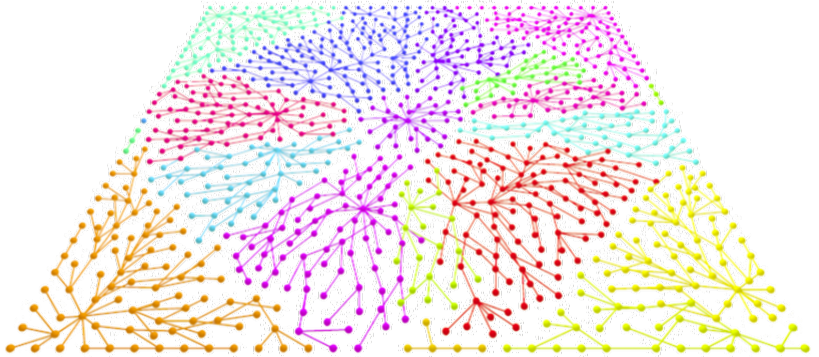
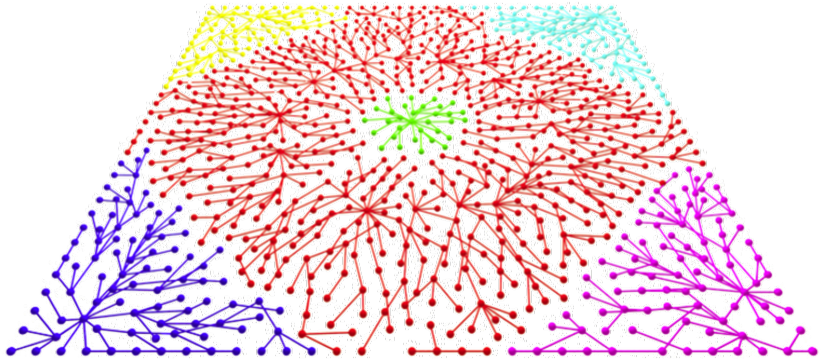
CS233, CME251: Geometric and Topological Data Analysis

Leonidas Guibas
Computer Science Department
Stanford University

Lecture 4
15 April 2020

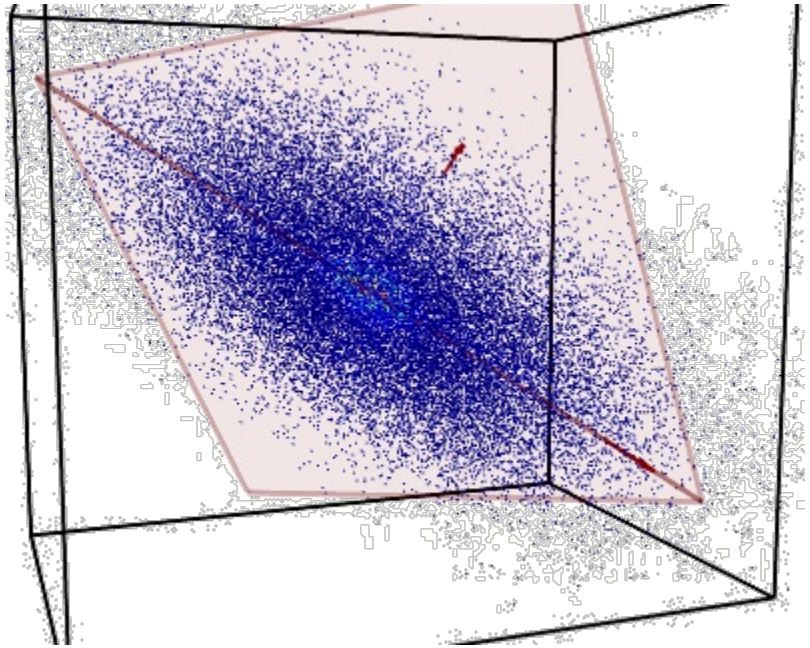




Last Time: Principal Components Analysis

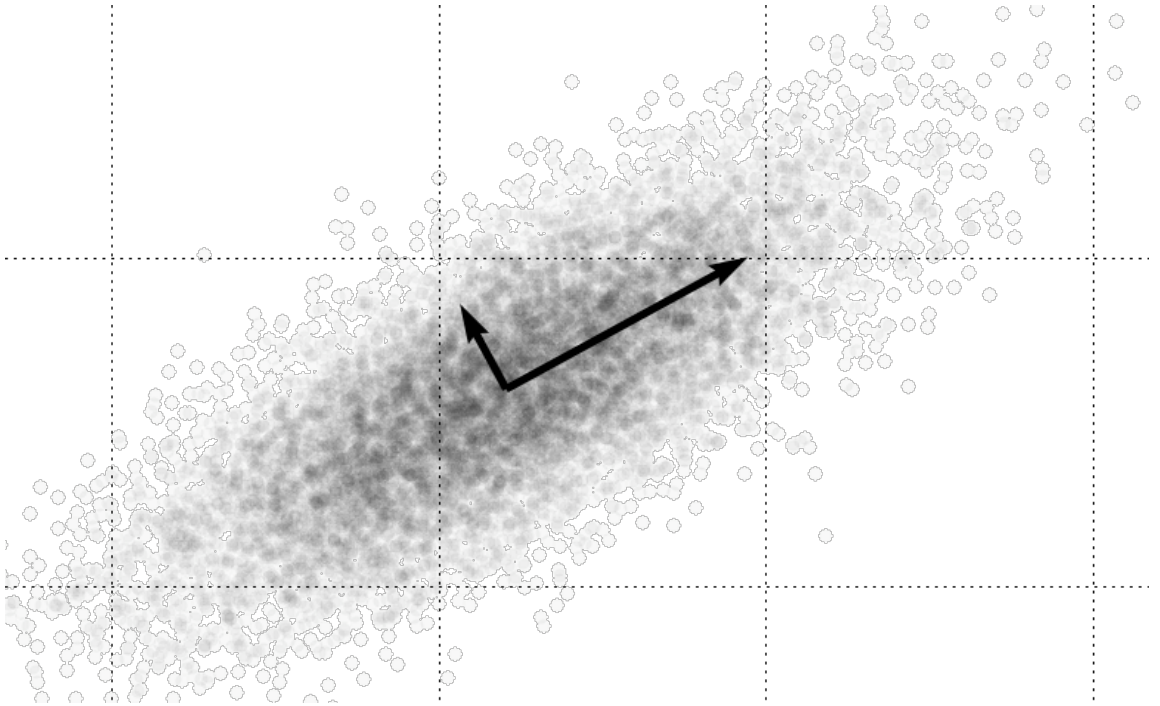
Principal Components Analysis (PCA)

- Introduced by Pearson (1901) and Hotelling (1933) to describe the variation in a set of multivariate data in terms of a set of uncorrelated variables.
- PCA looks for a single lower dimensional subspace that captures most of the variation in the data.
- Specifically, we aim to minimize the error introduced by projecting the data into this linear subspace.



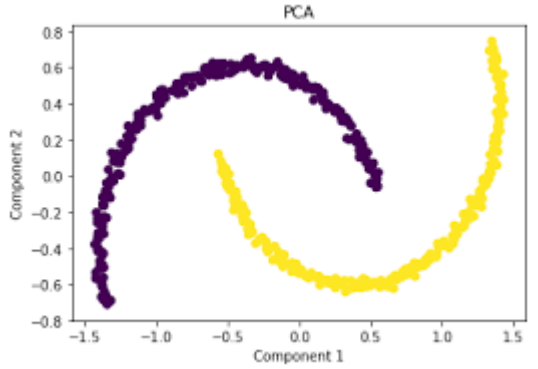
Last time: PCA, KPCA

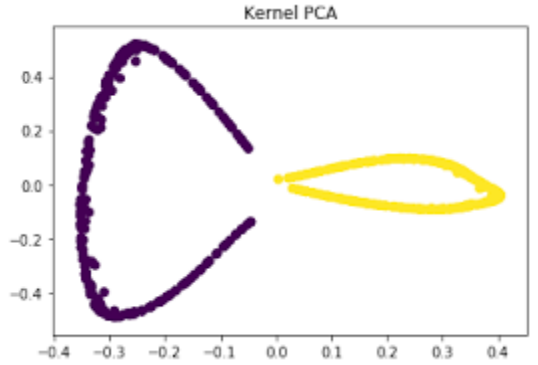
- Key result:
- Exploit spectral analysis of the covariance matrix *C* of the data
- For any integer *p*, the error-minimizing *p*-dimensional subspace is the one spanned by the first *p* eigenvectors of the covariance matrix



Kernel PCA (KPCA)

- Assumption behind PCA is that the data points **x** are well represented by a small-dimensional linear subspace
- Often this assumption does not hold …
- However, it may still be possible that a non-linear transformation $\phi(\mathbf{x})$ "linearizes" the data, but in a higher dimensional space -- then we can perform PCA in the space of $\phi(\mathbf{x})$
- Kernel PCA performs this "lifted" PCA; however, because of "kernel trick," it never computes the mapping $\phi(\mathbf{x})$ explicitly.





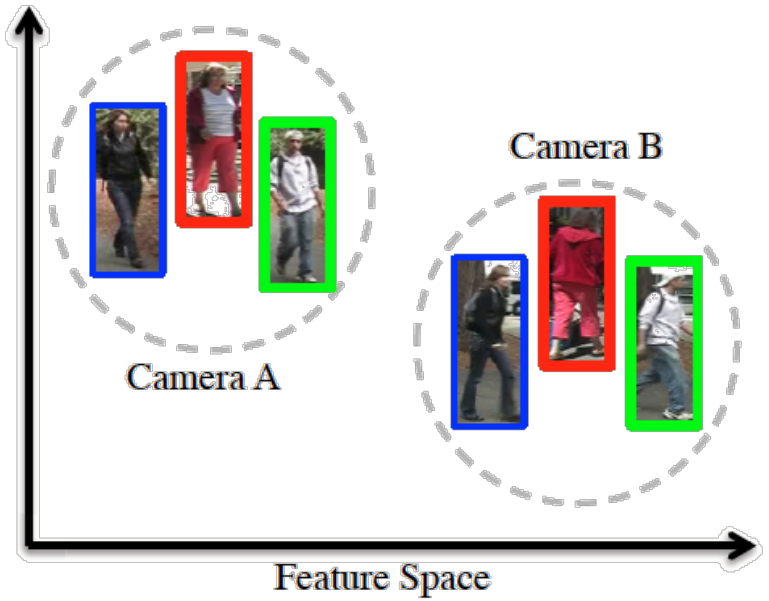
Today:
CCA & MDS

Canonical Correlation Analysis

Multi-dimensional Scaling

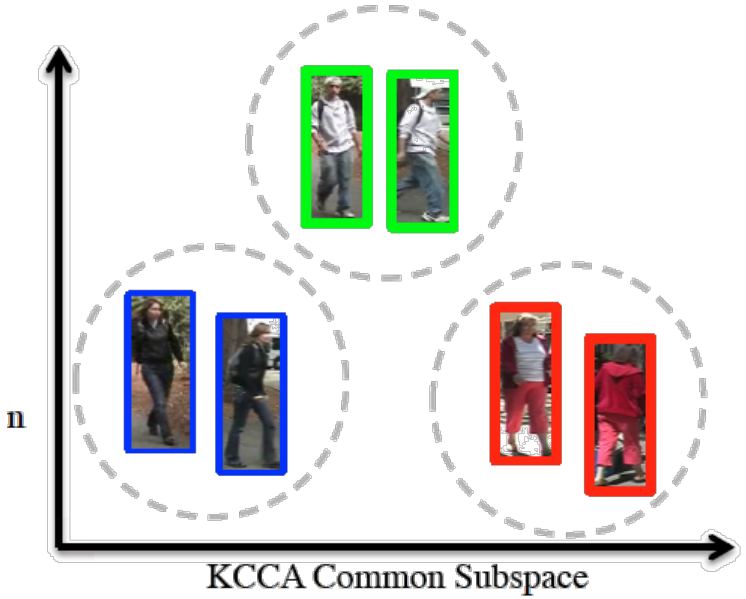
CCA: Shared Structure Across Different Data Sets

Different Views of the Same Data



Cluster by appearance similarity

Cluster by content similarity



Covariance and Correlation

Covariance and Correlation

- Pearson correlation coefficient between two random variables *X*, Y

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_x)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

$\mu_X = \bar{x}$
mean

σ_X
standard deviation

- Note that, by the Cauchy-Schwarz inequality

$$E[(X-\mu_x)(Y-\mu_Y)]^2 \leq E[(X-\mu_x)^2]E[(Y-\mu_Y)^2]$$

so

$$-1 \leq \rho_{X,Y} \leq 1$$

Covariance and Correlation

- Correlation measures a linear association between *X*, Y

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}$$

- Note that, *X*, Y independent, then corr(*X*,*Y*) = 0
  - but the opposite is not true

[*U* uniform in [0,2π], *X* = sin *U*, *Y* = cos *U*]

- High correlation not the same as causality

Empirical Correlation

- Empirical version, for *n* measurements x_i, y_i of *X* and *Y*

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{n s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

- We'll use centered versions, $\bar{x} = 0, \bar{y} = 0$

$$\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$$

$$= \frac{< \boldsymbol{x}, \boldsymbol{y} >}{||\boldsymbol{x}||||\boldsymbol{y}||}$$

Canonical Correlations for Two Sets of Variates

Canonical Correlations

- Canonical correlation analysis seeks a pair of linear transformations, one for each of the sets of variables *X*, *Y*, such that when the set of variables is transformed, the corresponding coordinates are maximally correlated.

- Consider projections of *X* and *Y*

$$\begin{aligned} \boldsymbol{x} \rightarrow & \quad < \boldsymbol{w}_x, \boldsymbol{x} > \\ \boldsymbol{y} \rightarrow & \quad < \boldsymbol{w}_y, \boldsymbol{y} > \end{aligned}$$

- So we get

$$\begin{aligned} \boldsymbol{S}_{x,\boldsymbol{w}_x} &= (< \boldsymbol{w}_x, \boldsymbol{x}_1 >, < \boldsymbol{w}_x, \boldsymbol{x}_2 >, \ldots, < \boldsymbol{w}_x, \boldsymbol{x}_n >) \\ \boldsymbol{S}_{y,\boldsymbol{w}_y} &= (< \boldsymbol{w}_y, \boldsymbol{y}_1 >, < \boldsymbol{w}_y, \boldsymbol{y}_2 >, \ldots, < \boldsymbol{w}_y, \boldsymbol{y}_n >) \end{aligned}$$

1st Canonical Correlation

- We choose w_x, w_y to maximize the correlation between these two vectors

$$\rho = (\rho_1 =) \quad \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} corr(S_x \boldsymbol{w}_x, S_y \boldsymbol{w}_y) = \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{< S_x \boldsymbol{w}_x, S_y \boldsymbol{w}_y >}{\|S_x \boldsymbol{w}_x\| \|S_y \boldsymbol{w}_y\|}$$

- Or, in empirical form, if we write $E[f(\boldsymbol{x}, \boldsymbol{y})] = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i, \boldsymbol{y}_i)$

We can re-write the correlation expression as

$$\rho = \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{E[< \boldsymbol{w}_x, \boldsymbol{x} >< \boldsymbol{w}_y, \boldsymbol{y} >]}{\sqrt{E[< \boldsymbol{w}_x, \boldsymbol{x} >^2] E[< \boldsymbol{w}_y, \boldsymbol{y} >^2]}}$$

Covariance Formulation

- Can re-write as

$$\rho = \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{E[\boldsymbol{w}_x^T \boldsymbol{x}\boldsymbol{y}^T \boldsymbol{w}_y]}{\sqrt{E[\boldsymbol{w}_x^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{w}_{\boldsymbol{x}}] E[\boldsymbol{w}_y^T \boldsymbol{y}\boldsymbol{y}^T \boldsymbol{w}_{\boldsymbol{y}}]}}$$

- so that

$$\rho = \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\boldsymbol{w}_x^T E[\boldsymbol{x}\boldsymbol{y}^T] \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^T E[\boldsymbol{x}\boldsymbol{x}^T] \boldsymbol{w}_{\boldsymbol{x}} \boldsymbol{w}_y^T E[\boldsymbol{y}\boldsymbol{y}^T] \boldsymbol{w}_{\boldsymbol{y}}}}$$

- If we now write

$$C(\boldsymbol{x}, \boldsymbol{y}) = E\left[\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}^T\right] = \begin{bmatrix} C_{\boldsymbol{x}\boldsymbol{x}} & C_{\boldsymbol{x}\boldsymbol{y}} \\ C_{\boldsymbol{y}\boldsymbol{x}} & C_{\boldsymbol{y}\boldsymbol{y}} \end{bmatrix} = C$$

- we get

$$\rho = \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\boldsymbol{w}_x^T C_{\boldsymbol{x}\boldsymbol{y}} \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^T C_{\boldsymbol{x}\boldsymbol{x}} \boldsymbol{w}_{\boldsymbol{x}} \boldsymbol{w}_y^T C_{\boldsymbol{y}\boldsymbol{y}} \boldsymbol{w}_{\boldsymbol{y}}}}$$

Solution by Eigenanalysis

- Note that the he expression $$\frac{\boldsymbol{w}_x^T C_{\boldsymbol{xy}} \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^T C_{\boldsymbol{xx}} \boldsymbol{w}_{\boldsymbol{x}} \boldsymbol{w}_y^T C_{\boldsymbol{yy}} \boldsymbol{w}_{\boldsymbol{y}}}}$$

  is invariant to re-scalings of w_x or w_y

- We can therefore solve the optimization problem

$$\max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \quad \boldsymbol{w}_x^T C_{\boldsymbol{xy}} \boldsymbol{w}_y$$

  subject to the constraints

$$\boldsymbol{w}_x^T C_{\boldsymbol{xx}} \boldsymbol{w}_x = 1 \text{ and } \boldsymbol{w}_y^T C_{\boldsymbol{yy}} \boldsymbol{w}_y = 1.$$

Lagrange Multipliers

$$L(\lambda, \boldsymbol{w}_x, \boldsymbol{w}_y) = \boldsymbol{w}_x^T C_{\boldsymbol{xy}} \boldsymbol{w}_y - \frac{\lambda_x}{2}(\boldsymbol{w}_x^T C_{\boldsymbol{xx}} \boldsymbol{w}_x - 1) - \frac{\lambda_y}{2}(\boldsymbol{w}_y^T C_{\boldsymbol{yy}} \boldsymbol{w}_y - 1)$$

- Setting derivatives of L to 0 w.r.t. w_x, w_y we get (after some manipulation)

$$\frac{\partial L}{\partial \boldsymbol{w}_x} = C_{\boldsymbol{xy}} \boldsymbol{w}_y - \lambda_x C_{\boldsymbol{xx}} \boldsymbol{w}_x = 0 \qquad \frac{\partial L}{\partial \boldsymbol{w}_y} = C_{\boldsymbol{yx}} \boldsymbol{w}_x - \lambda_y C_{\boldsymbol{yy}} \boldsymbol{w}_y = 0$$

multiply by w_x^T, subtract

multiply by w_y^T

$$0 = \boldsymbol{w}_x^T C_{xy} \boldsymbol{w}_y - \boldsymbol{w}_x^T \lambda_x C_{xx} \boldsymbol{w}_y - \boldsymbol{w}_y^T C_{yx} \boldsymbol{w}_x + \boldsymbol{w}_y^T \lambda_y C_{yy} \boldsymbol{w}_y$$

$$= \boldsymbol{w}_y^T \lambda_y C_{yy} \boldsymbol{w}_y - \boldsymbol{w}_x^T \lambda_x C_{xx} \boldsymbol{w}_y$$

Lagrange Multipliers

- Leading to $\lambda_x \boldsymbol{w}_x^T C_{\boldsymbol{xx}} \boldsymbol{w}_x = \lambda_y \boldsymbol{w}_y^T C_{\boldsymbol{yy}} \boldsymbol{w}_y \quad \text{and} \quad \lambda_x = \lambda_y (= \lambda)$
- Assuming C_{yy} is invertible, we can use the second derivative constraint above to get

$$\boldsymbol{w}_y = \frac{C_{\boldsymbol{yy}}^{-1} C_{\boldsymbol{yx}} \boldsymbol{w}_x}{\lambda}$$

$$\frac{\partial L}{\partial \boldsymbol{w}_x} = C_{\boldsymbol{xy}} \boldsymbol{w}_y - \lambda_x C_{\boldsymbol{xx}} \boldsymbol{w}_x = 0$$

- so now from the first derivate constraint we get

$$\frac{\partial L}{\partial \boldsymbol{w}_y} = C_{\boldsymbol{yx}} \boldsymbol{w}_x - \lambda_y C_{\boldsymbol{yy}} \boldsymbol{w}_y = 0$$

- or

$$\frac{\partial L}{\partial \boldsymbol{w}_x} = C_{\boldsymbol{xy}} \boldsymbol{w}_y - \lambda_x C_{\boldsymbol{xx}} \boldsymbol{w}_x = 0$$

Eigenvalue Problem

- A generalized eigenvalue problem

$$C_{\boldsymbol{xy}} C_{\boldsymbol{yy}}^{-1} C_{\boldsymbol{yx}} \boldsymbol{w}_x = \lambda^2 C_{\boldsymbol{xx}} \boldsymbol{w}_x$$

$$A\boldsymbol{x} = \lambda B\boldsymbol{x}$$

- Can symmetrically also get

$$C_{\boldsymbol{yx}} C_{\boldsymbol{xx}}^{-1} C_{\boldsymbol{xy}} \boldsymbol{w}_y = \lambda^2 C_{\boldsymbol{yy}} \boldsymbol{w}_y$$

- On can go back and forth between w_x and w_y

$$\boldsymbol{w}_y = \frac{C_{\boldsymbol{yy}}^{-1} C_{\boldsymbol{yx}} \boldsymbol{w}_x}{\lambda}$$

Solving the Eigenvalue Problem

- If C_{xx} is invertible, then can reduce to a standard symmetric eigenvalue problem

$$C_{\boldsymbol{xx}}^{-1} C_{\boldsymbol{xy}} C_{\boldsymbol{yy}}^{-1} C_{\boldsymbol{yx}} \boldsymbol{w}_x = \lambda^2 \boldsymbol{w}_x$$

- Numerically, all these inversions and multiplications lose precision
- Alternate approach:
  - C_{xx} and C_{yy} are symmetric positive definite; use complete Cholesky decomposition – R_{xx} lower triangular so that

$$C_{\boldsymbol{xx}} = R_{\boldsymbol{xx}} R_{\boldsymbol{xx}}^T$$

  - Now let $\boldsymbol{u}_x = R_{\boldsymbol{xx}}^T \boldsymbol{w}_x$
  - Can re-write

$$C_{\boldsymbol{xy}} C_{\boldsymbol{yy}}^{-1} C_{\boldsymbol{yx}} R_{\boldsymbol{xx}}^{-T} \boldsymbol{u}_x = \lambda^2 R_{\boldsymbol{xx}} \boldsymbol{u}_x$$

$$R_{\boldsymbol{xx}}^{-1} C_{\boldsymbol{xy}} C_{\boldsymbol{yy}}^{-1} C_{\boldsymbol{yx}} R_{\boldsymbol{xx}}^{-T} \boldsymbol{u}_x = \lambda^2 \boldsymbol{u}_x$$

For more on the numerics, see the Weenik article

Centered Variables, CCA vs PCA

- For centered variates, the covariance $C_{\boldsymbol{xy}} = \boldsymbol{x}^T \boldsymbol{y}$
- Canonical correlation analysis attempts to answer the question "which directions accounts for most of the covariance between the two data sets?" The goal is to find directions w_x, w_y so as to maximize

$$\boldsymbol{w}_x^T C_{\boldsymbol{xy}} \boldsymbol{w}_y = (\boldsymbol{x}\boldsymbol{w}_x)^T (\boldsymbol{y}\boldsymbol{w}_y)$$

- subject to

$$\|\boldsymbol{x}\boldsymbol{w}_x\| = 1, \|\boldsymbol{y}\boldsymbol{w}_y\| = 1$$

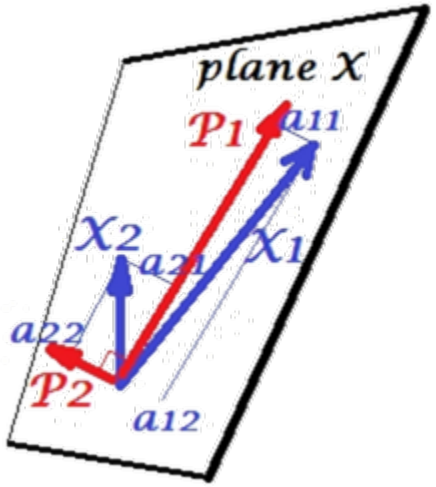
- In PCA we have a single variate and seek the direction that "maximizes the variance in the data"

$$\boldsymbol{w}_x^T C_{\boldsymbol{xx}} \boldsymbol{w}_x = (\boldsymbol{x}\boldsymbol{w}_x)^T (\boldsymbol{x}\boldsymbol{w}_x)$$

- subject to

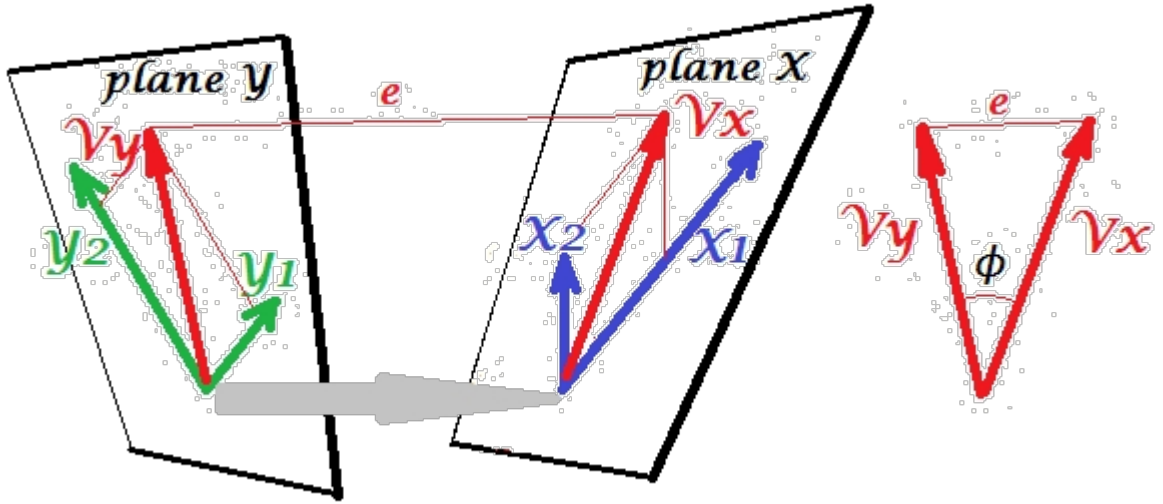
$$\|\boldsymbol{w}_x\| = 1$$

CCA vs PCA



PCA

CCA



Example 1

CCA Example 1: Scores Data

Example: $n = 88$ students took tests in each of 5 subjects: mechanics, vectors, algebra, analysis, statistics. (From Mardia et al. (1979) "Multivariate analysis".) Each test is out of 100 points

The tests on mechanics, vectors were closed book and those on algebra, analysis, statistics were open book. There's clearly some correlation between these two sets of scores:

```
     alg   ana   sta
mec 0.547 0.409 0.389
vec 0.610 0.485 0.436
```

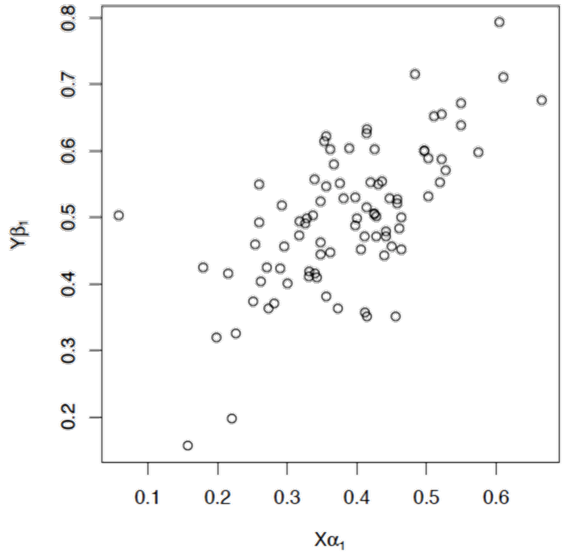
Canonical correlation analysis attempts to explain this phenomenon using the variables in each set jointly. Here X contains the closed book test scores and Y contains the open book test scores, so $X \in \mathbb{R}^{88\times 2}$ and $Y \in \mathbb{R}^{88\times 3}$

CCA Example 1: Scores Data

The first canonical directions (multiplied by 10^3):

$$\alpha_1 = \begin{pmatrix} 2.770 \\ 5.517 \end{pmatrix} \begin{matrix} \text{mec} \\ \text{vec} \end{matrix}, \quad \beta_1 = \begin{pmatrix} 8.782 \\ 0.860 \\ 0.370 \end{pmatrix} \begin{matrix} \text{alg} \\ \text{ana} \\ \text{sta} \end{matrix}$$

The first canonical correlation is $\rho_1 = 0.663$, and the variates:



Higher Order CCA

Higher-Order Canonical Correlates

- We defined the 1st canonical correlation ρ_1 through the projection vectors / directions

$$\boldsymbol{w}_x = \boldsymbol{w}_x^{(1)} \text{ and } \boldsymbol{w}_y = \boldsymbol{w}_y^{(1)}$$

- Given the first *k*-1 directions, the *k*-th canonical correlation is defined via vectors $\boldsymbol{w}_x^{(k)}$ and $\boldsymbol{w}_y^{(k)}$, so that we maximize

$$\max \quad (\boldsymbol{x}\boldsymbol{w}_x^{(k)})^T(\boldsymbol{y}\boldsymbol{w}_y^{(k)})$$

- but under orthogonality constraints to the previous directions

$$\|\boldsymbol{x}\boldsymbol{w}_x^{(k)}\| = 1, \|\boldsymbol{y}\boldsymbol{w}_y^{(k)}\| = 1$$

$$(\boldsymbol{x}\boldsymbol{w}_x^{(k)})^T(\boldsymbol{x}\boldsymbol{w}_x^{(j)}) = 0, j = 1, 2, \ldots, k-1$$

$$(\boldsymbol{y}\boldsymbol{w}_y^{(k)})^T(\boldsymbol{y}\boldsymbol{w}_y^{(j)}) = 0, j = 1, 2, \ldots, k-1$$

Equivalent Higher Order CCA Formulations

$$\max_{\mathbf{a}_i,\mathbf{b}_i} \sum_i \mathbf{a}_i^T \mathbf{X}\mathbf{Y}^T \mathbf{b}_i \quad \text{s.t.} \quad \mathbf{a}_i^T \mathbf{X}\mathbf{X}^T \mathbf{a}_j = \delta_{i=j} \forall j \le i, \quad \mathbf{b}_i^T \mathbf{Y}\mathbf{Y}^T \mathbf{b}_j = \delta_{i=j} \forall j \le i$$

$$\Leftrightarrow$$

$$\max_{\mathbf{A},\mathbf{B}} \operatorname{trace}\left(\mathbf{A}^T \mathbf{X}\mathbf{Y}^T \mathbf{B}\right) \quad \text{s.t.} \quad \mathbf{A}^T \mathbf{X}\mathbf{X}^T \mathbf{A} = \mathbf{I}_d, \quad \mathbf{B}^T \mathbf{Y}\mathbf{Y}^T \mathbf{B} = \mathbf{I}_d$$

$$\Updownarrow \qquad\qquad \Updownarrow$$

$$\begin{bmatrix} \mathbf{0} & \mathbf{X}\mathbf{Y}^T \\ \mathbf{Y}\mathbf{X}^T & \mathbf{0} \end{bmatrix} \begin{pmatrix} \mathbf{a}_i \\ \mathbf{b}_i \end{pmatrix} = \rho_i \begin{bmatrix} \mathbf{X}\mathbf{X}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{Y}\mathbf{Y}^T \end{bmatrix} \begin{pmatrix} \mathbf{a}_i \\ \mathbf{b}_i \end{pmatrix}$$

$$\Leftrightarrow$$

$$\min_{\mathbf{A},\mathbf{B}} \frac{1}{2} \left\| \mathbf{A}^T \mathbf{X} - \mathbf{B}^T \mathbf{Y} \right\|_F^2 \quad \text{s.t.} \quad \mathbf{A}^T \mathbf{X}\mathbf{X}^T \mathbf{A} = \mathbf{I}_d, \quad \mathbf{B}^T \mathbf{Y}\mathbf{Y}^T \mathbf{B} = \mathbf{I}_d$$

Principal Angles Between Subspaces

- Goal: capture geometric configuration of two subspaces with few and intuitive numbers
- Principal angles intuition
  - Measure angles between 'most similar' directions within subspaces
  - Capture relative 'orientation' of two subspaces
  - Recursive definitions with decreasing similarity
- Consider subspaces spanned by the cols of *X* and *Y* respectively

Definition

$$\cos\theta_i = \max_{\mathbf{u}_i \in \mathcal{Y}_1} \max_{\mathbf{v}_i \in \mathcal{Y}_2} \mathbf{u}_i^T \mathbf{v}_i$$

$$\text{s.t. } \|\mathbf{u}_i\|_2 = \|\mathbf{v}_i\|_2 = 1$$

$$\forall j < i : \mathbf{u}_i^T \mathbf{u}_j = \mathbf{v}_i^T \mathbf{v}_j = 0$$

$\Longleftrightarrow$

$$\max \text{ trace}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{A}^T \boldsymbol{X}^T \boldsymbol{Y} \boldsymbol{B})$$

such that

$$\boldsymbol{A}, \boldsymbol{B} \in \mathcal{O}_p$$

'align' them

CCA and Principal Angles

- CCA between RVs in column form $\rightarrow$ principal angles of row spaces spanned by data matrices

$$\rho_i = \cos\theta_i$$

- For i = 1, 2, …

- Note 1: The number of canonical directions/variates is

$$r = \min\{\mathrm{rank}(\boldsymbol{X}), \mathrm{rank}(\boldsymbol{Y})\}$$

- Note 2:
  - If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are orthogonal ($\boldsymbol{X}^T\boldsymbol{Y} = 0$) then all principal angles are 90^o and the corresponding canonical correlations are 0
  - If $\boldsymbol{X}$ and $\boldsymbol{Y}$ intersect in a d-dimensional subspace, then the first d principal angles are 0

CCA Computation via Sphering/Whitening

For any symmetric invertible matrix $A \in \mathbb{R}^{n \times n}$, there is a matrix $A^{1/2} \in \mathbb{R}^{n \times n}$, called the (symmetric) square root of A, such that $A^{1/2} A^{1/2} = A$

We write the inverse of $A^{1/2}$ as $A^{-1/2}$. Note $A^{-1/2} A A^{-1/2} = I$. (Why?)

Given centered matrices $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$, we define $V_X = X^T X \in \mathbb{R}^{p \times p}$ and $V_Y = Y^T Y \in \mathbb{R}^{q \times q}$. Then

$$\tilde{X} = X V_X^{-1/2} \in \mathbb{R}^{n \times p} \text{ and } \tilde{Y} = Y V_Y^{-1/2} \in \mathbb{R}^{n \times q}$$

are called the sphered versions of X and Y. Note that the sample covariance of $\tilde{X}$ and $\tilde{Y}$ is

$$\text{cov}(\tilde{X}) = I/n \text{ and } \text{cov}(\tilde{Y}) = I/n$$

Transformed Problem

As suggested by the previous slide, we will take $\tilde{X} = XV_X^{-1/2}$ and $\tilde{Y} = YV_Y^{-1/2}$, and we'll solve the problem

$$\tilde{\alpha}_1, \tilde{\beta}_1 = \underset{\|\tilde{X}\tilde{\alpha}\|_2=1,\, \|\tilde{Y}\tilde{\beta}\|_2=1}{\operatorname{argmax}} (\tilde{X}\tilde{\alpha})^T(\tilde{Y}\tilde{\beta})$$

Recall that then $\alpha_1 = V_X^{-1/2}\tilde{\alpha}_1$ and $\beta_1 = V_Y^{-1/2}\tilde{\beta}_1$.

So why is this simpler? Note that the constraint says

$$1 = (\tilde{X}\tilde{\alpha})^T(\tilde{X}\tilde{\alpha}) = \tilde{\alpha}^T V_X^{-1/2} X^T X V_X^{-1/2}\tilde{\alpha} = \tilde{\alpha}^T\tilde{\alpha}$$

i.e., $\|\tilde{\alpha}\|_2 = 1$. Similarly, $\|\tilde{\beta}\|_2 = 1$. Hence our problem can be rewritten as:

$$\tilde{\alpha}_1, \tilde{\beta}_1 = \underset{\|\tilde{\alpha}\|_2=1,\, \|\tilde{\beta}\|_2=1}{\operatorname{argmax}} \tilde{\alpha}^T M\tilde{\beta}$$

where $M = \tilde{X}^T\tilde{Y} = V_X^{-1/2} X^T Y V_Y^{-1/2} \in \mathbb{R}^{p\times q}$. The same is true for further directions

SVD to the Rescue

Now comes the singular value decomposition to the rescue (again!). Let $r = \min\{p, q\}$. Then we can decompose

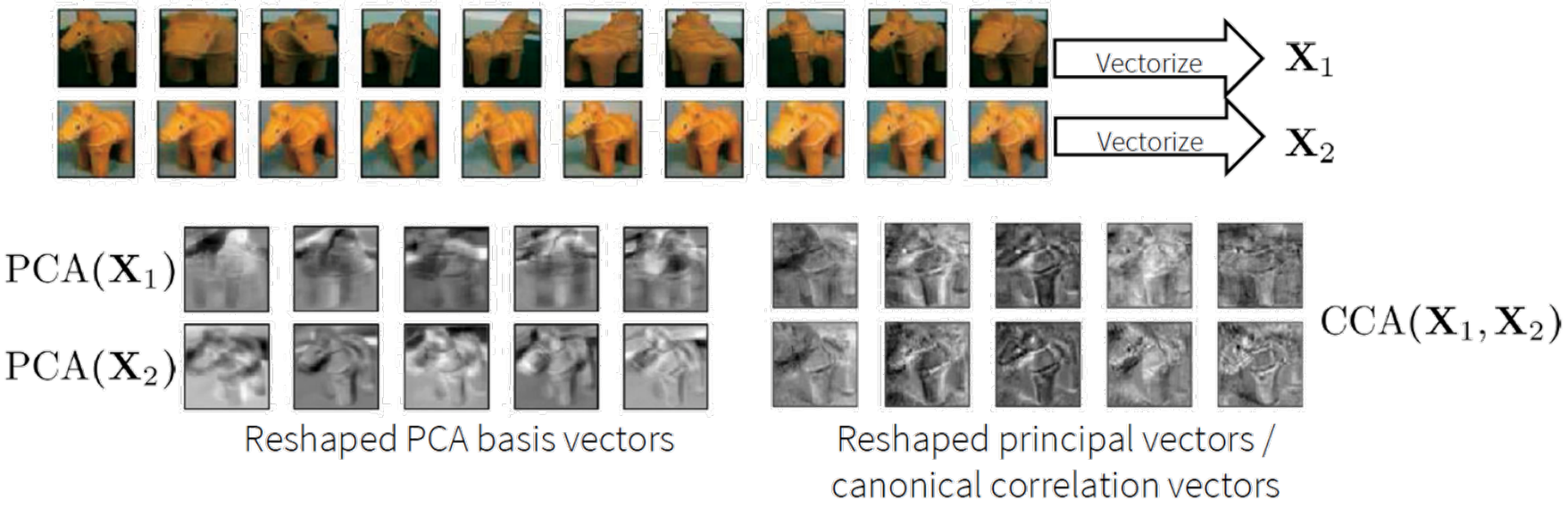
$$M = UDV^T$$

where $U \in \mathbb{R}^{p \times r}$, $V \in \mathbb{R}^{q \times r}$ have orthonormal columns, and $D = \text{diag}(d_1, \ldots d_r) \in \mathbb{R}^{r \times r}$ with $d_1 \geq \ldots \geq d_r \geq 0$. Further:

- The transformed canonical directions $\tilde{\alpha}_1, \ldots \tilde{\alpha}_r \in \mathbb{R}^p$ and $\tilde{\beta}_1, \ldots \tilde{\beta}_r \in \mathbb{R}^q$ are the columns of U and V, respectively
- The canonical directions $\alpha_1, \ldots \alpha_r \in \mathbb{R}^p$ and $\beta_1, \ldots \beta_r \in \mathbb{R}^q$ are the columns of $V_X^{-1/2}U$ and $V_Y^{-1/2}V$, respectively;
- the canonical variates $X\alpha_1, \ldots X\alpha_r \in \mathbb{R}^n$ and $Y\beta_1, \ldots Y\beta_r \in \mathbb{R}^n$ are the columns of $XV_X^{-1/2}U \in \mathbb{R}^{n \times r}$ and $YV_Y^{-1/2}V \in \mathbb{R}^{n \times r}$, respectively
- The canonical correlations $\rho_1 \geq \ldots \geq \rho_r$ are equal to $d_1 \geq \cdots \geq d_r$, the diagonal entries of D

More Examples

CCA Example 2: Object Recognition

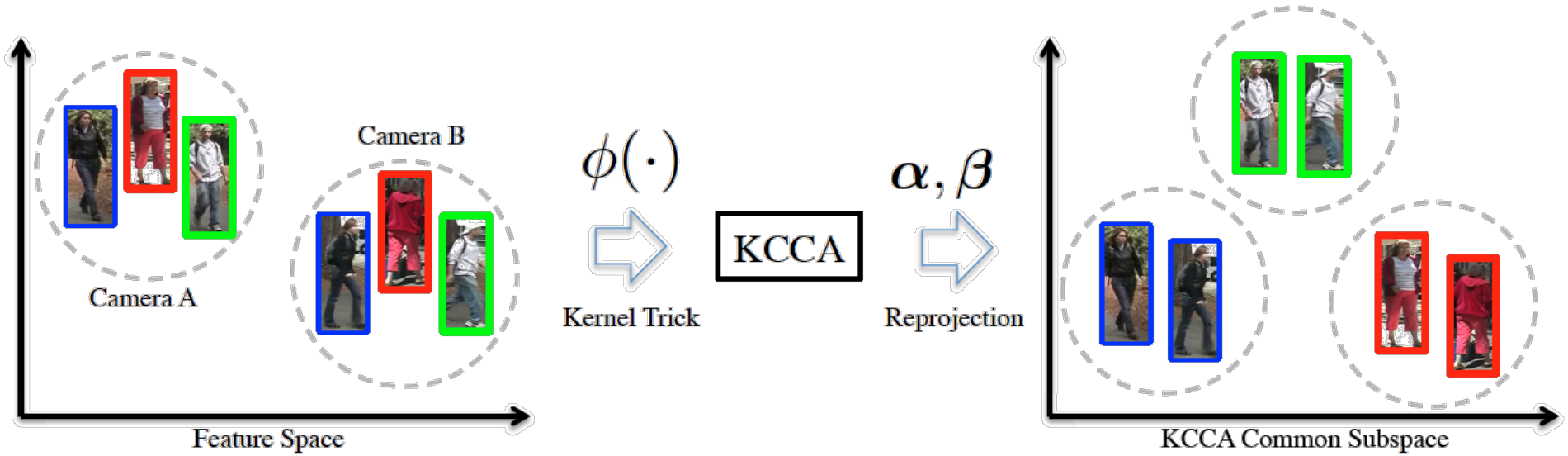


- "Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlation" Kim, Kittler, Cipolla [PAMI 07]
- Idea: Find low-dimensional subspace embedding s.t.
  - within-class CCA is maximized
  - Between-class CCA is minimized

CCA Example 3: KCCA for Matching People

- Lisanti, Giuseppe, Iacopo Masi, and Alberto Del Bimbo. "Matching people across camera views using kernel canonical correlation analysis." Proceedings of the International Conference on Distributed Smart Cameras. ACM, 2014.

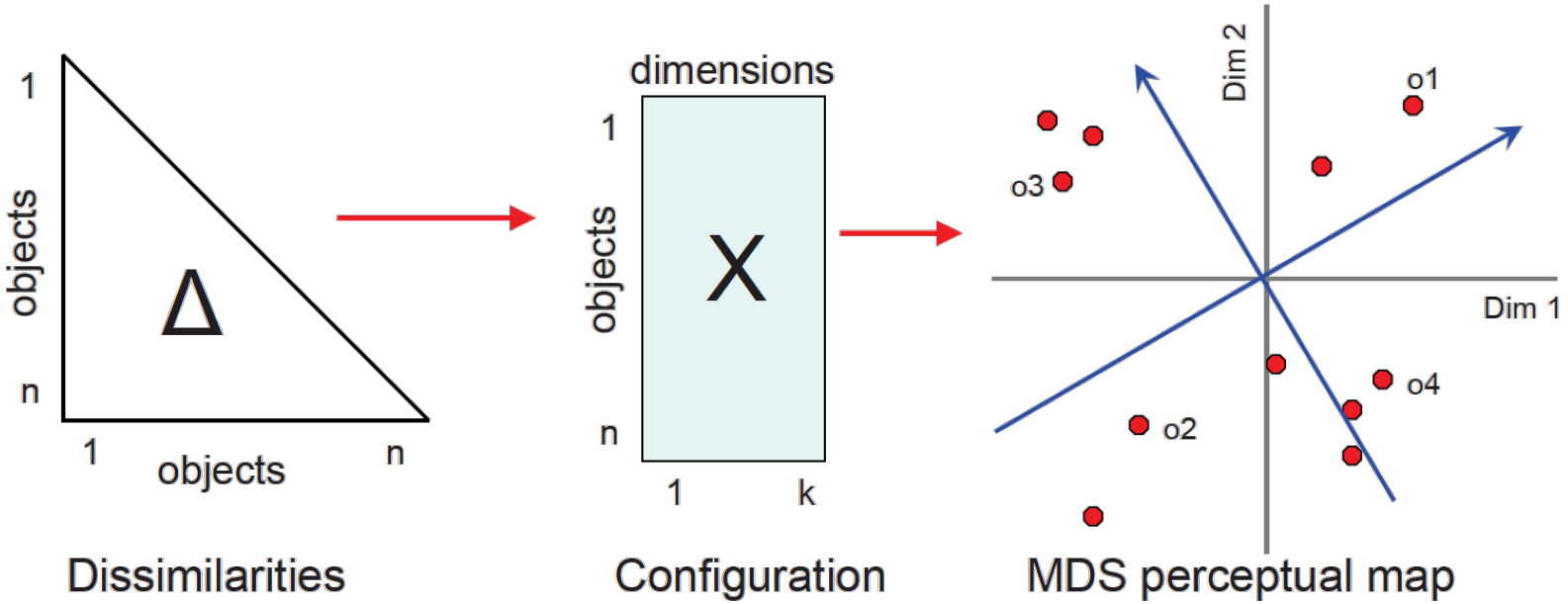
Common Subspace Through KCCA



Multidimensional Scaling

Multidimensional Scaling (MDS)

- A “distance preserving” embedding of the data into a Euclidean space
  - Sometimes distances are observed directly (e.g., similarity ratings)
  - Sometimes they can be calculated from a data table (e.g., Euclidean distances, correlations)



Formally (Metric MDS) …

- Given a (symmetric) matrix of pairwise "dis-similarities" between n objects / data sets

$$M = \left(\delta_{ij}\right)_{n\times n}$$

No need to satisfy the triangle inequality

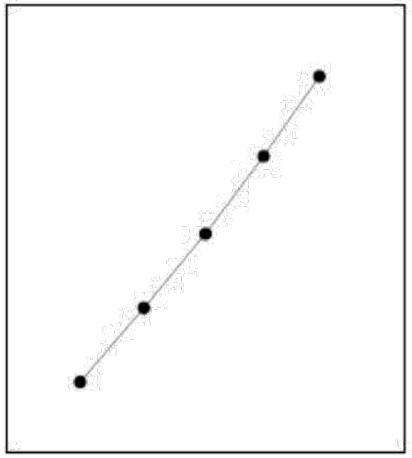
- Find n points in low-dimensional space R^d, so that their distance matrix is as close as possible to M

- Low d (=2,3) allows us to visualize the data directly

Distances and Dimensionality

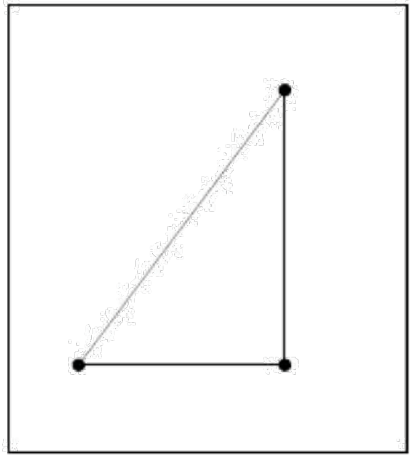
- How do distances/dissimilarities determine dimensionality?

$$D = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 2 & 3 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 1 \\ 4 & 3 & 2 & 1 & 0 \end{bmatrix}$$



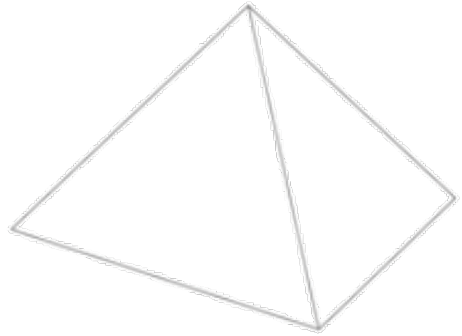
k=1

$$D = \begin{bmatrix} 0 & 3 & 4 \\ 3 & 0 & 5 \\ 4 & 5 & 0 \end{bmatrix}$$



k=2

$$D = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$



k=3

MDS: Motivating Example

- Travel times by train between French cities

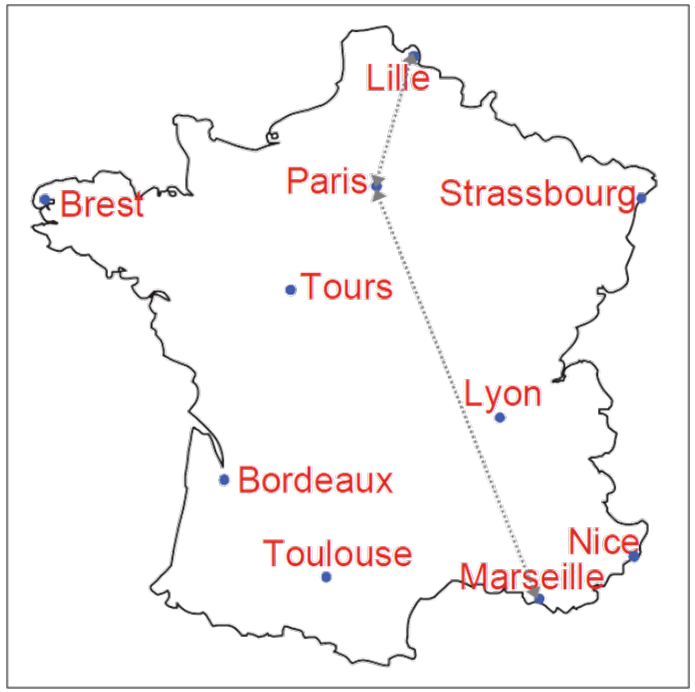
	Bor-deaux	Brest	Lille	Lyon	Mar-seille	Nice	Paris	Strassb ourg	Tou-louse	Tours
Bordeaux	0									
Brest	9:58	0								
Lille	6:39	7:11	0							
Lyon	8:05	7:11	4:52	0						
Marseille	5:47	8:49	6:12	1:35	0					
Nice	8:30	13:36	8:20	4:33	2:26	0				
Paris	2:59	4:17	1:04	2:01	3:00	5:52	0			
Strassbourg	8:08	10:16	6:54	4:36	7:04	11:15	4:01	0		
Toulouse	2:02	13:52	9:42	4:25	3:26	6:29	5:14	10:56	0	
Tours	2:36	5:38	4:17	4:21	5:13	9:04	1:13	6:03	6:06	0

- Considerations:
  - The recovered configuration (**X** = map) should be 2D
  - NSEW not relevant to distances– may have to rotate
  - Travel time not necessarily ~ map distance (TGV)
  - May need to consider other relations between dissimilarity and distance in MD space

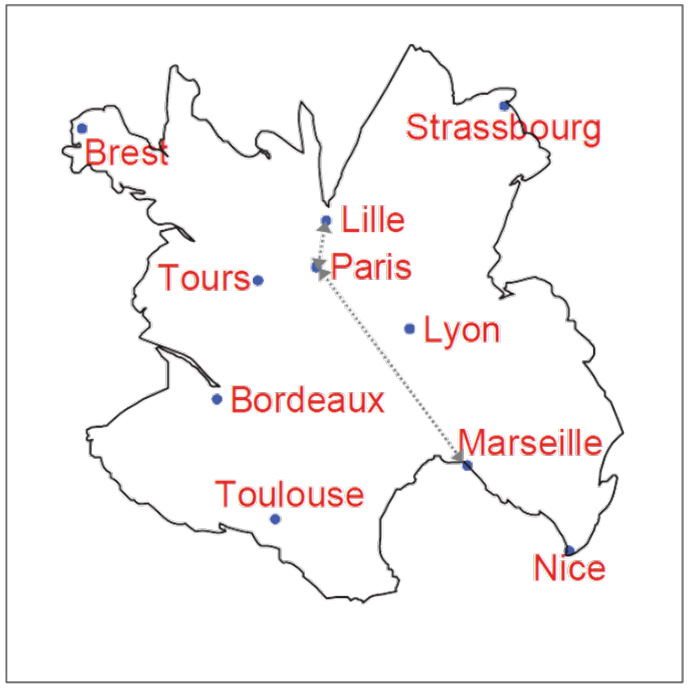
Source: Patrick Groenen, "Past, Present, and Future of Multidimensional Scaling", CARME, 2011

MDS: Motivating Example

- Travel times by train between French cities



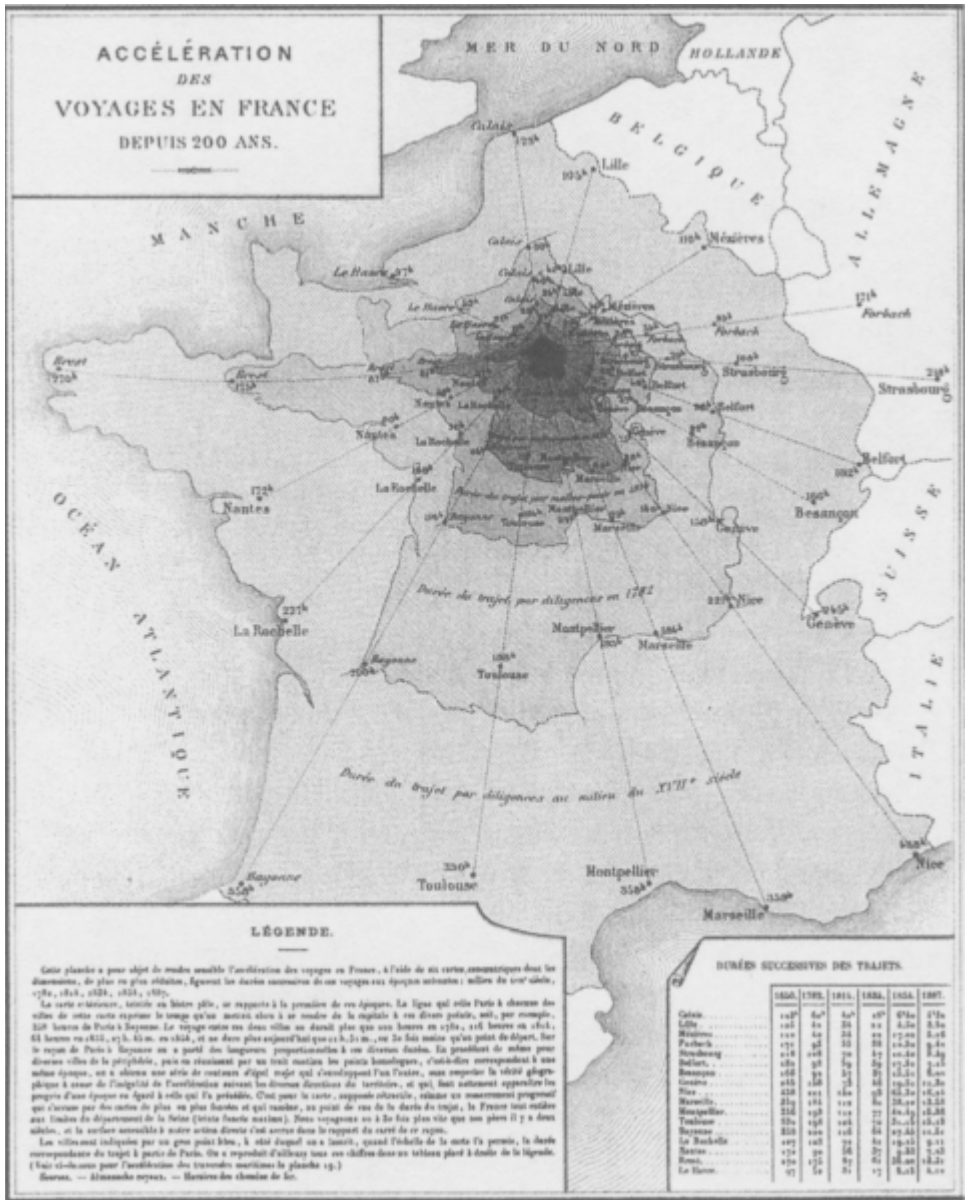
Actual map of France



MDS recovered map

Marseille, Lille, Lyon, ... closer to Paris in travel time (TGV)

MDS: Motivating Example



- Travel time only partially related to map distance

Emile Cheysson (1888) – anaphoric map, showing decrease in time to travel from Paris over 200 years

- Other considerations relating distance and dissimilarity

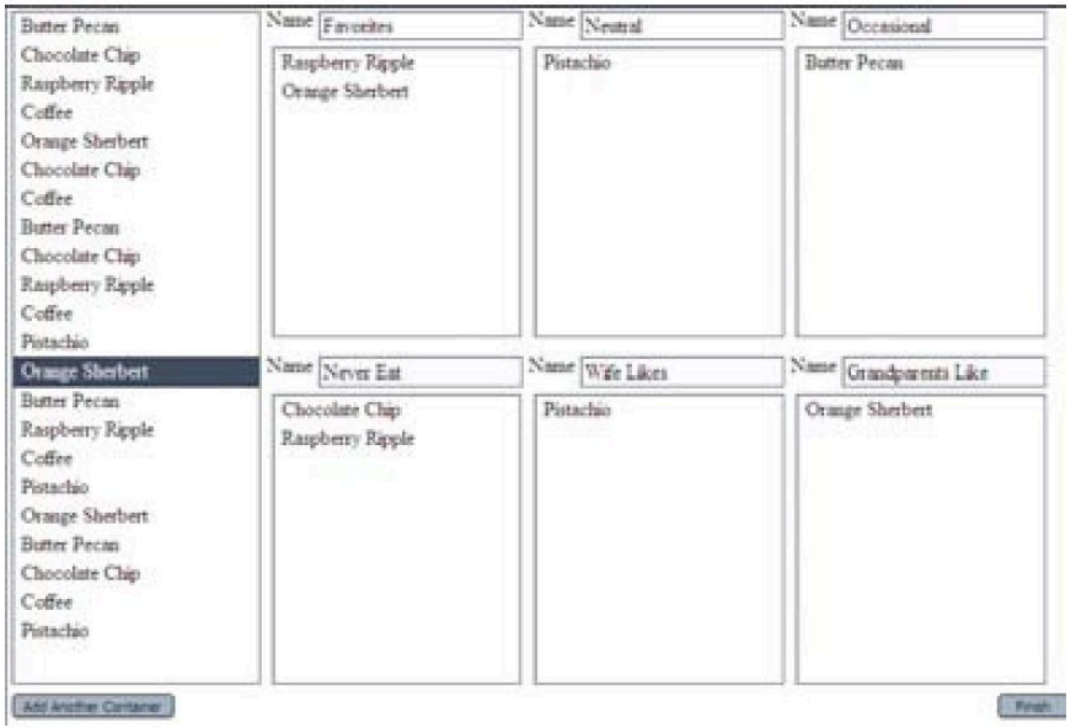
MDS Has Many Uses

- Psychology (perception, cognition)
- Political science (voting behavior, court decisions)
- Sociology (social network analysis)
- Archeology (artifact similarity)
- Biology/Chemistry (molecular structure, species analysis)
- Document retrieval & classification
- Graph layout
- Pattern recognition
- Dimension reduction
- …

Obtaining Dissimilarity Data

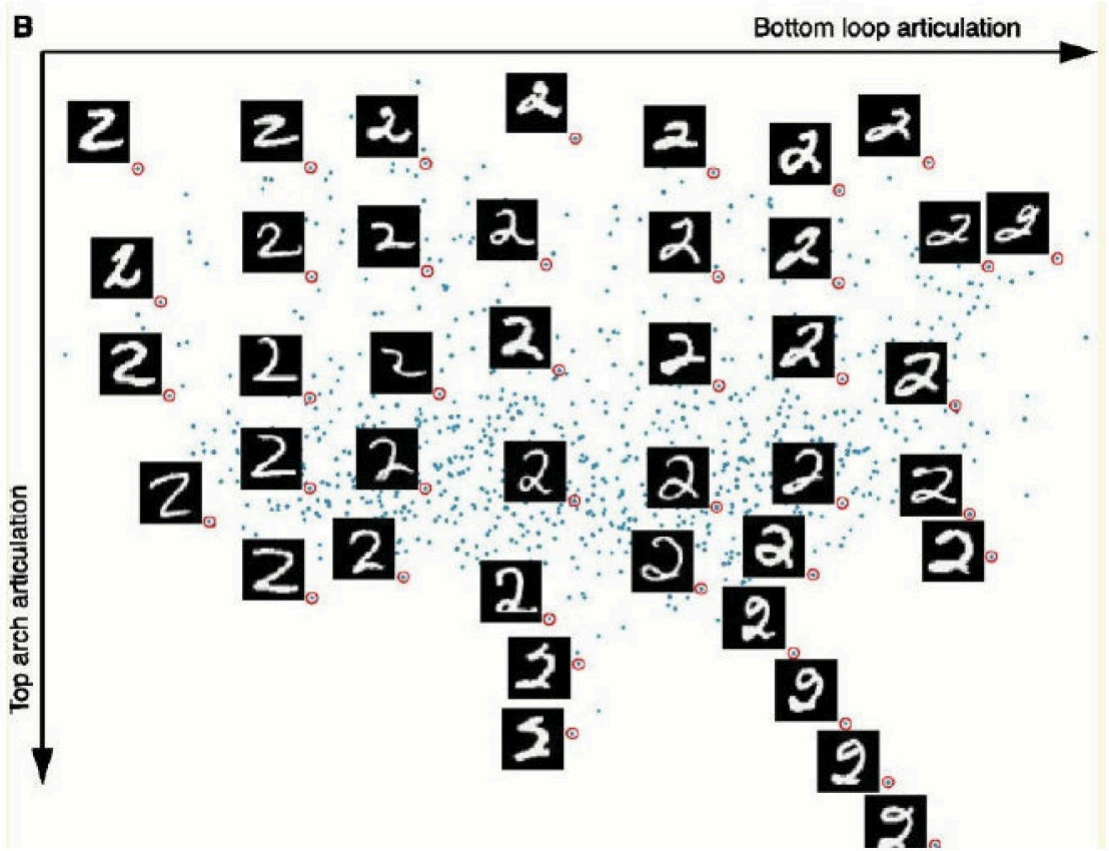
- Direct ratings (flavor comparisons)
- Confusion data (Morse dots and dashes)
- Co-occurrence data (Amazon recommendations)
- Sorting into groups

Sometimes distances are naturally defined – but sometimes we seek subjective dissimilarities



Can every distance matrix be realized in a Euclidean space?

Example: Pattern Recognition



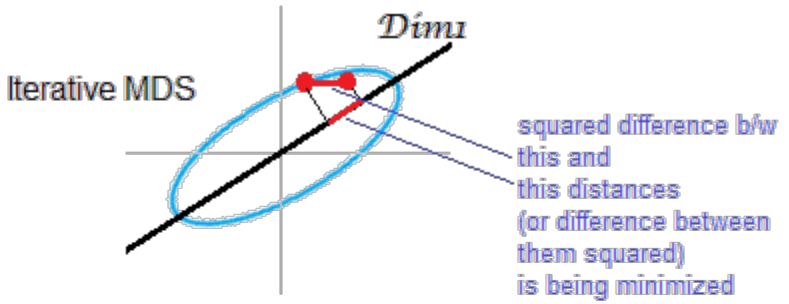
MDS of judged similarity of handwritten “2”s

Goal: determine features important in pattern recognition

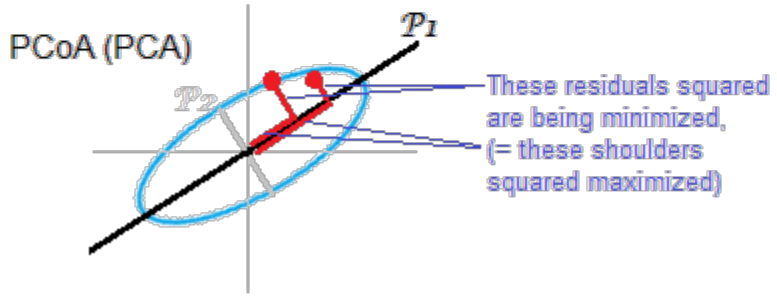
Classic Metric MDS

- Sometimes we can model our data as points in a high-dimensional Euclidean space – and we are looking for an embedding to a lower-dimensional space that preserves (absolute or relative) distances (in the high-d space) as much as possible.

- In this case the problem has a clean geometric solution.



Is this the same as PCA?



Classic Metric MDS

- To go from dimension D down to dimension d
- Given data $X \in R^{D \times n}$

$$X = \begin{pmatrix} | & | & & | \\ \mathbf{x_1} & \mathbf{x}_2 & \cdots & \mathbf{x}_n \\ | & | & & | \end{pmatrix} \quad \text{and} \quad M = \left(\text{dist}^2(\mathbf{x_i}, \mathbf{x_j})\right)_{n \times n}$$

- We look for X',

$$X' = \begin{pmatrix} | & & | \\ \mathbf{x_1}' & \cdots & \mathbf{x_n}' \\ | & & | \end{pmatrix} \in R^{d \times n}$$

- We can assume the $\mathbf{x}_i$' are centered

Classic Metric MDS

- So that we minimize $\| M' - M \|$ (related to the *stress* of the system)

- where $$M' = \left(\text{dist}^2(\mathbf{x_i}', \mathbf{x_j}') \right) = \left(\left\| \mathbf{x_i}' - \mathbf{x_j}' \right\|^2 \right) \in R^{n \times n}$$

- M' is the Euclidean distances matrix for points $\mathbf{x_i}'$.

$$\min \| M' - M \|$$

The Math Details

- Ideally we want $M' = \left(\left\| \mathbf{x}_i' - \mathbf{x}_j' \right\|^2 \right) = M$

$$\left(< \mathbf{x}_i' - \mathbf{x}_j', \ \mathbf{x}_i' - \mathbf{x}_j' > \right) = M$$

$$\left(\| \mathbf{x}_i' \|^2 + \| \mathbf{x}_j' \|^2 - 2 < \mathbf{x}_i', \mathbf{x}_j' > \right) = M$$

$$\begin{pmatrix} \| \mathbf{x}_1' \| & \| \mathbf{x}_1' \| & \dots & \| \mathbf{x}_1' \| \\ \| \mathbf{x}_2' \| & \| \mathbf{x}_2' \| & \dots & \| \mathbf{x}_2' \| \\ & & \vdots & \\ \| \mathbf{x}_n' \| & \| \mathbf{x}_n' \| & \dots & \| \mathbf{x}_n' \| \end{pmatrix} + \begin{pmatrix} \| \mathbf{x}_1' \| & \| \mathbf{x}_2' \| & & \| \mathbf{x}_n' \| \\ \| \mathbf{x}_1' \| & \| \mathbf{x}_2' \| & & \| \mathbf{x}_n' \| \\ \vdots & \vdots & \dots & \vdots \\ \| \mathbf{x}_1' \| & \| \mathbf{x}_2' \| & & \| \mathbf{x}_n' \| \end{pmatrix} - 2 \begin{pmatrix} — & \mathbf{x}_1' & — \\ & \vdots & \\ — & \mathbf{x}_n' & — \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{x}_1' & \cdots & \mathbf{x}_n' \\ | & & | \end{pmatrix}$$

want to get rid of these

$X'^T \qquad X'$

The Magic Matrix *J*

$$J = \begin{pmatrix} \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & -\frac{1}{n} & -\frac{1}{n} \\ \vdots & & \ddots & \vdots \\ -\frac{1}{n} & \cdots & -\frac{1}{n} & \frac{n-1}{n} \end{pmatrix}_{n\times n} = I - \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & 1 \\ \vdots & & \ddots & \vdots \\ 1 & \cdots & 1 & 1 \end{pmatrix} = I - \frac{1}{n}K$$

$$\begin{pmatrix} a & a & \cdots & a \end{pmatrix} \cdot J = 0$$

$$J \cdot \begin{pmatrix} b \\ b \\ \vdots \\ b \end{pmatrix} = 0$$

So We Get to The Gram Matrix

Cleaning the system:

$$\times J \Big/ \begin{pmatrix} \|\mathbf{x_1}'\| & \|\mathbf{x_1}'\| & \dots & \|\mathbf{x_1}'\| \\ \|\mathbf{x_2}'\| & \|\mathbf{x_2}'\| & \dots & \|\mathbf{x_2}'\| \\ & & \vdots & \\ \|\mathbf{x_n}'\| & \|\mathbf{x_n}'\| & \dots & \|\mathbf{x_n}'\| \end{pmatrix} + \begin{pmatrix} \|\mathbf{x_1}'\| & \|\mathbf{x_2}'\| & & \|\mathbf{x_n}'\| \\ \|\mathbf{x_1}'\| & \|\mathbf{x_2}'\| & & \|\mathbf{x_n}'\| \\ \vdots & \vdots & \dots & \vdots \\ \|\mathbf{x_1}'\| & \|\mathbf{x_2}'\| & & \|\mathbf{x_n}'\| \end{pmatrix} - 2X'^T X' = M \Big/ \times J$$

Note that $X'K = KX'^T = 0$, as X' is centered

$$J = I - \frac{1}{n}K$$

$$-2X'^T X' = JMJ$$

$$X'^T X' = -\tfrac{1}{2}JMJ =: B$$

$$X'^T X' = B$$

So from the distance matrix we can get the Gram (inner product) matrix.

And Finally Use the Spectral Hammer

We will use the spectral decomposition of B:

$$X'^T X' = B = \begin{pmatrix} | & & | \\ \mathbf{v_1} & \cdots & \mathbf{v_n} \\ | & & | \end{pmatrix} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{v_1} & \cdots & \mathbf{v_n} \\ | & & | \end{pmatrix}^T$$

$$X'^T X' = \underbrace{\begin{pmatrix} | & & | \\ \mathbf{v_1} & \cdots & \mathbf{v_d} \\ | & & | \end{pmatrix} \begin{pmatrix} | \\ \mathbf{v}_n \\ | \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_d} \end{pmatrix}}_{d \times d} \begin{pmatrix} \ddots & \\ & \sqrt{\lambda_n} \end{pmatrix}}_{X'^T} \underbrace{\begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_d} \end{pmatrix} \begin{pmatrix} \ddots & \\ & \sqrt{\lambda_n} \end{pmatrix} \begin{pmatrix} | & & | \\ \mathbf{v_1} & \cdots & \mathbf{v_d} \\ | & & | \end{pmatrix}^T \begin{pmatrix} | \\ \mathbf{v}_n \\ | \end{pmatrix}^T}_{X'}$$

$n \times d$

So We Get the *X'*

So we find X' by throwing away the last $n-d$ eigenvalues

$$X' = \begin{pmatrix} \sqrt{\lambda_1}\,\mathbf{v_1} \\ \cdots \quad \cdots \quad \cdots \\ \sqrt{\lambda_d}\,\mathbf{v_d} \end{pmatrix}_{d \times n}$$

For this *X'* : $$X' = \arg\min_{X'} \left\| X'^T X' - B \right\|_{L^2}$$

This choice minimizes the inner product (and distance) loss

$$\|A\|_{L^2} = \sqrt{\sum_{i,j} A_{ij}^2}$$

More General Metric MDS

- In general, we minimize directly the square loss on distances

$$\text{stress} = \mathcal{L}(\hat{d}_{ij}) = \left(\sum_{i<j} (\hat{d}_{ij} - f(d_{ij}))^2 / \sum d_{ij}^2 \right)^{\frac{1}{2}}$$

f monotonic

- Sammon mapping

$$\text{Sammon's stress}(\hat{d}_{ij}) = \frac{1}{\sum_{\ell<k} d_{\ell k}} \sum_{i<j} \frac{(\hat{d}_{ij} - d_{ij})^2}{d_{ij}}$$

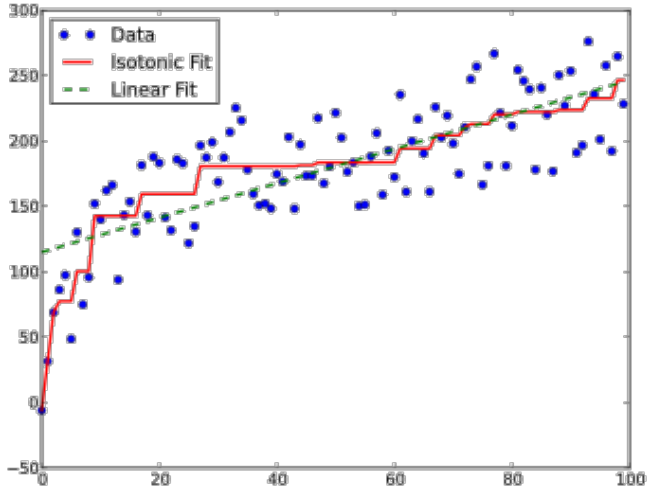
  - This weighting system normalizes the squared-errors in pairwise distances by using the distance in the original space. As a result, Sammon mapping preserves the small d_{ij}, giving them a greater degree of importance in the fitting procedure than for larger values of d_{ij}

Generally solved by gradient descent

Non-Metric MDS

- Sometimes all we can say is that (dis)similarity is *ordinally* related to distance in MD space (only ordering of distances matters, not the actual values)
- If we only have ordering information, we can use *monotone (isotonic) regression* to find "disparities" that are compatible with the ordering constraints of the dissimilarities (not unique)
- We can then use alternating least squares (ALS) by repeating
  - a monotone regression step, followed by
  - a metric stress reduction step

$$Stress(\hat{\mathbf{D}}, \mathbf{X}) = \left[\frac{\sum_{i<j} \left(\hat{d}_{ij} - d_{ij}(\mathbf{X}) \right)^2}{\sum_{i<j} d_{ij}^2(\mathbf{X})} \right]^{1/2}$$



The 1st Assignment

Assignment 1: PCA + CCA

- Problem 1 (PCA): Face Reconstruction/Recognition: Eigenfaces
- Problem 2 (CCA): Word Image Recognition

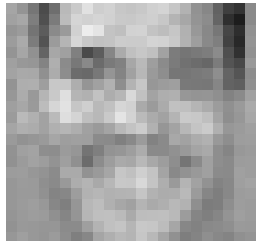
PCA

Face Recognition

- Photo organization
- Surveillance
- …
- Has to be coupled with face detection

Recognition: Embed!

- Treat image pixels as a long vector



$$\mathbf{x}$$

- Face recognition by nearest neighbor

$$\mathbf{y}_1 \cdots \mathbf{y}_n$$

$$k = \underset{k}{\operatorname{argmin}} \left\| \mathbf{y}_k{}^T - \mathbf{x} \right\|$$

Eigenfaces (PCA on Face Images)

- Compute covariance matrix of face images
- Derive the principal components ("eigenfaces")
  - *K* eigenvectors with largest eigenvalues
- Represent all face images in the dataset as linear combinations of eigenfaces
  - Perform nearest neighbor on these coefficients

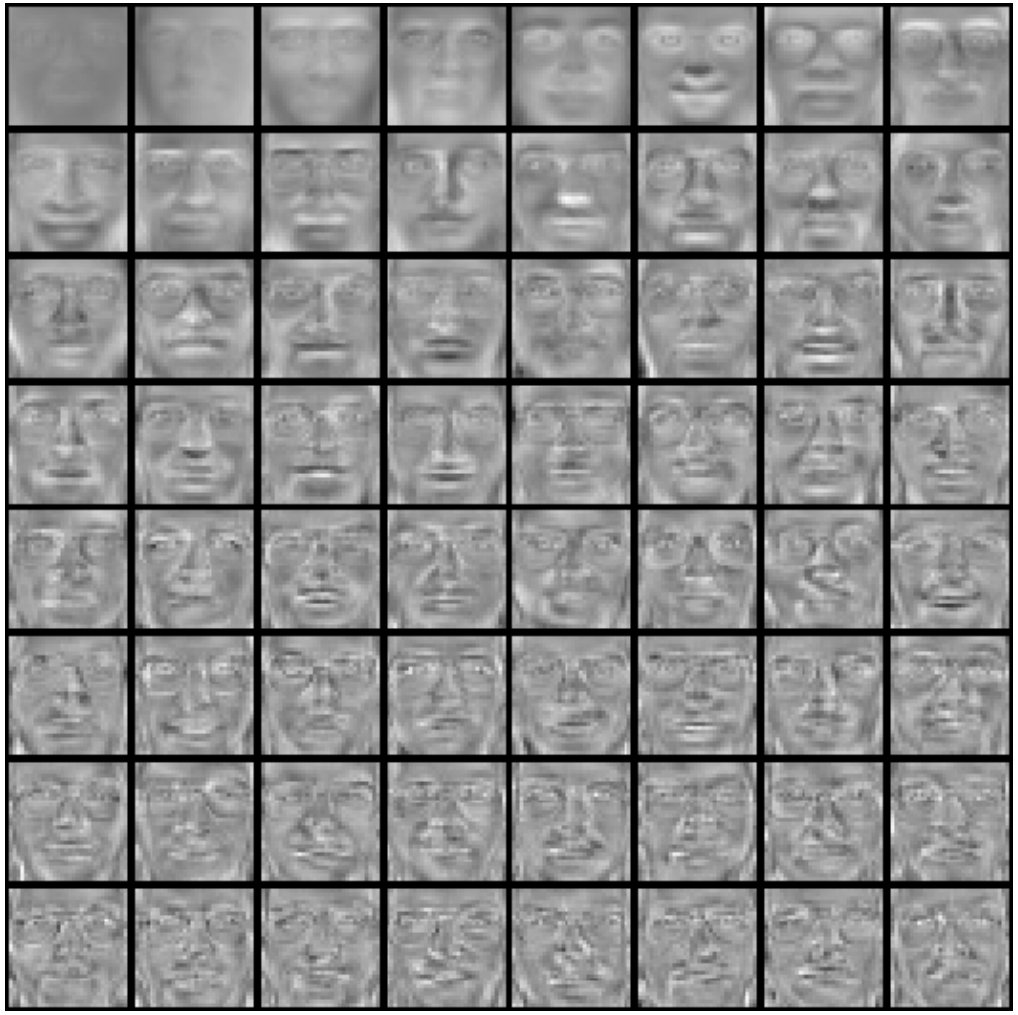
M. Turk and A. Pentland, Face Recognition using Eigenfaces, CVPR 1991

Training Images

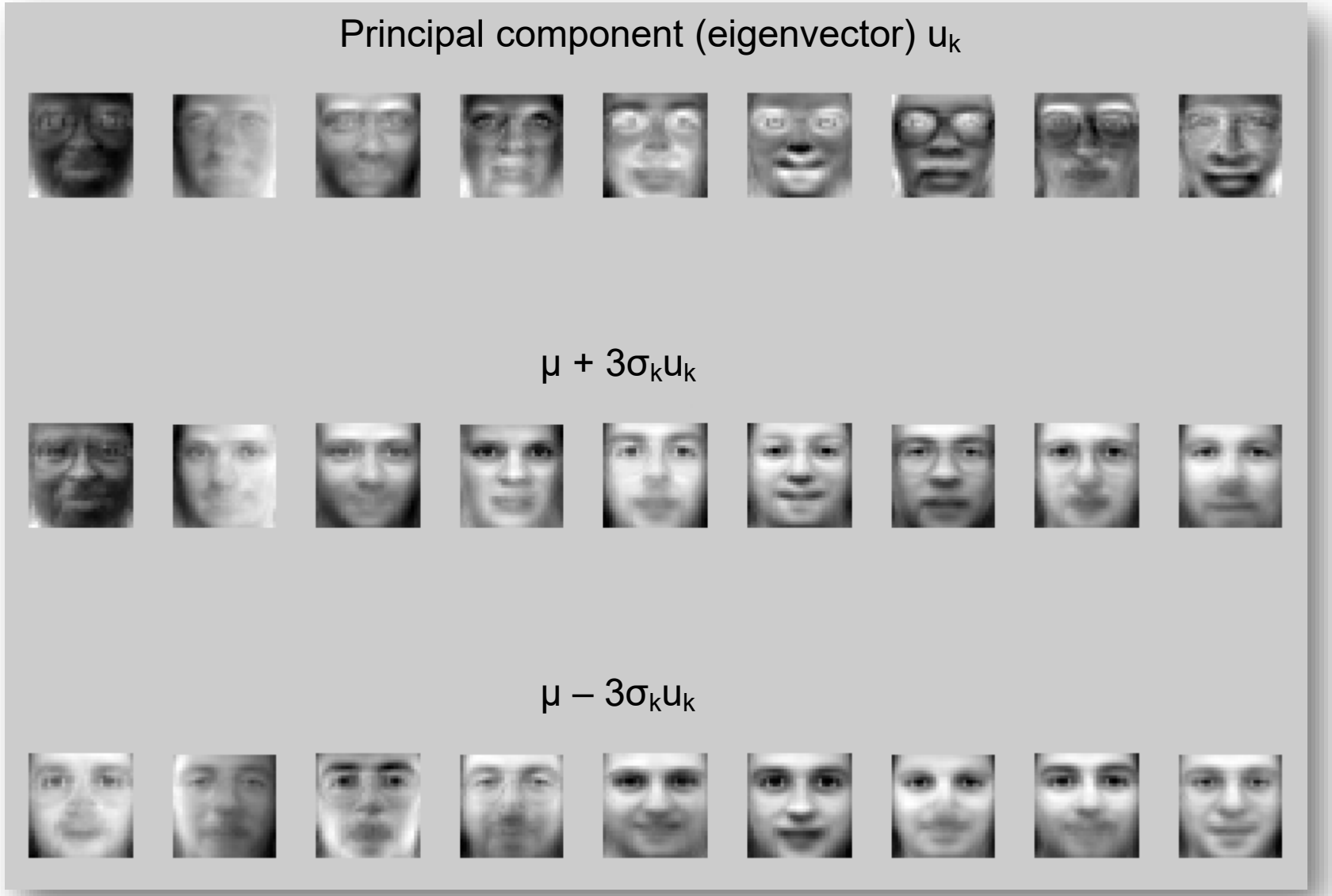
Eigenfaces

Mean: μ

Top eigenvectors: $u_1,\ldots u_k$



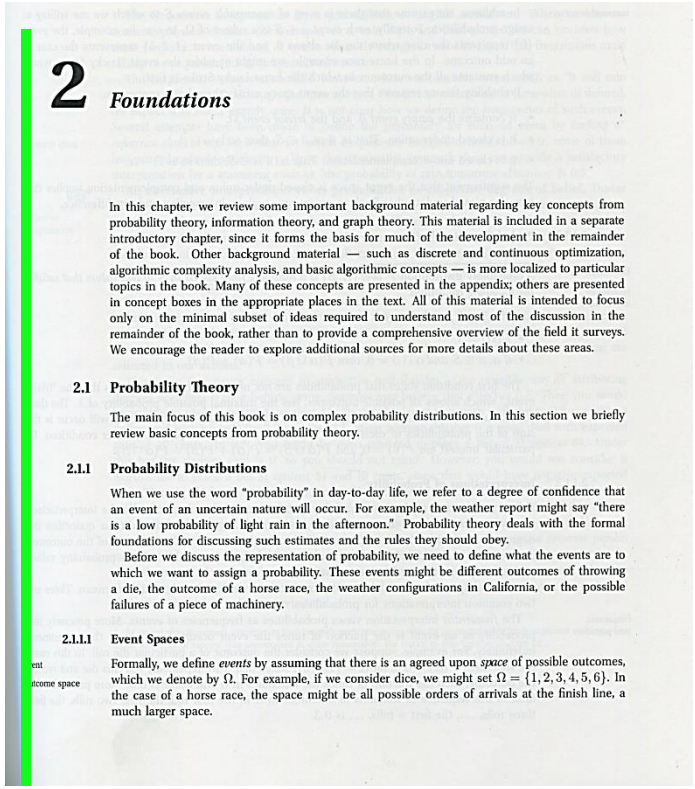
Eigenface Visualization



CCA

Camera Text Recognition

Scanned



2 *Foundations*

In this chapter, we review some important background material regarding key concepts from probability theory, information theory, and graph theory. This material is included in a separate introductory chapter, since it forms the basis for much of the development in the remainder of the book. Other background material — such as discrete and continuous optimization, algorithmic complexity analysis, and basic algorithmic concepts — is more localized to particular topics in the book. Many of these concepts are presented in the appendix; others are presented in concept boxes in the appropriate places in the text. All of this material is intended to focus only on the minimal subset of ideas required to understand most of the discussion in the remainder of the book, rather than to provide a comprehensive overview of the field it surveys. We encourage the reader to explore additional sources for more details about these areas.

2.1 Probability Theory

The main focus of this book is on complex probability distributions. In this section we briefly review basic concepts from probability theory.

2.1.1 Probability Distributions

When we use the word "probability" in day-to-day life, we refer to a degree of confidence that an event of an uncertain nature will occur. For example, the weather report might say "there is a low probability of light rain in the afternoon." Probability theory deals with the formal foundations for discussing such estimates and the rules they should obey.

Before we discuss the representation of probability, we need to define what the events are to which we want to assign a probability. These events might be different outcomes of throwing a die, the outcome of a horse race, the weather configurations in California, or the possible failures of a piece of machinery.

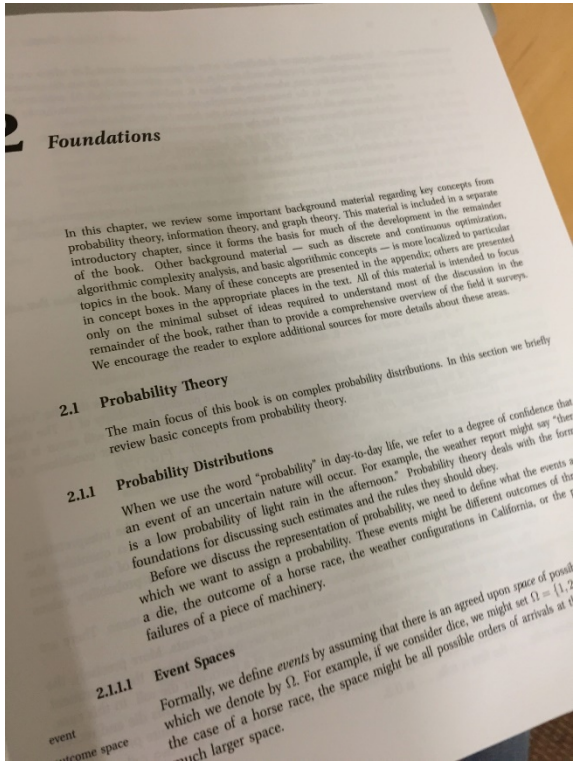
2.1.1.1 Event Spaces

ent
tcome space

Formally, we define *events* by assuming that there is an agreed upon *space* of possible outcomes, which we denote by Ω. For example, if we consider dice, we might set $\Omega = \{1, 2, 3, 4, 5, 6\}$. In the case of a horse race, the space might be all possible orders of arrivals at the finish line, a much larger space.

In this chapter, we review some important background material regarding key concepts from probability theory, information theory, and graph theory…

Camera



Foundations

In this chapter, we review some important background material regarding key concepts from probability theory, information theory, and graph theory. This material is included in a separate introductory chapter, since it forms the basis for much of the development in the remainder of the book. Other background material — such as discrete and continuous optimization, algorithmic complexity analysis, and basic algorithmic concepts — is more localized to particular topics in the book. Many of these concepts are presented in the appendix; others are presented in concept boxes in the appropriate places in the text. All of this material is intended to focus only on the minimal subset of ideas required to understand most of the discussion in the remainder of the book, rather than to provide a comprehensive overview of the field it surveys. We encourage the reader to explore additional sources for more details about these areas.

2.1 Probability Theory

The main focus of this book is on complex probability distributions. In this section we briefly review basic concepts from probability theory.

2.1.1 Probability Distributions

When we use the word "probability" in day-to-day life, we refer to a degree of confidence that an event of an uncertain nature will occur. For example, the weather report might say "there is a low probability of light rain in the afternoon." Probability theory deals with the formal foundations for discussing such estimates and the rules they should obey.

Before we discuss the representation of probability, we need to define what the events are to which we want to assign a probability. These events might be different outcomes of throwing a die, the outcome of a horse race, the weather configurations in California, or the possible failures of a piece of machinery.

2.1.1.1 Event Spaces

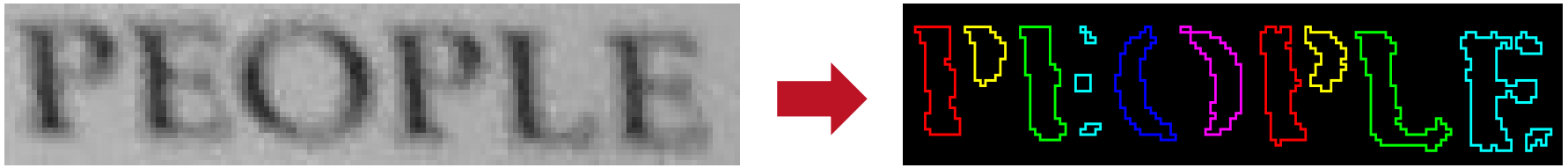
event
tcome space

Formally, we define *events* by assuming that there is an agreed upon *space* of possible outcomes, which we denote by Ω. For example, if we consider dice, we might set $\Omega = \{1, 2, 3, 4, 5, 6\}$. In the case of a horse race, the space might be all possible orders of arrivals at the finish line, a much larger space.

1n this chapter we FBViEW SUIHC important background mated-M regarding WY cncepn 5mm probability theory information theory and Efaph theory…

Why OCR Fails

- OCR reads text character by character, but character segmentation can be challenging.
  - View perspective distortions
  - Uneven lighting
  - Presence of noise and blur



Word Recognition via Image Retrieval

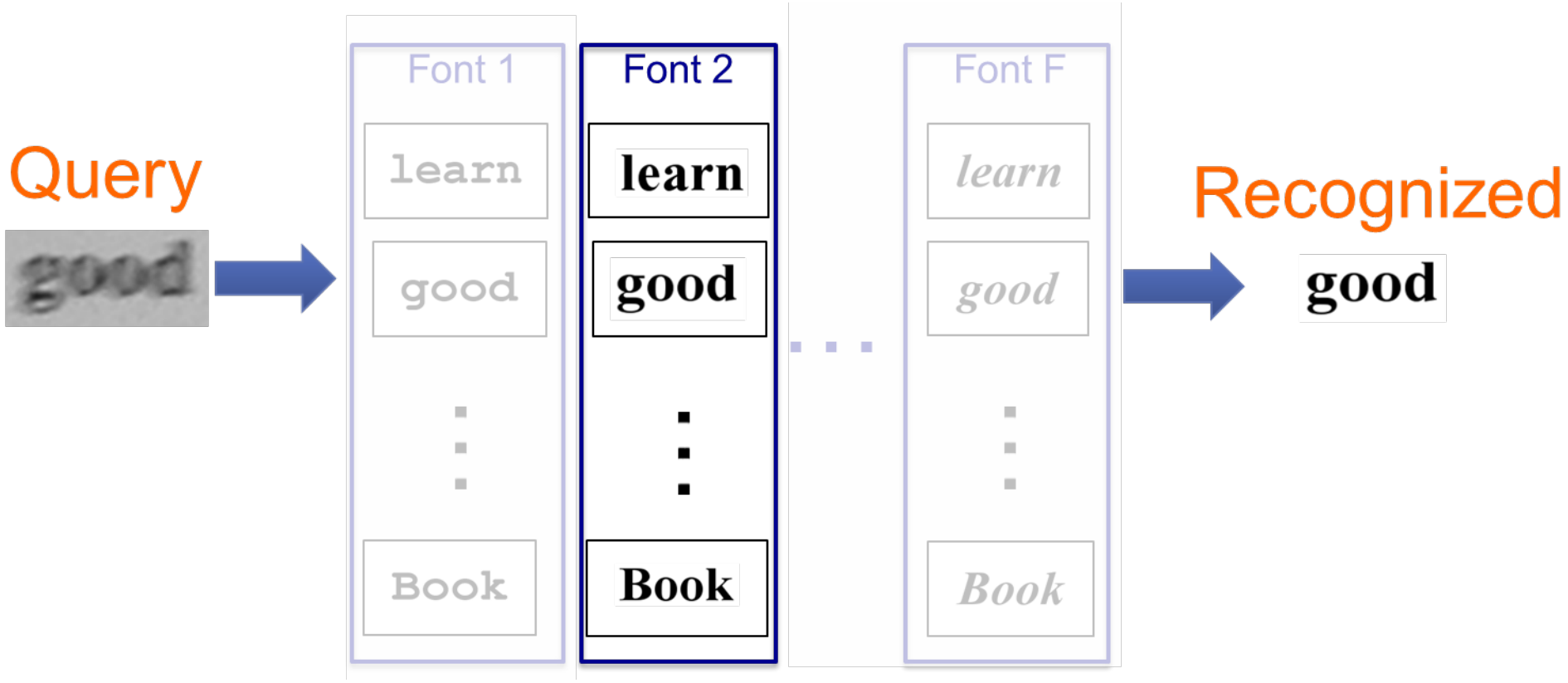
Query

Recognized

good

Huizhong Chen, Visual Word Recognition with Large-Scale Image Retrieval
Stanford EE Ph.D. Thesis, 2015

If Font Is Known, Task is Easier

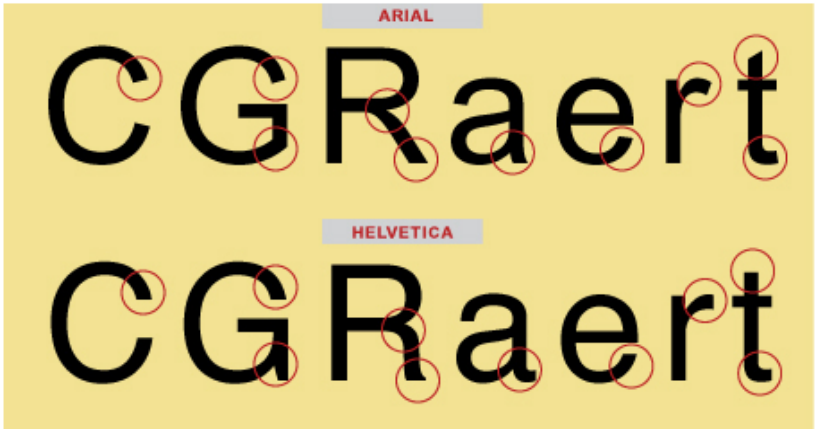


Inter-font Similarities

- Some fonts have visual similarities (e.g., Arial, Helvetica).

THE QUICK BROWN FOX JUMPS OVER THE LAZY DOG
THE QUICK BROWN FOX JUMPS OVER THE LAZY DOG

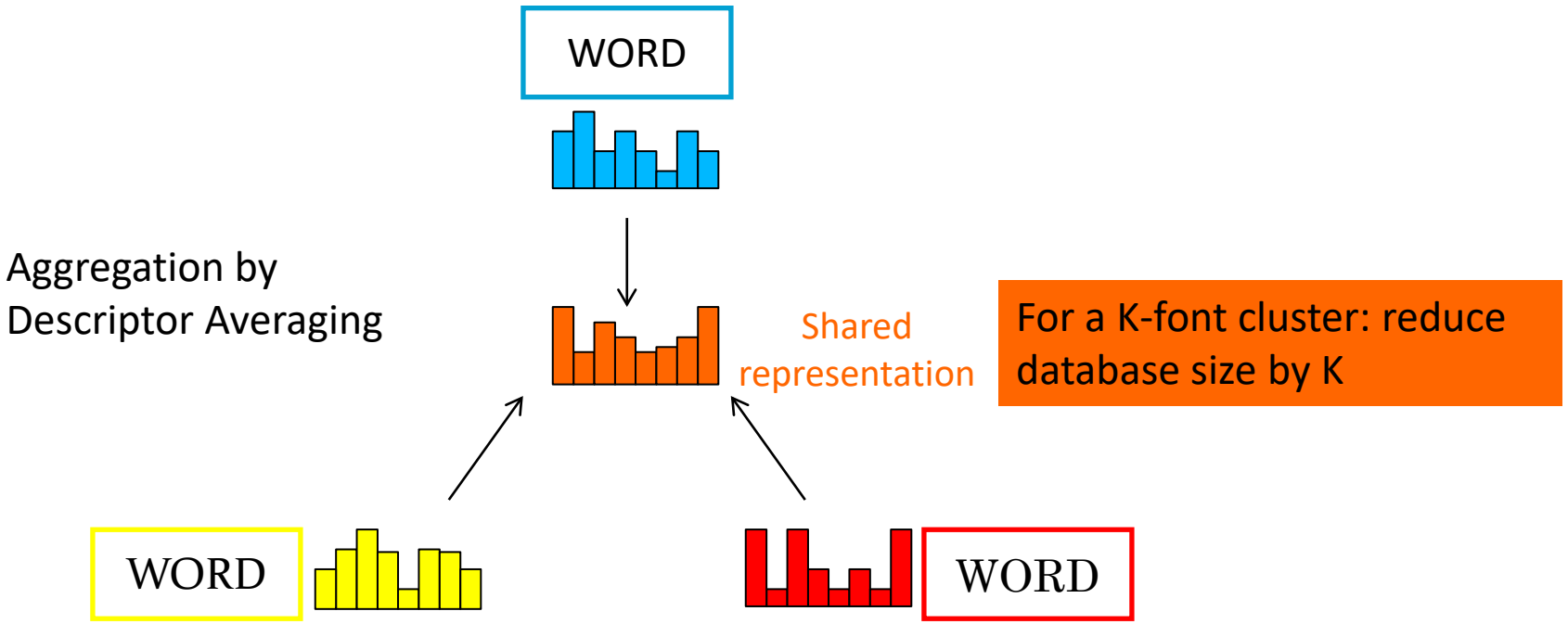
the quick brown fox jumps over the lazy dog
the quick brown fox jumps over the lazy dog



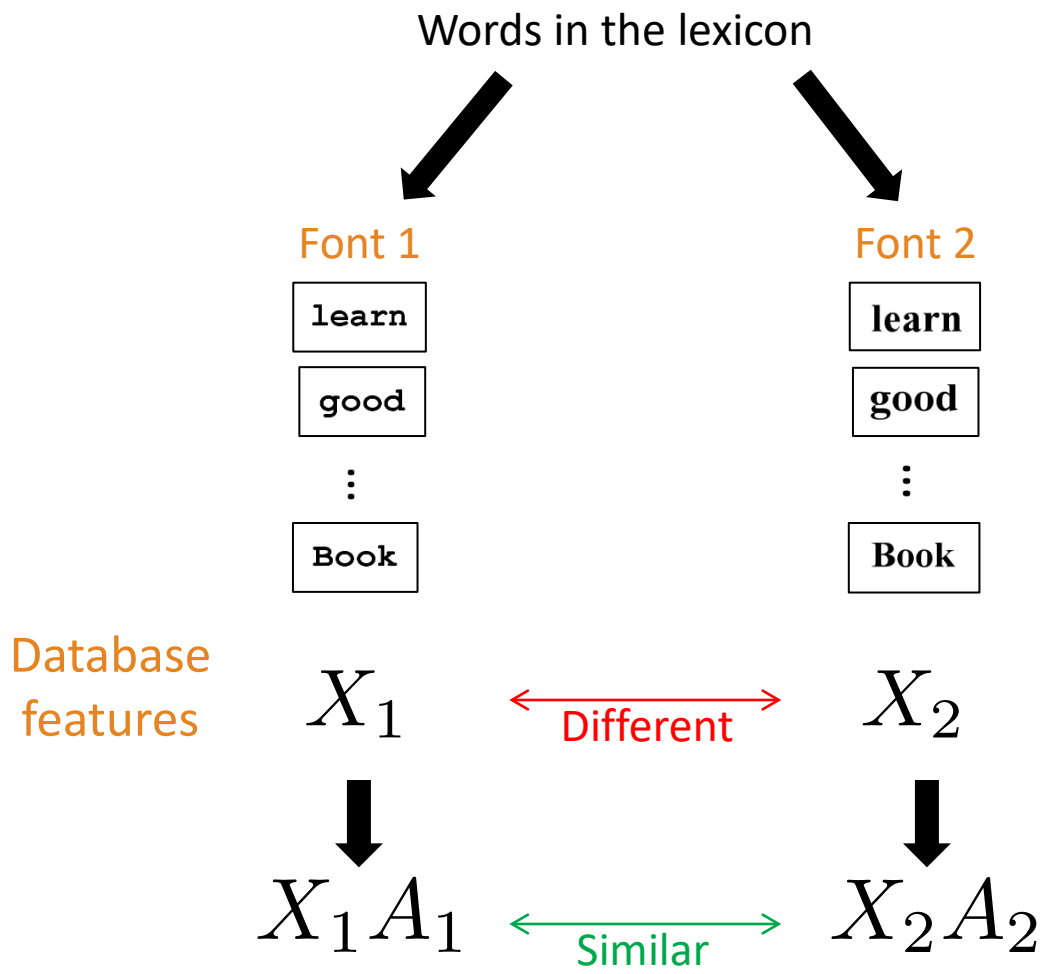
Source: www.fonts.com

Compact Data Base Representation

Font Clustering ⇨ Feature Aggregation



Motivation for CCA



Two View vs. Multi-View CCA

The End