Machine Learning

کاهش ابعاد

Dimensionality Reduction

یادگیری ماشین
(۱۳-۱۱-۸۰۵-۰۱)
فصل ششم

دانشگاه شهید بهشتی

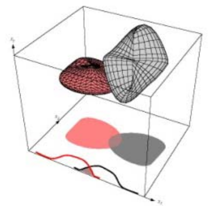
دانشکده‌ی مهندسی برق و کامپیوتر

پاییز ۱۳۹۳

احمد محمودی ازناوه

فهرست مطالب

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



کاهش 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 فرآیندی دسته‌بندی یا رگرسیون 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(ادامه...)

- «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 داده‌های آموزشی کم‌حجم باشد، «مقاوم‌تر» می‌باشند.
- «استخراج دانش»: با تعداد خصیصه‌های کم‌تر، در مورد داده‌ها و فرآیندهای مربوط به آن درک بهتری وجود خواهد داشت.
- «ساختار داده‌ها» هنگامی که تعداد خصیصه‌ها کم‌تر باشد، بهتر درک می‌شود، داده‌های پرت و غیرمعمول بهتر تشخیص داده می‌شود.

انتخاب–استخراج خصیصه

Feature Selection vs Extraction

- انتخاب خصیصه:
  - K خصیصه‌ی مهمتر (k<d) انتخاب می‌شود.
  - الگوریتم‌های انتخاب زیرمجموعه
- استخراج خصیصه:
  - K خصیصه‌ی جدید، انتخاب 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 مجوعه‌ی d-عضوی وجود دارد.
- جستجوی رو به جلو: **Forward search**
  - 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.

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

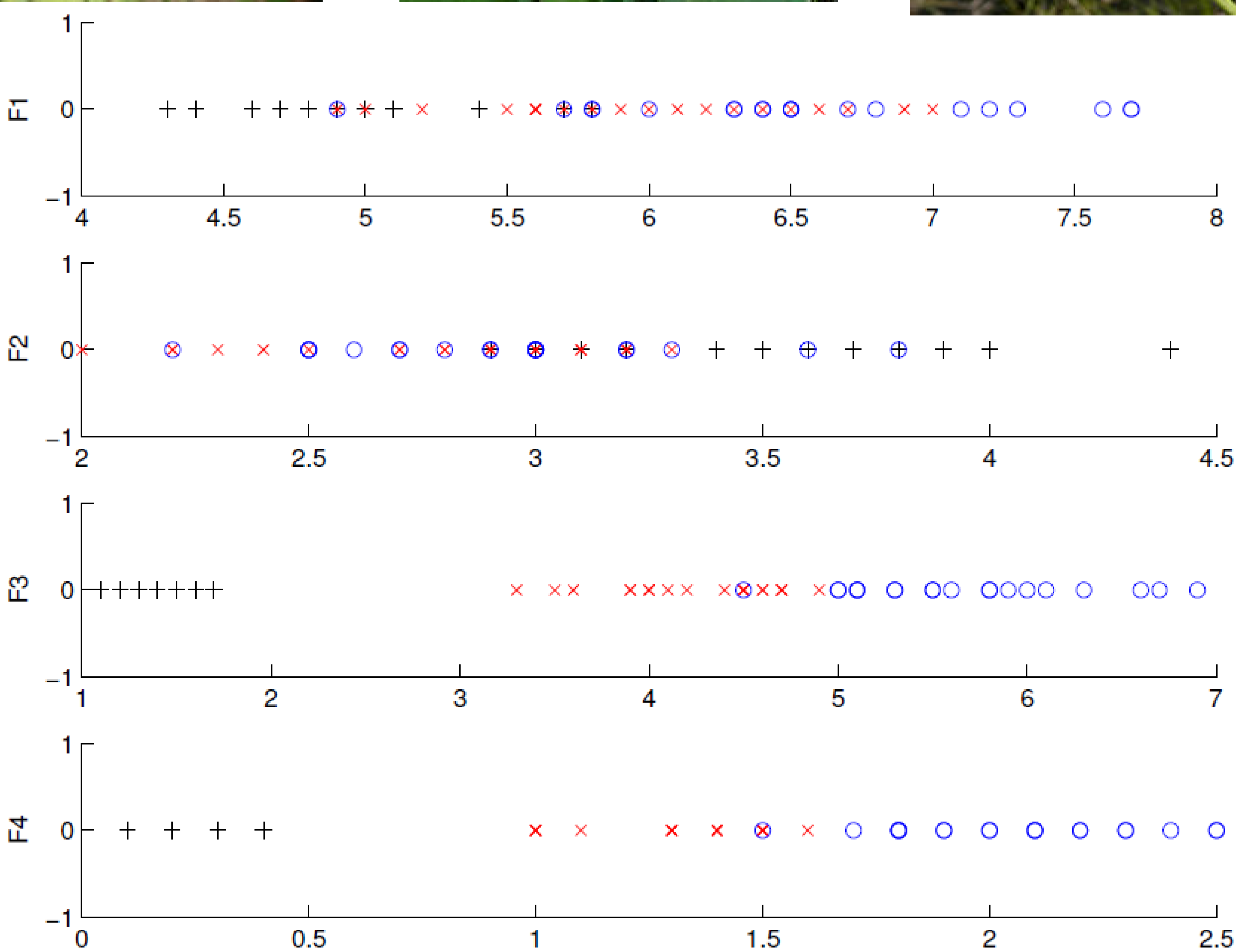
انتخاب زیرمجموعه (ادامه...)

Backward search

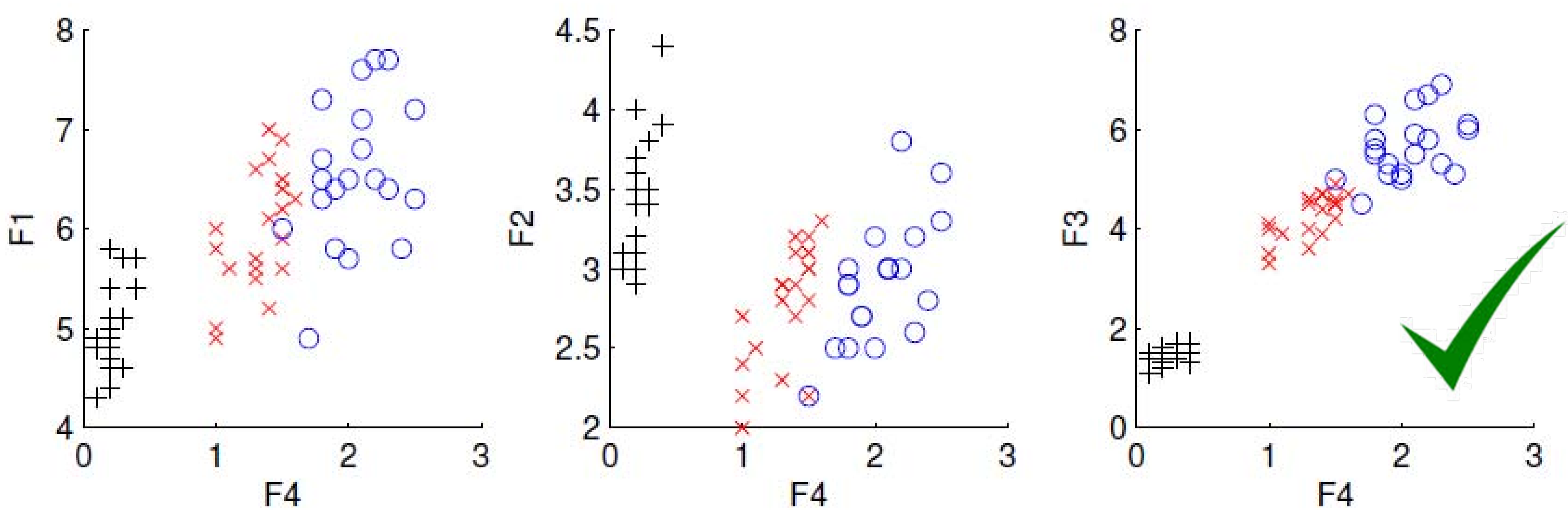
- جستجوی رو به عقب:
  - در گام نخست، مجموعه‌ی خصیصه‌ها، F در حالت اولیه تمامی خصیصه‌ها در نظر گرفته می‌شود.
  - در هر گام بدترین خصیصه از مجموعه‌ی خصیصه‌ها حذف می‌شود.
- هنگامی که تعداد خصیصه‌ها زیاد است، روش جستجوی رو به جلو ترجیح داده می‌شود.
- در کاربردهایی که یک خصیصه به تنهایی اطلاعات مفیدی ندارد،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

Iris data: Single feature

مثال



Iris data: Add one more feature to F4



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**:

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

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

تحلیل مؤلفه‌های اصلی(ادامه...)

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

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

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.

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 $Var(z_2)$ ماکزیمم شود، مشروط به متعامد بودن بر اولین مؤلفه‌ی اصلی و $||\mathbf{w}_2||=1$

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

تحلیل مؤلفه‌های اصلی (ادامه...)

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

- پس از مشتق گرفتن خواهیم داشت:

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

- با ضرب در $\mathbf{w}_1^T$

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

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

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

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

$$\Rightarrow \beta = 0$$

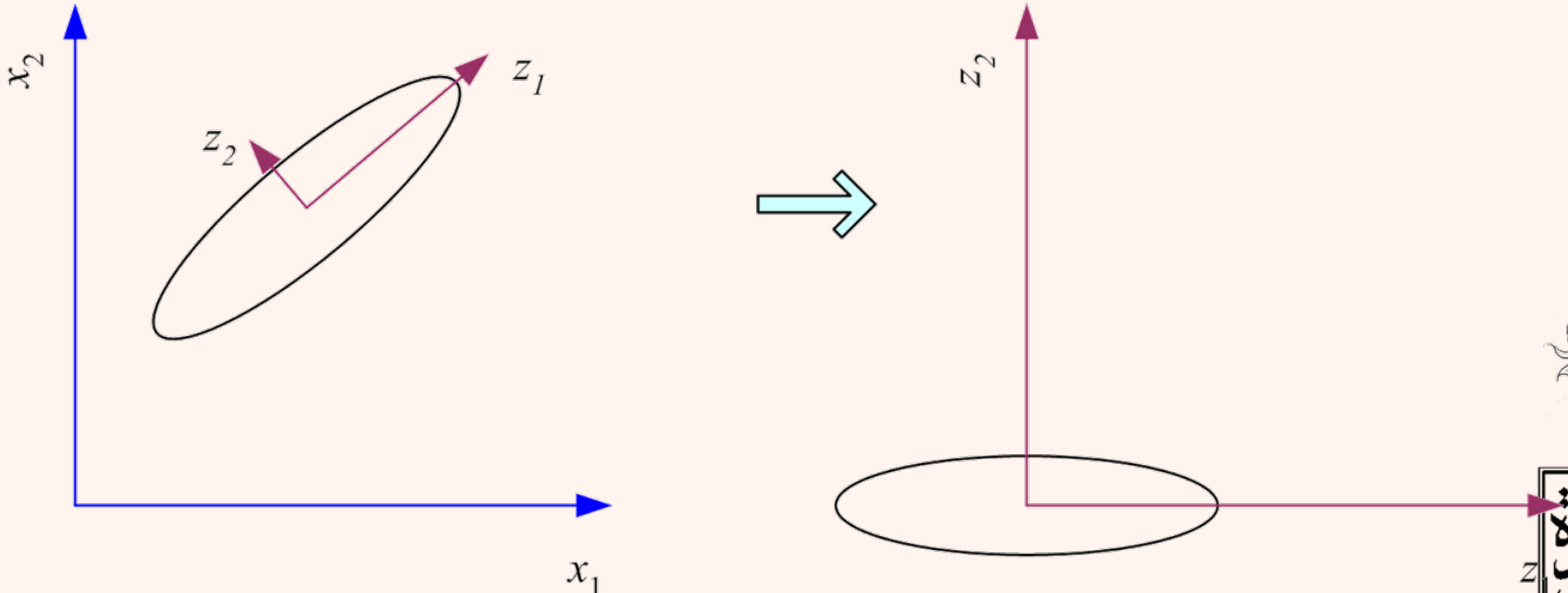
$$\Rightarrow \Sigma \mathbf{w}_2 = \alpha \mathbf{w}_2$$

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.

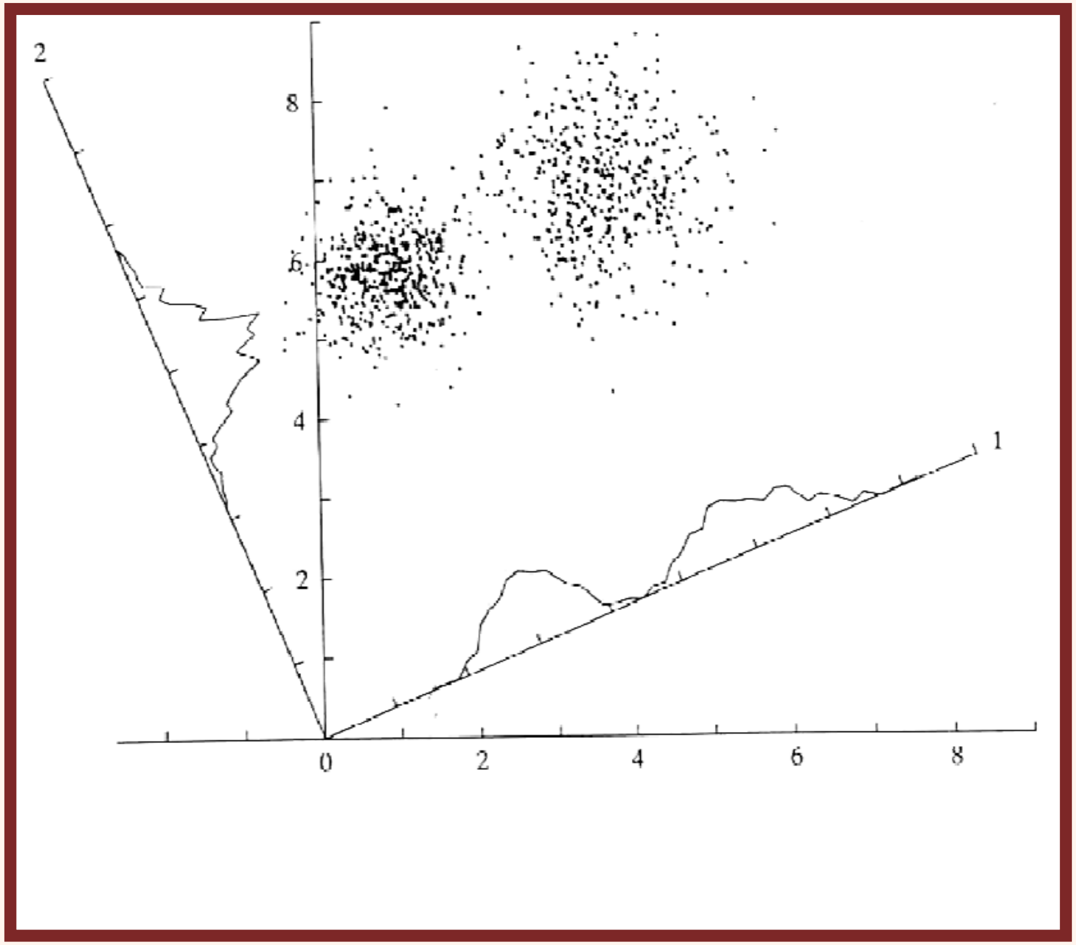
تحلیل مؤلفه‌های اصلی (ادامه...)

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

- ستون‌های $\mathbf{W}$، بردارهای ویژه‌ی ماتریس کوواریانس هستند.

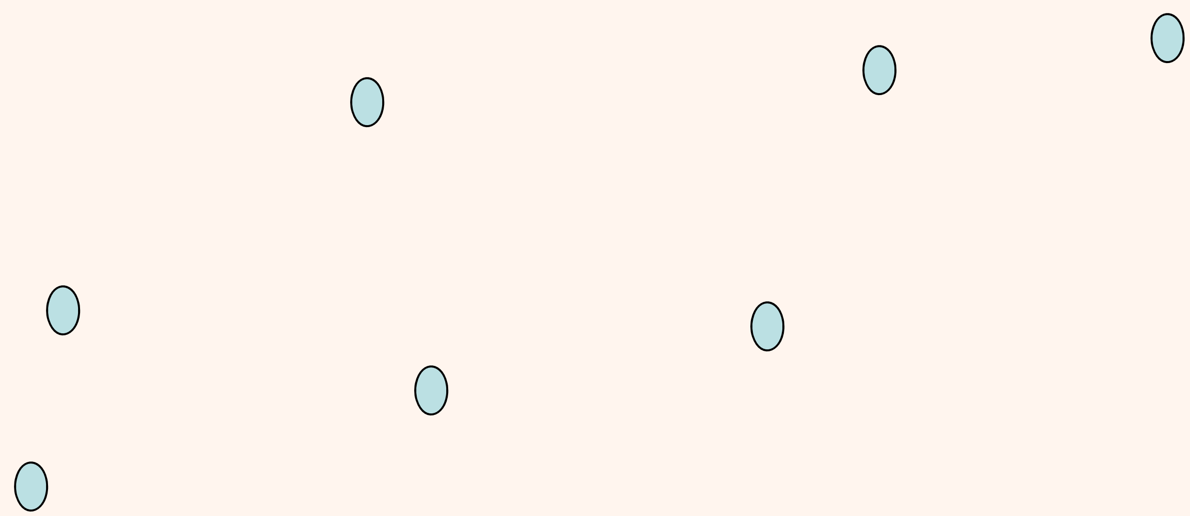


تحلیل مؤلفه‌های اصلی(ادامه...)



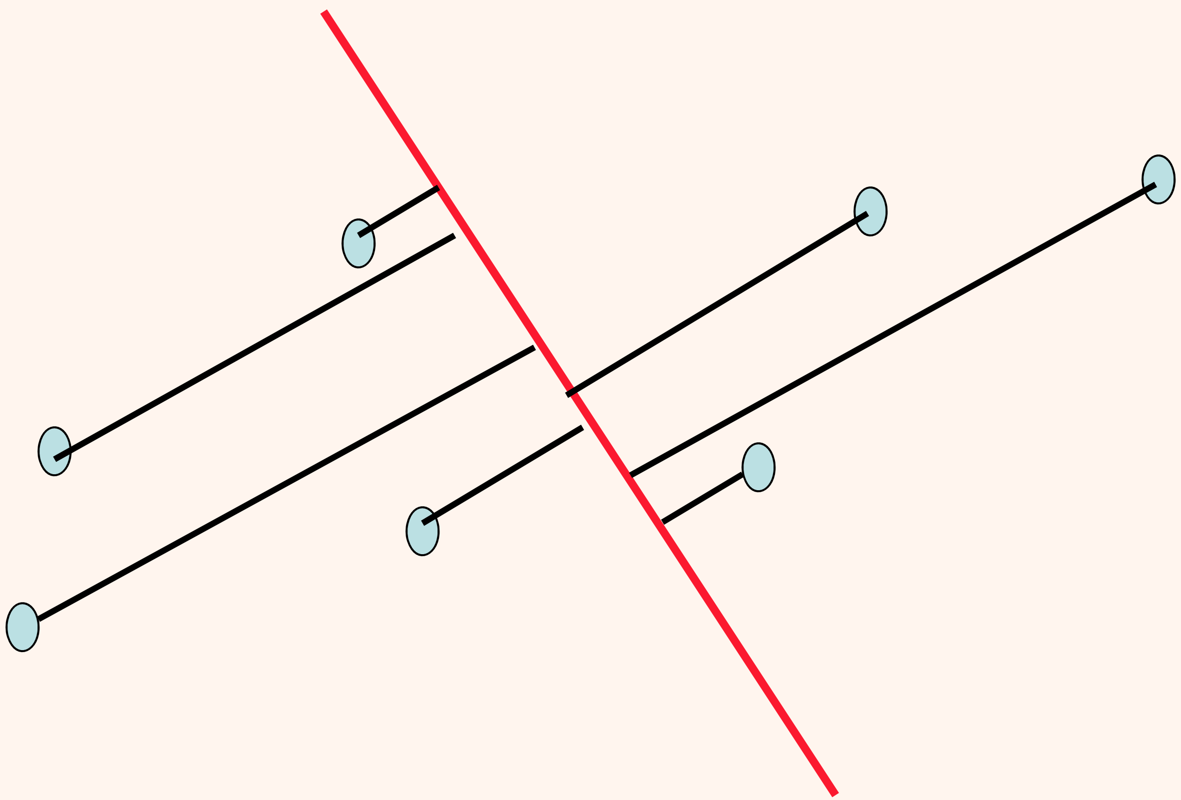
Haykin, S. Neural Networks: A Comprehensive Foundation,

تحلیل مؤلفه‌های اصلی(ادامه...)



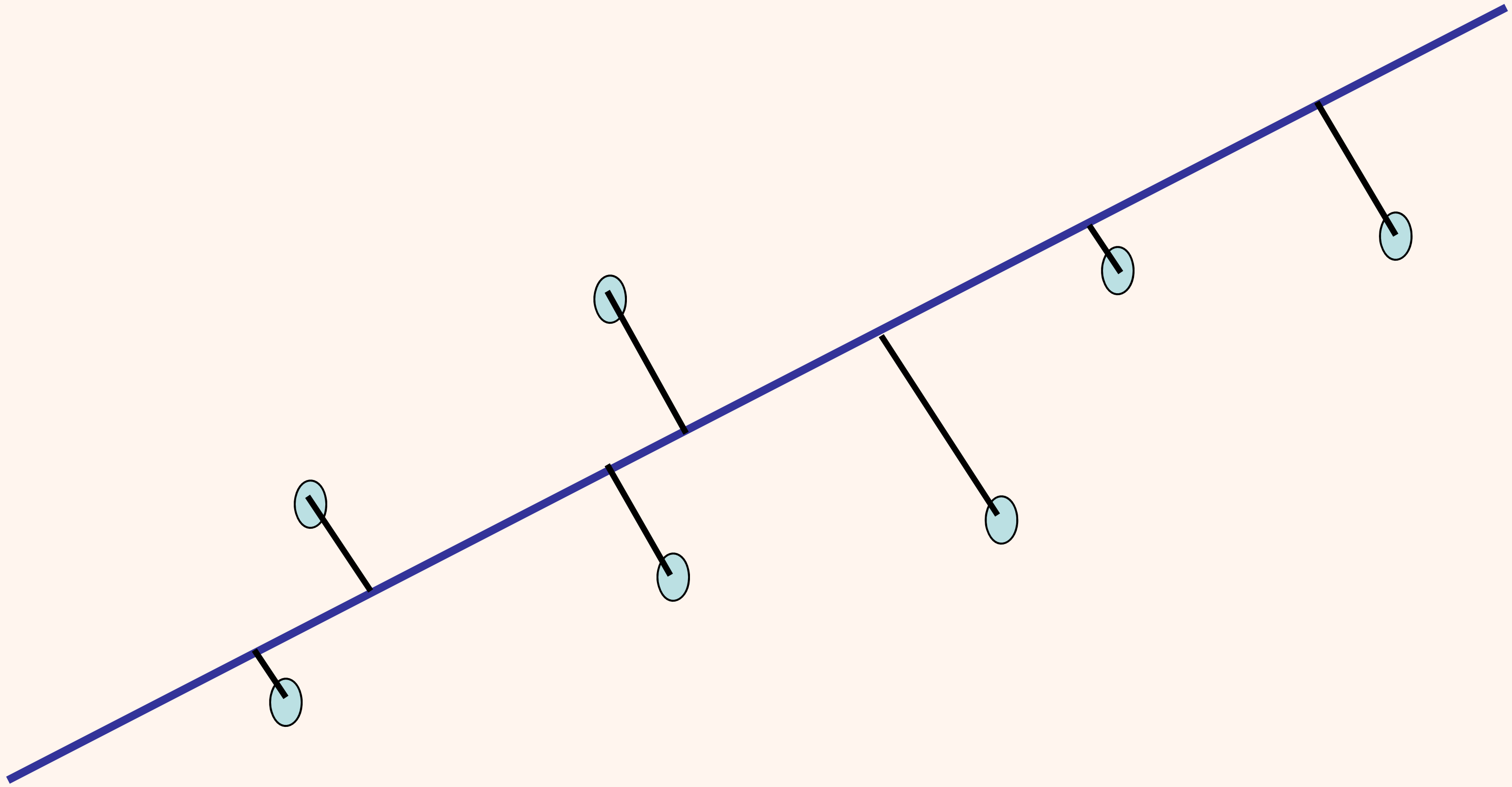
http://www.public.asu.edu/~jye02/

تحلیل مؤلفه‌های اصلی(ادامه...)



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

تحلیل مؤلفه‌های اصلی(ادامه...)



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

تحلیل مؤلفه‌های اصلی(ادامه...)

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

$$\mathbf{z} = \mathbf{W}^T \mathbf{x} \qquad \mathrm{Cov}(\mathbf{z}) = \mathbf{D}'$$

ماتریس قطری

- $C_{d \times d}$ 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...)

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

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

Spectral decomposition

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

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

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

$$\boxed{\mathrm{Cov}(\mathbf{z}) = \mathbf{D}}$$

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)

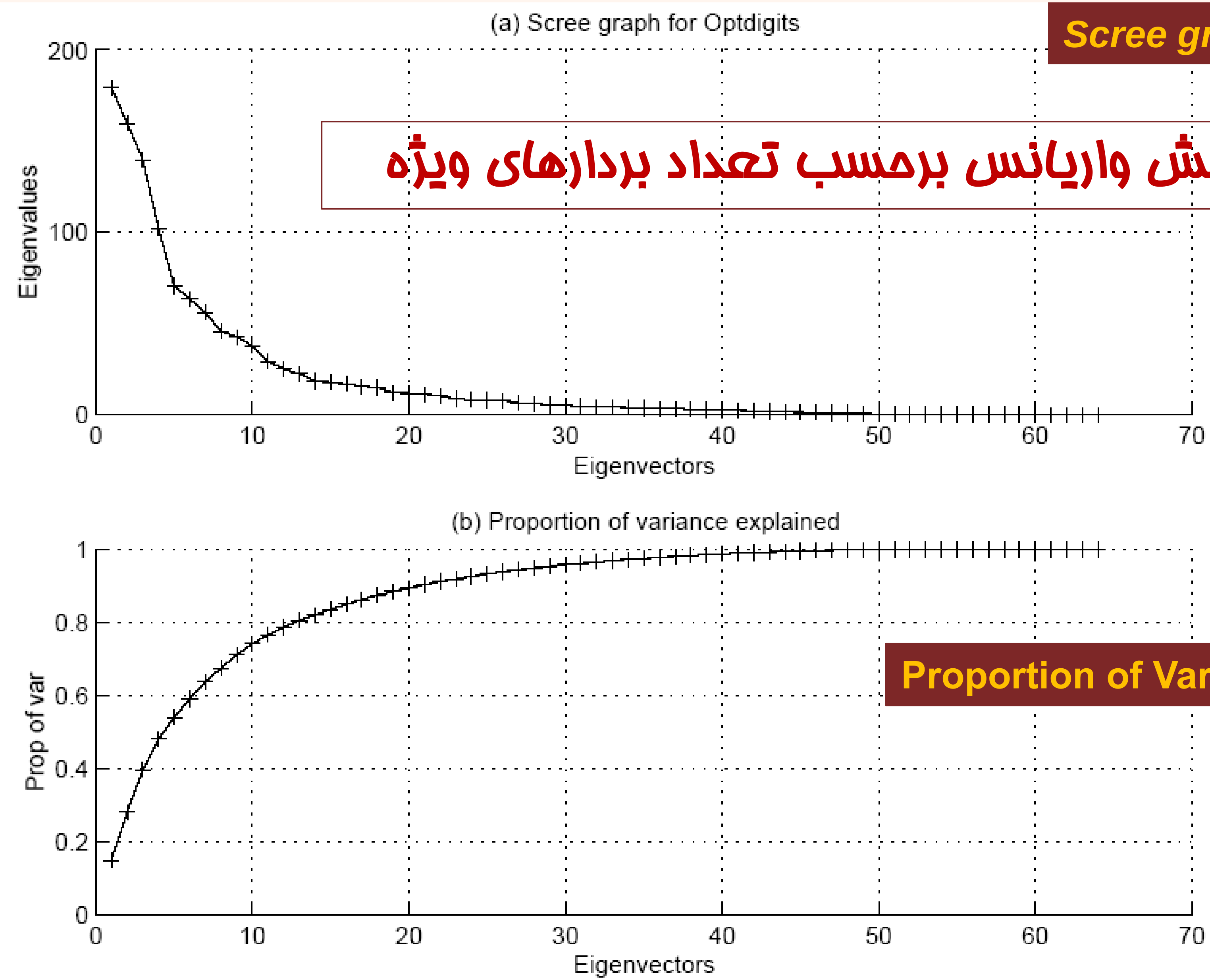
$$\frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_k+\cdots+\lambda_d}$$

PoV>0.9

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

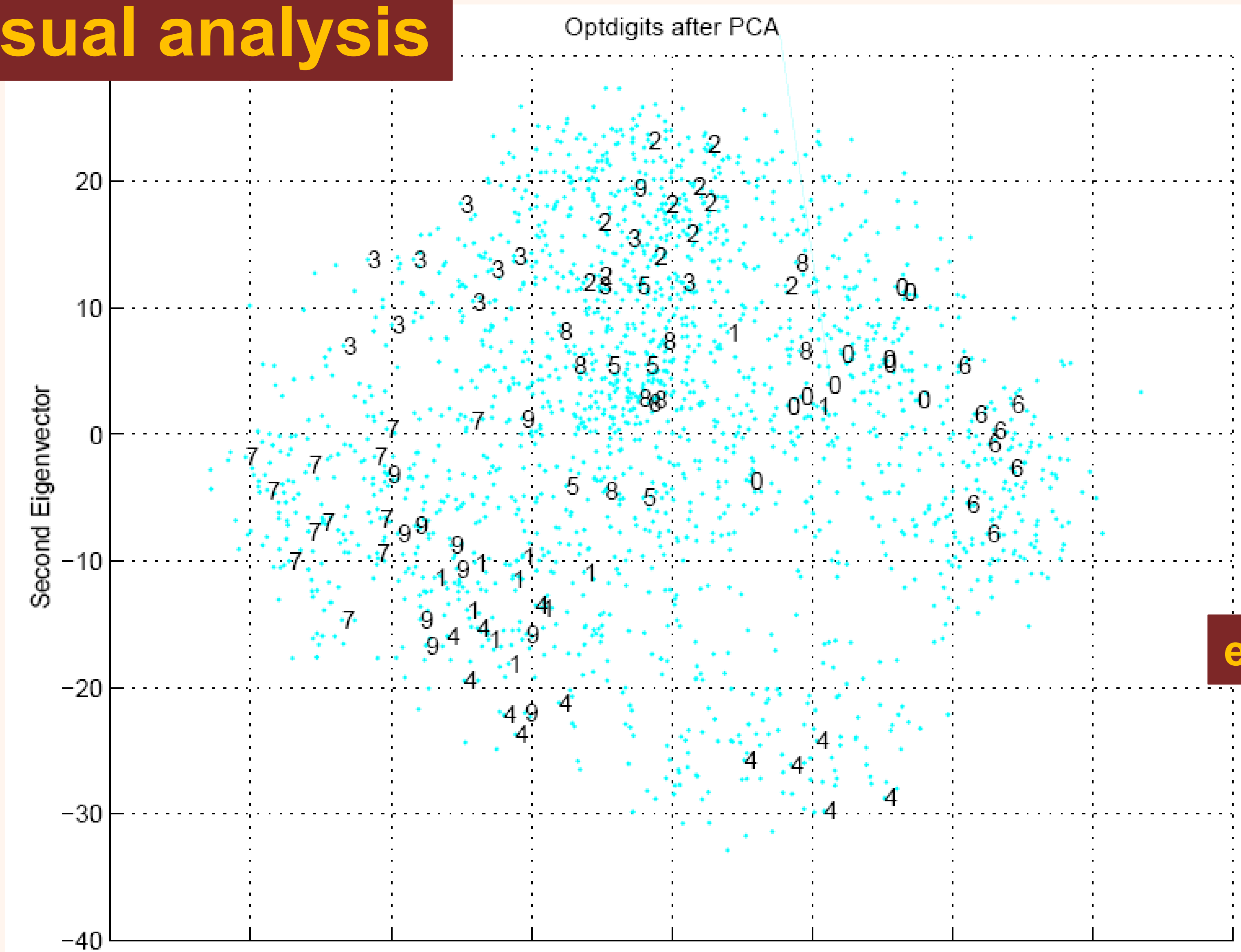
Scree graph

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.

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

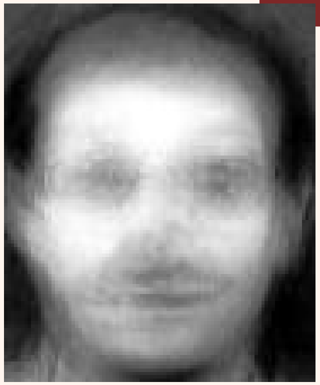
$$\sum_t \left\| \hat{\mathbf{x}}^t - \mathbf{x}^t \right\|$$

- در Karhunen-Loève expansion، برای هر کلاس یک ماتریس کوواریانس جداگانه در نظر گرفته می‌شود.
- در common principal components برای همه‌ی کلاس‌ها مؤلفه‌های اساسی یکسانی در نظر گرفته می‌شود، با این تفاوت که واریانس ابعاد متفاوت در نظر گرفته می‌شود.

$$S_i = \mathbf{C}\mathbf{D}_i\mathbf{C}^T$$

کاربرد PCA در شناسایی چهره

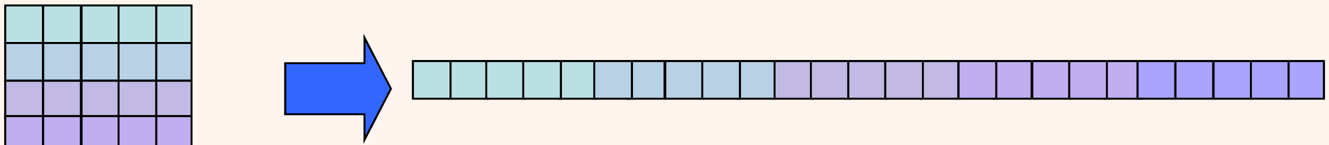
پایگاه داده‌ی ORL



میانگین چهره‌ها

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

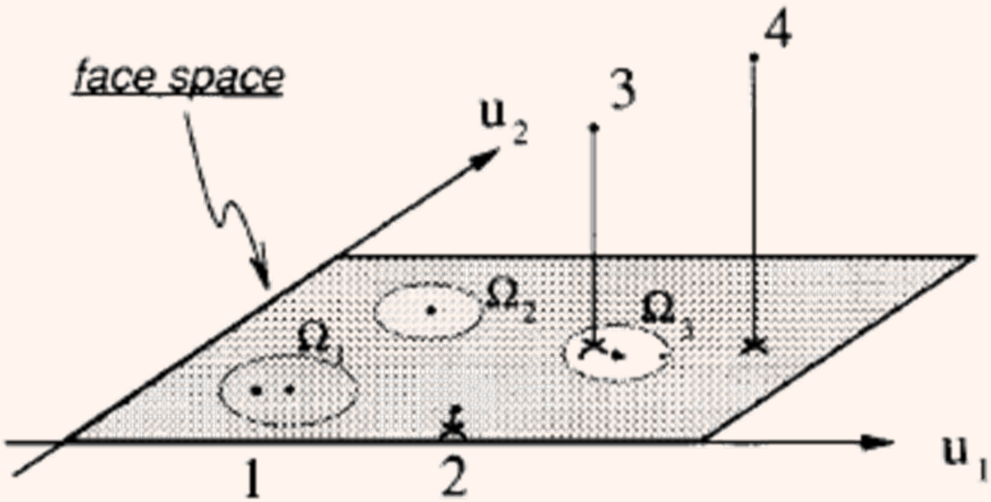
کاربرد PCA در شناسایی چهره(ادامه...)



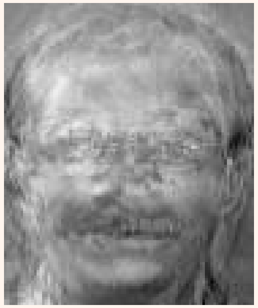
Eigenfaces

کاربرد PCA در شناسایی چهره(ادامه...)

تشخیص چهره



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



Feature Embedding

- ماتریس داده‌ها به صورت $\mathbf{X}_{N\times d}$ است.
- ماتریس کواریانس خصیصه‌ها $\mathbf{X}^T\mathbf{X}_{d\times d}$ می‌باشد، در نتیجه

$$\left(\mathbf{X}^T\mathbf{X}\right)\mathbf{w}_i = \lambda_i \mathbf{w}_i$$

- با ضرب طرفین در $\mathbf{X}$

$$\left(\mathbf{X}\mathbf{X}^T\right)\mathbf{X}\mathbf{w}_i = \lambda \mathbf{X}\mathbf{w}_i$$

- در نتیجه $\mathbf{X}\mathbf{w}_\mathrm{i}$ بردار ویژه‌ی $\mathbf{X}\mathbf{X}^T_{N\times N}$ با مقدار ویژه‌ی λ_i است.
- این ماتریس شباهت دوبه‌دو نمونه‌ها را نشان می‌دهد.

Feature Embedding

- در این حالت بردار ویژه، مختصات نمونه‌ها در راستای $\mathbf{w}_i$ را نشان می‌دهد.

$$\left(\mathbf{X}\mathbf{X}^T\right)\mathbf{X}\mathbf{w}_i = \lambda\mathbf{X}\mathbf{w}_i$$

- ثابت می‌شود، رتبه‌ی ماتریس حداکثر $\min(d, N)$ می‌باشد.
- برای یک پایگاه داده حاوی چهل تصویر ۲۵۶×۲۵۶
  – ماتریس کواریانس خصیصه‌ها ۶۵۵۳۶×۶۵۵۳۶ خواهد بود.
  – در حالی که ماتریس کواریانس نمونه‌ها ۴۰×۴۰ می‌باشد.

Factor Analysis

تحلیل عاملی

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

latent factors

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

factor loadings

noise sources

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

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

تحلیل عاملی(ادامه...)

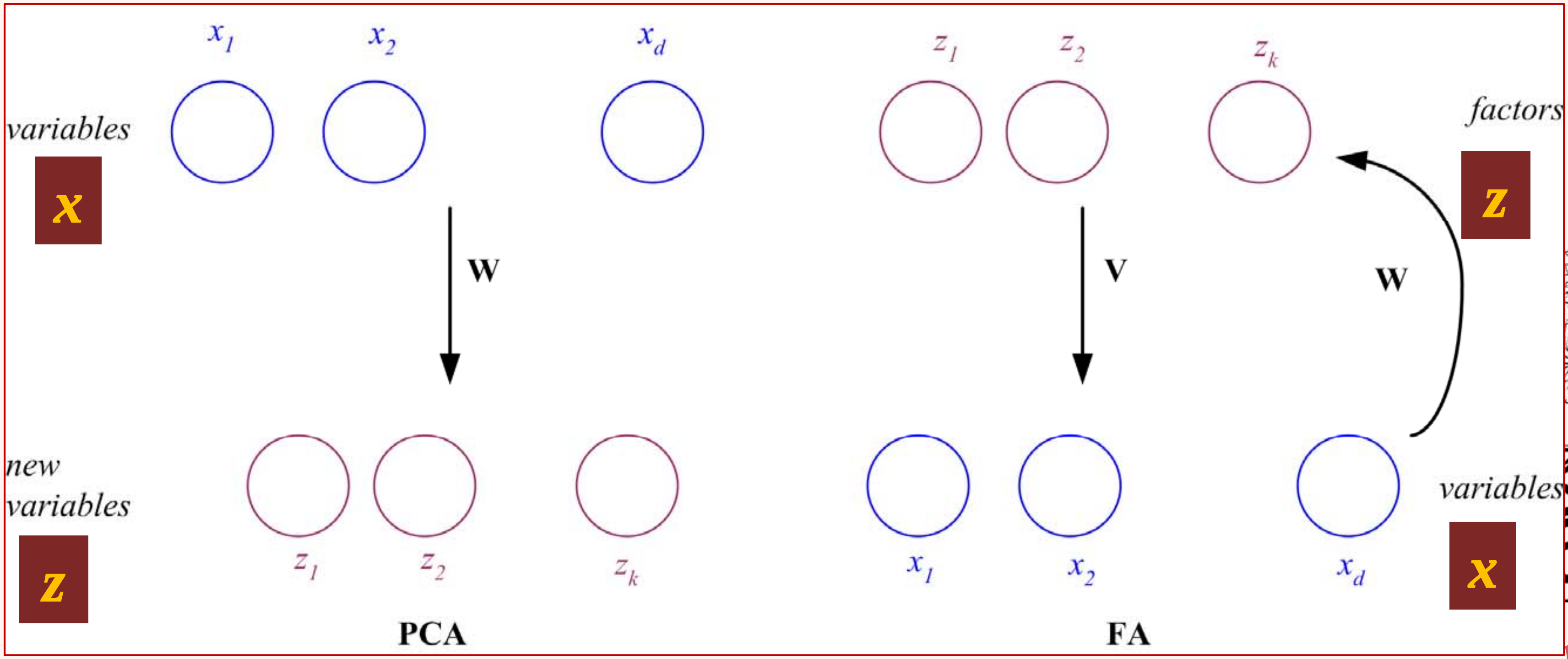
- در واقع این طور فرض می‌شود که مجموعه متغیرهایی که همبستگی بالایی با یکدیگر دارند و همبستگی آن‌ها با سایر متغیرها پایین است، دارای عوامل مشترکی هستند. با استفاده از تحلیل عاملی متغیرها بدین‌ترتیب «خوشه‌بندی» می‌شوند.

Factor clusters

- تحلیل عاملی مانند PCA، «بی‌نظارت» است.

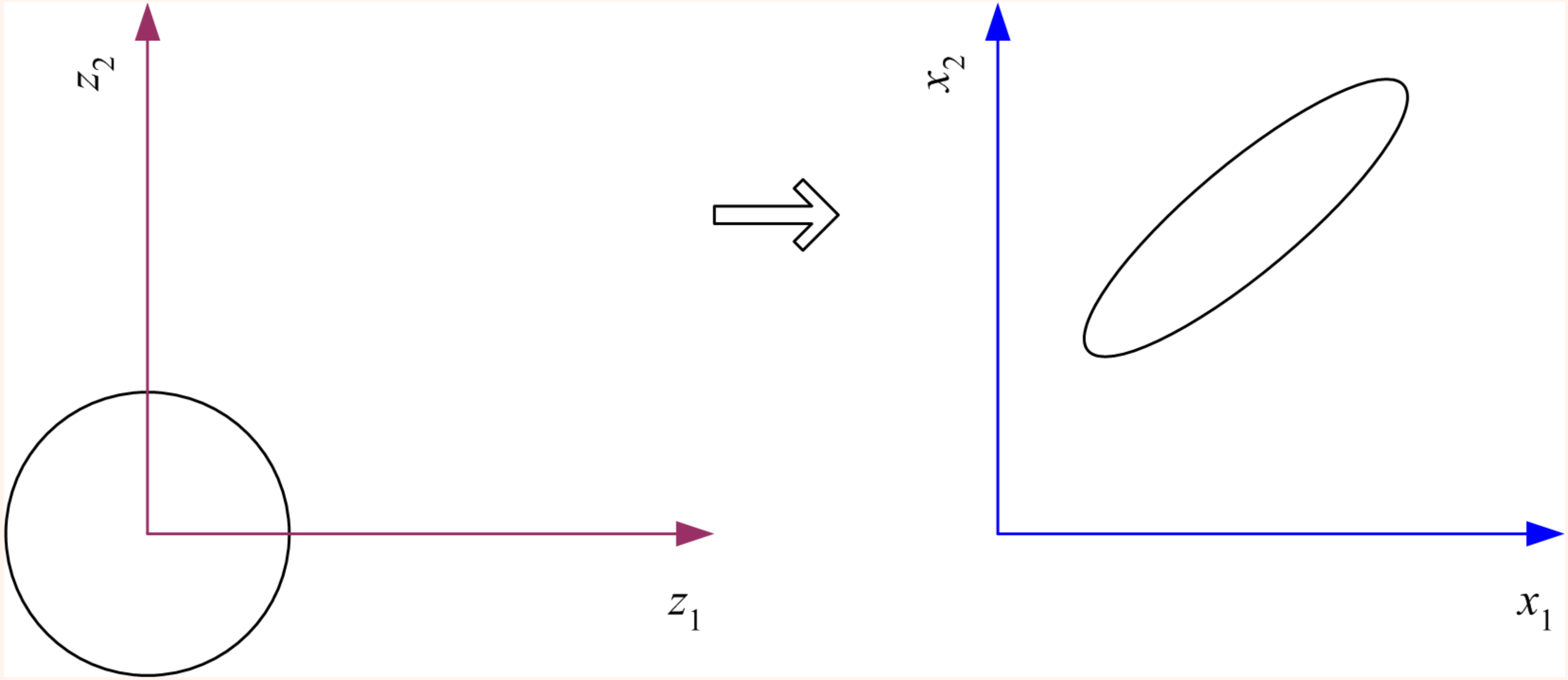
PCA vs FA

- PCA From $\boldsymbol{x}$ to $\boldsymbol{z}$ $\boldsymbol{z} = \mathbf{W}^T(\boldsymbol{x} - \boldsymbol{\mu})$
- FA From $\boldsymbol{z}$ to $\boldsymbol{x}$ $\boldsymbol{x} - \boldsymbol{\mu} = \mathbf{V}\boldsymbol{z} + \boldsymbol{\varepsilon}$



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، μ=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

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

واریانس مربوط به عوامل مشترک

تحلیل عاملی(ادامه...)

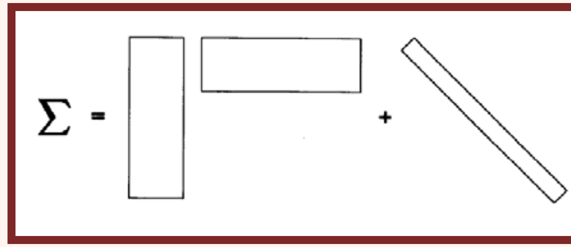
$$\Sigma = \mathrm{Cov}(\mathbf{x}) = \mathrm{Cov}(\mathbf{V}\mathbf{z} + \boldsymbol{\varepsilon})$$

$$= \mathrm{Cov}(\mathbf{V}\mathbf{z}) + \mathrm{Cov}(\boldsymbol{\varepsilon})$$

$$= \mathbf{V}\mathrm{Cov}(\mathbf{z})\mathbf{V}^T + \boldsymbol{\psi} \qquad \boxed{\mathrm{Cov}(\mathbf{z}) = \mathbf{I}}$$

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

ماتریس قطری



ALVIN C. RENCHER

با فرض داشتن دو عامل

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

- 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:

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

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

loading، همبستگی متغیرها با فاکتورها را نشان می‌دهند

تحلیل عاملی(ادامه...)

Principal Component Method

- با در اختیار داشتن تخمین ماتریس کواریانس

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

- در صورتی که از $\boldsymbol{\psi}$ صرفنظر کنیم:

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

- با تجزیه‌ی طیفی S

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

- در نتیجه

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

- و مقادیر ψ_i

$$\psi_i = s_i^2 - \sum_{j=1}^{k} v_{ij}^2 \qquad \boxed{\mathrm{Var}(x_i) = v_{i1}^2 + v_{i2}^2 + \ldots + v_{ik}^2 + \psi_i}$$

تحلیل عاملی(ادامه...)

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

$$(\mathbf{V}T)(\mathbf{V}T)^T = \mathbf{VTT}^T\mathbf{V}^T = \mathbf{VV}^T = \mathbf{S}$$

- بدین‌ترتیب مشاهده می‌شود که حل به دست آمده یکتا نیست.
  - ضرب در یک ماترس متعامد فاصله از مبدا را تغییر نمی‌دهد. تنها باعث چرخش محورها می‌شود.
  - بدین‌ترتیب می‌توان با این کار مناسب‌ترین فاکتورها را یافت.

کاهش بعد با استفاده از تحلیل عوامل

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

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

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

- 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

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

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

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

$$\boxed{\begin{aligned} \mathbf{x}_{d\times 1} &= \mathbf{V}_{d\times k} \mathbf{z}_{k\times 1} \\ \mathbf{x}_{d\times 1} \mathbf{z}^T_{1\times k} &= \mathbf{V}_{d\times k} \\ \frac{\mathbf{X}_{d\times N} \mathbf{Z}^T_{N\times k}}{N-1} &= \mathbf{V}_{d\times k} \end{aligned}}$$

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

رگرسیون چندمتغیره‌ی خطی

$$\mathbf{X} = \begin{bmatrix} 1 & x_1^1 & x_2^1 & \cdots & x_d^1 \\ 1 & x_1^2 & x_2^2 & \cdots & x_d^2 \\ \vdots & & & & \\ 1 & x_1^N & x_2^N & \cdots & x_d^N \end{bmatrix}, \mathrm{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}, \quad \mathbf{r} = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^N \end{bmatrix}$$

$$\mathbf{A} = \left(\mathbf{D}^T\mathbf{D}\right) \qquad \mathrm{y} = \mathbf{D}^T\mathbf{r}$$

$$\mathbf{w} = \left(\mathbf{D}^T\mathbf{D}\right)^{-1}\mathbf{D}^T\mathbf{r}$$

$$X^T X W = X^T r \Rightarrow W = \left(X^T X\right)^{-1} X^T r$$

این مدل شبیه به مدلی است که برای رگرسیون چند جمله‌ای تک متغیره داشتیم.

$x_1=x,\ x_2=x^2,\ x_3=x^3,\ \ldots.\ x_k=x^k$

برای رگیرسیون چندجمله‌ای و چند متغیره نیز می‌توانیم به صورت مشابه عمل کنیم:

$z_1=x_1,\ z_2=x_2,\ z_3=x_1^2,\ z_4=x_2^2,\ z_5=x_1x_2$

تجزیه‌ی مقدارهای تکین

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

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

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

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

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

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

تجزیه‌ی مقدارهای تکین

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

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

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

MULTIDIMENSIONAL SCALING

- در این شیوه هدف کاهش ابعاد به نحوی است که حتی‌المقدور فاصله‌ی بین نمونه‌ها حفظ شود.
  - در صورتی که داشته باشیم:

d_{ij} در فضای اصلی؛ d-بعدی **فاصله‌ی بین نمونه‌ی i و j**

δ_{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 s و r

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

ماتریس فاصله‌ها به صورت روبرو تعریف می‌شود:

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

با بازنویسی روبط خواهیم داشت:

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

$$\boxed{b_{rs} = \sum_{j=1}^{d} x_j^r x_j^s}$$

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

ماتریس B به صورت زیر تعریف می‌شود:

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

ادامه

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

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

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

یا

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

در صورتی که T به صورت زیر تعریف شود:

$$T = \sum_{j=1}^{d} b_{tt} = \sum_t \sum_j \left(x_j^t\right)^2$$

خواهیم داشت:

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

و تعاریف زیر:

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

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

$$d_{\bullet s}^2 = \frac{1}{N} \sum_s d_{rs}^2 \quad = \frac{1}{N}\mathbf{JD}$$

$$d_{r\bullet}^2 = \frac{1}{N} \sum_r d_{rs}^2 \quad = \frac{1}{N}\mathbf{DJ}$$

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

در نتیجه:

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

ادامه

Classical Solution(Principal coordinate analysis)

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

نمایش به صورت ماتریسی

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

در نتیجه خواهیم داشت:

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

در صورتی که بتوان Z را به گونه‌ای یافت که:

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

می‌توان گفت که مساله حل شده است!

Classical Solution(Principal coordinate analysis)

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

با تجزیه‌ی طیفی ماتریس **Z**

$$\mathbf{B} = C\Lambda C^{T} = \left(C\Lambda^{1/2}\right)\left(\Lambda^{1/2}C^{T}\right)$$

$$= \left(C\Lambda^{1/2}\right)\left(C\Lambda^{1/2}\right)^{T}$$

در صورتی که رتبه‌ی ماتریس **Z** برابر با k باشد:

$$Z = \left(C\Lambda^{1/2}\right) \qquad Z = \left(C_{N\times N}\Lambda_{N\times k}^{1/2}\right)$$

داده‌ها به فضای k-بعدی نگاشت شده‌اند، در صورتی که رتبه‌ی 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

$$\mathbf{D} = \begin{bmatrix} 0 & 8 & 8 & 8 & 8 \\ 8 & 0 & 16 & 32 & 16 \\ 8 & 16 & 0 & 16 & 32 \\ 8 & 32 & 16 & 0 & 16 \\ 8 & 16 & 32 & 16 & 0 \end{bmatrix} \qquad \mathbf{B} = \left(\mathbf{I} - \frac{1}{N}\mathbf{J}\right)\left(-\frac{1}{2}\mathbf{D}\right)\left(\mathbf{I} - \frac{1}{N}\mathbf{J}\right)$$

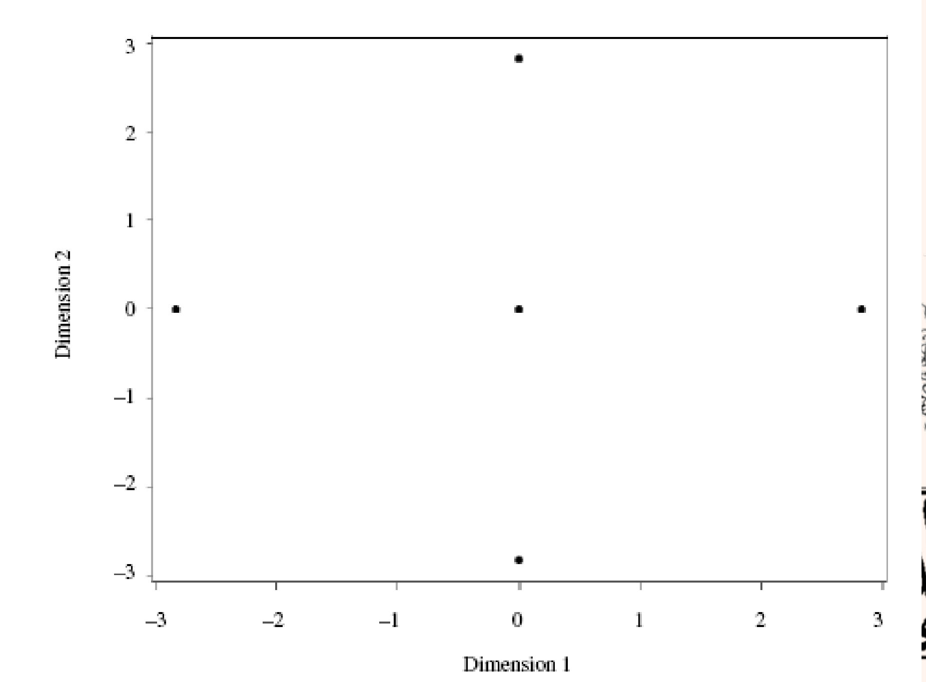
مثال (ادامه...)

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

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

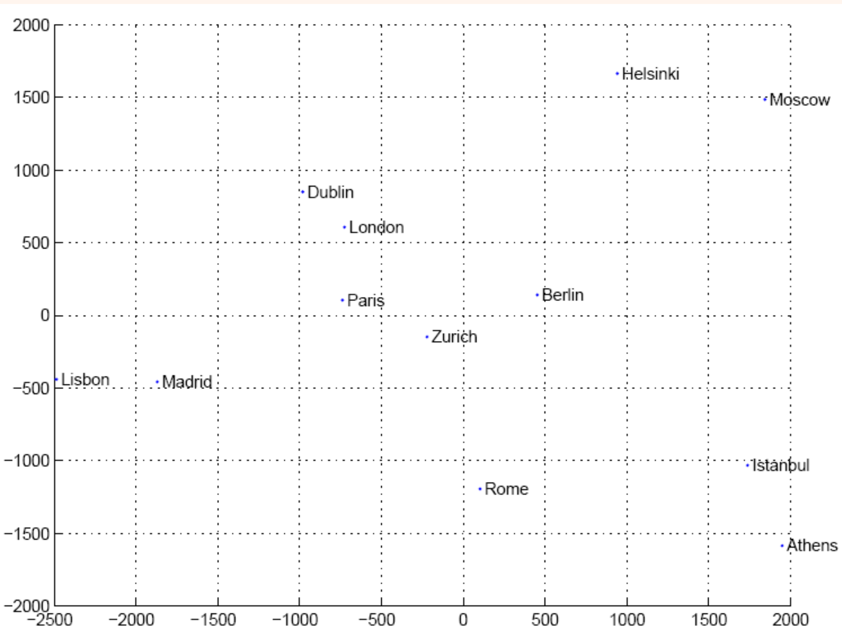
رتبه‌ی ماتریس دو است

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



Rencher, A. C. (2003). Methods of Multivariate Analysis, Wiley.

Map of Europe by MDS



به صورت کلی می‌توان به این مساله به صورت رگرسیون نگاه کرد: $z = g(x \mid \theta)$
و از شیوه‌های غیرخطی بهره جست.

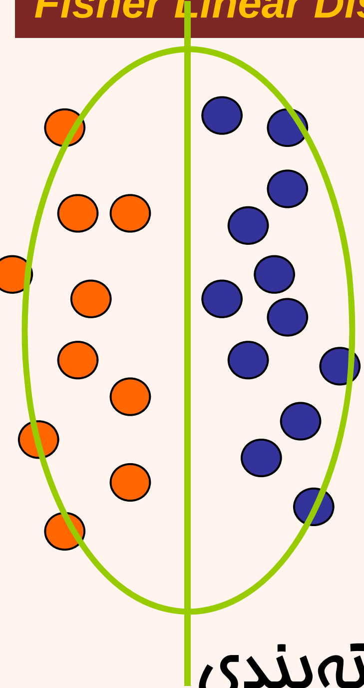
Map from CIA – The World Factbook: http://www.cia.gov/

تحلیل تفکیک خطی

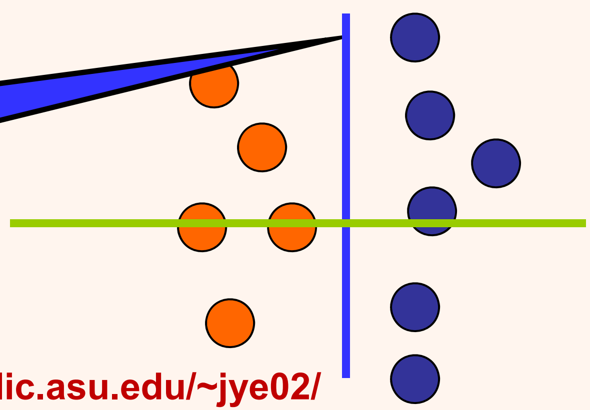
Fisher Linear Discriminant Analysis

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

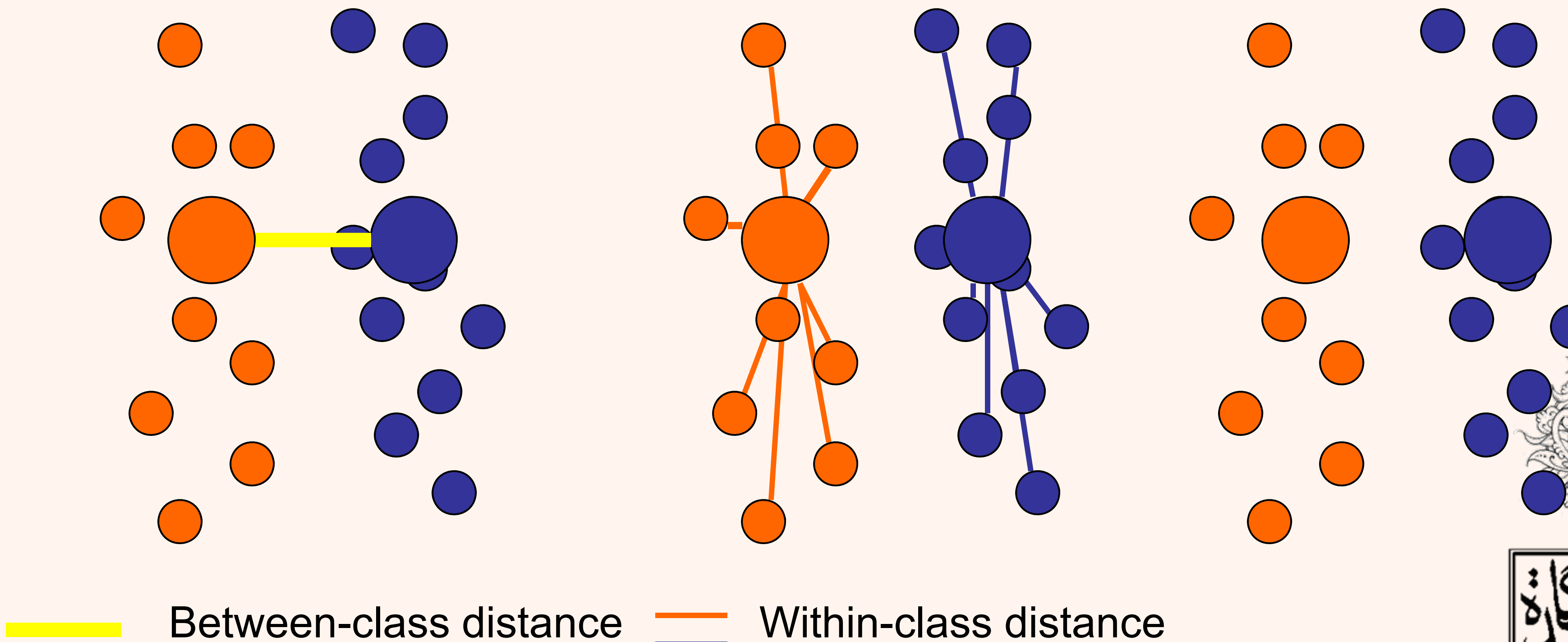
در این راستا دو کلاس بدون خطا طبقه‌بندی می‌شوند

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...)

دسته‌بندی دو کلاسه

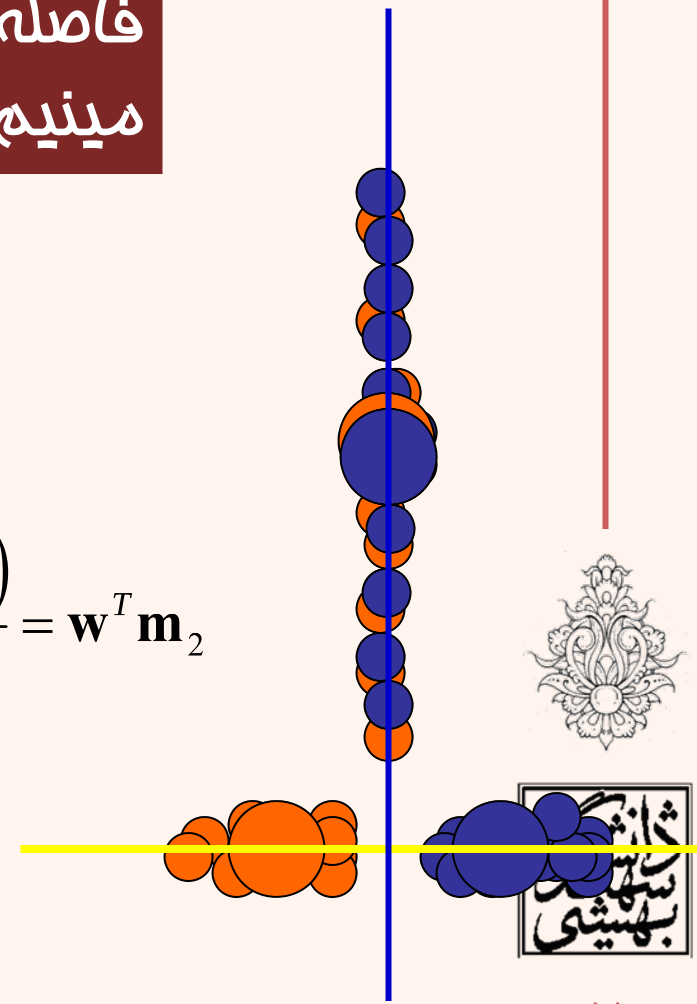
در LDA، نگاشت به گونه‌ای انجام می‌شود که فاصله‌ی بین دو کلاس حداکثر شده و فاصله‌ی نمونه‌های متعلق به یک کلاس مینیمم گردد.

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

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

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

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



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

تحلیل تفکیک خطی (ادامه...)

دسته‌بندی دو کلاسه

$$(m_1 - m_2)^2 = \left(\mathbf{w}^T\mathbf{m}_1 - \mathbf{w}^T\mathbf{m}_2\right)^2$$

$$= \mathbf{w}^T\boxed{(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T}\mathbf{w}$$

$$= \mathbf{w}^T\mathbf{S}_B\mathbf{w}$$

Between class scatter matrix

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

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

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

$$= \sum_t \mathbf{w}^T\mathbf{S}_1\mathbf{w}r^t$$

Class scatter matrix for C1

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

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

$$\boxed{\mathbf{S}_W = S_1 + S_2}$$

Total whitin Class scatter

تحلیل تفکیک خطی (ادامه...)

- هدف ماکزیمیم کردن رابطه‌ی زیر است:

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

- با مشتق گرفتن:

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

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

یادآوری: جداساز خطی

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

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

بدین‌ترتیب، برای کلاس نرمال، LDA جداساز بهینه است.

دسته‌بندی برای بیش از دو کلاس

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

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

Within-class scatter

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

Between-class scatter:

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

- 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.

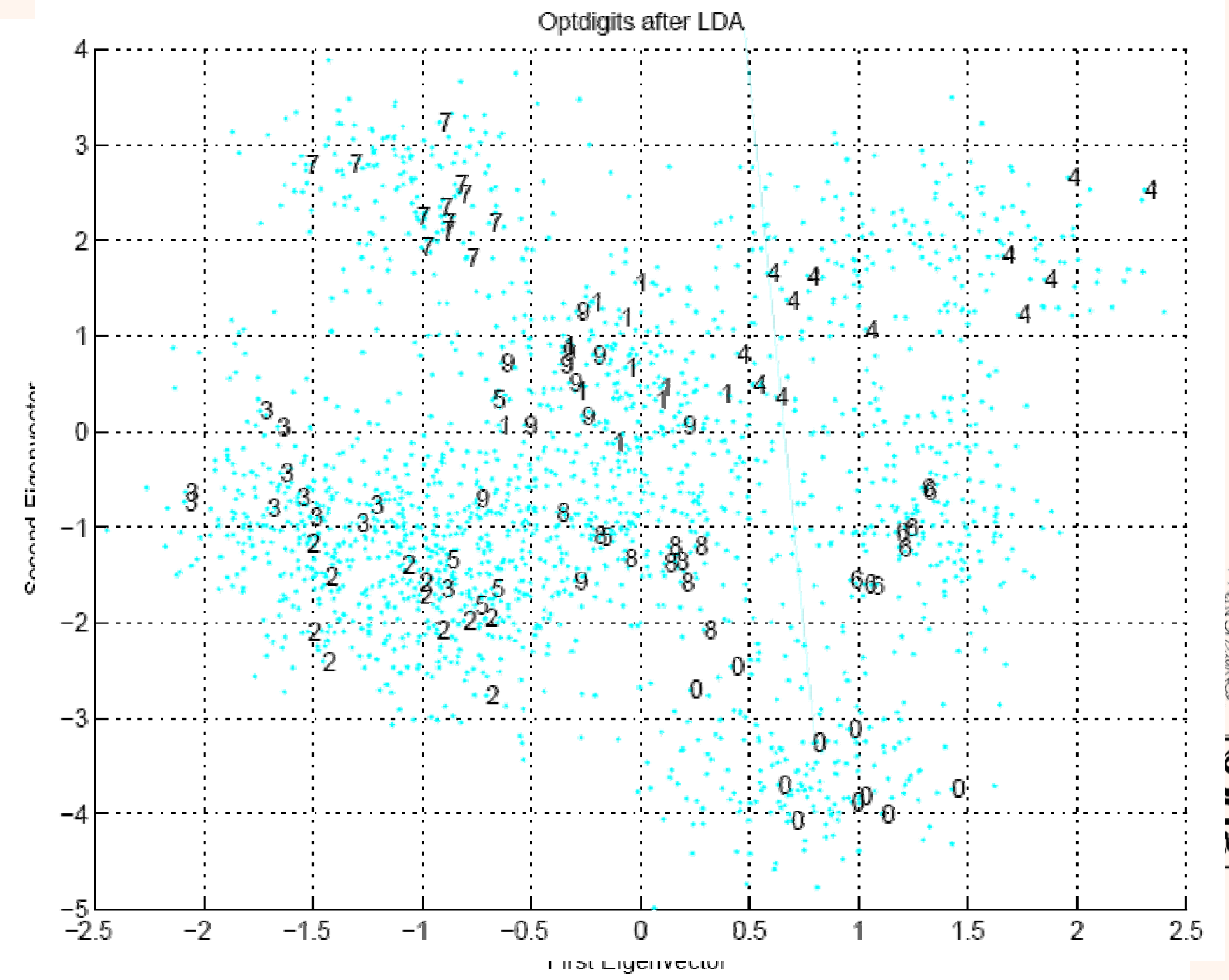
دسته‌بندی برای بیش از دو کلاس

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

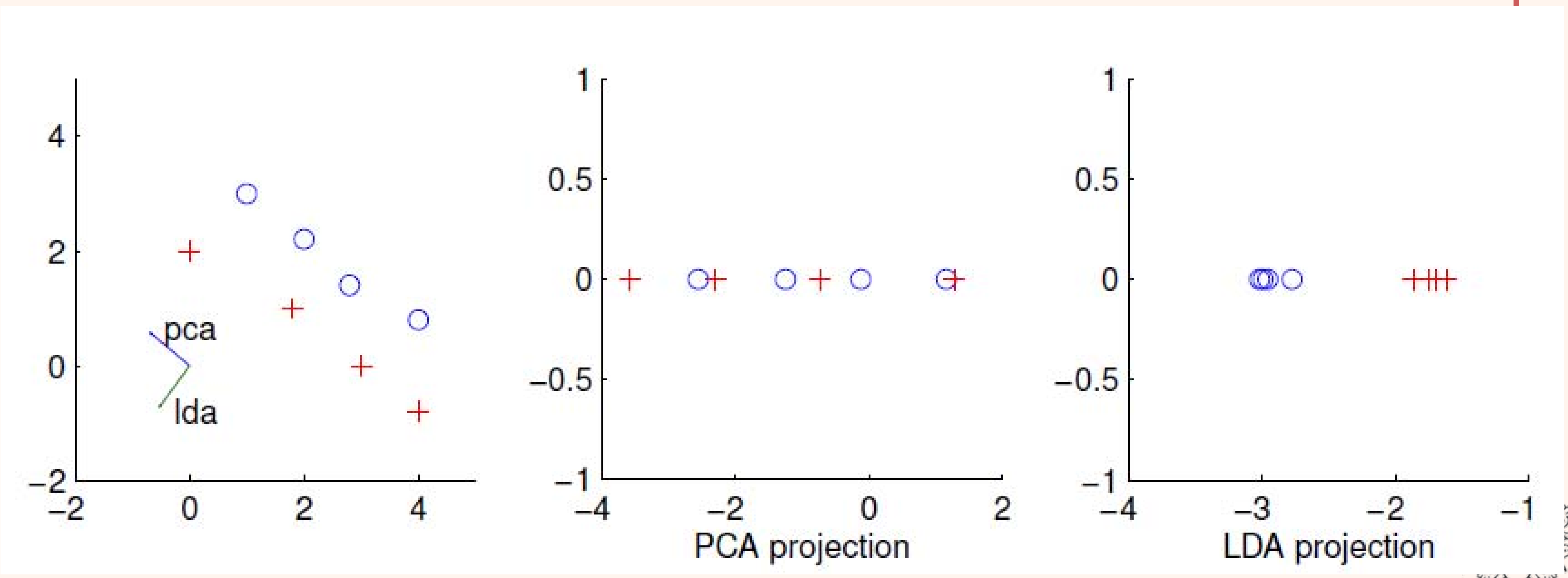
$$J(\mathbf{W}) = \frac{\left|\mathbf{W}^T \mathbf{S}_B \mathbf{W}\right|}{\left|\mathbf{W}^T \mathbf{S}_W \mathbf{W}\right|}$$

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

مثال


Optdigits after LDA
Second Eigenvector
First Eigenvector

PCA vs LDA



کاربردهای LDA

- Face recognition
  - Belhumeour *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

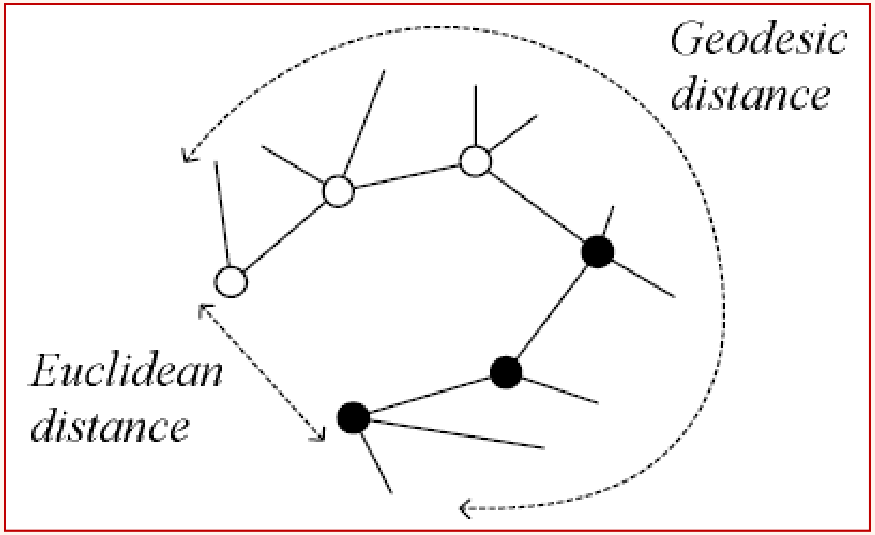
روش های غیرخطی

- در این شیوه از «geodesic distance» استفاده می‌شود.
  - برای نقاط نزدیک، فاصله‌ی اقلیدسی به عنوان تقریب مورد استفاده قرار می‌گیرد.

$$\left\| x^r - x^s \right\| \le \varepsilon$$

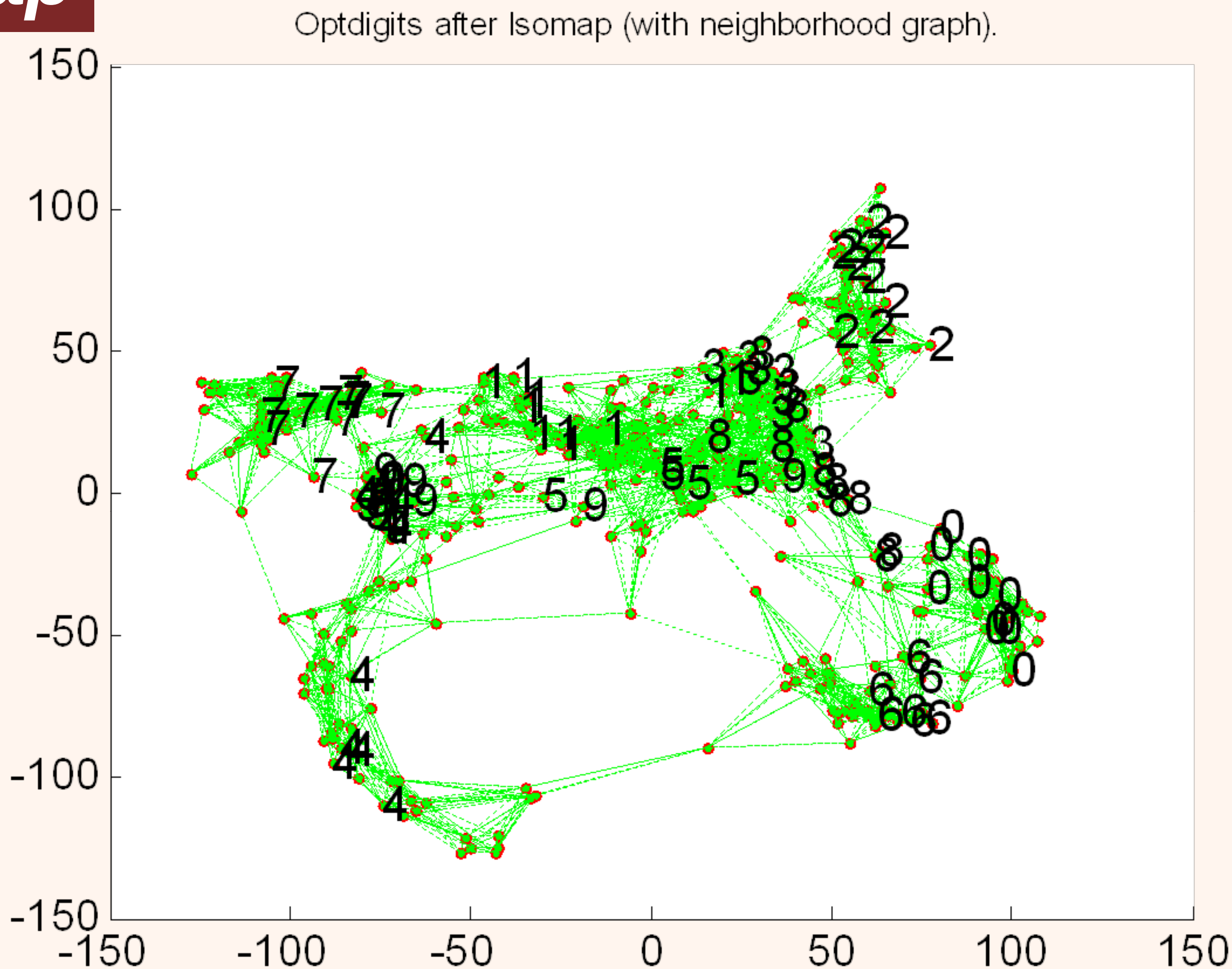
  - برای نقاط دورتر، مجموع فواصل نقاط همسایه مورد استفاده قرار می‌گیرد.
  - با تشکیل گراف فاصله‌ها، می‌توان تقریبی از فاصله‌ی نقاط روی یک «خمینه» به دست آورد.

Isomap



- پس از محاسبه‌ی فواصل می‌توان از MDS برای کاهش ابعاد استفاده کرد.
- با افزایش نقاط، ضمن افزایش دقت، پیچیدگی محاسباتی نیز افزایش می‌یابد، در این شرایط می‌توان به جای همه‌ی «نقاط برجسته» استفاده کرد.
- یکی از مشکلات این روش این است که با افزودن نقاط جدید، الگوریتم باید از نو اجرا شود.

Isomap



Matlab source from http://web.mit.edu/cocosci/isomap/isomap.html