Fusion and Inference from Multiple Data Sources in Commensurate Space

Zhiliang Ma

Department of Applied Mathematics & Statistics
Johns Hopkins University

QMDNS 2010
George Mason University, Fairfax, Virginia
May 25, 2010

Fusion and Inference from Multiple Data Sources in Commensurate Space

Zhiliang Ma

Department of Applied Mathematics & Statistics
Johns Hopkins University

Carey E. Priebe & David J. Marchette

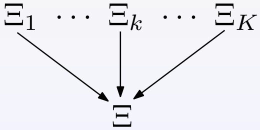
QMDNS 2010
George Mason University, Fairfax, Virginia
May 25, 2010

Data Fusion

Given data of different types (e.g. measurements, graphs, images) or data measured under different conditions (e.g. photos taken under indoor lighting and outdoor lighting), we wish to perform inference using all the data.

1. Inferential tasks can be done using each data type separately
$\Longrightarrow$ Fusion in joint space: $\Xi_1 \times \cdots \times \Xi_K$

2. Inferential tasks cannot be done in any single space alone, and require information from other spaces
$\Longrightarrow$ Fusion in commensurate space:



Formulation

		Condition				
		1	$\cdots$	k	$\cdots$	K
	1	$\boldsymbol{x}_{11}$	$\cdots$	$\boldsymbol{x}_{1k}$	$\cdots$	$\boldsymbol{x}_{1K}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Object	i	$\boldsymbol{x}_{i1}$	$\cdots$	$\boldsymbol{x}_{ik}$	$\cdots$	$\boldsymbol{x}_{iK}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	n	$\boldsymbol{x}_{n1}$	$\cdots$	$\boldsymbol{x}_{nk}$	$\cdots$	$\boldsymbol{x}_{nK}$

- n objects, each measured under K different conditions
- $\boldsymbol{x}_{i1} \sim \cdots \sim \boldsymbol{x}_{ik} \sim \cdots \sim \boldsymbol{x}_{iK}$ denotes K matched features vectors representing a single object
- $\boldsymbol{x}_{ik} \in \Xi_k$

Formulation

$$
\begin{array}{cc|ccccc}
 & & & & \textbf{Condition} & & \\
 & & 1 & \cdots & k & \cdots & K \\
 & 1 & \boldsymbol{x}_{11} & \cdots & \boldsymbol{x}_{1k} & \cdots & \boldsymbol{x}_{1K} \\
 & \vdots & \vdots & & \vdots & & \vdots \\
\textbf{Object} & i & \boldsymbol{x}_{i1} & \cdots & \boldsymbol{x}_{ik} & \cdots & \boldsymbol{x}_{iK} \\
 & \vdots & \vdots & & \vdots & & \vdots \\
 & n & \boldsymbol{x}_{n1} & \cdots & \boldsymbol{x}_{nk} & \cdots & \boldsymbol{x}_{nK}
\end{array}
$$

- n objects, each measured under K different conditions
- $\boldsymbol{x}_{i1} \sim \cdots \sim \boldsymbol{x}_{ik} \sim \cdots \sim \boldsymbol{x}_{iK}$ denotes K matched features vectors representing a single object
- $\boldsymbol{x}_{ik} \in \Xi_k$

Example

Indoor lighting versus outdoor lighting for face recognition $(K = 2)$

The Matching Problem

K new measurements $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k, \ldots, \boldsymbol{z}_K$

Problem 1

Are the K new measurements matched feature vectors representing a single object measured under K conditions?

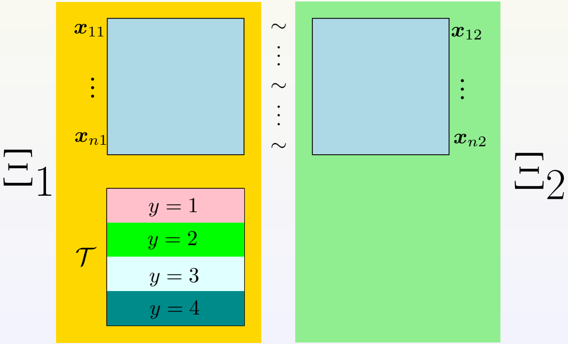
The Classification Problem

Given additional m labeled measurements under first condition:

$$\mathcal{T} = \{(\boldsymbol{z}_{i1}, y_i) \in \Xi_1 \times \{1, \ldots, C\},\ i = 1, \ldots, m\}$$

Problem 2

Create classifiers $g_k(\cdot; \mathcal{T})$ for determining the class labels of new objects measured under condition $k,\ k = 2, \ldots, K$.



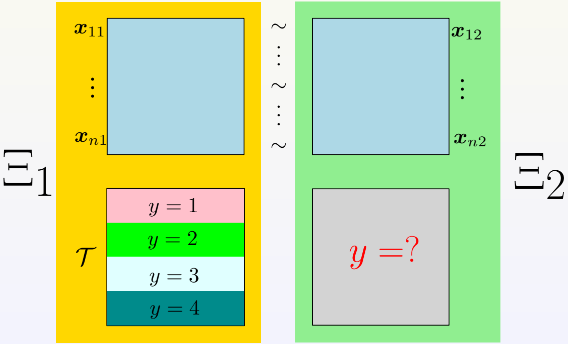
The Classification Problem

Given additional m labeled measurements under first condition:

$$\mathcal{T} = \{(\boldsymbol{z}_{i1}, y_i) \in \Xi_1 \times \{1, \ldots, C\},\ i = 1, \ldots, m\}$$

Problem 2

Create classifiers $g_k(\cdot; \mathcal{T})$ for determining the class labels of new objects measured under condition k, $k = 2, \ldots, K$.



Solution

Assume all measurements are collected under the same condition

1. How to determine whether K new measurements, $\{\boldsymbol{z}_i \in \Xi, i = 1, \ldots, K\}$, represent the same object?

$$\sum_{i<j} d(\boldsymbol{z}_i, \boldsymbol{z}_j)$$

2. Given $\mathcal{T} = \{(\boldsymbol{z}_i, y_i) \in \Xi \times \{1, \ldots, C\},\ i = 1, \ldots, m\}$, how to determine the class label of a new observation $\boldsymbol{z}_{\text{new}} \in \Xi$?

$$g(\cdot) : \Xi \to \{1, \ldots, C\}, \quad \hat{y}_{\text{new}} = g(\boldsymbol{z}_{\text{new}})$$

Solution

~~Assume all measurements are collected under the same condition~~

1. How to determine whether K new measurements, $\{\boldsymbol{z}_i \in \Xi, i = 1, \ldots, K\}$, represent the same object?

$$\sum_{i<j} d(\boldsymbol{z}_i, \boldsymbol{z}_j)$$

2. Given $\mathcal{T} = \{(\boldsymbol{z}_i, y_i) \in \Xi \times \{1, \ldots, C\},\ i = 1, \ldots, m\}$, how to determine the class label of a new observation $\boldsymbol{z}_{\text{new}} \in \Xi$?

$$g(\cdot) : \Xi \to \{1, \ldots, C\}, \quad \hat{y}_{\text{new}} = g(\boldsymbol{z}_{\text{new}})$$

The Key

The key is to transform measurements collected under different conditions into one commensurate space so that (transformed) feature vectors are comparable.

Wikipedia Documents

- Wikipedia is a free, multilingual encyclopedia project
- 13 million articles (2.9 million in the English Wikipedia) have been written collaboratively by volunteers around the world
- A Wikipedia document has information regarding
  - Title
  - Unique ID number
  - Text - the content of the article
  - Internal links - links to other (related) articles
  - External links - links to other content on the web or elsewhere
  - Language links - links to “the same” page in other languages.

Two Wikipedias: English and French

Consider a subset of English and French Wikipedias that are 1-1 correspondent $(n = 1380)$

	Language	
	$\boldsymbol{E}$	$\boldsymbol{F}$
Topic 1	$\boldsymbol{x}_{11}$	$\boldsymbol{x}_{12}$
$\vdots$	$\vdots$	$\vdots$
Topic $\boldsymbol{i}$	$\boldsymbol{x}_{i1}$	$\boldsymbol{x}_{i2}$
$\vdots$	$\vdots$	$\vdots$
Topic $\boldsymbol{n}$	$\boldsymbol{x}_{n1}$	$\boldsymbol{x}_{n2}$

Feature Representation versus Dissimilarity Representation

- A document lives in high-dimensional space. E.g., an English corpus of size 2000, x is more than 250,000 dimensional!

Feature Representation versus Dissimilarity Representation

- A document lives in high-dimensional space. E.g., an English corpus of size 2000, $\boldsymbol{x}$ is more than 250,000 dimensional!
- Feature Representation: $\mathbf{X}$ $2000 \times 250,000$
  Dissimilarity Representation: $\mathbf{D}$ 2000×2000

Feature Representation versus Dissimilarity Representation

- A document lives in high-dimensional space. E.g., an English corpus of size 2000, $\boldsymbol{x}$ is more than 250,000 dimensional!
- Feature Representation: $\mathbf{X}$ $2000 \times 250,000$
  Dissimilarity Representation: $\mathbf{D}$ 2000×2000
- Define the *dissimilarity* between two documents $\boldsymbol{a}$, $\boldsymbol{b}$ as

$$\delta(\boldsymbol{a}, \boldsymbol{b}) = 1 - s(\boldsymbol{a}, \boldsymbol{b}) = 1 - \cos \angle(\boldsymbol{a}, \boldsymbol{b}) = 1 - \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\| \|\boldsymbol{b}\|}$$

- English: $\mathbf{D}_E = [\delta(\boldsymbol{x}_{i1}, \boldsymbol{x}_{j1})]_{n \times n}$
  French: $\mathbf{D}_F = [\delta(\boldsymbol{x}_{i2}, \boldsymbol{x}_{j2})]_{n \times n}$

Multidimensional Scaling

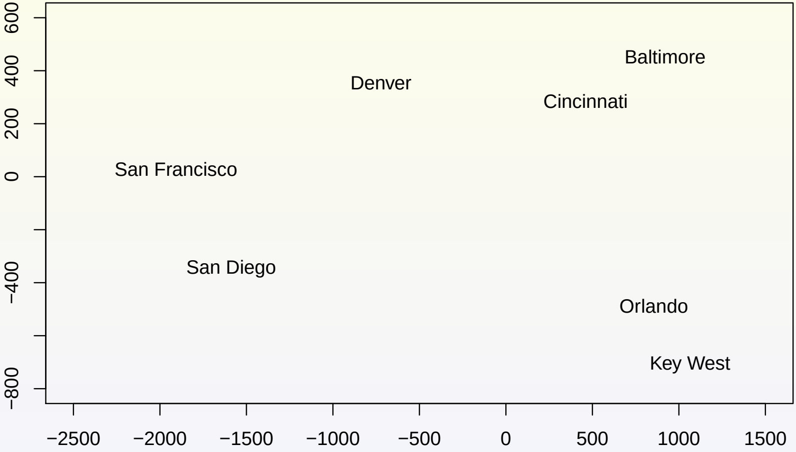
- Multidimensional scaling (MDS): $\mathbf{D} \mapsto \tilde{\mathbf{X}} \in \mathbb{R}^{n \times p}$
- An example of criterion: $\sum_{i<j} \left[\delta_{ij} - d_{ij}(\tilde{\mathbf{X}})\right]^2$

Multidimensional Scaling

- Multidimensional scaling (MDS): $\mathbf{D} \mapsto \tilde{\mathbf{X}} \in \mathbb{R}^{n \times p}$
- An example of criterion: $\sum_{i<j} \left[\delta_{ij} - d_{ij}(\tilde{\mathbf{X}})\right]^2$
- An example: seven cities I have visited

Miles	Baltimore	Cincinnati	Denver	Key West	Orlando	San Diego
Cincinnati	507					
Denver	1641	1173				
Key West	1199	1146	2065			
Orlando	903	878	1806	309		
San Diego	2634	2161	1087	2690	2431	
San Francisco	2864	2374	1235	3067	2808	554

Multidimensional Scaling



Out-of-Sample Embedding

Definition

Out-of-sample embedding techniques insert additional points into previously constructed configurations

An example of criterion: $\sum_{i=1}^{n} \left[\delta_{\mathsf{new},i} - d_i(\tilde{\boldsymbol{z}}_{\mathsf{new}}, \tilde{\mathbf{X}}) \right]^2$

Out-of-Sample Embedding

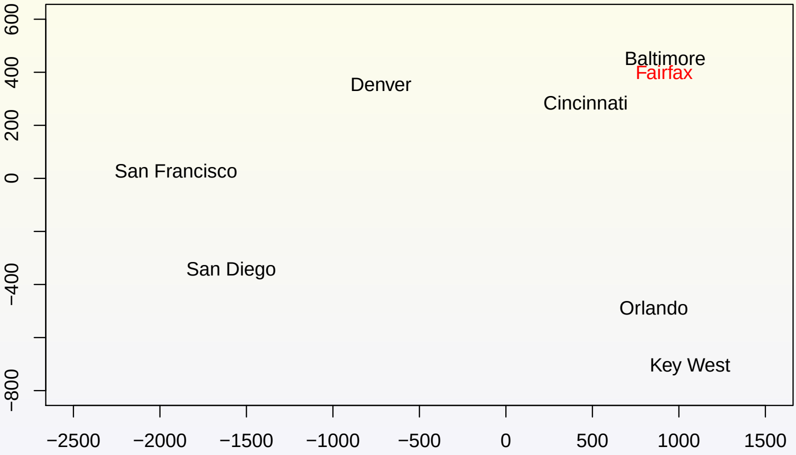
Definition

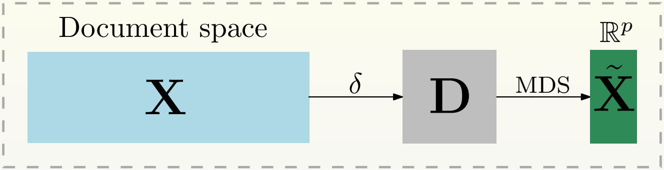
Out-of-sample embedding techniques insert additional points into previously constructed configurations

An example of criterion: $\sum_{i=1}^{n} \left[\delta_{\text{new},i} - d_i(\tilde{z}_{\text{new}}, \tilde{\mathbf{X}}) \right]^2$

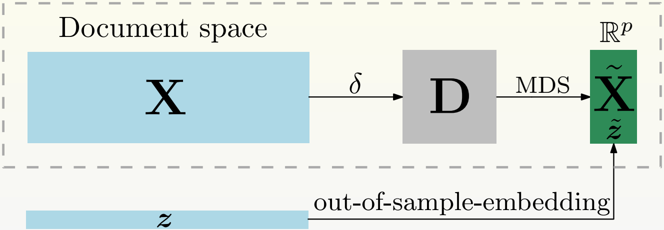
Miles	Fairfax
Baltimore	48
Cincinnati	471
Denver	1616
Key West	1158
Orlando	864
San Diego	2615
San Francisco	2857

Out-of-Sample Embedding





Document space
X
δ
D
MDS
X̃
ℝ^p



Document space
X
δ
D
MDS
ℝ^p
X̃
z̃
z
out-of-sample-embedding

The Matching Problem — Hypothesis Testing

- Given n pairs of matched documents: $E \sim F$
- Two new documents: $\boldsymbol{z}_1$ (English) and $\boldsymbol{z}_2$ (French)
- Goal: determine whether a match is present between $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ (i.e., whether they are on the same topic)

$$H_0 : \boldsymbol{z}_1 \sim \boldsymbol{z}_2 \text{ versus } H_A : \boldsymbol{z}_1 \nsim \boldsymbol{z}_2$$

(we control the probability of missing a true match)

Procrustes Approach

- Multidimensional scaling

$$\mathbf{D}_E \to \tilde{\mathbf{X}}_E,\ \mathbf{D}_F \to \tilde{\mathbf{X}}_F$$

- Procrustes$(\tilde{\mathbf{X}}_E, \tilde{\mathbf{X}}_F)$ yields Q^*, i.e.,

$$Q^* = \underset{Q^T Q = I}{\arg\min} \|\tilde{\mathbf{X}}_E - \tilde{\mathbf{X}}_F Q\|_F^2$$

- $\tilde{\mathbf{X}}_E$ and $\tilde{\mathbf{X}}_F Q^*$ defines the commensurate space
- Out-of-sample embedding: $\boldsymbol{z}_1 \mapsto \tilde{\boldsymbol{z}}_E,\ \boldsymbol{z}_2 \mapsto \tilde{\boldsymbol{z}}_F$
- $\|\tilde{\boldsymbol{z}}_E - Q^* \tilde{\boldsymbol{z}}_F\| > c$ yields “reject”

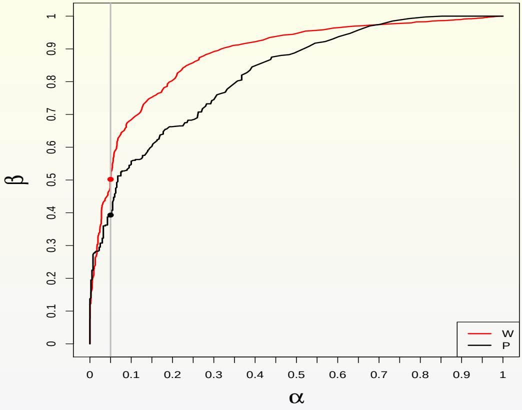
Our W approach

- Impute the dissimilarities between E and F by $\mathbf{W} = (\mathbf{D}_E + \mathbf{D}_F)/2$ to obtain one omnibus dissimilarity matrix $\mathbf{M}$

$$\mathbf{M} = \begin{bmatrix} \mathbf{D}_E & \mathbf{W} \\ \mathbf{W} & \mathbf{D}_F \end{bmatrix}$$

- Embed $\mathbf{M}$ as $2n$ points in $\mathbb{R}^p$: $\tilde{\mathbf{X}}_E, \ \tilde{\mathbf{X}}_F$
- $\tilde{\mathbf{X}}_E$ and $\tilde{\mathbf{X}}_F$ defines the commensurate space ($\|\tilde{\mathbf{X}}_E - \tilde{\mathbf{X}}_F\|$ is considered in embedding $\mathbf{M}$)
- Out-of-sample embedding: $\boldsymbol{z}_1 \mapsto \tilde{\boldsymbol{z}}_E, \ \boldsymbol{z}_2 \mapsto \tilde{\boldsymbol{z}}_F$
- $\|\tilde{\boldsymbol{z}}_E - \tilde{\boldsymbol{z}}_F\| > c$ yields "reject"

ROC Curves — β against α



- $\alpha = 0.05$, $\hat{\beta}_W = 0.502$
  correctly eliminating 50.2% of the false matches via W approach
- $\alpha = 0.05$, $\hat{\beta}_P = 0.393$

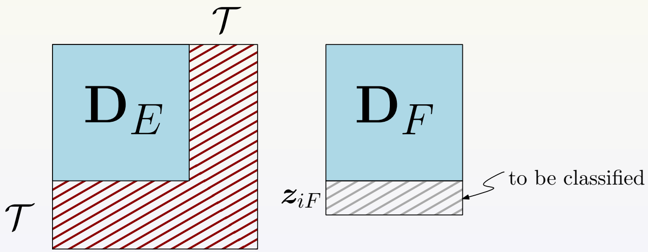
$\beta_{\alpha=0.05}$

$$\hat{\beta}_W = 0.502,\ \hat{\beta}_P = 0.393$$

- $H_0 : \hat{\beta}_W = \hat{\beta}_P$ versus $H_A : \hat{\beta}_W > \hat{\beta}_P$
- Wilcoxon signed-rank tests on the powers based on 200 Monte Carlo replicates
- p-value $= 0.0001$, indicating that the W approach is statistically significantly better than P approach in documents matching.

Classification

- Given n pairs of matched documents: $E \sim F \Longrightarrow \mathbf{D}_E,\ \mathbf{D}_F$
- Given additional m labeled English documents: $\mathcal{T} = \{(\boldsymbol{z}_{iE}, y_i),\ i = 1, \ldots, m\}$, where $y_i \in \{1, 2, 3, 4, 5\}$
- Determine the class label of new French documents $\boldsymbol{z}_{iF}$

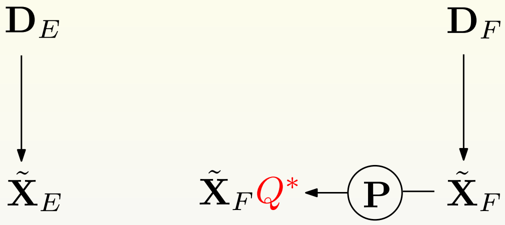


Procrustes Approach

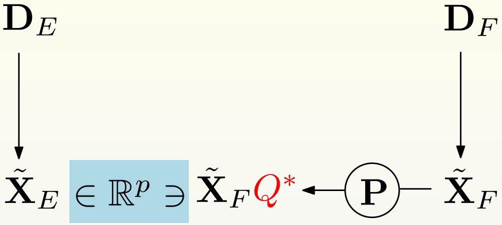
$$\mathbf{D}_E \longrightarrow \tilde{\mathbf{X}}_E$$

$$\mathbf{D}_F \longrightarrow \tilde{\mathbf{X}}_F$$

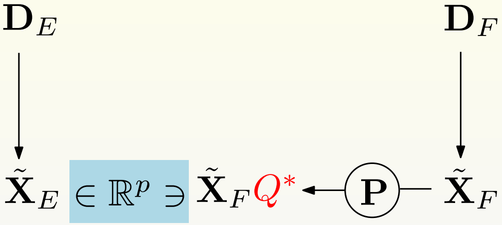
Procrustes Approach



Procrustes Approach



Procrustes Approach



- Out-of-sample embedding: $\boldsymbol{z}_{iE} \mapsto \tilde{\boldsymbol{z}}_{iE}$, $\boldsymbol{z}_{iF} \mapsto \tilde{\boldsymbol{z}}_{iF}$
- Create classifier g based on $\{(\tilde{\boldsymbol{z}}_{1E}, y_1), \ldots, (\tilde{\boldsymbol{z}}_{mE}, y_m)\}$
- $\hat{y}_{iF} = g(Q^* \tilde{\boldsymbol{z}}_{iF})$

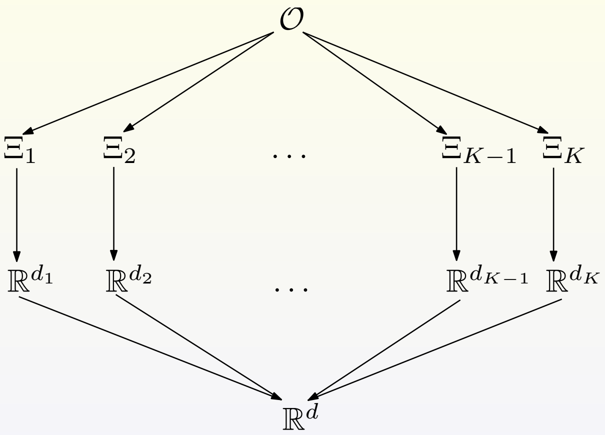
W approach

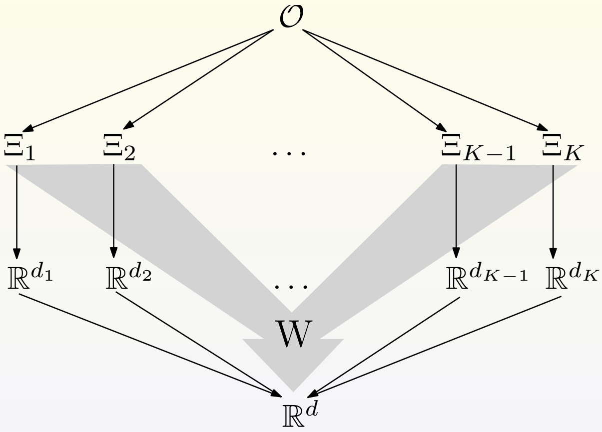
- Embed $\mathbf{M} = \begin{bmatrix} \mathbf{D}_E & \mathbf{W} \\ \mathbf{W} & \mathbf{D}_F \end{bmatrix}$ to obtain $\tilde{\mathbf{X}}_E$ and $\tilde{\mathbf{X}}_F$,

  where $\mathbf{W} = (\mathbf{D}_E + \mathbf{D}_F)/2$
- $\tilde{\mathbf{X}}_E$ and $\tilde{\mathbf{X}}_F$ are in the same (commensurate) space
- Out-of-sample embedding: $\boldsymbol{z}_{iE} \mapsto \tilde{\boldsymbol{z}}_{iE}$, $\boldsymbol{z}_{iF} \mapsto \tilde{\boldsymbol{z}}_{iF}$
- Create classifier g based on $\{(\tilde{\boldsymbol{z}}_{1E}, y_1), \ldots, (\tilde{\boldsymbol{z}}_{mE}, y_m)\}$
- $\hat{y}_{iF} = g(\tilde{\boldsymbol{z}}_{iF})$

Classification Error Rate

- Probability of misclassification: $\hat{L}_P = 0.496, \ \hat{L}_W = 0.279$
- McNemar's Chi-squared test results in p-value ≈ 0
- W approach is statistically significantly better than P approach in classification


$\mathcal{O}$
Ξ_1
Ξ_2
$\cdots$
Ξ_{K-1}
Ξ_K
$\mathbb{R}^{d_1}$
$\mathbb{R}^{d_2}$
$\cdots$
$\mathbb{R}^{d_{K-1}}$
$\mathbb{R}^{d_K}$
$\mathbb{R}^d$



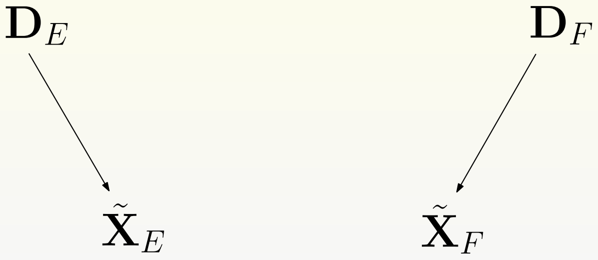
$\mathcal{O}$
Ξ_1
Ξ_2
$\cdots$
Ξ_{K-1}
Ξ_K
$\mathbb{R}^{d_1}$
$\mathbb{R}^{d_2}$
$\cdots$
$\mathbb{R}^{d_{K-1}}$
$\mathbb{R}^{d_K}$
W
$\mathbb{R}^d$

Thank you!!!

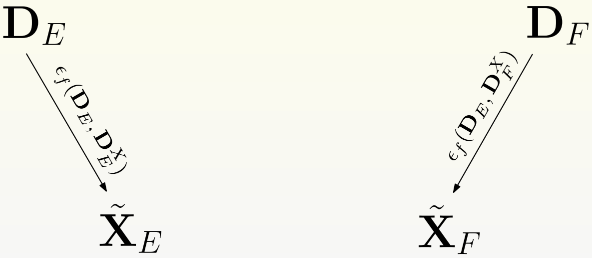
"*The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data*"

— J.W. Tukey

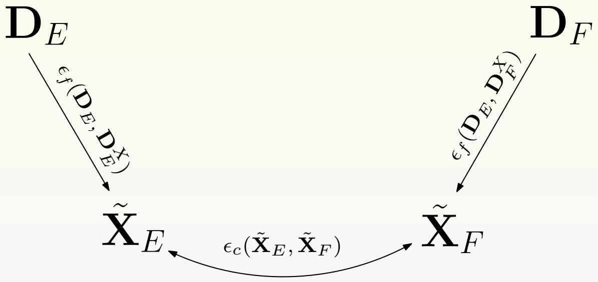
Fidelity and Commensurability



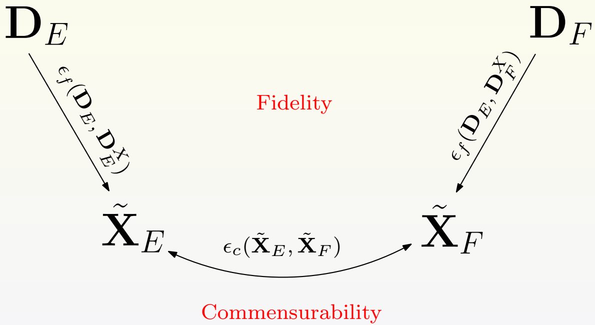
Fidelity and Commensurability



Fidelity and Commensurability



Fidelity and Commensurability



Separate Optimization versus Joint Optimization

Definition

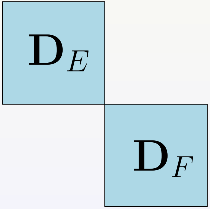
- $\epsilon_{fL} \overset{\text{def.}}{=} \epsilon_f(\mathbf{D}_L, \mathbf{D}_L^X) = \frac{1}{n^2}\|\mathbf{D}_L - \mathbf{D}_L^X\|_F^2,\ L \in \{E, F\}$
- $\epsilon_c \overset{\text{def.}}{=} \epsilon_c(\mathbf{X}_E, \mathbf{X}_F) = \frac{1}{n}\|\mathbf{X}_E - \mathbf{X}_F\|_F^2$

Separate Optimization versus Joint Optimization

Definition

- $\epsilon_{fL} \stackrel{\text{def.}}{=} \epsilon_f(\mathbf{D}_L, \mathbf{D}_L^X) = \dfrac{1}{n^2}\|\mathbf{D}_L - \mathbf{D}_L^X\|_F^2,\ L \in \{E, F\}$
- $\epsilon_c \stackrel{\text{def.}}{=} \epsilon_c(\mathbf{X}_E, \mathbf{X}_F) = \dfrac{1}{n}\|\mathbf{X}_E - \mathbf{X}_F\|_F^2$

- Procrustes approach:
  1. minimizes ϵ_{fE} and ϵ_{fF}
  2. minimizes ϵ_c



Separate Optimization versus Joint Optimization

Definition

- $\epsilon_{fL} \overset{\text{def.}}{=} \epsilon_f(\mathbf{D}_L, \mathbf{D}_L^X) = \frac{1}{n^2}\|\mathbf{D}_L - \mathbf{D}_L^X\|_F^2,\ L \in \{E, F\}$
- $\epsilon_c \overset{\text{def.}}{=} \epsilon_c(\mathbf{X}_E, \mathbf{X}_F) = \frac{1}{n}\|\mathbf{X}_E - \mathbf{X}_F\|_F^2$

- Procrustes approach:
  1. minimizes ϵ_{fE} and ϵ_{fF}
  2. minimizes ϵ_c

- W approach minimizes

  $\epsilon_{fE} + \epsilon_{fF} + \epsilon_c$