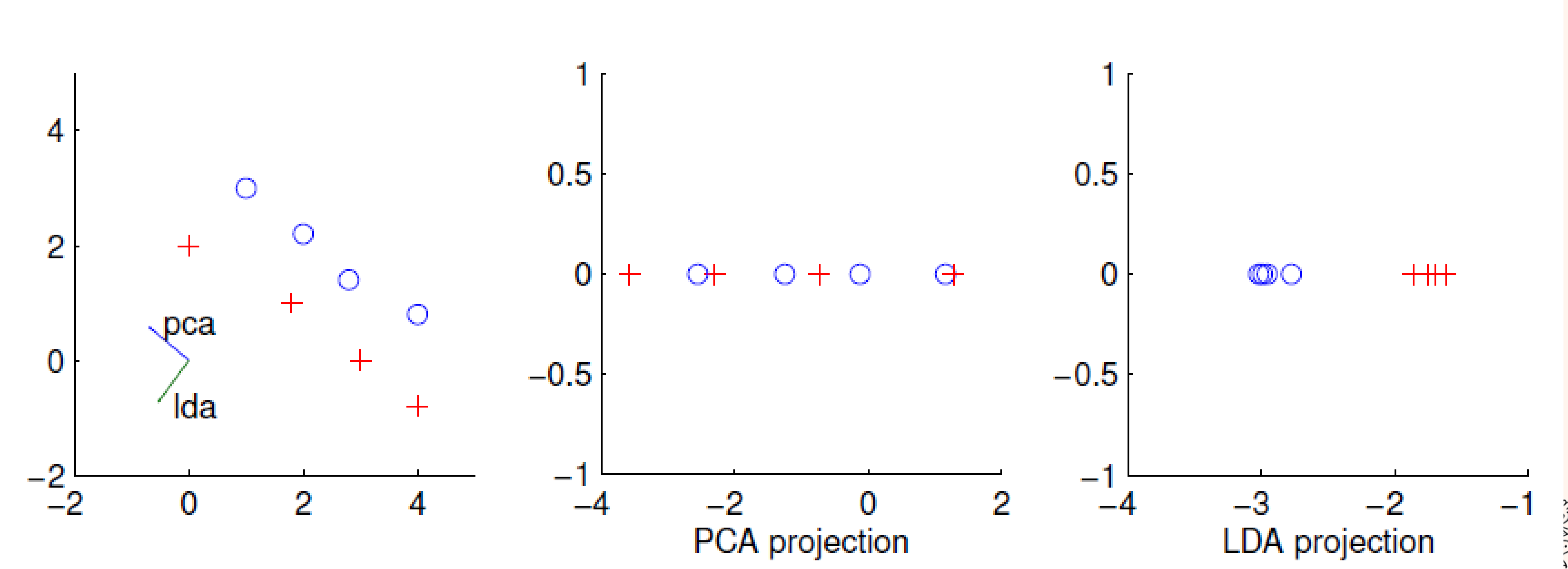




مثال


Optdigits after LDA
Second Eigenvector
First Eigenvector

PCA vs LDA



تجزیه‌ی مقدارهای تکین

- با استفاده از SVD، یک ماتریس به سه ماتریس تجزیه می‌شود:

$$\mathbf{X}_{N\times d} = \mathbf{V}_{N\times N}\mathbf{A}_{N\times d}\mathbf{W}^T_{d\times d}$$

- $\mathbf{V}$ شامل بردارهای ویژه‌ی XX^T می‌باشد، $\mathbf{W}$ شامل بردارهای ویژه‌ی X^TX است و A مقادیر ویژه را در k عنصر قطری خود دارد.

$$\mathbf{XX}^T = \left(\mathbf{VAW}^T\right)\left(\mathbf{VAW}^T\right)^T = \mathbf{VAW}^T\mathbf{WA}^T\mathbf{V}^T = \mathbf{VEV}^T$$

$$\mathbf{X}^T\mathbf{X} = \left(\mathbf{VAW}^T\right)^T\left(\mathbf{VAW}^T\right) = \mathbf{WA}^T\mathbf{V}^T\mathbf{VAW}^T = \mathbf{WDW}^T$$

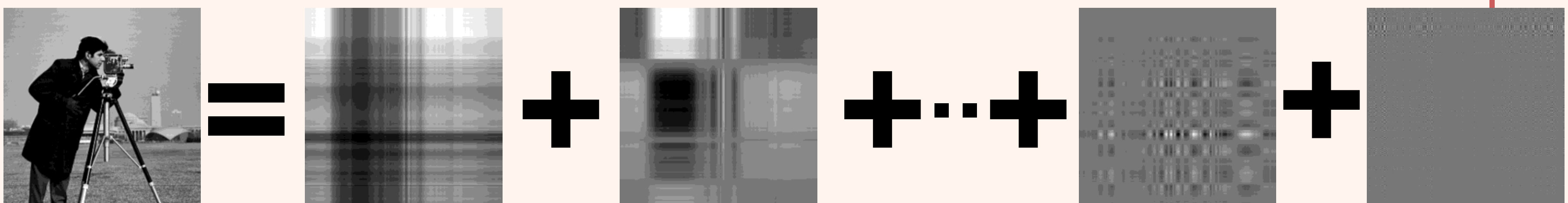
$$\mathbf{E} = \mathbf{AA}^T \qquad \mathbf{D} = \mathbf{A}^T\mathbf{A}$$

تجزیه‌ی مقدارهای تکین

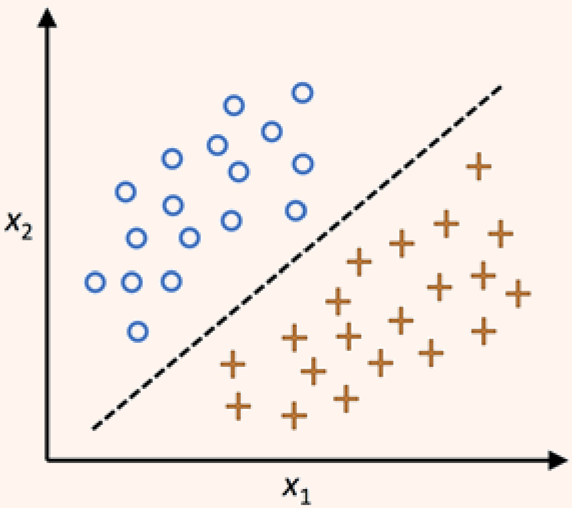
$$\boxed{\mathbf{X}_{N\times d}=\mathbf{V}_{N\times N}\mathbf{A}_{N\times d}\mathbf{W}_{d\times d}^{T}}$$

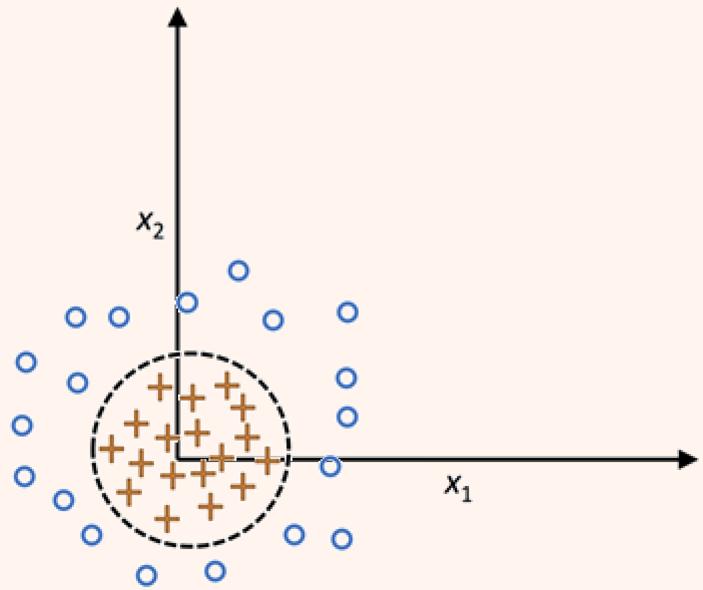
$$\mathbf{X}=a_1\mathbf{v}_1\mathbf{w}_1^T+\ldots+a_k\mathbf{v}_\mathbf{k}\mathbf{w}_k^T$$

$$\mathbf{X}=\sum_{i=1}^{k}a_k\mathbf{v}_k\mathbf{w}_k^T$$



- زمانی که داده‌ها جدایی‌پذیر خطی نیستند، استفاده از روش‌ها خطی برای کاهش بعد موجب از دست رفتن اطلاعات کلاس‌ها می‌شود.





در این حالت نیز می‌توان داده‌ها را به فضایی با ابعاد بالاتر نگاشت کرد به گونه‌ای که در آن فضا به صورت خطی قابل جداسازی باشند و پس از آن از روش‌های سنتی کاهش بعد بهره برد.

KPCA

Kernel principal component analysis

- در این حالت نیز می‌توان از kernel trick استفاده کرد.
  - با توجه به این که در Feature embedding از ماتریس تشابه(ضرب داخلی) نمونه‌های آموزشی استفاده می‌شود با استفاده از توابع هسته‌ی مناسب به راحتی و بدون مواجه شدن با نگاشت به فضایی دیگر، کاهش ابعاد به شیوه‌ی غیرخطی قابل انجام خواهد بود.

$$\phi : \mathbb{R}^d \to \mathbb{R}^k \quad (k \gg d)$$

توابع کرنل ضرب داخلی دو بردار ویژگی در فضایی با ابعاد بالا را بدون نگاشت با آن فضا به دست می‌آورد:

$$k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \phi\left(\mathbf{x}^{(i)}\right)^T \phi\left(\mathbf{x}^{(j)}\right)$$

توابع کرنل متعارف:

Radial Basis function(RBF)

$$\kappa\left(\boldsymbol{x}^{(i)},\boldsymbol{x}^{(j)}\right)=\exp\left(-\gamma\left\|\boldsymbol{x}^{(i)}-\boldsymbol{x}^{(j)}\right\|^{2}\right)$$

$$\kappa\left(\boldsymbol{x}^{(i)},\boldsymbol{x}^{(j)}\right)=\left(\boldsymbol{x}^{(i)T}\boldsymbol{x}^{(j)}+\theta\right)^{p}$$

polynomial kernel

Hyperbolic Tangent Kernel

$$\kappa\left(\boldsymbol{x}^{(i)},\boldsymbol{x}^{(j)}\right)=\tanh\left(\eta\boldsymbol{x}^{(i)T}\boldsymbol{x}^{(j)}+\theta\right)$$

حذف میانگین

$$k_{ij} = \phi_i^T \phi_j$$

- باید توجه داشت که لازم است نمونه‌ها در فضای جدید به گونه‌ای نرمال شود که میانگین آن‌ها صفر باشد:

$$k'_{ij} = \left(\phi_i - \frac{1}{N}\sum_m \phi_m\right)^T \left(\phi_j - \frac{1}{N}\sum_n \phi_n\right)$$

$$k'_{ij} = \phi_i^T \phi_j - \phi_i^T \left[\frac{1}{N}\sum_n \phi_n\right] - \left[\frac{1}{N}\sum_m \phi_m^T\right]\phi_j + \left[\frac{1}{N}\sum_n \phi_n\right]\left[\frac{1}{N}\sum_m \phi_m^T\right]$$

$$k'_{ij} = k_{ij} - \left[\frac{1}{N}\sum_n k_{in}\right] - \left[\frac{1}{N}\sum_m k_{mj}\right] + \frac{1}{N^2}\sum_m \sum_n k_{mn}$$

$$\mathbf{1}_{N[i,j]} = [1/N]_{i,j}$$

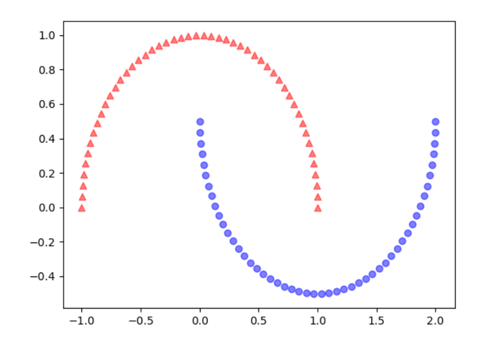
$$\boxed{\mathbf{K}' = \mathbf{K} - \mathbf{1}_N \mathbf{K} - \mathbf{K}\mathbf{1}_N + \mathbf{1}_N \mathbf{K}\mathbf{1}_N}$$

مثال گام به گام

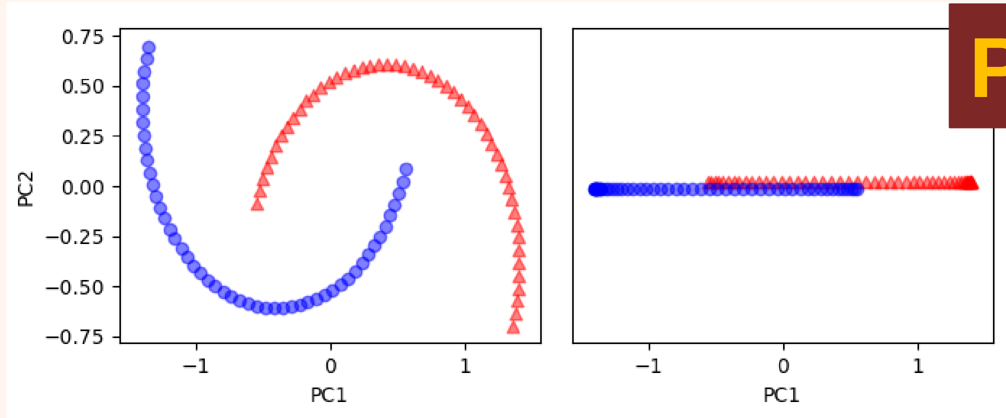
```
def rbf_kernel_pca(X, gamma, n_components):

    sq_dists = pdist(X, 'sqeuclidean')
    mat_sq_dists = squareform(sq_dists)
    K = exp(-gamma * mat_sq_dists)

    N = K.shape[0]
    one_n = np.ones((N, N)) / N
    K = K - one_n.dot(K) - K.dot(one_n) + one_n.dot(K).dot(one_n)


 eigvals, eigvecs = eigh(K)
    eigvals, eigvecs = eigvals[::-1], eigvecs[:, ::-1]



    X_pc = np.column_stack((eigvecs[:, i]
                            for i in range(n_components)))



    return X_pc
```

$$\kappa\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right) = \exp\left(-\gamma \left\| \boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)} \right\|^2\right)$$

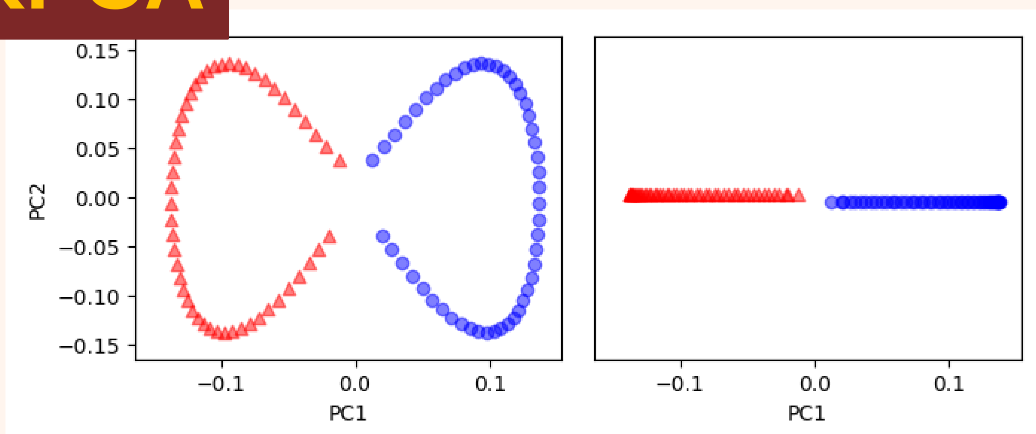
$$\mathbf{K}' = \mathbf{K} - \mathbf{1}_n \mathbf{K} - \mathbf{K}\mathbf{1}_n + \mathbf{1}_n \mathbf{K} \mathbf{1}_n$$



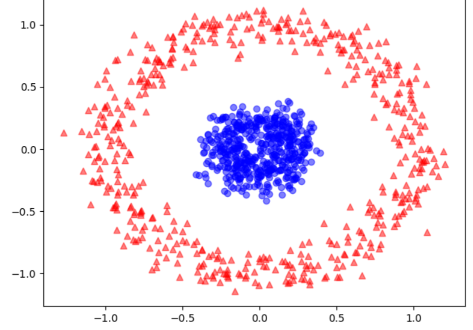
PCA



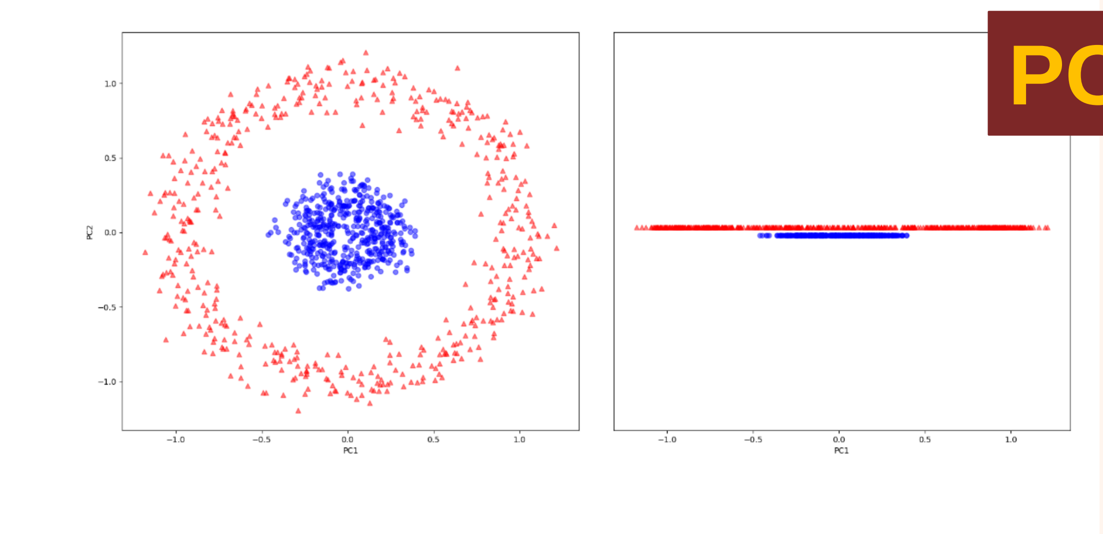
KPCA



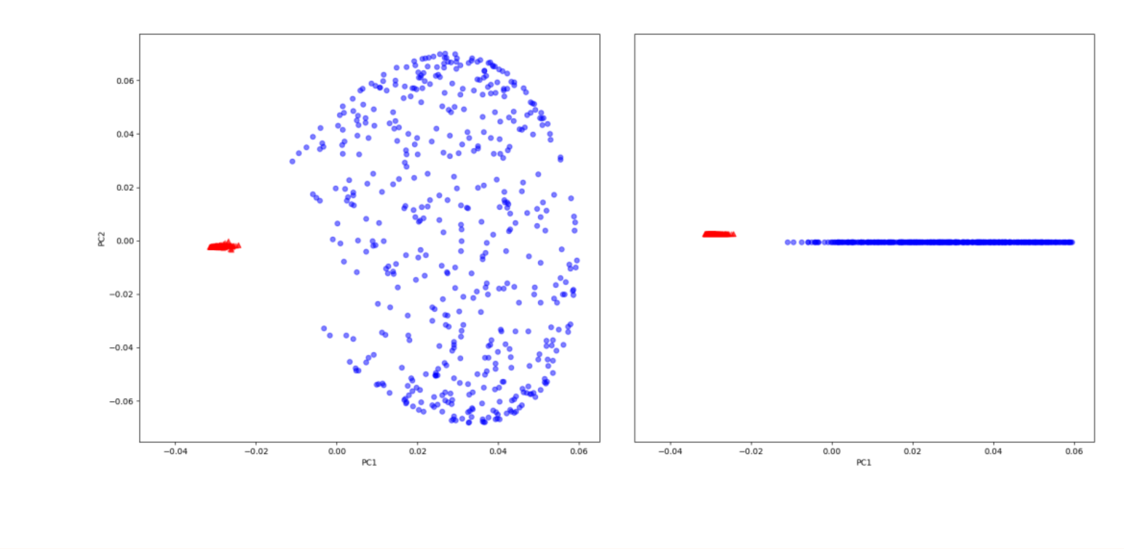

```
from sklearn.decomposition import KernelPCA

X, y = make_moons(n_samples=100, random_state=123)
scikit_kpca = KernelPCA(n_components=2, kernel='rbf', gamma=15)
X_skernpca = scikit_kpca.fit_transform(X)
```


1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

PCA


1.0
0.5
0.0
−0.5
−1.0
PC2
−1.0
−0.5
0.0
0.5
1.0
PC1
−1.0
−0.5
0.0
0.5
1.0
PC1

KPCA


0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
PC2
−0.04
−0.02
0.00
0.02
0.04
0.06
PC1
−0.04
−0.02
0.00
0.02
0.04
0.06
PC1