

مقدمه‌ای بر  
یادگیری ماشین  
(۰۱-۸۰۵-۱۱-۱۳)

بخش نهم  
کاهش ابعاد

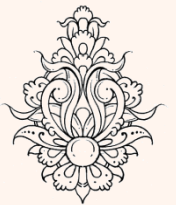
Dimensionality  
Reduction



دانشگاه شهید بهشتی  
پژوهشکده‌ی فضای مجازی  
بهار ۱۳۹۶  
احمد محمودی ازناوه

# فهرست مطالب

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



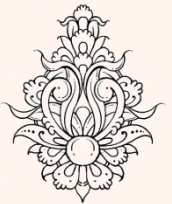
# نکبت ابعاد!

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

## Curse of dimensionality

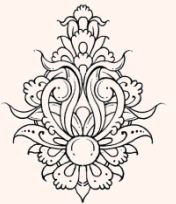
بزرگ شدن ابعاد معادل افزایش حجم و تنگی داده (sparsity) است که این قضیه برای کلیه کاربردهایی که در آن معناداری آماری (statistical significance) اهمیت دارد، می‌تواند مشکل ساز شود.

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

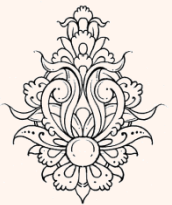


# مزایای کاهش ابعاد (ادامه...)

- «استفراجه 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 (فصیصه)

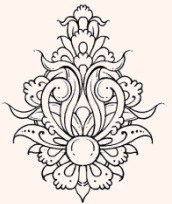
## Feature Selection vs Extraction

- انتخاب فصیصه:

- $K$  فصیصه‌ی مهم‌تر ( $k < d$ ) انتخاب می‌شود.
- الگوریتم‌های انتخاب زیرمجموعه

- استخراج فصیصه:

- $K$  فصیصه‌ی جدید، استخراج می‌شود.
- نگاشت از فضای  $n$ -بعدی به فضای  $k$ -بعدی
- روش‌های استخراج فصیصه 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 مجموعه  $d$ -عضوی وجود دارد، بررسی تمام حالات به جز زمانی که  $d$  کوچک باشد، امکان‌پذیر نیست.

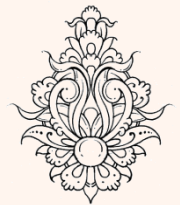
## Forward search

- جستجوی رو به جلو:

- در گام نخست، مجموعه‌ی خصیصه‌ها،  $F$  در حالت اولیه  $\emptyset$  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.

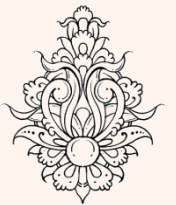


$$j = \operatorname{argmin}_j E(F \cup x_j)$$
$$\text{Add } x_j \text{ to } F \text{ if } E(F \cup x_j) < E(F)$$

# 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)



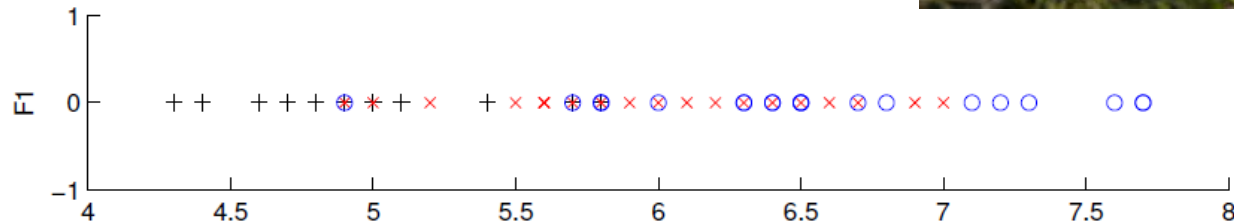


# Iris data: Single feature

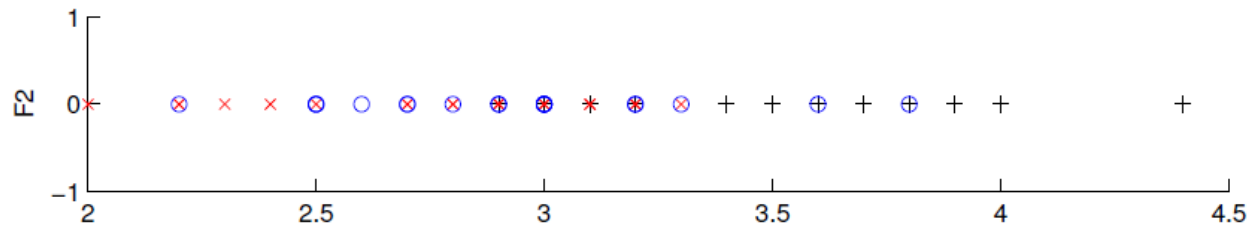
مثال



0.76



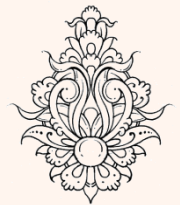
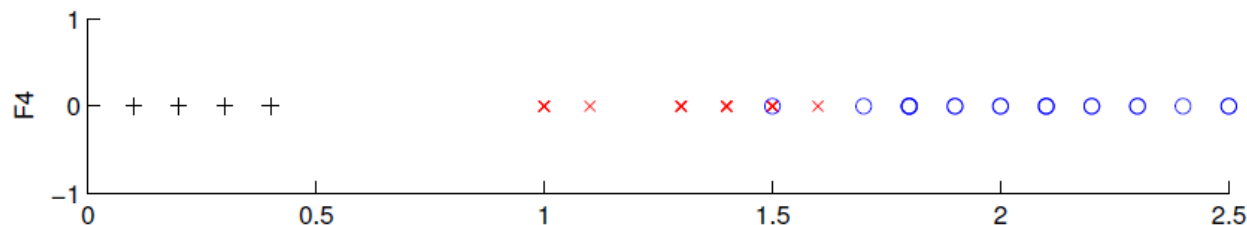
0.57



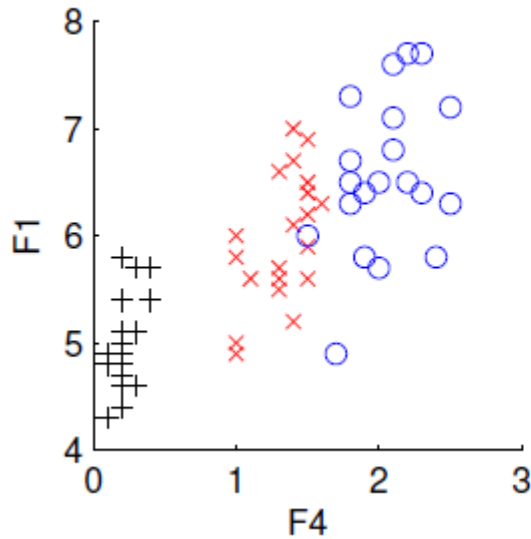
0.92



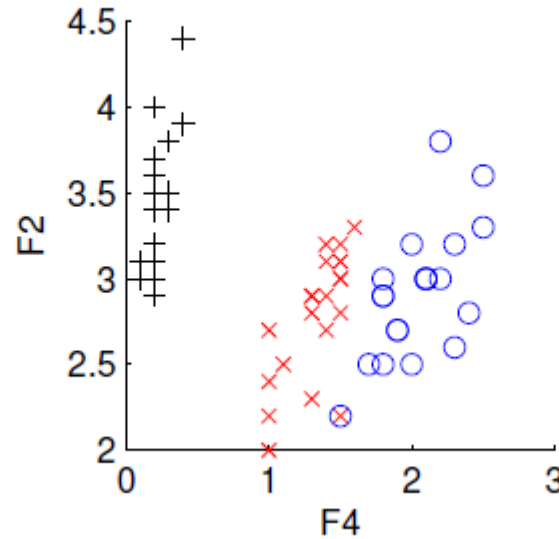
0.94



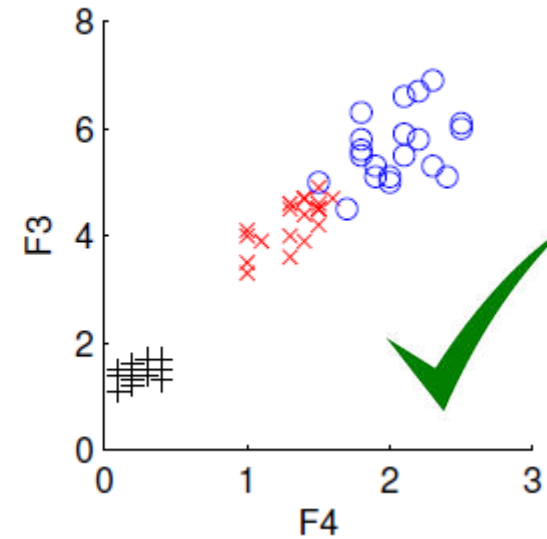
# Iris data: Add one more feature to F4



0.87



0.94

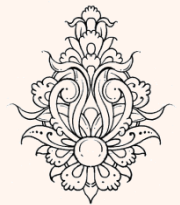


0.96

در صورت اضافه کردن فمحصه‌ی بعدی نتایج افت می‌کند!

در بسیاری موارد انتخاب فمحصه‌ها به نوع دسته‌بند بستگی دارد.

در صورت کوچک بودن پایگاه داده، فمحصه‌ی انتخاب شده، می‌تواند به نمونه‌ی 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.

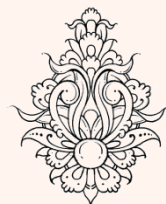
– نگاشت  $x$  در راستای  $w$ :

$$z = w^T x$$

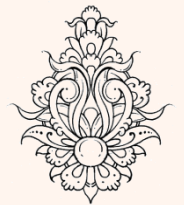
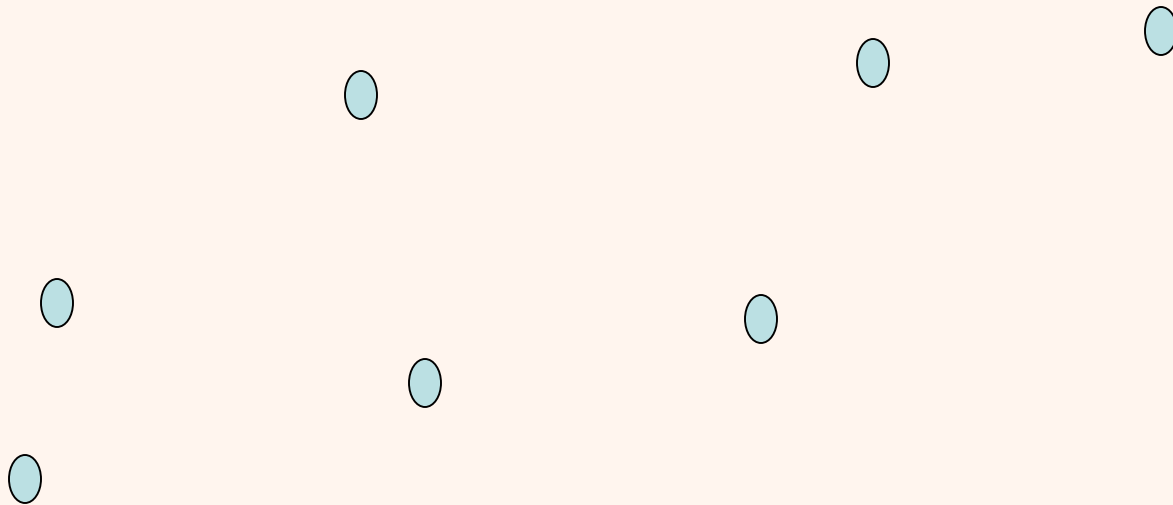
- این راستا به گونه‌ای انتخاب می‌شود که  $\text{Var}(z)$  **ماکزیمم** شود، راستایی که داده در امتداد آن بیشترین تغییرات را داشته باشد.

– این مسأله باعث می‌شود، تفاوت نمونه‌ها آشکارتر شود.

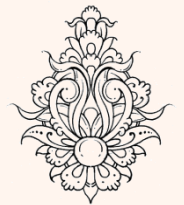
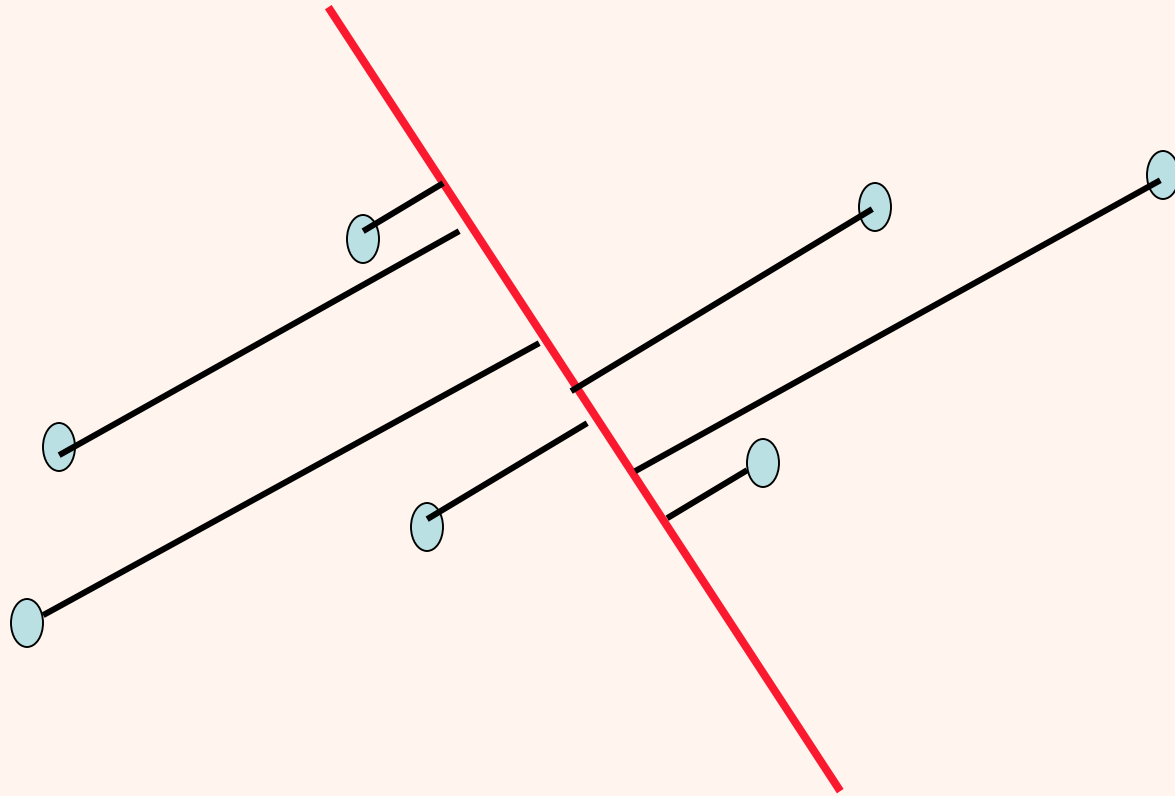
- این شیوه‌ی کاهش بعد به صورت «بی‌نظارت» است.



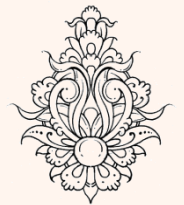
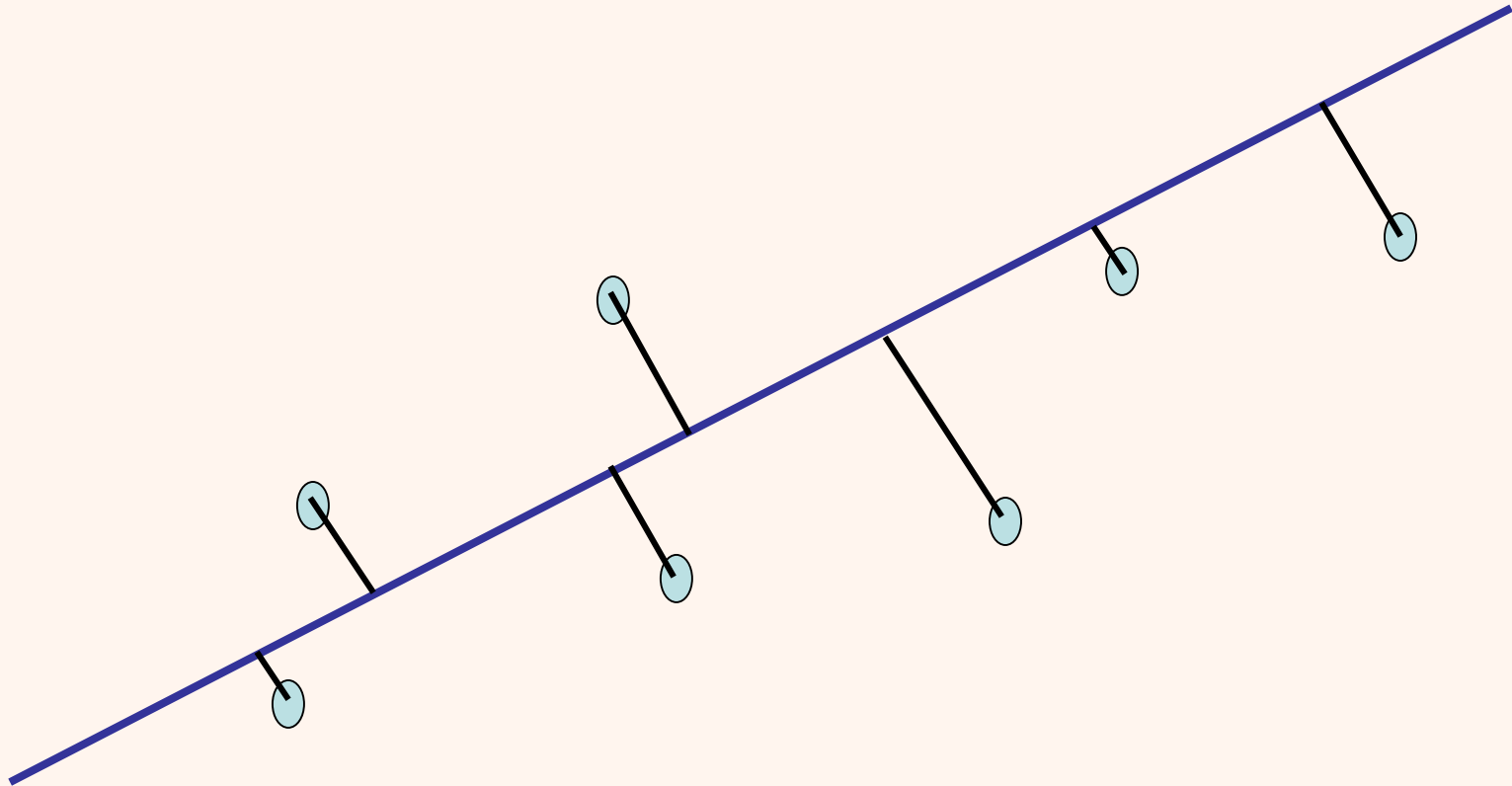
# تحلیل مؤلفه‌های اصلی (ادامه...)



# تحلیل مؤلفه‌های اصلی (ادامه...)



# تحلیل مؤلفه‌های اصلی (ادامه...)



# تحلیل مؤلفه‌های اصلی (ادامه...)

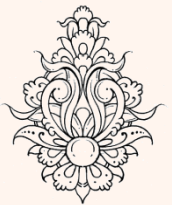
- در راستای  $w$ ، پراکندگی داده ماکزیمم می‌شود:

$$\begin{aligned}\text{Var}(z) &= \text{Var}(w^T x) = E[(w^T x - w^T \mu)^2] \\ &= E[(w^T x - w^T \mu)(w^T x - w^T \mu)] \\ &= E[w^T (x - \mu)(x - \mu)^T w] \\ &= w^T E[(x - \mu)(x - \mu)^T] w = w^T \Sigma w\end{aligned}$$

where  $\text{Cov}(x) = \Sigma$

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

$$||w|| = 1$$



# تحلیل مؤلفه‌های اصلی (ادامه...)

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

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

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

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

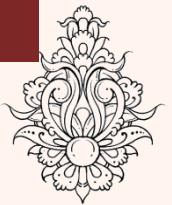
- در نتیجه

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

در نتیجه  $\mathbf{w}_1$  یکی از بردارهای ویژه‌ی ماتریس  $\Sigma$  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.

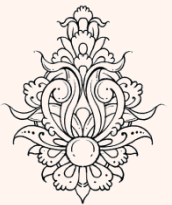




# تحلیل مؤلفه‌های اصلی (ادامه...)

- 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.
- برای یافتن دومین مؤلفه‌ی اصلی ( $w_2$ )، باید  $\text{Var}(z_2)$  ماکزیمم شود، مشروط به متعامد بودن بر اولین مؤلفه‌ی اصلی و  $\|w_2\|=1$

$$\max_{w_2} w_2^T \Sigma w_2 - \alpha (w_2^T w_2 - 1) - \beta (w_2^T w_1 - 0)$$



# تحلیل مؤلفه‌های اصلی (ادامه...)

$$\max_{\mathbf{w}_2} \mathbf{w}_2^T \Sigma \mathbf{w}_2 - \alpha (\mathbf{w}_2^T \mathbf{w}_2 - 1) - \beta (\mathbf{w}_2^T \mathbf{w}_1 - 0)$$

• پس از مشتق گرفتن خواهیم داشت:

$$2\Sigma \mathbf{w}_2 - 2\alpha \mathbf{w}_2 - \beta \mathbf{w}_1 = 0$$

• با ضرب در  $\mathbf{w}_1^T$

$$\mathbf{w}_1^T \Sigma \mathbf{w}_2 - 2\alpha \mathbf{w}_1^T \mathbf{w}_2 - \beta \mathbf{w}_1^T \mathbf{w}_1 = 0$$

$$\mathbf{w}_1^T \mathbf{w}_2 = 0$$

$$\mathbf{w}_1^T \Sigma \mathbf{w}_2 - \beta \mathbf{w}_1^T \mathbf{w}_1 = 0$$

$$\mathbf{w}_1^T \Sigma \mathbf{w}_2 = \mathbf{w}_2^T \Sigma \mathbf{w}_1 = \lambda_1 \mathbf{w}_2^T \mathbf{w}_1 = 0$$

→  $\beta = 0$

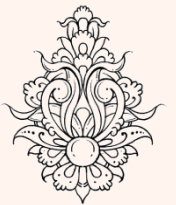
→  $\Sigma \mathbf{w}_2 = \alpha \mathbf{w}_2$

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...)

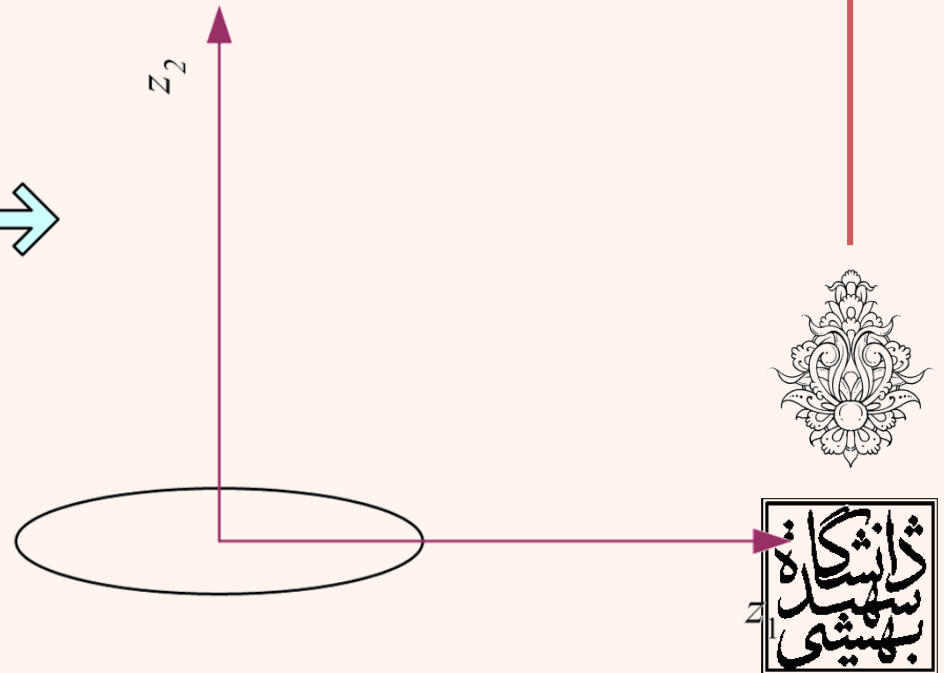
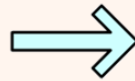
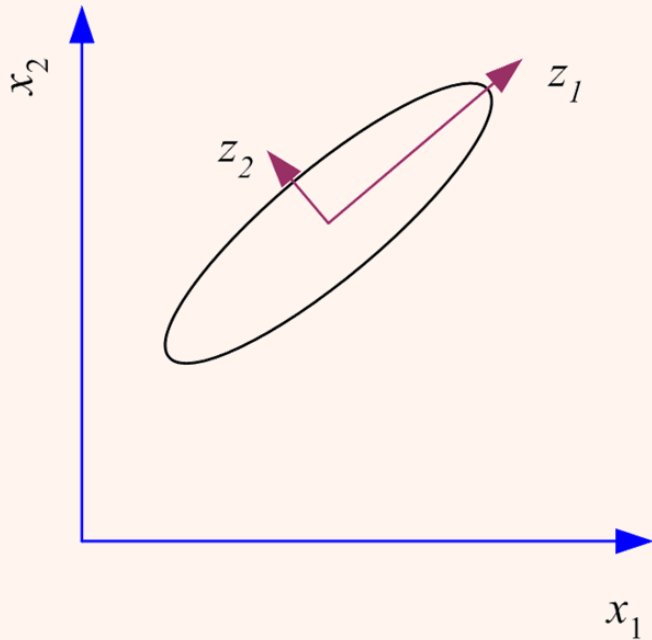
- در صورتی که ماتریس متقارن باشد، بردارهای ویژه‌ی آن متعامد هستند.
- در صورتی که ماتریس positive definite باشند، مقادیر ویژه همگی مثبت خواهند بود.
- در صورتی که ماتریس singular باشد، به اندازه‌ی rank ماتریس مقادیر ویژه غیرصفر خواهیم داشت.



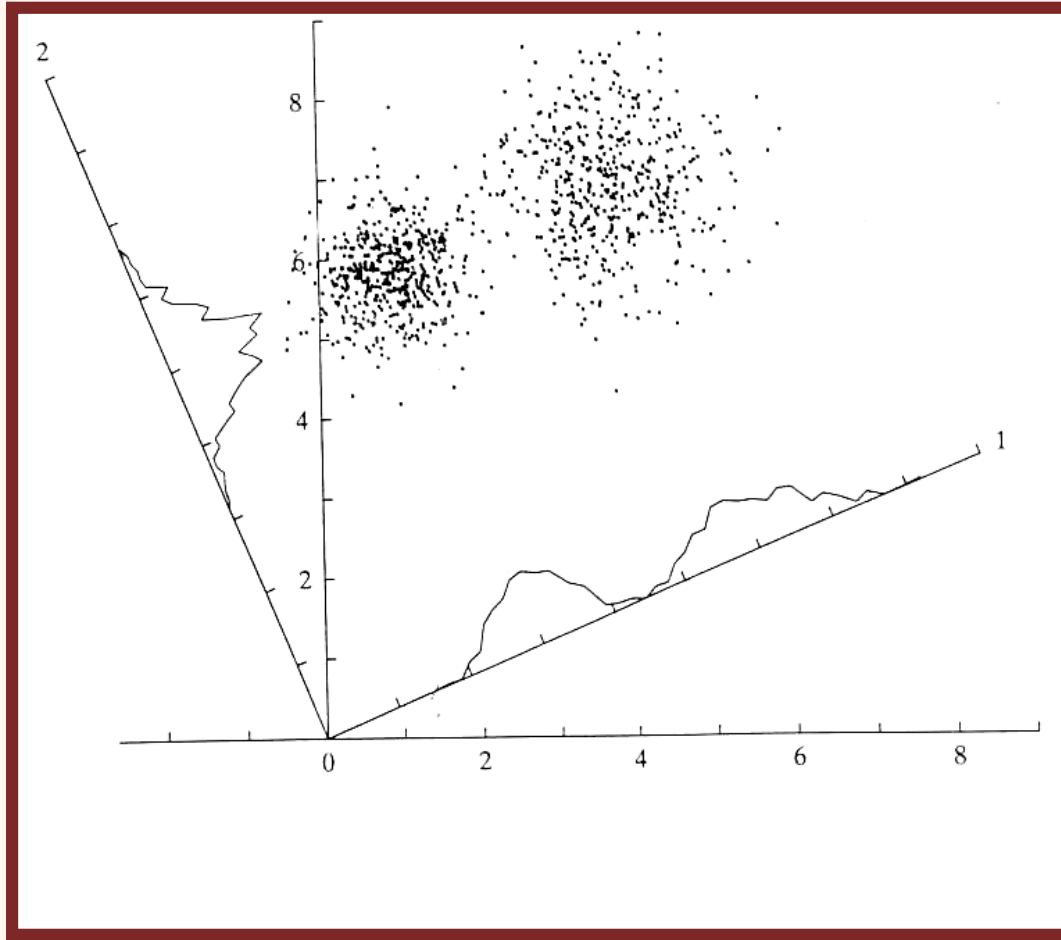
# تحلیل مؤلفه‌های اصلی (ادامه...)

$$z = W^T(x - m)$$

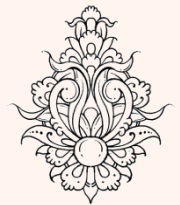
- ستون‌های  $W$ ، بردارهای ویژه‌ی ماتریس کواریانس هستند.



# تحلیل مؤلفه‌های اصلی (ادامه...)



**Haykin, S. Neural Networks: A Comprehensive Foundation,**



# تحلیل مؤلفه‌های اصلی (ادامه...)

- از زاویه‌ی دیگری نیز می‌توان به این مسأله نگاه کرد؛ هدف یافتن ماتریس تبدیلی است که داده‌های را به گونه‌ای نگاشت کند که در فضای جدید «ناهمبسته» باشند.

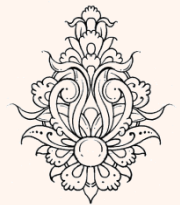
$$\mathbf{z} = \mathbf{W}^T \mathbf{x} \quad \text{Cov}(\mathbf{z}) = \mathbf{D} \quad \text{ماتریس قطری}$$

- $\mathbf{C}_{d \times d}$  ماتریسی است که ستون‌هایش بردارهای ویژه‌ی ماتریس کواریانس است:

$$\mathbf{C}^T \mathbf{C} = \mathbf{I}$$

$$\mathbf{S} = \mathbf{S} \mathbf{C} \mathbf{C}^T$$

ادامه



# تحلیل مؤلفه‌های اصلی (ادامه...)

$$\begin{aligned} \mathbf{S} &= \mathbf{S}\mathbf{C}\mathbf{C}^T \\ &= \mathbf{S}[\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_d]\mathbf{C}^T \\ &= [\mathbf{S}\mathbf{c}_1, \mathbf{S}\mathbf{c}_2, \dots, \mathbf{S}\mathbf{c}_d]\mathbf{C}^T \\ &= [\lambda_1\mathbf{c}_1, \lambda_2\mathbf{c}_2, \dots, \lambda_d\mathbf{c}_d]\mathbf{C}^T \\ &= \lambda_1\mathbf{c}_1\mathbf{c}_1^T + \lambda_2\mathbf{c}_2\mathbf{c}_2^T + \dots + \lambda_d\mathbf{c}_d\mathbf{c}_d^T \\ &= \mathbf{C}\mathbf{D}\mathbf{C}^T \end{aligned}$$

ماتریس قطری که عناصر روی قطر اصلی مقادیر ویژه‌ی ماتریس کواریانس هستند.

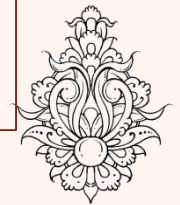
## Spectral decomposition of $\mathbf{S}$

$$\mathbf{C}^T\mathbf{S}\mathbf{C} = \mathbf{D}$$

$$\mathbf{z} = \mathbf{W}^T\mathbf{x}, \quad \text{Cov}(\mathbf{z}) = \mathbf{W}^T\mathbf{S}\mathbf{W}$$

$$\mathbf{W} = \mathbf{C}$$

$$\text{Cov}(\mathbf{z}) = \mathbf{D}$$



# 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.

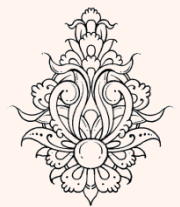
– در این حالت  $K$  مؤلفه‌ی پرارزش انتخاب می‌شوند، با فرض آن که مقادیر ویژه به صورت صعودی مرتب شده باشند.

**Proportion of Variance (PoV)**

**PoV > 0.9**

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

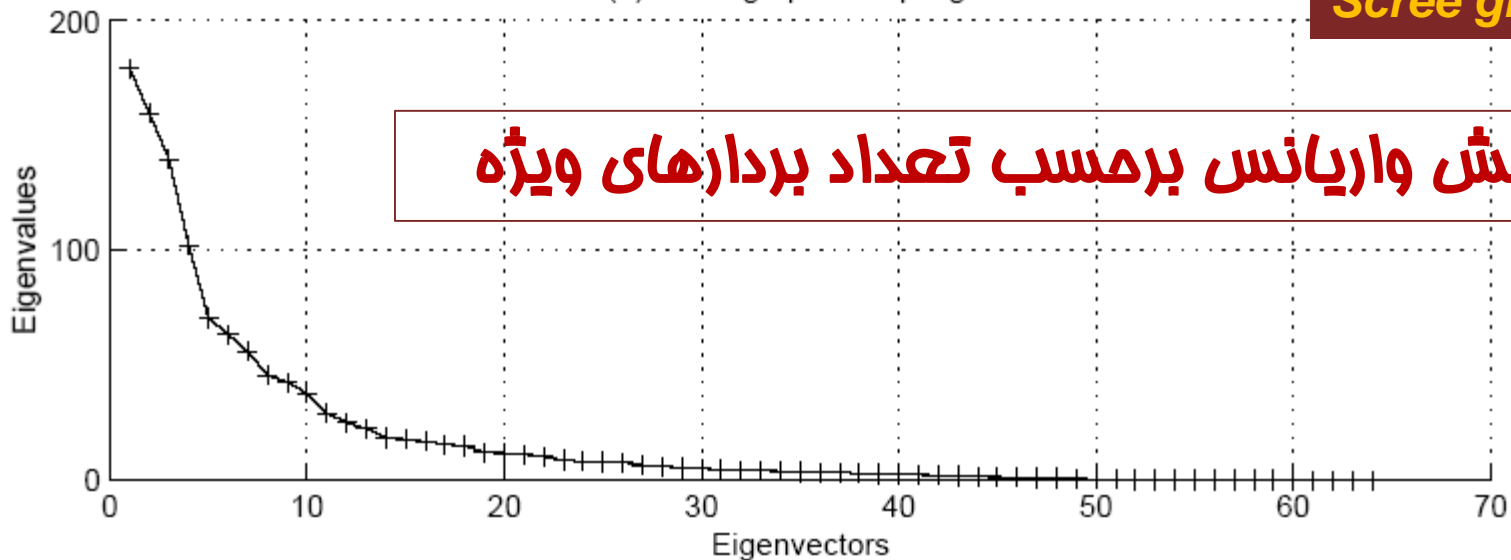
– در کاربردهای نظیر پردازش تصویر یا صوت، معمولاً کاهش ابعاد قابل توجه است.





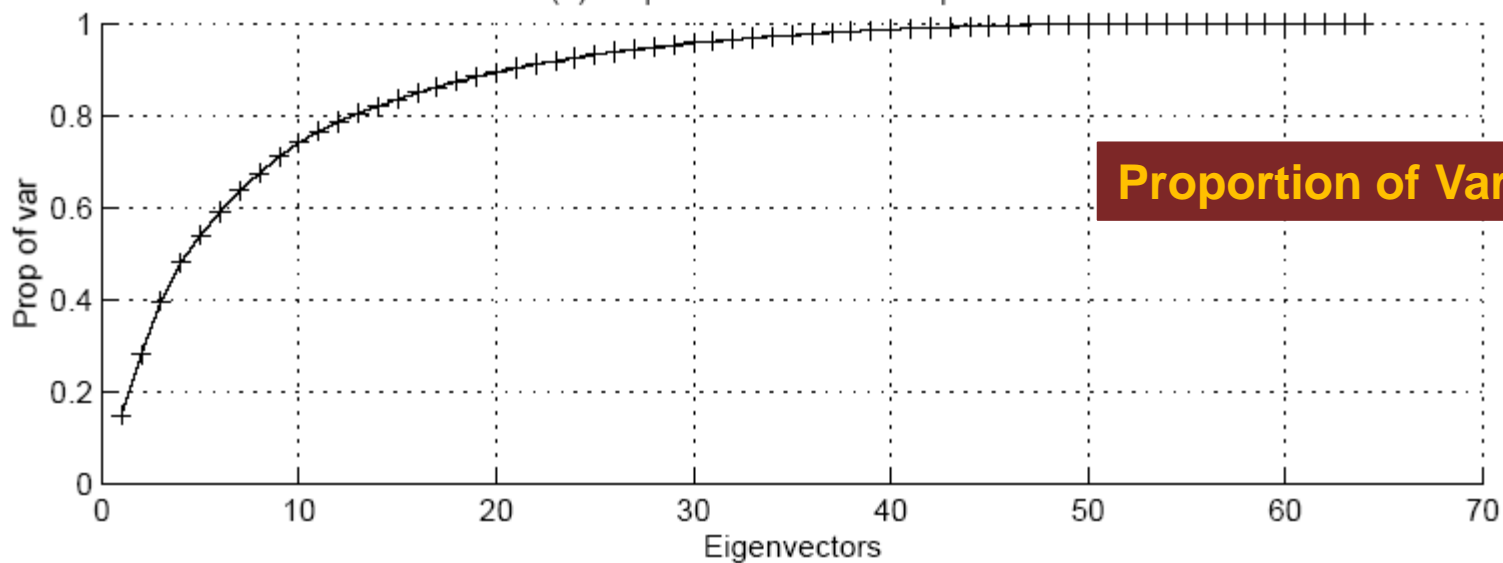
(a) Scree graph for Optdigits

Scree graph



نمایش واریانس بر حسب تعداد بردارهای ویژه

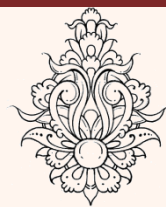
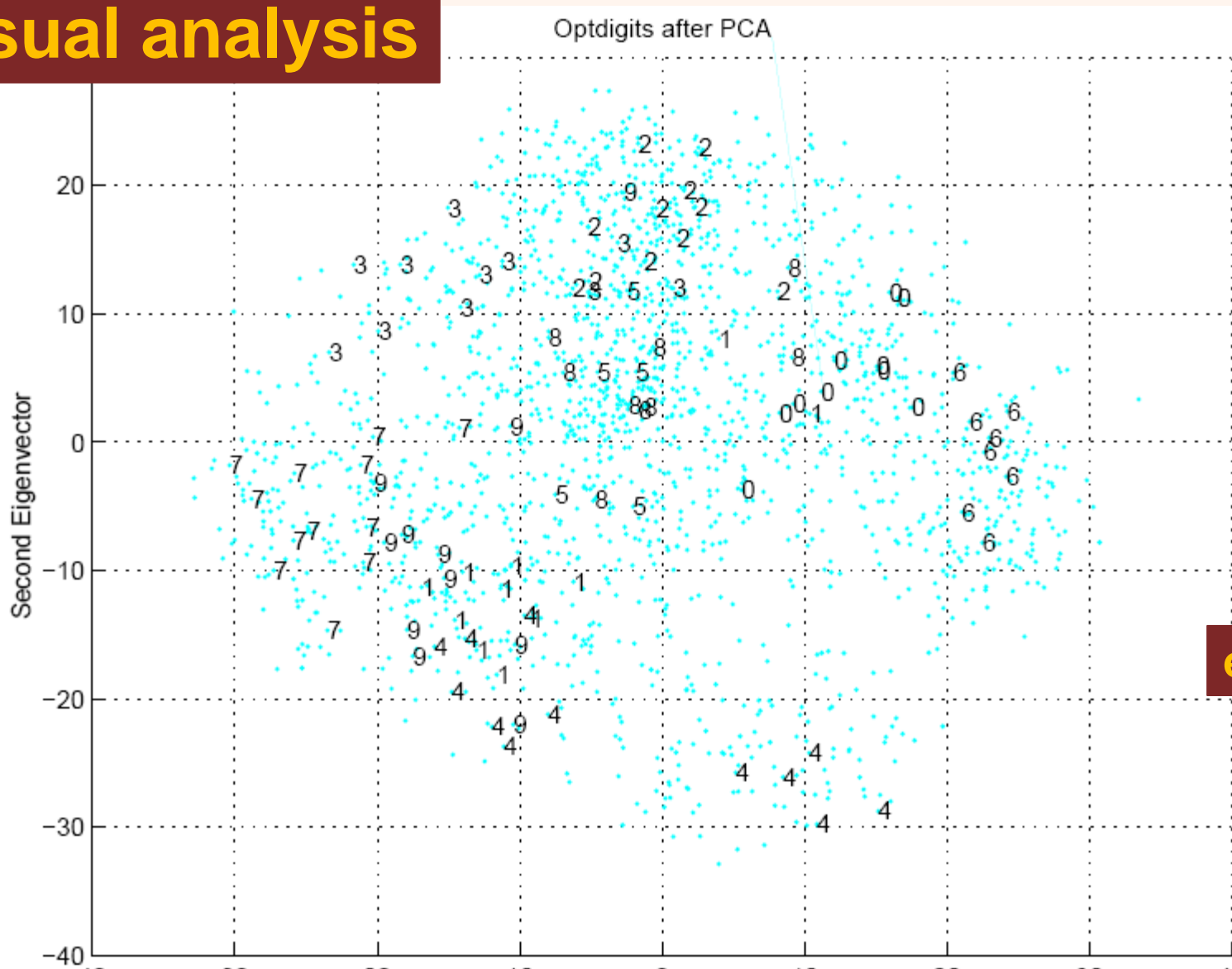
(b) Proportion of variance explained



Proportion of Variance (PoV)



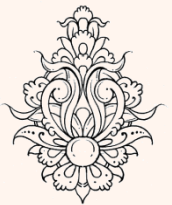
# Visual analysis



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

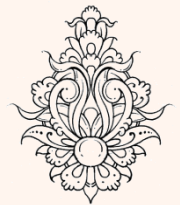
## چند نکته

- 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 تخییرات 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 کواریانس است.
- در میان تمام بردارهای متعامد، PCA کم‌ترین میزان خطا را دارد.  
$$\sum_t \|\hat{\mathbf{x}}^t - \mathbf{x}^t\|$$
**Reconstruction error**
- Hotelling transform<sup>t</sup> و Karhunen-Loève expansion نام‌های دیگری هستند برای مفاهیم مشابه به کار می‌روند.



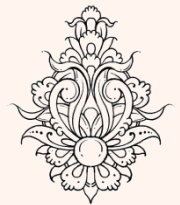
# کاربرد PCA در شناسایی چهره



پایگاه داده‌ی ORL

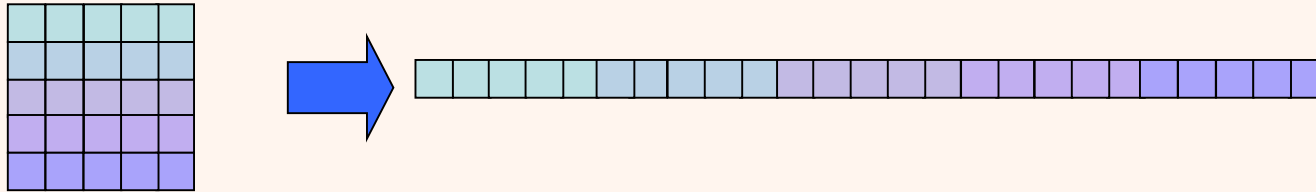


میانگین چهره‌ها

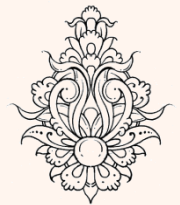
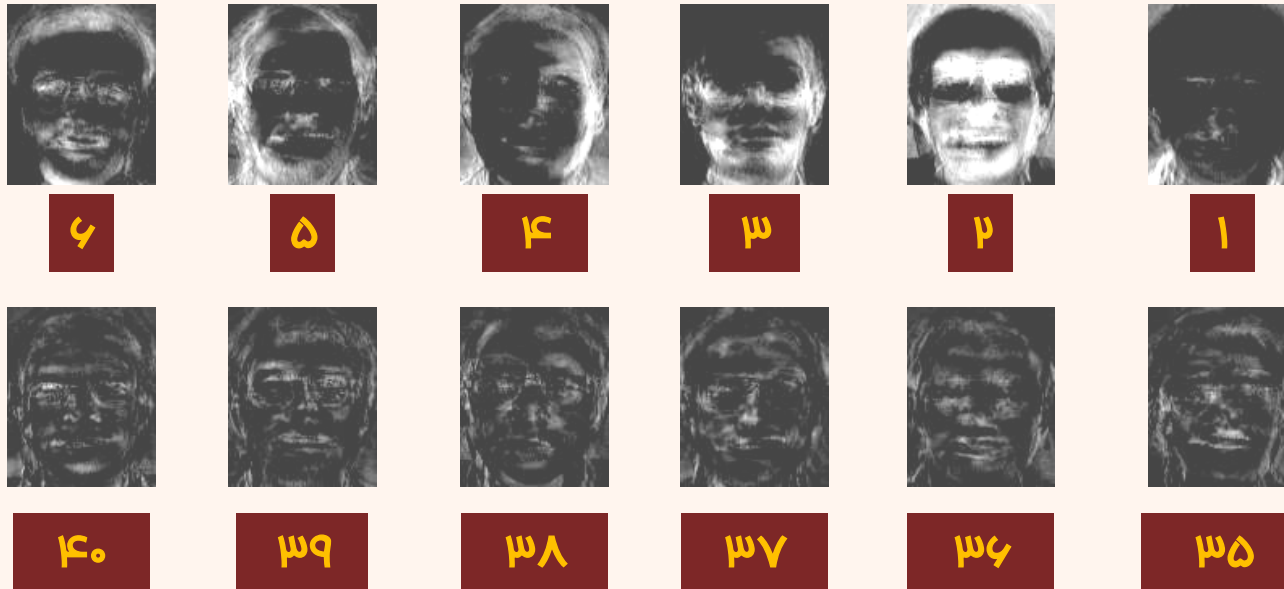


*M. Turk, A. Pentland, "Eigenfaces for Recognition", Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71-86, 1991.*

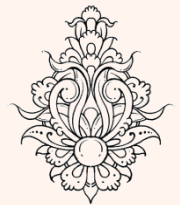
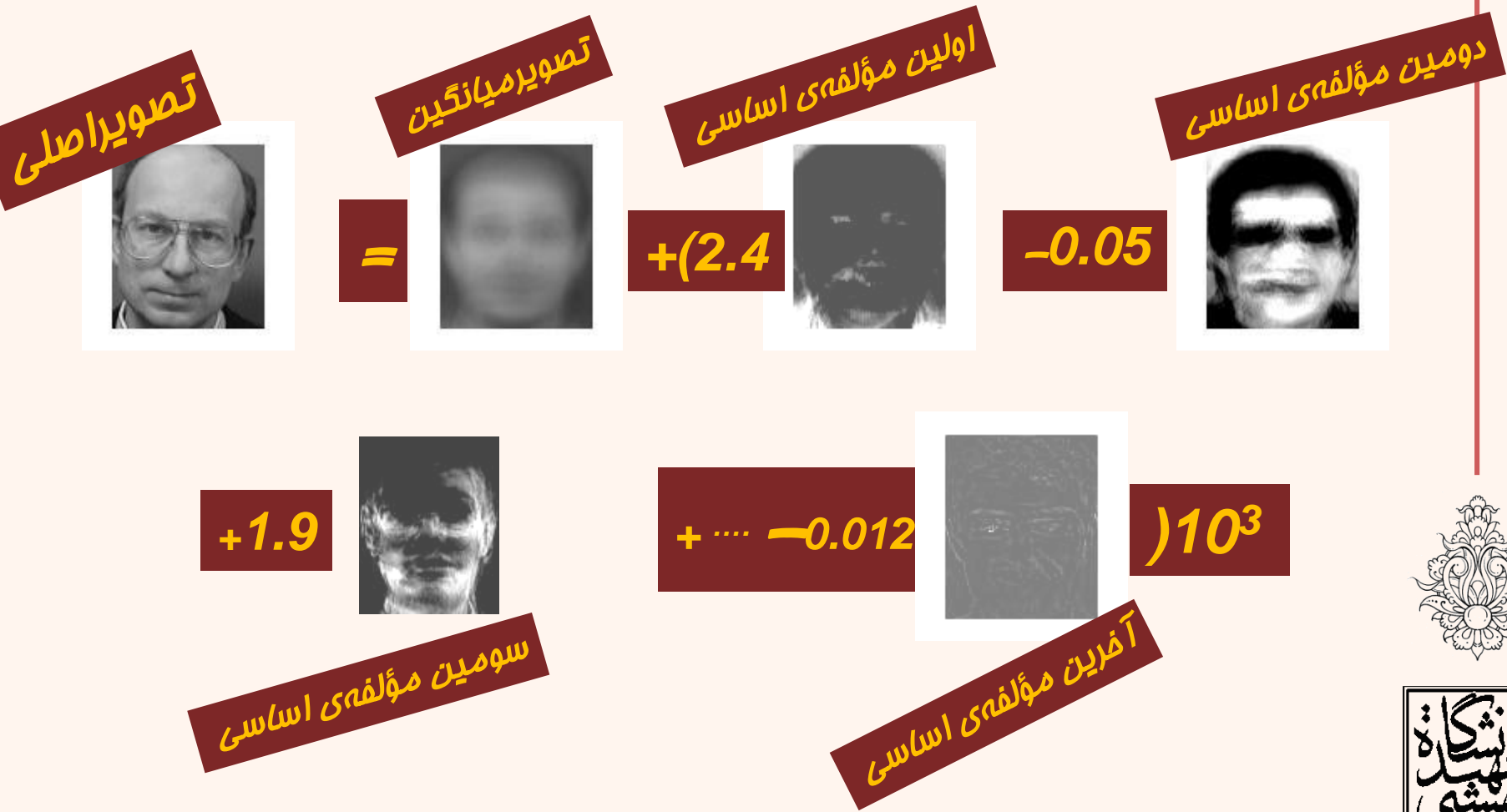
# کاربرد PCA در شناسایی چهره (ادامه...)



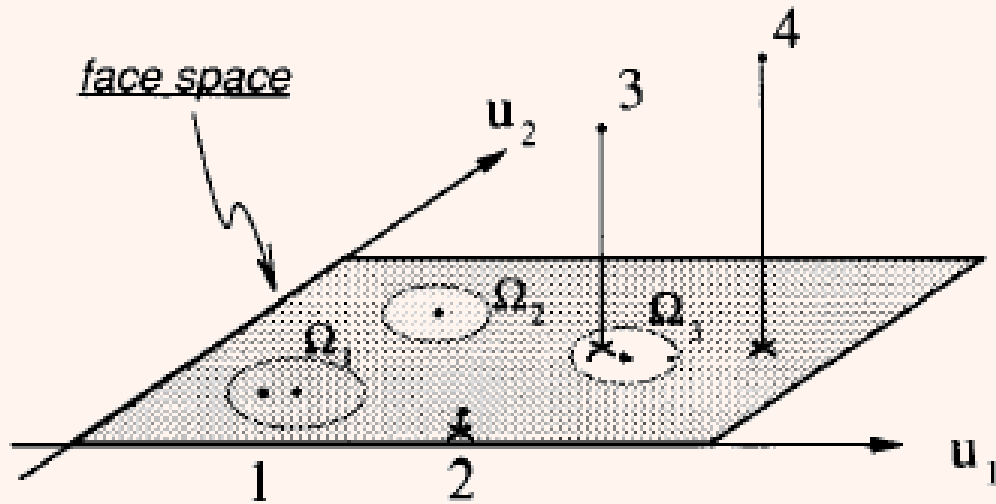
## Eigenfaces



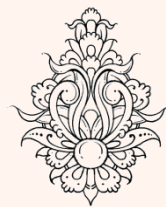
# کاربرد PCA در شناسایی چهره (ادامه...)



# تشخیص چهره



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

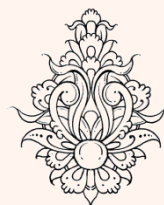




# Feature Embedding

- ماتریس داده‌ها به صورت  $X_{N \times d}$  است.
- ماتریس کواریانس خصیصه‌ها  $X^T X_{d \times d}$  می‌باشد، در نتیجه
- با ضرب طرفین در  $X$
- در نتیجه  $X w_i$  بردار ویژهی  $XX^T_{N \times N}$  با مقدار ویژهی  $\lambda_i$  است.
- در این حالت بردار ویژه، مختصات نمونه‌ها در راستای  $w_i$  را نشان می‌دهد.

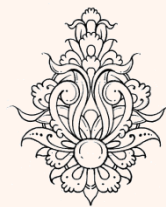
$$(XX^T)Xw_i = \lambda Xw_i$$



# Feature Embedding

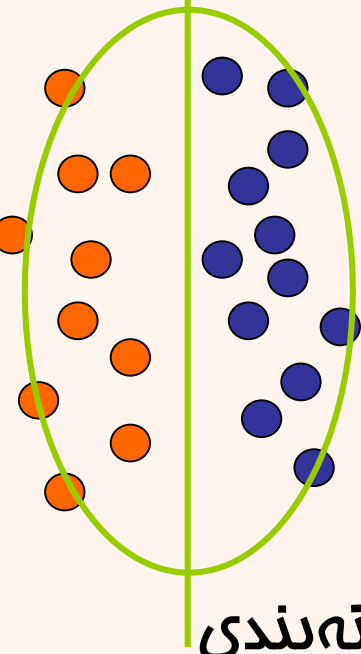
- ثابت می‌شود، رتبه‌ی ماتریس کواریانس حداکثر  $\min(d, N)$  می‌باشد.
- برای یک پایگاه داده حاوی چهل تصویر  $۲۵۶ \times ۲۵۶$ 
  - ماتریس کواریانس خصیصه‌ها  $۶۵۵۳۶ \times ۶۵۵۳۶$  خواهد بود.
  - در حالی که ماتریس شباهت نمونه‌ها  $۴۰ \times ۴۰$  می‌باشد.
  - این ماتریس شباهت دو به دو نمونه‌ها را نشان می‌دهد، از این نظر می‌توان گفت این شیوه داده‌ها را در یک فضای کابعدی به گونه‌ای قرار می‌دهد که فاصله‌ی بین آن‌ها حفظ می‌شود.

$$XX^T = VEV^T$$





## دسته‌بندی دو کلاسه



• آیا PCA برای دسته‌بندی مناسب است؟

- راستای نگاشت بر اساس واریانس، انتخاب می‌شود.

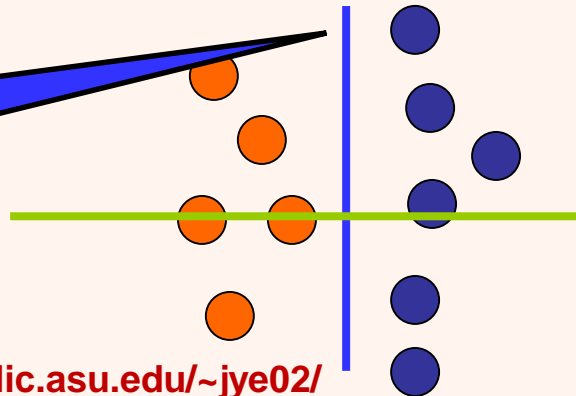
- در این میان ممکن است اطلاعات دسته‌ها از بین بروند.

• 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

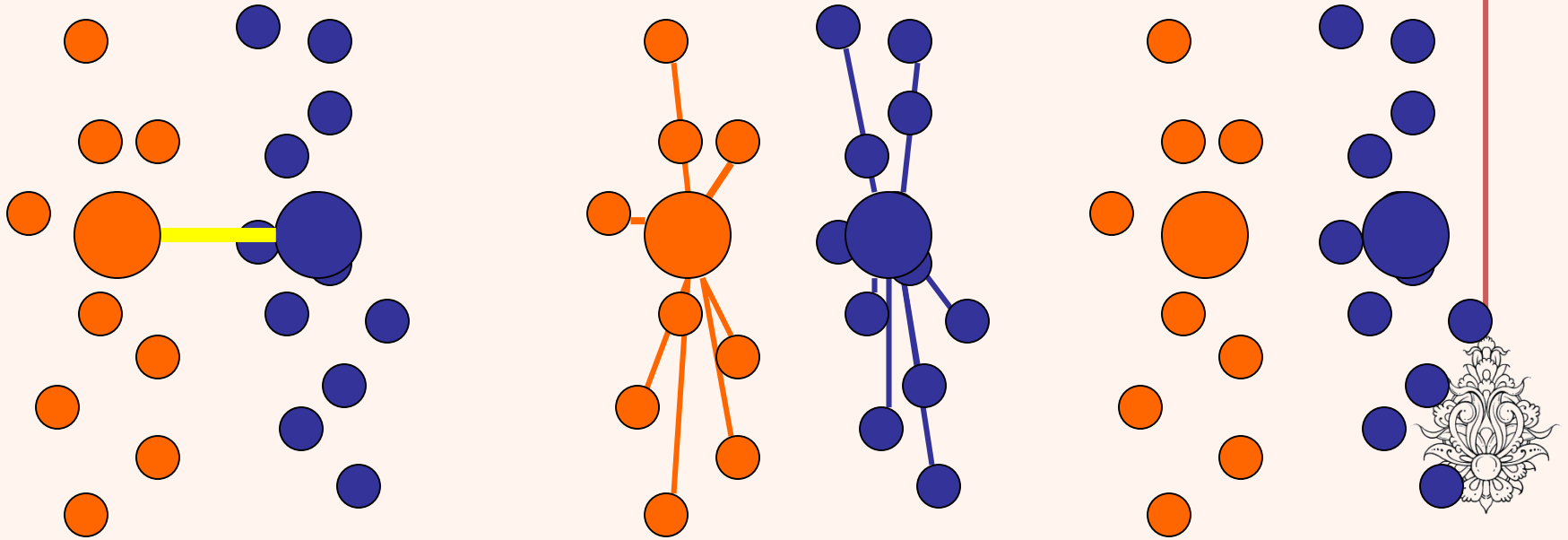


در این راستا دو کلاس بدون خطا طبقه‌بندی می‌شوند

# کاهش ابعاد برای دسته‌بندی

دسته‌بندی دو کلاس

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



Between-class distance

Within-class distance





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

# تحلیل تفکیک خطی (ادامه...)

دسته‌بندی دو کلاس

$$\begin{aligned} (m_1 - m_2)^2 &= (\mathbf{w}^T \mathbf{m}_1 - \mathbf{w}^T \mathbf{m}_2)^2 \\ &= \mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{w} \\ &= \mathbf{w}^T \mathbf{S}_B \mathbf{w} \end{aligned}$$

Between class scatter matrix

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$$

$$\begin{aligned} s_1^2 &= \sum_t (\mathbf{w}^T \mathbf{x}^t - m_1)^2 r^t \\ &= \sum_t \mathbf{w}^T (\mathbf{x}^t - \mathbf{m}_1)(\mathbf{x}^t - \mathbf{m}_1)^T \mathbf{w} r^t \\ &= \mathbf{w}^T \mathbf{S}_1 \mathbf{w} r^t \end{aligned}$$

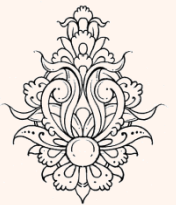
Class scatter matrix for C1

$$\mathbf{S}_1 = \sum_t r^t (\mathbf{x}^t - \mathbf{m}_1)(\mathbf{x}^t - \mathbf{m}_1)^T$$

$$s_1^2 + s_2^2 = \mathbf{w}^T \mathbf{S}_W \mathbf{w}$$

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

Total within Class scatter



# تحلیل تفکیک خطی (ادامه...)

دسته‌بندی دو کلاس

- هدف ماکزیمیم کردن رابطه‌ی زیر است:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} = \frac{|\mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2)|^2}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

- با مشتق گرفتن:

$$\frac{\mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2)}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} 2 \left( (\mathbf{m}_1 - \mathbf{m}_2) - \frac{\mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2)}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \mathbf{S}_W \mathbf{w} \right) = 0$$

Scalar

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

یادآوری: جداساز خطی

$$\mathbf{w} = \Sigma^{-1} (\mu_1 - \mu_2)$$

when  $p(\mathbf{x} | C_i) \sim \mathcal{N}(\mu_i, \Sigma)$

بدین ترتیب، برای کلاس  
نرمال، LDA جداساز بهینه  
است.

دانشگاه  
تهران  
پیشین

# دسته‌بندی برای بیش از دو کلاس

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

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

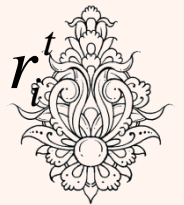
$$\mathbf{S}_W = \sum_{i=1}^K \mathbf{S}_i \quad \mathbf{S}_i = \sum_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^T$$

Within-class scatter

Between-class scatter:

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

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



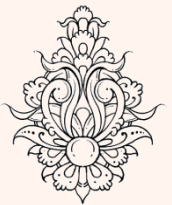


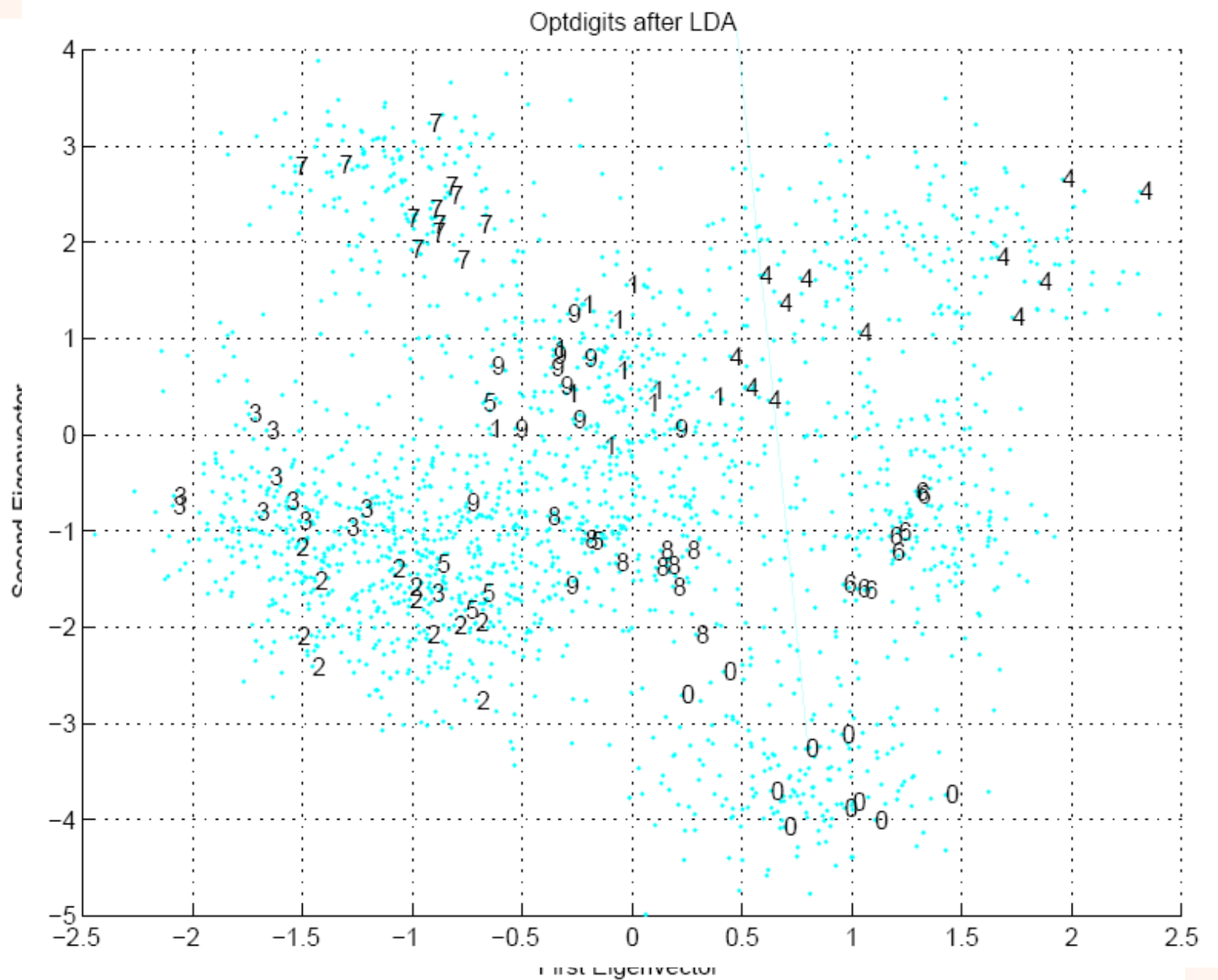
# دسته‌بندی برای بیش از دو کلاس

- 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.

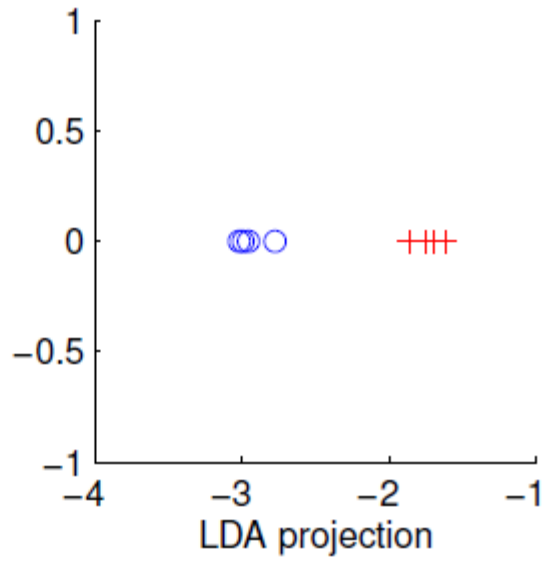
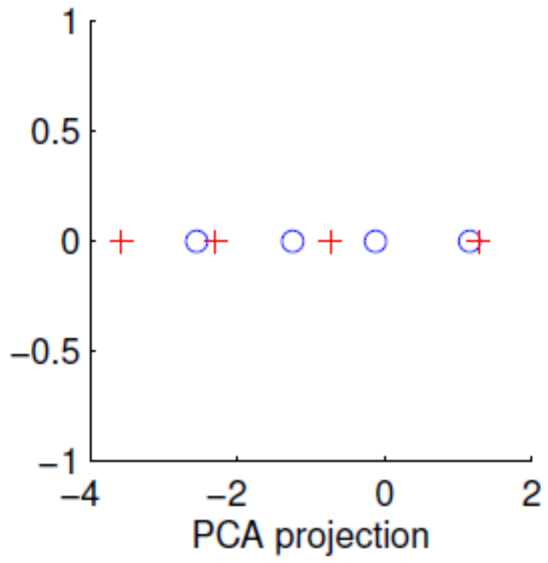
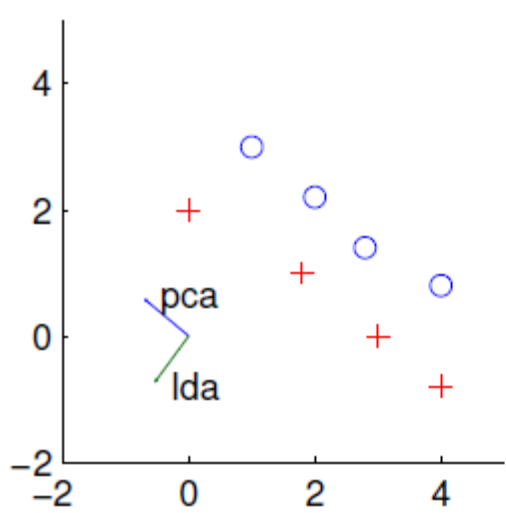
$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|}$$

- در این حالت پاسخ، بردارهای ویژه متناظر با بزرگ‌ترین مقادیر ویژه‌ی ماتریس  $S_W^{-1} S_B$  خواهد بود.



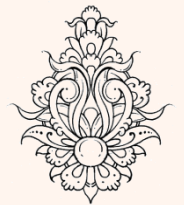


# PCA vs LDA



# کاربردهای LDA

- Face recognition
  - Belhumeur *et al.*, PAMI'97
- Image retrieval
  - Swets and Weng, PAMI'96
- Gene expression data analysis
  - Dudoit *et al.*, JASA'02; Ye *et al.*, TCBB'04
- Protein expression data analysis
  - Lilien *et al.*, Comp. Bio.'03
- Text mining
  - Park *et al.*, SIMAX'03; Ye *et al.*, PAMI'04
- Medical image analysis
  - Dundar, SDM'05



- در «**تحلیل عاملی**» فرض می‌کنیم که یک مجموعه‌ای «**عامل مخفی**» وجود دارد (z) که ترکیب آن‌ها متغیرها (x) را می‌سازد.

latent factors

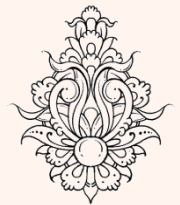
$$x_i - \mu_i = v_{i1}z_1 + v_{i2}z_2 + \dots + v_{ik}z_k + \varepsilon_i$$

factor loadings

noise sources

$$E[z_j] = 0, \text{Var}(z_j) = 1, \text{Cov}(z_i, z_j) = 0, i \neq j,$$

$$E[\varepsilon_i] = \psi_i, \text{Cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j, \text{Cov}(\varepsilon_i, z_j) = 0,$$

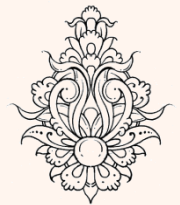


# تحلیل عاملی (ادامه...)

- در واقع این طور فرض می‌شود که مجموعه متغیرهایی که همبستگی بالایی با یکدیگر دارند و همبستگی آنها با سایر متغیرها پایین است، دارای عوامل مشترکی هستند. بدین ترتیب با استفاده از تحلیل عاملی متغیرها «فوشه‌بندی» می‌شوند.

**Factor clusters**

- تحلیل عاملی مانند PCA، «بی‌نظارت» است.

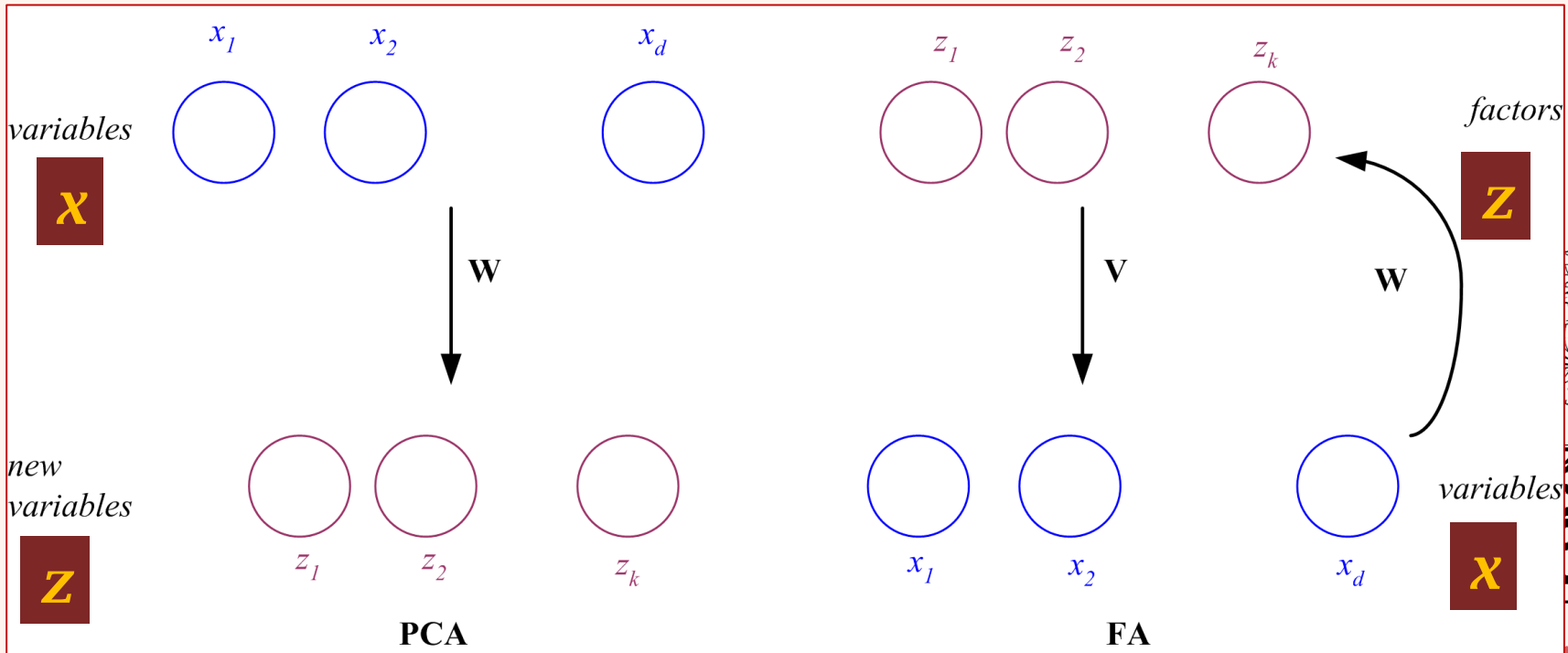


# PCA vs FA

- PCA From  $\mathbf{x}$  to  $\mathbf{z}$
- FA From  $\mathbf{z}$  to  $\mathbf{x}$

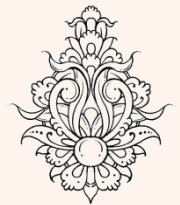
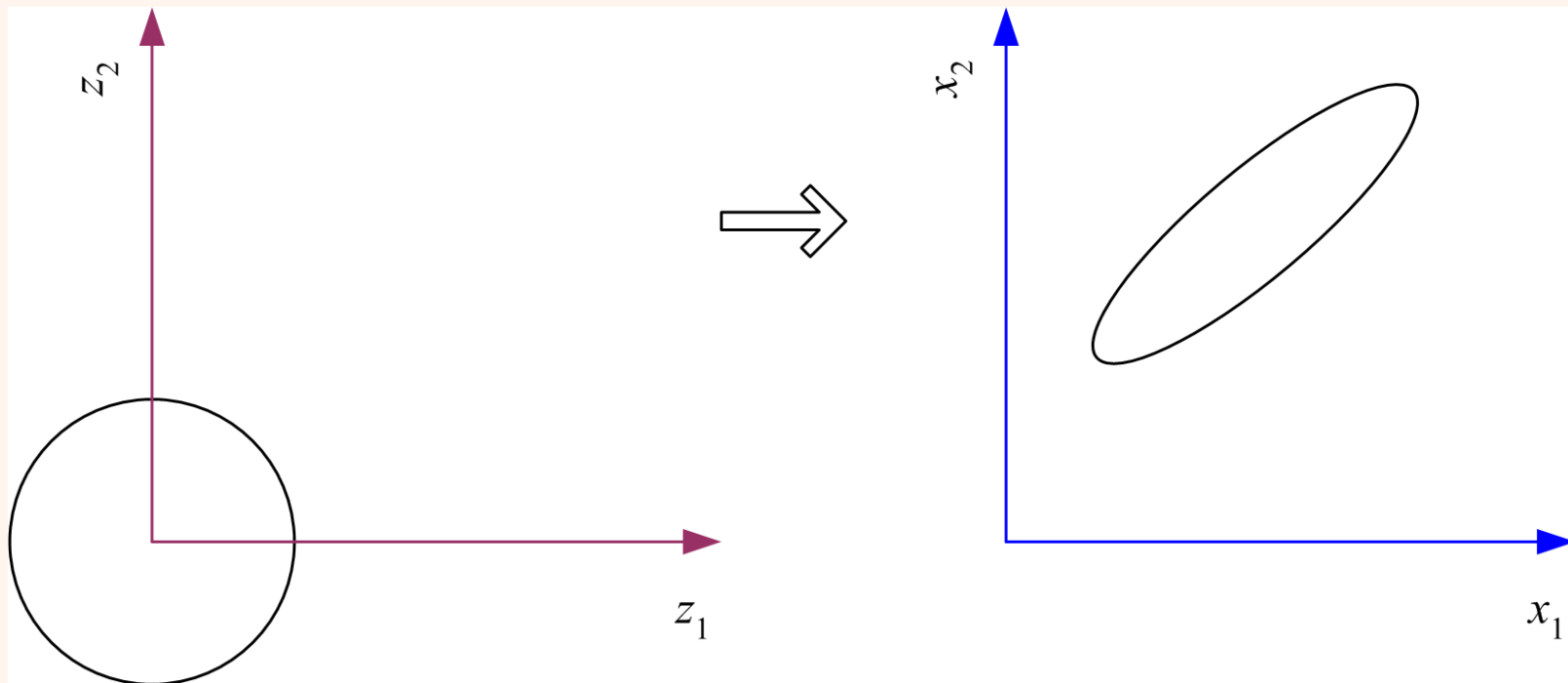
$$\mathbf{z} = \mathbf{W}^T(\mathbf{x} - \boldsymbol{\mu})$$

$$\mathbf{x} - \boldsymbol{\mu} = \mathbf{V}\mathbf{z} + \boldsymbol{\varepsilon}$$



# تحلیل عاملی (ادامه...)

- در تحلیل عاملی، عوامل (پس از چرخش، تغییر مقیاس و انتقال) متغیرها را می‌سازند.





# تحلیل عاملی (ادامه...)

$$x_i - \mu_i = v_{i1}z_1 + v_{i2}z_2 + \dots + v_{ik}z_k + \varepsilon_i$$

$$x_i = \sum_{j=1}^k v_{ij}z_j + \varepsilon_i$$

$$\mathbf{x} - \boldsymbol{\mu} = \mathbf{V}\mathbf{z} + \boldsymbol{\varepsilon}$$

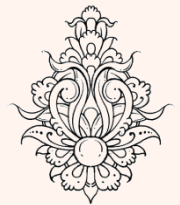
بدون لطمه به کلیت مسأله در ادامه فرض می‌کنیم،  $\mu = 0$

$$\mathbf{x}_{d \times 1} = \mathbf{V}_{d \times k} \mathbf{z}_{k \times 1} + \boldsymbol{\varepsilon}_{d \times 1}$$

واریانس مربوط به  $x_i$

$$\text{Var}(x_i) = v_{i1}^2 + v_{i2}^2 + \dots + v_{ik}^2 + \psi_i$$

واریانس مربوط به عوامل مشترک



# تحلیل عاملی (ادامه...)

$$\Sigma = \text{Cov}(\mathbf{x}) = \text{Cov}(\mathbf{Vz} + \boldsymbol{\varepsilon})$$

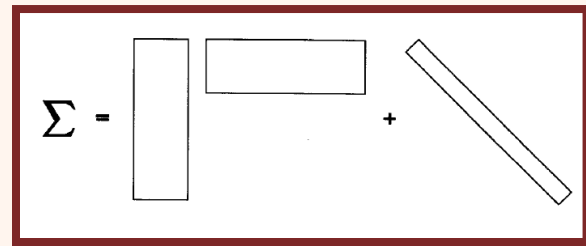
$$= \text{Cov}(\mathbf{Vz}) + \text{Cov}(\boldsymbol{\varepsilon})$$

$$= \mathbf{V}\text{Cov}(\mathbf{z})\mathbf{V}^T + \boldsymbol{\Psi}$$

$$\text{Cov}(\mathbf{z}) = \mathbf{I}$$

$$= \mathbf{V}\mathbf{V}^T + \boldsymbol{\Psi}$$

ماتریس قطری



ALVIN C. RENCHER

با فرض داشتن دو عامل

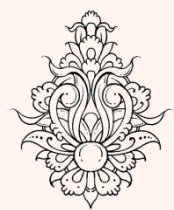
$$\text{Cov}(x_1, x_2) = v_{11}v_{21} + v_{12}v_{22}$$

• در صورتی که کواریانس دو متغیر بالا باشد، به این معناست که از طریق یک عامل مشترک به هم مرتبط هستند و در نتیجه برای هر دو ضریب مربوط به آن عامل بالا خواهد بود.

• همچنین داریم:

$$\text{Cov}(x_1, z_2) = \text{Cov}(v_{12}z_2, z_2) = v_{12} \text{Var}(z_2) = v_{12}$$

$$\text{Cov}(\mathbf{x}, \mathbf{z}) = \mathbf{V}$$



loading، همبستگی متغیرها با فاکتورها را نشان می‌دهند

# تحلیل عاملی (ادامه...)

## Principal Component Method

- با در اختیار داشتن تخمین ماتریس کواریانس

$$\mathbf{S} = \mathbf{V}\mathbf{V}^T + \psi$$

- در صورتی که از  $\psi$  صرفنظر کنیم:

$$\mathbf{S} = \mathbf{V}\mathbf{V}^T$$

- با تجزیه طیفی  $\mathbf{S}$

$$\mathbf{S} = \mathbf{C}\mathbf{D}\mathbf{C}^T = \mathbf{C}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{C}^T = (\mathbf{C}\mathbf{D}^{1/2})(\mathbf{C}\mathbf{D}^{1/2})^T$$

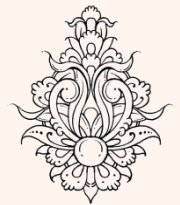
- در نتیجه

$$\mathbf{V} = \mathbf{C}\mathbf{D}^{1/2}$$

- و مقادیر  $\psi_i$

$$\psi_i = s_i^2 - \sum_{j=1}^k v_{ij}^2$$

$$\text{Var}(x_i) = v_{i1}^2 + v_{i2}^2 + \dots + v_{ik}^2 + \psi_i$$



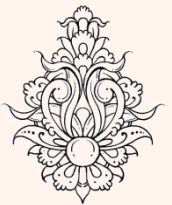
# تحلیل عاملی (ادامه...)

- در صورتی که  $V$  در یک ماتریس متعامد (مانند  $T$ ) ضرب شود:  $(TT^T=I)$

$$(VT)(VT)^T = VTT^T V^T = VV^T = S$$

- بدین ترتیب مشاهده می‌شود که حل به دست آمده یکتا نیست.

- ضرب در یک ماتریس متعامد فاصله از مبدا را تغییر نمی‌دهد. تنها باعث چرخش محورها می‌شود.
- بدین ترتیب می‌توان با این کار مناسب‌ترین فاکتورها را یافت.



# کاهش بعد با استفاده از تحلیل عوامل

$$z_j = \sum_{i=1}^d w_{ji} x_i + \varepsilon_j, \quad j=1, \dots, k$$

$$\mathbf{z}^t = \mathbf{W}^T \mathbf{x}^t + \boldsymbol{\varepsilon}, \quad \forall t = 1, \dots, N$$

$$\left(\mathbf{z}^t\right)^T = \left(\mathbf{x}^t\right)^T \mathbf{W} + \boldsymbol{\varepsilon}^T, \quad \forall t = 1, \dots, N$$

• برای همه  $N$  نمونه

$$\mathbf{Z}_{N \times k} = \mathbf{X}_{N \times d} \mathbf{W}_{d \times k} + \boldsymbol{\Xi}_{N \times k}$$

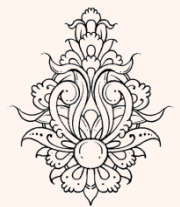
• شبیه مسأله‌ی رگرسیون خطی چند متغیره با چند خروجی

$$\mathbf{W} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{Z}$$

$$= \left(\frac{\mathbf{X}^T \mathbf{X}}{N-1}\right)^{-1} \left(\frac{\mathbf{X}^T \mathbf{Z}}{N-1}\right)$$

$$= \mathbf{S}^{-1} \mathbf{V}$$

$$\mathbf{Z} = \mathbf{X} \mathbf{W} = \mathbf{X} \mathbf{S}^{-1} \mathbf{V}$$



## تجزیه‌ی مقادارهای تکین

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

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

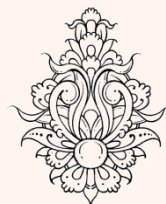
- $\mathbf{V}$  شامل بردارهای ویژه‌ی  $\mathbf{X}\mathbf{X}^T$  می‌باشد،  $\mathbf{W}$  شامل بردارهای ویژه‌ی  $\mathbf{X}^T\mathbf{X}$  است و  $\mathbf{A}$  مقادیر ویژه را در  $k$  عنصر قطری خود دارد.

$$\mathbf{X}\mathbf{X}^T = (\mathbf{V}\mathbf{A}\mathbf{W}^T)(\mathbf{V}\mathbf{A}\mathbf{W}^T)^T = \mathbf{V}\mathbf{A}\mathbf{W}^T\mathbf{W}\mathbf{A}^T\mathbf{V}^T = \mathbf{V}\mathbf{E}\mathbf{V}^T$$

$$\mathbf{X}^T\mathbf{X} = (\mathbf{V}\mathbf{A}\mathbf{W}^T)^T(\mathbf{V}\mathbf{A}\mathbf{W}^T) = \mathbf{W}\mathbf{A}^T\mathbf{V}^T\mathbf{V}\mathbf{A}\mathbf{W}^T = \mathbf{W}\mathbf{D}\mathbf{W}^T$$

$$\mathbf{E} = \mathbf{A}\mathbf{A}^T$$

$$\mathbf{D} = \mathbf{A}^T\mathbf{A}$$

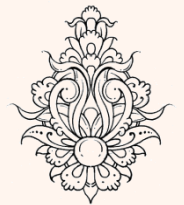
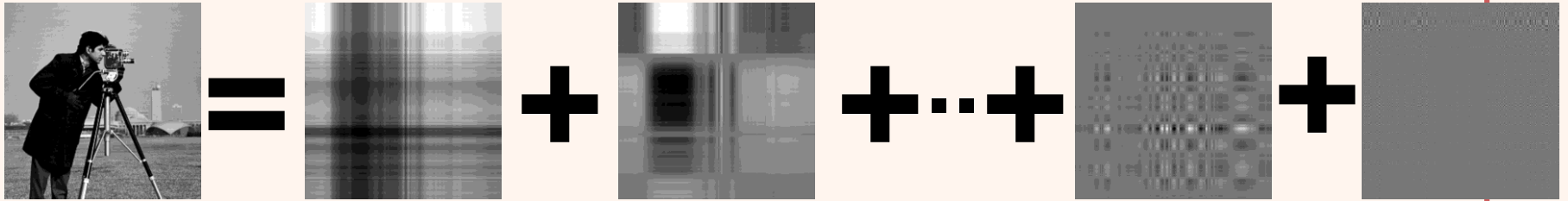


$$\mathbf{X}_{N \times d} = \mathbf{V}_{N \times N} \mathbf{A}_{N \times d} \mathbf{W}_{d \times d}^T$$

# تجزیه‌ی مقدارهای تکین

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

$$\mathbf{X} = \sum_{i=1}^k a_i \mathbf{v}_i \mathbf{w}_i^T$$



# MULTIDIMENSIONAL SCALING

• در این شیوه هدف نگاشت (کاهش ابعاد) به نحوی است که متریال المقذور فاصله بین نمونه‌ها حفظ شود.

– در صورتی که داشته باشیم:

$d_{ij}$

در فضای اصلی؛  $d$ -بعدی

فاصله‌ی بین نمونه‌ی  $i$  و  $j$

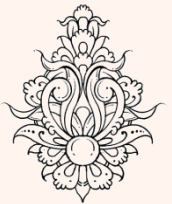
$\delta_{ij}$

پس از کاهش بعد؛  $k$ -بعدی ( $K < d$ )

– کاهش بعد به نحوی صورت پذیرد که  $\delta_{ij} \cong d_{ij}$

– به طور کلی این کار به دو صورت انجام می‌پذیرد:

- Metric MDS
- Nonmetric MDS





# Classical Solution(Principal coordinate analysis)

فاصله‌ی بین دو نمونه‌ی  $r$  و  $s$

$$d_{rs}^2 = (\mathbf{x}^r - \mathbf{x}^s)(\mathbf{x}^r - \mathbf{x}^s)^T = \|\mathbf{x}^r - \mathbf{x}^s\|^2 = \sum_{j=1}^d (x_j^r - x_j^s)^2$$

$$\mathbf{D} = [d_{ij}^2]$$

ماتریس فاصله‌ها به صورت روبرو تعریف می‌شود:

با بازنویسی روابط خواهیم داشت:

$$d_{rs}^2 = \sum_{j=1}^d (x_j^r - x_j^s)^2 = \sum_{j=1}^d (x_j^r)^2 + \sum_{j=1}^d (x_j^s)^2 - 2 \sum_{j=1}^d x_j^r x_j^s$$
$$b_{rs} = \sum_{j=1}^d x_j^r x_j^s$$

$$d_{rs}^2 = b_{rr} + b_{ss} - 2b_{rs}$$

ماتریس  $B$  به صورت زیر تعریف می‌شود:

$$B = [b_{ij}] = \mathbf{X}\mathbf{X}^T$$

قیدی برای مساله در نظر گرفته می‌شود(بدون لطمه به کلیت):

$$\sum_{t=1}^N x_j^t = 0$$

یادگیری ماشین



# Classical Solution(Principal coordinate analysis)

$$d_{rs}^2 = b_{rr} + b_{ss} - 2b_{rs}$$

یا  $b_{rs} = \frac{1}{2}(b_{rr} + b_{ss} - d_{rs}^2)$

در صورتی که T به صورت زیر تعریف شود:

$$T = \sum_{t=1}^N b_{tt} = \sum_t \sum_j (x_j^t)^2$$

خواهیم داشت:

$$\sum_r d_{rs}^2 = T + Nb_{ss} \quad \sum_s d_{rs}^2 = Nb_{rr} + T \quad \sum_r \sum_s d_{rs}^2 = 2NT$$

$d_{\bullet s}^2 = \frac{1}{N} \sum_r d_{rs}^2 = \frac{1}{N} \mathbf{D} \mathbf{J}$  و تعاریف زیر:

$$\mathbf{D} = [d_{ij}^2]$$

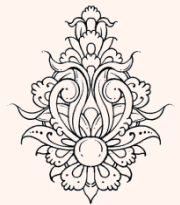
$d_{r \bullet}^2 = \frac{1}{N} \sum_s d_{rs}^2 = \frac{1}{N} \mathbf{J} \mathbf{D}$

$$\mathbf{J} = [1]$$

$d_{\bullet \bullet}^2 = \frac{1}{N^2} \sum_r \sum_s d_{rs}^2 = \frac{1}{N^2} \mathbf{J} \mathbf{D} \mathbf{J}$

در نتیجه:

$$b_{rs} = \frac{1}{2}(d_{r \bullet}^2 + d_{\bullet s}^2 - d_{\bullet \bullet}^2 - d_{rs}^2)$$



$$b_{rs} = \frac{1}{2} (d_{r\bullet}^2 + d_{\bullet s}^2 - d_{\bullet\bullet}^2 - d_{rs}^2)$$

نمایش به صورت ماتریسی

$$[\mathbf{B}]_{ij} = \frac{1}{2} \left( \left[ \frac{1}{N} \mathbf{D}\mathbf{J} \right]_{ij} + \left[ \frac{1}{N} \mathbf{J}\mathbf{D} \right]_{ij} - \left[ \frac{1}{N^2} \mathbf{D}\mathbf{J}\mathbf{D} \right]_{ij} - [\mathbf{D}]_{ij} \right)$$

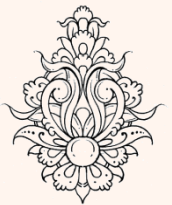
در نتیجه خواهیم داشت:

$$\mathbf{X}_{N \times d} \mathbf{X}_{d \times N}^T = \mathbf{B} = \left( \mathbf{I} - \frac{1}{N} \mathbf{J} \right) \left( -\frac{1}{2} \mathbf{D} \right) \left( \mathbf{I} - \frac{1}{N} \mathbf{J} \right)$$

در صورتی که بتوان  $\mathbf{Z}$  را به گونه‌ای یافت که:

$$\mathbf{B} = \mathbf{Z}_{N \times k} \mathbf{Z}_{k \times N}^T$$

می‌توان گفت که مسأله حل شده است!



$$\mathbf{B} = \mathbf{Z}_{N \times k} \mathbf{Z}_{k \times N}^T$$

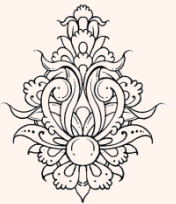
با تجزیه طیفی ماتریس  $\mathbf{B}$

$$\begin{aligned} \mathbf{B} &= \mathbf{C} \mathbf{\Lambda} \mathbf{C}^T = (\mathbf{C} \mathbf{\Lambda}^{1/2}) (\mathbf{\Lambda}^{1/2} \mathbf{C}^T) \\ &= (\mathbf{C} \mathbf{\Lambda}^{1/2}) (\mathbf{C} \mathbf{\Lambda}^{1/2})^T \end{aligned}$$

در صورتی که رتبه ماتریس  $\mathbf{\Lambda}$  برابر با  $k$  باشد:

$$\mathbf{Z} = (\mathbf{C} \mathbf{\Lambda}^{1/2}) \quad \mathbf{Z} = (\mathbf{C}_{N \times N} \mathbf{\Lambda}_{N \times k}^{1/2})$$

داده‌ها به فضای  $k$ -بعدی نگاشت شده‌اند، در صورتی که رتبه ماتریس بیشتر باشد، با حذف ابعاد متناظر با مقادیر ویژه‌ی کمتر تقریب مناسب به دست خواهد آمد.



# مثال

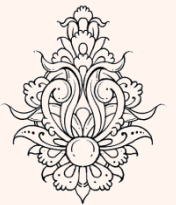
• فاصله‌ی اقلیدسی پنج نمونه به صورت زیر است:

$$\begin{bmatrix} 0 & 2\sqrt{2} & 2\sqrt{2} & 2\sqrt{2} & 2\sqrt{2} \\ 2\sqrt{2} & 0 & 4 & 4\sqrt{2} & 4 \\ 2\sqrt{2} & 4 & 0 & 4 & 4\sqrt{2} \\ 2\sqrt{2} & 4\sqrt{2} & 4 & 0 & 4 \\ 2\sqrt{2} & 4 & 4\sqrt{2} & 4 & 0 \end{bmatrix}$$

• در این صورت

$$\mathbf{D} = \begin{bmatrix} 0 & 8 & 8 & 8 & 8 \\ 8 & 0 & 16 & 32 & 16 \\ 8 & 16 & 0 & 16 & 32 \\ 8 & 32 & 16 & 0 & 16 \\ 8 & 16 & 32 & 16 & 0 \end{bmatrix}$$

$$\mathbf{B} = \left( \mathbf{I} - \frac{1}{N} \mathbf{J} \right) \left( -\frac{1}{2} \mathbf{D} \right) \left( \mathbf{I} - \frac{1}{N} \mathbf{J} \right)$$



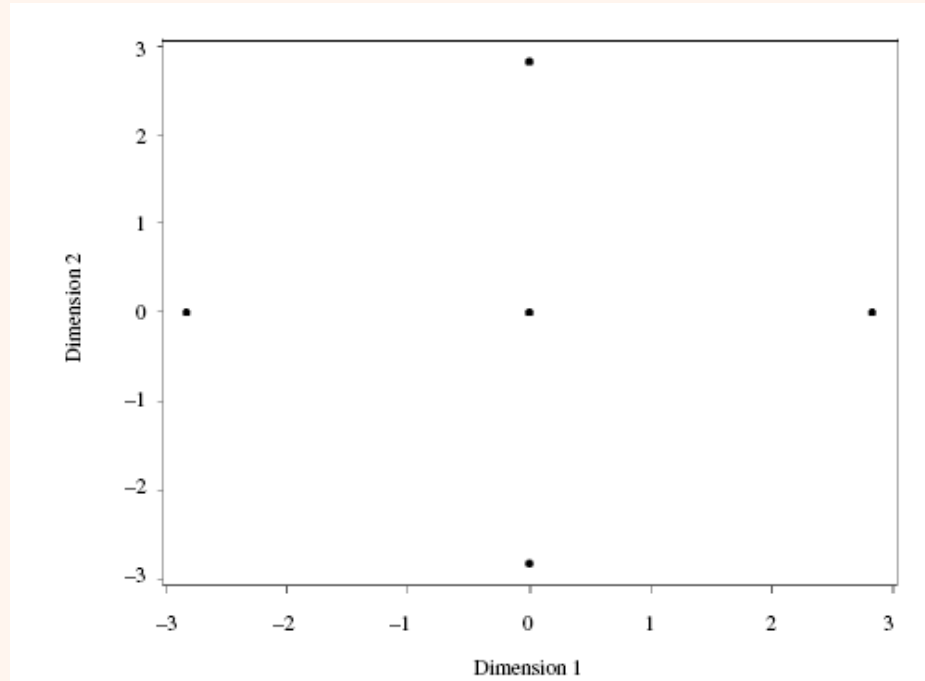
## مثال (ادامه...)

$$\mathbf{B} = \left( \mathbf{I} - \frac{1}{N} \mathbf{J} \right) \left( -\frac{1}{2} \mathbf{D} \right) \left( \mathbf{I} - \frac{1}{N} \mathbf{J} \right)$$

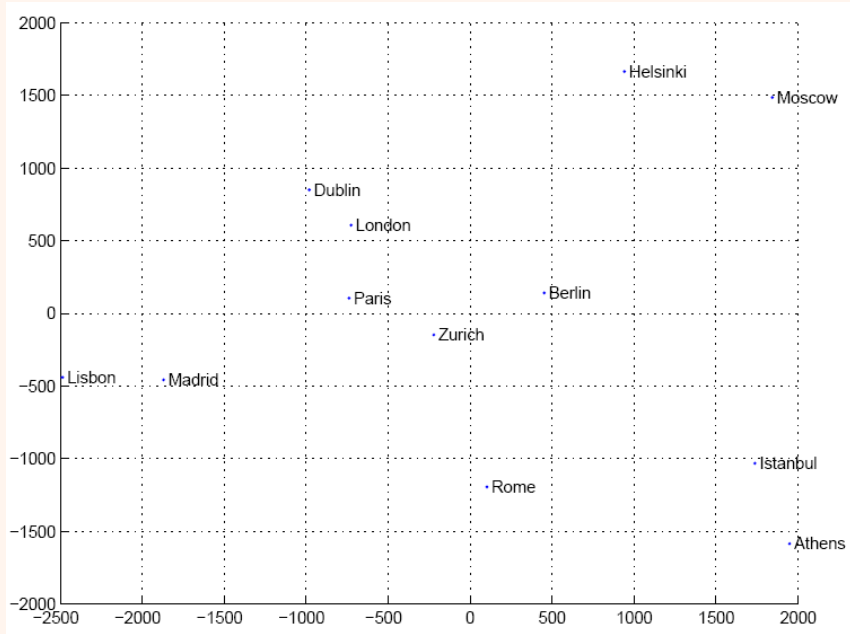
رتبه‌ی ماتریس دو است

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 8 & 0 & -8 & 0 \\ 0 & 0 & 8 & 0 & -8 \\ 0 & -8 & 0 & 8 & 0 \\ 0 & 0 & -8 & 0 & 8 \end{bmatrix}$$

$$\mathbf{Z} = \begin{bmatrix} 0 & 0 \\ 2\sqrt{2} & 0 \\ 0 & 2\sqrt{2} \\ -2\sqrt{2} & 0 \\ 0 & -2\sqrt{2} \end{bmatrix}$$



# Map of Europe by MDS



تراشک  
بهشتی

Map from CIA – The World Factbook: <http://www.cia.gov/>