

# Combining Dissimilarity Representations in Embedding Product Space

Zhiliang Ma

Department of Applied Mathematics & Statistics  
Johns Hopkins University

May 22, 2008  
Interface, Durham, NC

# Combining Dissimilarity Representations in Embedding Product Space

Zhiliang Ma

Department of Applied Mathematics & Statistics  
Johns Hopkins University



Carey E.  
Priebe



Adam  
Cardinal-  
Stakenas



Youngser  
Park

May 22, 2008  
Interface, Durham, NC

# Dissimilarity Representations

## Definition

A *dissimilarity measure* is a function  $\delta : \Xi \times \Xi \rightarrow \mathcal{R}_+ \cup \{0\}$  with:

1. Positivity:  $\delta(x_1, x_2) \geq 0$
2. Symmetry:  $\delta(x_1, x_2) = \delta(x_2, x_1)$
3. Reflexivity:  $\delta(x, x) = 0$

# Dissimilarity Representations

## Definition

A *dissimilarity measure* is a function  $\delta : \Xi \times \Xi \rightarrow \mathcal{R}_+ \cup \{0\}$  with:

1. Positivity:  $\delta(x_1, x_2) \geq 0$
2. Symmetry:  $\delta(x_1, x_2) = \delta(x_2, x_1)$
3. Reflexivity:  $\delta(x, x) = 0$
4. Triangle Inequality:  $\delta(x_1, x_3) \leq \delta(x_1, x_2) + \delta(x_2, x_3)$

# Dissimilarity Representations

## Definition

A *dissimilarity measure* is a function  $\delta : \Xi \times \Xi \rightarrow \mathcal{R}_+ \cup \{0\}$  with:

1. Positivity:  $\delta(x_1, x_2) \geq 0$
2. Symmetry:  $\delta(x_1, x_2) = \delta(x_2, x_1)$
3. Reflexivity:  $\delta(x, x) = 0$
4. Triangle Inequality:  ~~$\delta(x_1, x_3) \leq \delta(x_1, x_2) + \delta(x_2, x_3)$~~

# Dissimilarity Representations

## Definition

A *dissimilarity measure* is a function  $\delta : \Xi \times \Xi \rightarrow \mathcal{R}_+ \cup \{0\}$  with:

1. Positivity:  $\delta(x_1, x_2) \geq 0$
2. Symmetry:  $\delta(x_1, x_2) = \delta(x_2, x_1)$
3. Reflexivity:  $\delta(x, x) = 0$
4. Triangle Inequality:  ~~$\delta(x_1, x_3) \leq \delta(x_1, x_2) + \delta(x_2, x_3)$~~

A *dissimilarity representation* for a set of  $n$  objects is expressed as a symmetric, nonnegative and hollow matrix  $\Delta$ .

# Why Using Dissimilarity Representations

## Why Using Dissimilarity Representations

- Effective features are hard to extract from the raw or pre-processed data, while pairwise comparisons between objects can be directly derived. For example, graphs, shapes, images, spectra and etc.

## Why Using Dissimilarity Representations

- Effective features are hard to extract from the raw or pre-processed data, while pairwise comparisons between objects can be directly derived. For example, graphs, shapes, images, spectra and etc.
- Unable to access to feature data, the only observation is the interpoint comparisons. For example the brain shape comparisons (LDDMM). There are many such examples in Psychology, too.

# How Do People Deal with Dissimilarities?

– [R.P.W. Duin's three approaches]

# How Do People Deal with Dissimilarities?

– [R.P.W. Duin's three approaches]

- The *neighborhood-based approach* interprets dissimilarity values as neighborhood relations

# How Do People Deal with Dissimilarities?

– [R.P.W. Duin's three approaches]

- The *neighborhood-based approach* interprets dissimilarity values as neighborhood relations
- The *dissimilarity space approach* defines a representation set  $R = \{p_1, \dots, p_r\}$ , and interprets dissimilarities from a point to each element of the representation set as features of this point

# How Do People Deal with Dissimilarities?

– [R.P.W. Duin's three approaches]

- The *neighborhood-based approach* interprets dissimilarity values as neighborhood relations
- The *dissimilarity space approach* defines a representation set  $R = \{p_1, \dots, p_r\}$ , and interprets dissimilarities from a point to each element of the representation set as features of this point
- The *embedding approach* embeds dissimilarity matrix into  $\mathcal{R}^d$ , in such a way that the interpoint distances,  $\|x_i - x_j\|_2$ , approximate the dissimilarities,  $\delta_{ij}$

# Examples of Dissimilarity Measures

for a set of objects

Euclidean distance  $\delta_{ij} = \left( \sum_k (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}}$

Mahalanobis distance  $\delta_{ij} = \left( (\mathbf{x}_i - \mathbf{x}_j)^T \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j) \right)^{\frac{1}{2}}$

Minkowski metric  $\delta_{ij} = \left( \sum_k |x_{ik} - x_{jk}|^\lambda \right)^{\frac{1}{\lambda}}$

# Examples of Dissimilarity Measures

for a set of objects

Euclidean distance  $\delta_{ij} = \left( \sum_k (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}}$

Mahalanobis distance  $\delta_{ij} = \left( (\mathbf{x}_i - \mathbf{x}_j)^T \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j) \right)^{\frac{1}{2}}$

Minkowski metric  $\delta_{ij} = \left( \sum_k |x_{ik} - x_{jk}|^\lambda \right)^{\frac{1}{\lambda}}$

Other dissimilarity measures: LDDMM, Hausdorff distance, angular separation, and etc.

# Examples of Dissimilarity Measures

for a set of objects

Euclidean distance  $\delta_{ij} = \left( \sum_k (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}}$

Mahalanobis distance  $\delta_{ij} = \left( (\mathbf{x}_i - \mathbf{x}_j)^T \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j) \right)^{\frac{1}{2}}$

Minkowski metric  $\delta_{ij} = \left( \sum_k |x_{ik} - x_{jk}|^\lambda \right)^{\frac{1}{\lambda}}$

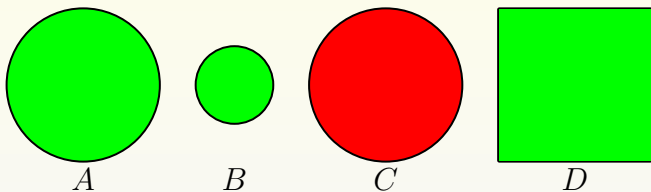
Other dissimilarity measures: LDDMM, Hausdorff distance, angular separation, and etc.

## Question

Is it beneficial to combine multiple dissimilarities?

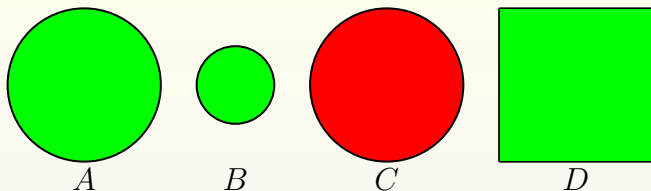
# Combining Dissimilarity Matrices

– A Toy Example (4 - class classification problem)



# Combining Dissimilarity Matrices

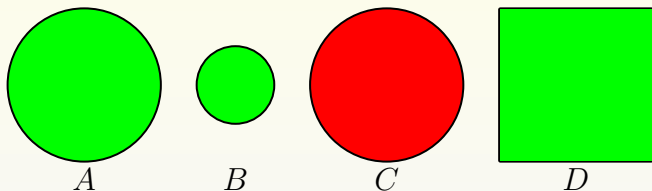
– A Toy Example (4 - class classification problem)



$$\delta_1(i, j) = I\{i \text{ and } j \text{ have different } \textit{shape}\}$$

# Combining Dissimilarity Matrices

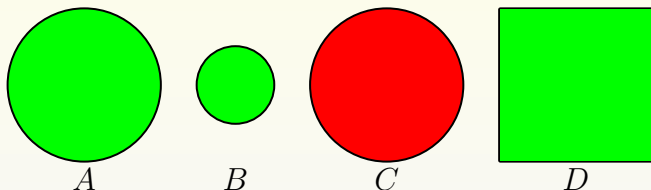
– A Toy Example (4 - class classification problem)



$$\delta_1(i, j) = I\{i \text{ and } j \text{ have different } \textit{shape}\} \quad \Rightarrow \{ABC, D\}$$

# Combining Dissimilarity Matrices

– A Toy Example (4 - class classification problem)



$$\delta_1(i, j) = I\{i \text{ and } j \text{ have different } \textit{shape}\}$$

$$\Rightarrow \{ABC, D\}$$

$$\delta_2(i, j) = I\{i \text{ and } j \text{ have different } \textit{color}\}$$

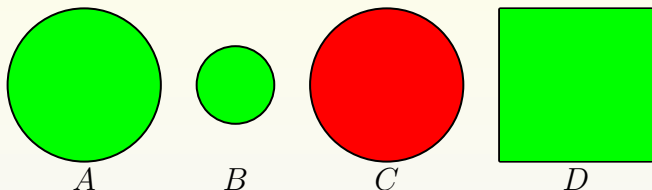
$$\Rightarrow \{ABD, C\}$$

$$\delta_3(i, j) = I\{i \text{ and } j \text{ have different } \textit{size}\}$$

$$\Rightarrow \{ACD, B\}$$

# Combining Dissimilarity Matrices

– A Toy Example (4 - class classification problem)



$$\delta_1(i, j) = I\{i \text{ and } j \text{ have different } \textit{shape}\} \quad \Rightarrow \{ABC, D\}$$

$$\delta_2(i, j) = I\{i \text{ and } j \text{ have different } \textit{color}\} \quad \Rightarrow \{ABD, C\}$$

$$\delta_3(i, j) = I\{i \text{ and } j \text{ have different } \textit{size}\} \quad \Rightarrow \{ACD, B\}$$

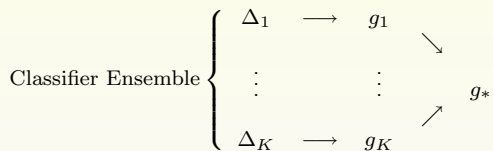
$$\delta_*(i, j) = \delta_1(i, j) + \delta_2(i, j) + \delta_3(i, j) \quad \Rightarrow \{A, B, C, D\}$$

# Combining Dissimilarity Matrices

– Three ways

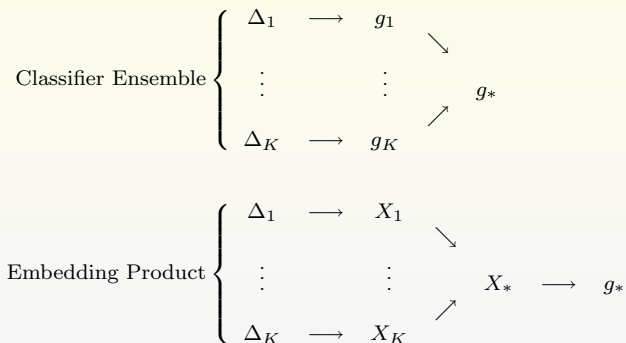
# Combining Dissimilarity Matrices

– Three ways



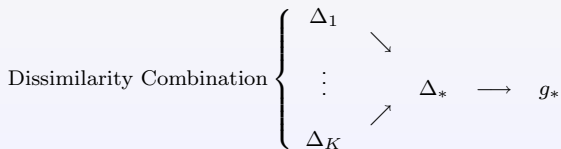
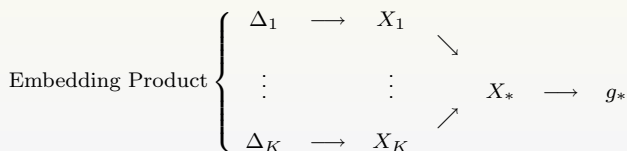
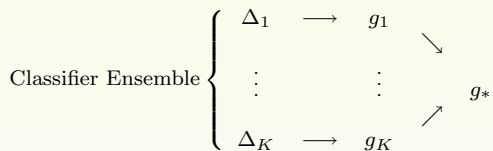
# Combining Dissimilarity Matrices

– Three ways



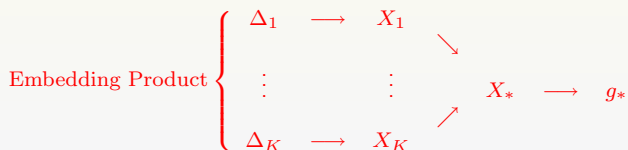
# Combining Dissimilarity Matrices

– Three ways



# Combining Dissimilarity Matrices

– Three ways



# Combining Dissimilarities Through Embedding Product

– Miller et al.(2007)

# Combining Dissimilarities Through Embedding Product

– Miller et al.(2007)

- Embedding:  $\Delta_1 \rightarrow X_1 \in \mathcal{R}^{p_1}$ ,  $\Delta_2 \rightarrow X_2 \in \mathcal{R}^{p_2}$

# Combining Dissimilarities Through Embedding Product

– Miller et al.(2007)

- Embedding:  $\Delta_1 \rightarrow X_1 \in \mathcal{R}^{p_1}$ ,  $\Delta_2 \rightarrow X_2 \in \mathcal{R}^{p_2}$
- $X = [X_1 \ X_2] \in \mathcal{R}^{p_1+p_2}$ , train a classifier  $g$  and results classification error  $\hat{L}_{p_1,p_2} \triangleq P[g(X) \neq Y]$

# Combining Dissimilarities Through Embedding Product

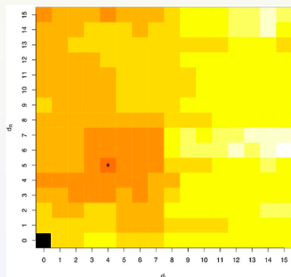
– Miller et al.(2007)

- Embedding:  $\Delta_1 \rightarrow X_1 \in \mathcal{R}^{p_1}$ ,  $\Delta_2 \rightarrow X_2 \in \mathcal{R}^{p_2}$
- $X = [X_1 \ X_2] \in \mathcal{R}^{p_1+p_2}$ , train a classifier  $g$  and results classification error  $\hat{L}_{p_1,p_2} \triangleq P[g(X) \neq Y]$
- $(p_1^*, p_2^*) = \arg \min_{p_1=0:d_1, p_2=0:d_2} \hat{L}_{p_1,p_2}$

# Combining Dissimilarities Through Embedding Product

– Miller et al.(2007)

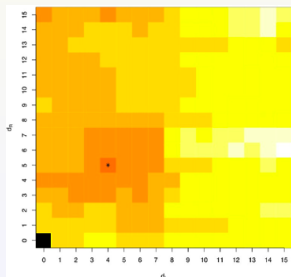
- Embedding:  $\Delta_1 \rightarrow X_1 \in \mathcal{R}^{p_1}$ ,  $\Delta_2 \rightarrow X_2 \in \mathcal{R}^{p_2}$
- $X = [X_1 \ X_2] \in \mathcal{R}^{p_1+p_2}$ , train a classifier  $g$  and results classification error  $\hat{L}_{p_1,p_2} \triangleq P[g(X) \neq Y]$
- $(p_1^*, p_2^*) = \arg \min_{p_1=0:d_1, p_2=0:d_2} \hat{L}_{p_1,p_2}$



# Combining Dissimilarities Through Embedding Product

– Miller et al.(2007)

- Embedding:  $\Delta_1 \rightarrow X_1 \in \mathcal{R}^{p_1}$ ,  $\Delta_2 \rightarrow X_2 \in \mathcal{R}^{p_2}$
- $X = [X_1 \ X_2] \in \mathcal{R}^{p_1+p_2}$ , train a classifier  $g$  and results classification error  $\hat{L}_{p_1,p_2} \triangleq P[g(X) \neq Y]$
- $(p_1^*, p_2^*) = \arg \min_{p_1=0:d_1, p_2=0:d_2} \hat{L}_{p_1,p_2}$



Need to search  $\prod_{k=1}^K (d_k + 1) - 1$  times. When  $K > 2$  ??

# Combining Dissimilarities Through Embedding Product

– Principal Components Analysis

# Combining Dissimilarities Through Embedding Product

– Principal Components Analysis

- Embedding:  $\Delta_i \rightarrow X_i \in \mathcal{R}^{d_i}$

# Combining Dissimilarities Through Embedding Product

– Principal Components Analysis

- Embedding:  $\Delta_i \rightarrow X_i \in \mathcal{R}^{d_i}$
- $\tilde{X} = [X_1, X_2, \dots, X_K] \in \mathcal{R}^d, d = \sum d_k$

# Combining Dissimilarities Through Embedding Product

## – Principal Components Analysis

- Embedding:  $\Delta_i \rightarrow X_i \in \mathcal{R}^{d_i}$
- $\tilde{X} = [X_1, X_2, \dots, X_K] \in \mathcal{R}^d$ ,  $d = \sum d_k$
- $X = PCA(\tilde{X}) \in \mathcal{R}^d \implies X(p)$ ,  $p^* = \arg \min_{p=1:d} \hat{L}_p$

# Combining Dissimilarities Through Embedding Product

## – Principal Components Analysis

- Embedding:  $\Delta_i \rightarrow X_i \in \mathcal{R}^{d_i}$
- $\tilde{X} = [X_1, X_2, \dots, X_K] \in \mathcal{R}^d$ ,  $d = \sum d_k$
- $X = PCA(\tilde{X}) \in \mathcal{R}^d \implies X(p)$ ,  $p^* = \arg \min_{p=1:d} \hat{L}_p$

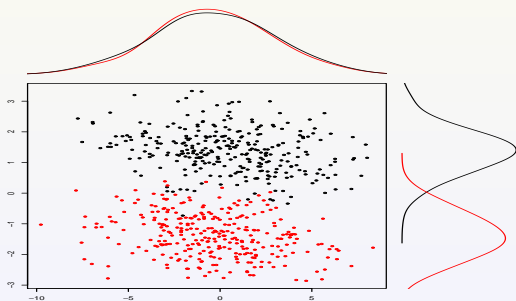
Need to search  $\sum_{k=1}^K d_k$  times

# Combining Dissimilarities Through Embedding Product

– Principal Components Analysis

- Embedding:  $\Delta_i \rightarrow X_i \in \mathcal{R}^{d_i}$
- $\tilde{X} = [X_1, X_2, \dots, X_K] \in \mathcal{R}^d$ ,  $d = \sum d_k$
- $X = PCA(\tilde{X}) \in \mathcal{R}^d \implies X(p)$ ,  $p^* = \arg \min_{p=1:d} \hat{L}_p$

Need to search  $\sum_{k=1}^K d_k$  times



Problem: parallel cigars!

# Combining Dissimilarities Through Embedding Product

–  $J$  function, a model selection & dimension reduction method

- $\tilde{X} \triangleq [X_1, X_2, \dots, X_K], \Sigma = \pi\Sigma_1 + (1 - \pi)\Sigma_0$

# Combining Dissimilarities Through Embedding Product

–  $J$  function, a model selection & dimension reduction method

- $\tilde{X} \triangleq [X_1, X_2, \dots, X_K]$ ,  $\Sigma = \pi\Sigma_1 + (1 - \pi)\Sigma_0$
- $X = P\tilde{X}$ , where  $P$  satisfies  $\Sigma = P\Lambda P^T$

# Combining Dissimilarities Through Embedding Product

–  $J$  function, a model selection & dimension reduction method

- $\tilde{X} \triangleq [X_1, X_2, \dots, X_K]$ ,  $\Sigma = \pi\Sigma_1 + (1 - \pi)\Sigma_0$
- $X = P\tilde{X}$ , where  $P$  satisfies  $\Sigma = P\Lambda P^T$
- Let  $\mu_c = E[X|Y = c] \in \mathcal{R}^d$ , where  $c = 0, 1$ , and let  $\sigma_i = \Lambda_{ii}$

# Combining Dissimilarities Through Embedding Product

–  $J$  function, a model selection & dimension reduction method

- $\tilde{X} \triangleq [X_1, X_2, \dots, X_K]$ ,  $\Sigma = \pi\Sigma_1 + (1 - \pi)\Sigma_0$
- $X = P\tilde{X}$ , where  $P$  satisfies  $\Sigma = P\Lambda P^T$
- Let  $\mu_c = E[X|Y = c] \in \mathcal{R}^d$ , where  $c = 0, 1$ , and let  $\sigma_i = \Lambda_{ii}$
- Define  $J$  function for the  $i$ th dimension:

$$J_i \triangleq \frac{|\mu_{1i} - \mu_{0i}|}{\sigma_i}, \quad i = 1, \dots, d$$

# Combining Dissimilarities Through Embedding Product

–  $J$  function, a model selection & dimension reduction method

- $\tilde{X} \triangleq [X_1, X_2, \dots, X_K]$ ,  $\Sigma = \pi\Sigma_1 + (1 - \pi)\Sigma_0$
- $X = P\tilde{X}$ , where  $P$  satisfies  $\Sigma = P\Lambda P^T$
- Let  $\mu_c = E[X|Y = c] \in \mathcal{R}^d$ , where  $c = 0, 1$ , and let  $\sigma_i = \Lambda_{ii}$
- Define  $J$  function for the  $i$ th dimension:

$$J_i \triangleq \frac{|\mu_{1i} - \mu_{0i}|}{\sigma_i}, \quad i = 1, \dots, d$$

- Reorder the dimensions of  $X$  according the  $J$  values (from largest to smallest) and use the first  $p$  dimensions

$$X \implies X_j \implies X_j(p), \quad p^* = \arg \min_{p=1:d} \hat{L}_p$$

## Combining Dissimilarities Through Embedding Product

–  $J$  function, a model selection & dimension reduction method

- $\tilde{X} \triangleq [X_1, X_2, \dots, X_K]$ ,  $\Sigma = \pi\Sigma_1 + (1 - \pi)\Sigma_0$
- $X = P\tilde{X}$ , where  $P$  satisfies  $\Sigma = P\Lambda P^T$
- Let  $\mu_c = E[X|Y = c] \in \mathcal{R}^d$ , where  $c = 0, 1$ , and let  $\sigma_i = \Lambda_{ii}$
- Define  $J$  function for the  $i$ th dimension:

$$J_i \triangleq \frac{|\mu_{1i} - \mu_{0i}|}{\sigma_i}, \quad i = 1, \dots, d$$

- Reorder the dimensions of  $X$  according the  $J$  values (from largest to smallest) and use the first  $p$  dimensions

$$X \implies X_j \implies X_j(p), \quad p^* = \arg \min_{p=1:d} \hat{L}_p$$

Need to search  $\sum_{k=1}^K d_k$  times!    cheating vs non-cheating ??

# Combining Dissimilarities Through Embedding Product

## – Theorem

### Theorem

Let  $(X, Y) \sim F_{XY}$ , where  $X \in \mathcal{R}^d$ ,  $Y \sim \text{Bernoulli}(\pi)$ ,  $F_{X|Y=c} = N(\boldsymbol{\mu}_c, \Sigma)$ . Let  $f : \mathcal{R}^d \rightarrow \mathcal{R}^p$  be any projection function, where  $p < d$ . And let  $f_j$  be the projection function deduced from the above  $J$ -function procedure. If  $L_{f(X)}^*$  and  $L_{f_j(X)}^*$  denote the Bayes error probabilities for  $(f(X), Y)$  and  $(f_j(X), Y)$ , then

$$L_{f(X)}^* \geq L_{f_j(X)}^*.$$

# Combining Dissimilarities Through Embedding Product

– Sketch of Proof - assumptions

Without loss of generality, we assume

# Combining Dissimilarities Through Embedding Product

– Sketch of Proof - assumptions

Without loss of generality, we assume

- $\Sigma$  is invertible. Since if it is not, we can project  $X$  onto a lower dimensional space where the covariance matrix of  $X$  is non-singular without loss of information

# Combining Dissimilarities Through Embedding Product

– Sketch of Proof - assumptions

Without loss of generality, we assume

- $\Sigma$  is invertible. Since if it is not, we can project  $X$  onto a lower dimensional space where the covariance matrix of  $X$  is non-singular without loss of information
- $\Sigma = I_d$ . Since we can always find a matrix  $A$  such that  $\tilde{X} = AX \sim N(A\mu_c, I_d)$ , and any function of  $X$  can be written as  $f(X) = f(A^{-1}\tilde{X}) \triangleq \tilde{f}(\tilde{X})$
- the dimensions of  $X$  are ordered according to the values of  $|\mu_1 - \mu_0|$  (largest to smallest)

# Combining Dissimilarities Through Embedding Product

– Sketch of Proof

- $f_j(X) = T_j X \sim N(T_j \boldsymbol{\mu}_c, I_p)$ , and  $T_j = [I_p \mid \mathbf{0}_{p \times (d-p)}]$

# Combining Dissimilarities Through Embedding Product

## – Sketch of Proof

- $f_j(X) = T_j X \sim N(T_j \boldsymbol{\mu}_c, I_p)$ , and  $T_j = [I_p \mid \mathbf{0}_{p \times (d-p)}]$
- For any other  $f(X) = TX$ , find a matrix  $B$  such that  $BTX \sim N(BT \boldsymbol{\mu}_c, I_p)$

# Combining Dissimilarities Through Embedding Product

– Sketch of Proof

- $f_j(X) = T_j X \sim N(T_j \boldsymbol{\mu}_c, I_p)$ , and  $T_j = [I_p \mid \mathbf{0}_{p \times (d-p)}]$
- For any other  $f(X) = TX$ , find a matrix  $B$  such that  $BTX \sim N(BT \boldsymbol{\mu}_c, I_p)$
- $\|T_j \boldsymbol{\mu}_1 - T_j \boldsymbol{\mu}_0\| \geq \|BT \boldsymbol{\mu}_1 - BT \boldsymbol{\mu}_0\|$

# Combining Dissimilarities Through Embedding Product

## – Sketch of Proof

- $f_j(X) = T_j X \sim N(T_j \boldsymbol{\mu}_c, I_p)$ , and  $T_j = [I_p \mid \mathbf{0}_{p \times (d-p)}]$
- For any other  $f(X) = TX$ , find a matrix  $B$  such that  $BTX \sim N(BT \boldsymbol{\mu}_c, I_p)$
- $\|T_j \boldsymbol{\mu}_1 - T_j \boldsymbol{\mu}_0\| \geq \|BT \boldsymbol{\mu}_1 - BT \boldsymbol{\mu}_0\|$
- $L_{f_j(X)}^* \leq L_{Bf(X)}^* = L_{f(X)}^*$  □

# A Simulated Experiment

– Experiment Design

$$\left[ \begin{array}{c|c} X_1 & X_2 \\ \hline 2n \times 40 & 2n \times 40 \end{array} \right]$$

# A Simulated Experiment

– Experiment Design

$$\left[ \begin{array}{c|c} X_1 & X_2 \\ \hline 2n \times 40 & 2n \times 40 \end{array} \right]$$

$$X_1 \sim \frac{1}{2}N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1) + \frac{1}{2}N(-\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$$

# A Simulated Experiment

## – Experiment Design

$$\left[ \begin{array}{c|c} X_1 & X_2 \\ \hline 2n \times 40 & 2n \times 40 \end{array} \right]$$

$$X_1 \sim \frac{1}{2}N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1) + \frac{1}{2}N(-\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$$

$$X_2 \sim \frac{1}{2}N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2) + \frac{1}{2}N(-\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$$

# A Simulated Experiment

## – Experiment Design

$$\left[ \begin{array}{c|c} X_1 & X_2 \\ \hline 2n \times 40 & 2n \times 40 \end{array} \right] \quad \begin{array}{l} X_1 \sim \frac{1}{2}N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1) + \frac{1}{2}N(-\boldsymbol{\mu}, \boldsymbol{\Sigma}_1) \\ X_2 \sim \frac{1}{2}N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2) + \frac{1}{2}N(-\boldsymbol{\mu}, \boldsymbol{\Sigma}_2) \end{array}$$

$$\boldsymbol{\mu} = (1, 1, 1, 1, 1, 0, \dots, 0)^T \in \mathcal{R}^{40}$$

# A Simulated Experiment

## - Experiment Design

$$\left[ \begin{array}{c|c} X_1 & X_2 \\ \hline 2n \times 40 & 2n \times 40 \end{array} \right] \quad \begin{array}{l} X_1 \sim \frac{1}{2}N(\boldsymbol{\mu}, \Sigma_1) + \frac{1}{2}N(-\boldsymbol{\mu}, \Sigma_1) \\ X_2 \sim \frac{1}{2}N(\boldsymbol{\mu}, \Sigma_2) + \frac{1}{2}N(-\boldsymbol{\mu}, \Sigma_2) \end{array}$$

$$\boldsymbol{\mu} = (1, 1, 1, 1, 1, 0, \dots, 0)^T \in \mathcal{R}^{40}$$

$$\Sigma_1 = \begin{pmatrix} 1 & & & \\ & 2 & & \\ & & \ddots & \\ & & & 40 \end{pmatrix} \quad \Sigma_2 = \left( \frac{\sqrt{i}\sqrt{j}}{2^{|i-j|}} \right)_{ij}$$

# A Simulated Experiment

## - Experiment Design

$$\left[ \begin{array}{c|c} X_1 & X_2 \\ \hline 2n \times 40 & 2n \times 40 \end{array} \right] \quad \begin{array}{l} X_1 \sim \frac{1}{2}N(\boldsymbol{\mu}, \Sigma_1) + \frac{1}{2}N(-\boldsymbol{\mu}, \Sigma_1) \\ X_2 \sim \frac{1}{2}N(\boldsymbol{\mu}, \Sigma_2) + \frac{1}{2}N(-\boldsymbol{\mu}, \Sigma_2) \end{array}$$

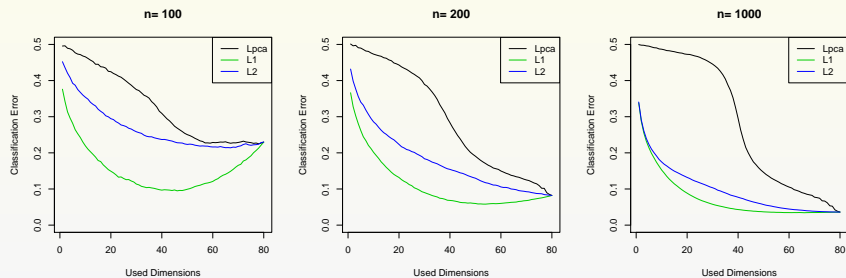
$$\boldsymbol{\mu} = (1, 1, 1, 1, 1, 0, \dots, 0)^T \in \mathcal{R}^{40}$$

$$\Sigma_1 = \begin{pmatrix} 1 & & & \\ & 2 & & \\ & & \ddots & \\ & & & 40 \end{pmatrix} \quad \Sigma_2 = \left( \frac{\sqrt{i}\sqrt{j}}{2^{|i-j|}} \right)_{ij}$$

training on  $n$  points, testing the other  $n$  points

# A Simulated Experiment

## – Results



**Figure:** Using both training and testing labels  $\implies L_1$ , using only training labels  $\implies L_2$ . These plots depict that (1)  $L_1 < L_2 < L_{pca}$  for all  $p < d$ ; (2) the difference between  $L_1$  and  $L_2$  decreases as the sample size  $n$  increases.

# Tiger Data

## – Image/Caption Fusion

- 140,577 *images & captions* were collected from Yahoo! Photos website [Jeff Solka and his team].

# Tiger Data

## – Image/Caption Fusion

- 140,577 *images & captions* were collected from Yahoo! Photos website [Jeff Solka and his team].
- 1,600 pairs were selected using query word “tiger” on captions.

# Tiger Data

## – Image/Caption Fusion

- 140,577 *images & captions* were collected from Yahoo! Photos website [Jeff Solka and his team].
- 1,600 pairs were selected using query word “tiger” on captions.
- They were labeled manually based only on captions:

label	#
animal tiger	148
Detroit Tigers baseball team	145
Tiger Woods the golfer	897
Tamil Tigers soldiers of Sri Lanka	330
Leicester Tigers rugby team	48
others	32

# Tiger Data

## – Image/Caption Fusion

- 140,577 *images & captions* were collected from Yahoo! Photos website [Jeff Solka and his team].
- 1,600 pairs were selected using query word “tiger” on captions.
- They were labeled manually based only on captions:

label	#
animal tiger	148
Detroit Tigers baseball team	145
Tiger Woods the golfer	897
Tamil Tigers soldiers of Sri Lanka	330
Leicester Tigers rugby team	48
others	32

- Two class problem: “Tiger Woods” and “Tamil Tigers”.

# “Tiger” Dissimilarity Matrices

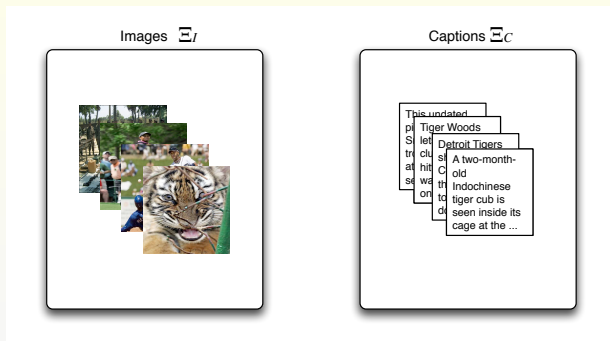


Figure: Conceptual depiction of the “tiger data set”

- *image* features: first & second order pixel derivatives [Jain].
- *caption* features: Mutual Information [Lin & Pantel].
- **dissimilarities**:  $1 - (\text{Random Forest proximity})$  [Breiman].

## $J$ function procedure

- $n = 1227$ ,  $n_0 = 897$ ,  $n_1 = 330$ ;  $\Delta_c, \Delta_i \in M_n(\mathcal{R})$

## $J$ function procedure

- $n = 1227$ ,  $n_0 = 897$ ,  $n_1 = 330$ ;  $\Delta_c, \Delta_i \in M_n(\mathcal{R})$
- Embedding:  $\Delta_c \rightarrow X_c \in \mathcal{R}^{30}$ ,  $\Delta_i \rightarrow X_i \in \mathcal{R}^{30}$

## $J$ function procedure

- $n = 1227$ ,  $n_0 = 897$ ,  $n_1 = 330$ ;  $\Delta_c, \Delta_i \in M_n(\mathcal{R})$
- Embedding:  $\Delta_c \rightarrow X_c \in \mathcal{R}^{30}$ ,  $\Delta_i \rightarrow X_i \in \mathcal{R}^{30}$
- randomly selecting (without replacement) 600, 300 observations as training and developing sets, respectively and using the rest as test set

## $J$ function procedure

- $n = 1227$ ,  $n_0 = 897$ ,  $n_1 = 330$ ;  $\Delta_c, \Delta_i \in M_n(\mathcal{R})$
- Embedding:  $\Delta_c \rightarrow X_c \in \mathcal{R}^{30}$ ,  $\Delta_i \rightarrow X_i \in \mathcal{R}^{30}$
- randomly selecting (without replacement) 600, 300 observations as training and developing sets, respectively and using the rest as test set
- calculate  $J$  (using all labels vs using only training labels)

$$\hat{J}_i = \frac{|\hat{\mu}_{0i} - \hat{\mu}_{1i}|}{\hat{\sigma}_i}, \quad i = 1, \dots, 60$$

## $J$ function procedure

- $n = 1227$ ,  $n_0 = 897$ ,  $n_1 = 330$ ;  $\Delta_c, \Delta_i \in M_n(\mathcal{R})$
- Embedding:  $\Delta_c \rightarrow X_c \in \mathcal{R}^{30}$ ,  $\Delta_i \rightarrow X_i \in \mathcal{R}^{30}$
- randomly selecting (without replacement) 600, 300 observations as training and developing sets, respectively and using the rest as test set
- calculate  $J$  (using all labels vs using only training labels)

$$\hat{J}_i = \frac{|\hat{\mu}_{0i} - \hat{\mu}_{1i}|}{\hat{\sigma}_i}, \quad i = 1, \dots, 60$$

- modified  $J$  (for using only training labels)

$$J_i^m = w \cdot \frac{\hat{\sigma}_i}{\sum \hat{\sigma}_i} + (1 - w) \cdot \frac{\hat{J}_i}{\sum \hat{J}_i}$$

## $J$ function procedure

- $n = 1227$ ,  $n_0 = 897$ ,  $n_1 = 330$ ;  $\Delta_c, \Delta_i \in M_n(\mathcal{R})$
- Embedding:  $\Delta_c \rightarrow X_c \in \mathcal{R}^{30}$ ,  $\Delta_i \rightarrow X_i \in \mathcal{R}^{30}$
- randomly selecting (without replacement) 600, 300 observations as training and developing sets, respectively and using the rest as test set
- calculate  $J$  (using all labels vs using only training labels)

$$\hat{J}_i = \frac{|\hat{\mu}_{0i} - \hat{\mu}_{1i}|}{\hat{\sigma}_i}, \quad i = 1, \dots, 60$$

- modified  $J$  (for using only training labels)

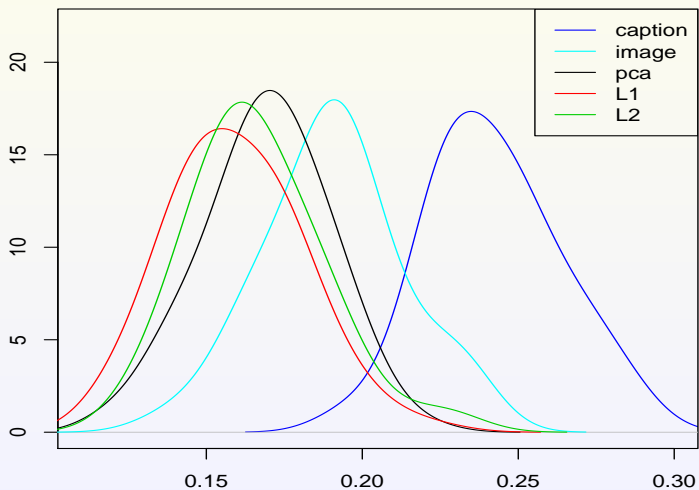
$$J_i^m = w \cdot \frac{\hat{\sigma}_i}{\sum \hat{\sigma}_i} + (1 - w) \cdot \frac{\hat{J}_i}{\sum \hat{J}_i}$$

- using developing set to tune  $w$  and  $p$

# Tiger Data

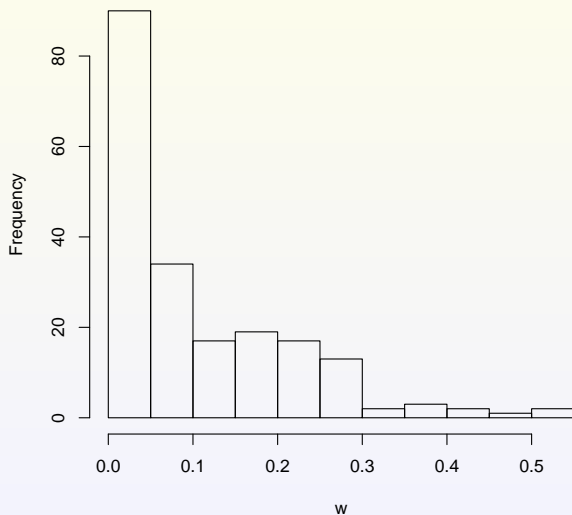
– Results

## Density Estimate of Classification Error



# Tiger Data

– Histogram of  $w$





“The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data” – J.W. Tukey