Manifold Matching: Joint Optimization of Fidelity & Commensurability

Carey E. Priebe
Department of Applied Mathematics & Statistics
Johns Hopkins University

September 23, 2010
\((X, Y, Z) \sim F_{X, Y, Z}\)
\(Y : \rho \rightarrow \sum \frac{J}{\bar{J}}\)
\(Z : \rho \rightarrow \{0, 1\}\)
\(X|Z=2 : \rho \rightarrow \sum_{2}\)

\(\tilde{g}_2 \sim_{X-Z-2} \rightarrow \bar{J}\)
\(\tilde{g}_2' \sim_{X-Z-2} \rightarrow \bar{J}\)

- Available training data:
  \(\mathcal{T}_0 = \{(X_i, Y_i) : J, Z_i = 0\}\)
  \(\mathcal{T}_0' = \{(X_i, Y_i) : \bar{J}, Z_i = 0\}\)
  \(\mathcal{T}_1 = \{(X_i, Y_i) : J, Z_i = 1\}\)

- Data to be classified:
  \((X, Y, Z) = 1\)

\(\tilde{g}_0 \xrightarrow{\sim} \bar{J}\)  \(\tilde{g}_1 \xrightarrow{\sim} \bar{J}\)

- No training data \(\mathcal{T}_2\), available at this type (\(Z = 1\)) for these classes (\(Y \in \bar{J}\))

\[(1) \quad \tilde{g}_1 = \tilde{g}_0 \circ \omega \Rightarrow \tilde{g}_1 (\hat{\theta}(X), \tilde{\mathcal{T}}_0)\]

- LDA approach (implicitly) identifies portions \(P_{\theta}(Z)\) for which \(\hat{\theta}\) is a satisfactory estimate

\[(2) \quad \tilde{g}_0 : \tilde{\mathcal{T}}_0 = \tilde{g}_0 : \hat{\theta}\)

- Known: estimate

\[(3) \quad \tilde{g}_1 : \tilde{\mathcal{T}}_1 = \tilde{g}_1 : \hat{\theta}\)

Unknown \& needed
Classifier Technology and the Illusion of Progress

David J. Hand

Abstract. A great many tools have been developed for supervised classification, ranging from early methods such as linear discriminant analysis through to modern developments such as neural networks and support vector machines. A large number of comparative studies have been conducted in attempts to establish the relative superiority of these methods. This paper argues that these comparisons often fail to take into account important aspects of real problems, so that the apparent superiority of more sophisticated methods may be something of an illusion. In particular, simple methods typically yield performance almost as good as more sophisticated methods, to the extent that the difference in performance may be swamped by other sources of uncertainty that generally are not considered in the classical supervised classification paradigm.

Key words and phrases: Supervised classification, error rate, misclassification rate, simplicity, principle of parsimony, population drift, selectivity bias, flat maximum effect, problem uncertainty, empirical comparisons.
Collaborators

David J. Marchette
Zhiliang Ma
Sancar Adali
Donniell E. Fishkind
&c.

Support: AFOSR, NSSEFF, ONR, HLTCOE, ASEE
Problem Formulation

Given $x_{i1} \sim \cdots \sim x_{ik} \sim \cdots \sim x_{iK}, \ i = 1, \ldots, n$
Problem Formulation

Given $x_{i1} \sim \cdots \sim x_{ik} \sim \cdots \sim x_{iK}$, $i = 1, \ldots, n$

- $n$ objects are each measured under $K$ different conditions
- $x_{i1} \sim \cdots \sim x_{ik} \sim \cdots \sim x_{iK}$ denotes $K$ matched feature vectors representing a single object $O_i$
- $x_{ik} \in \Xi_k$
Problem Formulation

Given $x_{i1} \sim \cdots \sim x_{ik} \sim \cdots \sim x_{iK}$, $i = 1, \ldots, n$

- $n$ objects are each measured under $K$ different conditions
- $x_{i1} \sim \cdots \sim x_{ik} \sim \cdots \sim x_{iK}$ denotes $K$ matched feature vectors representing a single object $O_i$
- $x_{ik} \in \Xi_k$
- $K$ new measurements \( \{y_k\}_{k=1}^K, y_k \in \Xi_k \)
Problem Formulation

Given $x_{i1} \sim \cdots \sim x_{ik} \sim \cdots \sim x_{iK}$, $i = 1, \ldots, n$

- $n$ objects are each measured under $K$ different conditions
- $x_{i1} \sim \cdots \sim x_{ik} \sim \cdots \sim x_{iK}$
  denotes $K$ matched feature vectors representing a single object $O_i$
- $x_{ik} \in \Xi_k$
- $K$ new measurements $\{y_k\}_{k=1}^K$, $y_k \in \Xi_k$

Question
Are $\{y_k\}_{k=1}^K$ matched feature vectors representing a single object measured under $K$ conditions?
## Hypotheses

Each space $\Xi_k$ comes with a dissimilarity $\delta_k$, yielding dissimilarity matrices $\Delta_1, \ldots, \Delta_K$.

Given new measurements $\{y_k\}_{k=1}^K$ we can obtain within-condition dissimilarities $\delta_k(y_k, x_{ik})$, $i = 1, \ldots, n$, $k = 1, \ldots, K$.

Goal ($K = 2$): determine whether $y_1$ and $y_2$ are a match $H_0$: $y_1 \sim y_2$ versus $H_A$: $y_1 \not\sim y_2$ (we control the probability of missing a true match).

<table>
<thead>
<tr>
<th>Object $O_1$</th>
<th>$x_{11}$</th>
<th>$\sim$</th>
<th>$\cdots$</th>
<th>$\sim$</th>
<th>$x_{1K}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>Object $O_n$</td>
<td>$x_{n1}$</td>
<td>$\sim$</td>
<td>$\cdots$</td>
<td>$\sim$</td>
<td>$x_{nK}$</td>
</tr>
</tbody>
</table>
Hypotheses

\[ \Xi_1 \quad \cdots \quad \Xi_K \]

Object \( O_1 \) \( \mathbf{x}_{11} \) \( \sim \cdots \sim \) \( \mathbf{x}_{1K} \)

\[ \vdots \]

Object \( O_n \) \( \mathbf{x}_{n1} \) \( \sim \cdots \sim \) \( \mathbf{x}_{nK} \)

- Each space \( \Xi_k \) comes with a dissimilarity \( \delta_k \), yielding dissimilarity matrices \( \Delta_1, \cdots, \Delta_K \)
Hypotheses

\[ \Xi_1 \quad \cdots \quad \Xi_K \]

Object \( O_1 \) \( \mathbf{x}_{11} \sim \cdots \sim \mathbf{x}_{1K} \)

\vdots \quad \vdots \quad \vdots \quad \vdots

Object \( O_n \) \( \mathbf{x}_{n1} \sim \cdots \sim \mathbf{x}_{nK} \)

- Each space \( \Xi_k \) comes with a dissimilarity \( \delta_k \), yielding dissimilarity matrices \( \Delta_1, \cdots, \Delta_K \)
- Given new measurements \( \{y_k\}_{k=1}^K \), we can obtain within-condition dissimilarities

\[ \delta_k(y_k, \mathbf{x}_{ik}), \quad i = 1, \ldots, n, \quad k = 1, \ldots, K \]
Hypotheses

Each space $\Xi_k$ comes with a dissimilarity $\delta_k$, yielding dissimilarity matrices $\Delta_1, \ldots, \Delta_K$

Given new measurements $\{y_k\}_{k=1}^K$ we can obtain within-condition dissimilarities

$$\delta_k(y_k, x_{ik}), \ i = 1, \ldots, n, \ k = 1, \ldots, K$$

Goal ($K = 2$): determine whether $y_1$ and $y_2$ are a match
Hypotheses

\[ \Xi_1 \cdots \Xi_K \]

Object \( O_1 \quad x_{11} \sim \cdots \sim x_{1K} \)

\vdots

Object \( O_n \quad x_{n1} \sim \cdots \sim x_{nK} \)

- Each space \( \Xi_k \) comes with a dissimilarity \( \delta_k \), yielding dissimilarity matrices \( \Delta_1, \cdots, \Delta_K \)

- Given new measurements \( \{y_k\}_{k=1}^K \)
  we can obtain within-condition dissimilarities

\[ \delta_k(y_k, x_{ik}), \; i = 1, \ldots, n, \; k = 1, \ldots, K \]

- Goal (\( K = 2 \)): determine whether \( y_1 \) and \( y_2 \) are a match

\[ H_0 : y_1 \sim y_2 \; \text{versus} \; H_A : y_1 \sim y_2 \]

(we control the probability of missing a true match)
what are these “conditions” and what does it mean to be “matched”

- let condition be language for a text document, and “matched” mean “on the same topic”
- let condition be modality for a photo, and “matched” mean “of the same person”
  - indoor lighting vs outdoor lighting
  - two cameras of different quality
  - passport photos and airport surveillance photos
- let condition 1 be wiki text document and condition 2 be wiki hyperlink structure
- let condition 1 be text document and condition 2 be photo
- ... or just a single space with multiple dissimilarities
The English is clear enough to lorry drivers — but the Welsh reads “I am not in the office at the moment. Send any work to be translated.”

<http://news.bbc.co.uk/2/hi/uk_news/wales/7702913.stm>
Conditional distributions are induced by maps $\pi_k$ from “object space” $\Xi$

\[ \Xi \xrightarrow{\pi_1} \Xi_1 \quad \ldots \quad \Xi \xrightarrow{\pi_K} \Xi_K \]

Conditional spaces $\Xi_k$ are not commensurate
Manifold Matching I

Conditional distributions are induced by maps $\pi_k$ from "object space" $\Xi$

$$
\begin{array}{c}
\Xi \\
\pi_1 \quad \ldots \quad \pi_K \\
\Xi_1 \quad \Xi_K \\
\exists \phi? \\
\end{array}
$$

Conditional spaces $\Xi_k$ are *not* commensurate
Dirichlet Setting

Let $S^p$ be the standard $p$-simplex in $\mathbb{R}^{p+1}$

Let $\Xi_1 = S^p$ and $\Xi_2 = S^p$
(but the fact that the two spaces are the same is unknown to the algorithms ...)

Let $\alpha_i \sim iid \text{ Dirichlet}(1)$ represent $n$ “objects” or “topics”
Let $X_{ik} \sim iid \text{ Dirichlet}(r\alpha_i + 1)$ represent $K$ languages (WCHs)
Dirichlet Setting

Let $S^p$ be the standard $p$-simplex in $\mathbb{R}^{p+1}$

Let $\Xi_1 = S^p$ and $\Xi_2 = S^p$
(but the fact that the two spaces are the same is unknown to the algorithms ...)

Let $\alpha_i \sim iid \text{ Dirichlet}(1)$ represent $n$ “objects” or “topics”
Let $X_{ik} \sim iid \text{ Dirichlet}(r\alpha_i + 1)$ represent $K$ languages (WCHs)

- $r$ controls “what it means to be matched”
  (document variability & translation quality analogy)
Experimental Results

Projection_200LSA_3LDA_32to3_all_pick7_13_15

Enron topic-conditional WCHs
Matched points are used to define maps $\rho_k$ to the same space $\mathcal{X}$ (with distance $d$)

\[ \Xi \quad \pi_1 \quad \ldots \quad \pi_K \quad \Xi_1 \quad \ldots \quad \Xi_K \quad \rho_1 \quad \ldots \quad \rho_K \quad \mathcal{X} \]

Reject for $d(\tilde{y}_1, \tilde{y}_2)$ “large”
Matched points are used to define maps $\rho_k$ to the same space $\mathcal{X}$ (with distance $d$)

$$
\Xi_1 \xrightarrow{\pi_1} \cdots \xrightarrow{\pi_K} \Xi_K
$$

$$
\Xi_1 \xrightarrow{\rho_1} \cdots \xrightarrow{\rho_K} \Xi_K
$$

$\mathcal{X} = \mathbb{R}^d$

Reject for $d(\tilde{y}_1, \tilde{y}_2)$ “large”
Multidimensional scaling yields high-dimensional embeddings: $\Delta_1 \mapsto X'_1$ and $\Delta_2 \mapsto X'_2$

Canonical correlation finds $U_1 : X'_1 \mapsto X_1$ and $U_2 : X'_2 \mapsto X_2$ to maximize correlation

Out-of-sample embedding: $y_1 \mapsto y'_1$, $y_2 \mapsto y'_2$

Both $\tilde{y}_1 = U_1^T y'_1$ and $\tilde{y}_2 = U_2^T y'_2$ are in $\mathbb{R}^d$ with same coordinate system (i.e., they are commensurate)

Reject for $d(\tilde{y}_1, \tilde{y}_2)$ “large”
procrustes \circ mds

- Multidimensional scaling yields low-dimensional embeddings: \( \Delta_1 \mapsto X_1 \) and \( \Delta_2 \mapsto X_2 \)
- Procrustes\((X_1, X_2)\) yields
  \[
  Q^* = \arg\min_{Q^TQ=I} \|X_1 - X_2Q\|_F
  \]
- Out-of-sample embedding: \( y_1 \mapsto \tilde{y}_1, \ y_2 \mapsto \tilde{y}'_2 \)
- Both \( \tilde{y}_1 \) and \( \tilde{y}_2 = Q^*\tilde{y}'_2 \) are in \( \mathbb{R}^d \)
  with same coordinate system (i.e., they are commensurate)
- Reject for \( d(\tilde{y}_1, \tilde{y}_2) \) “large”
fidelity & commensurability

Fidelity is how well the mapping preserves original dissimilarities; our within-condition \textit{fidelity error} is given by

\[
\epsilon_{f_k} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} (d(\tilde{x}_{ik}, \tilde{x}_{jk}) - \delta_k(x_{ik}, x_{jk}))^2.
\]

Commensurability is how well the mapping preserves matchedness; our between-condition \textit{commensurability error} is given by

\[
\epsilon_{c_{k_1k_2}} = \frac{1}{n} \sum_{1 \leq i \leq n} (d(\tilde{x}_{ik_1}, \tilde{x}_{ik_2}) - \delta_{k_1k_2}(x_{ik_1}, x_{ik_2}))^2.
\]

Alas, \(\delta_{k_1k_2}\) does not exist; however, our story seems to suggest that it might be reasonable to let \(\delta_{k_1k_2}(x_{ik_1}, x_{ik_2}) = 0\) for all \(i, k_1, k_2\).

NB: There is also between-condition \textit{separability error} given by

\[
\epsilon_{s_{k_1k_2}} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} (d(\tilde{x}_{ik_1}, \tilde{x}_{jk_2}) - \delta_{k_1k_2}(x_{ik_1}, x_{jk_2}))^2.
\]
Methodological Comparison

- **canonical correlation** optimizes commensurability
  *without regard for fidelity*

- **procrustes ◦ mds** optimizes fidelity
  *without regard for commensurability*
Methodological Comparison

- **canonical correlation** optimizes commensurability 
  *without regard for fidelity*

- **procrustes ◦ mds** optimizes fidelity  
  *without regard for commensurability*

- compare: **joint optimization of fidelity & commensurability** . . .
Omnibus Embedding Approach

\[
\begin{align*}
M^{2n \times 2n} &= \begin{bmatrix}
\Delta_1^{n \times n} & W^{n \times n} \\
W^T & \Delta_2^{n \times n}
\end{bmatrix}
\end{align*}
\]

- Under “matched” assumption, impute dissimilarities \( \delta_{12}(x_{i1}, x_{j2}) \) to obtain an omnibus dissimilarity matrix \( M \).
- Embed \( M \) as \( 2n \) points in \( \mathbb{R}^d \).
- Let \( u_{i1} = \delta_1(y_1, x_{i1}) \) and \( v_{i2} = \delta_2(y_2, x_{i2}) \).
- Under \( H_0 : y_1 \sim y_2 \), impute \( v_{i1} = \delta_{12}(y_1, x_{i2}) \) and \( u_{i2} = \delta_{12}(y_2, x_{i1}) \).
- Out-of-sample embedding of \((u_{1}^T, v_{1}^T)^T\) and \((u_{2}^T, v_{2}^T)^T\) yields \( \tilde{y}_1 \) and \( \tilde{y}_2 \).
Simulation results indicate that
joint optimization of fidelity & commensurability
via omnibus embedding approach
is (for this case)
superior to canonical correlation and procrustes•mds
Spurious Correlation Phenomenon

Let $\Xi_k = S^p \times S^q$;
$S^p$ encodes “signal” and $S^q$ encodes “noise”

On $S^p$, let $\alpha_i \sim iid \text{ Dirichlet}(1)$ and $X_{ik}^1 \sim iid \text{ Dirichlet}(r\alpha_i + 1)$
(signal, as before)

On $S^q$, let $X_{ik}^2 \sim iid \text{ Dirichlet}(1)$
(pure noise)

For $c \in [0, 1]$, let $X_{ik} = [(1 - c)X_{ik}^1, cX_{ik}^2]$
Incommensurability Phenomenon I
Incommensurability Phenomenon II: Grassmann & Hausdorff

Dirichlet

Scale & Polarity

Procrustes

$\|M_1 - M_2 P_S\|$

Sqrt of Commensurability Error

$\|M_1 - M_2 Q_l\|$

Sqrt of Commensurability Error
Unitarily Invariant Metrics on the Grassmann Space

Li Qiu
Department of Electrical & Electronic Engineering
Hong Kong University of Science & Technology
Clear Water Bay, Kowloon, Hong Kong, China

Yanxia Zhang
Department of Applied Mathematics
Beijing Institute of Technology
Beijing 100081 China

Chi-Kwong Li
Department of Mathematics
College of William and Mary
Williamsburg, Virginia 23187-8795 USA

Abstract
Let $G_{m,n}$ be the Grassmann space of $m$-dimensional subspaces of $\mathbb{F}^n$. Denote by $\theta_1(\mathcal{X},\mathcal{Y}), \ldots, \theta_m(\mathcal{X},\mathcal{Y})$ the canonical angles between subspaces $\mathcal{X}, \mathcal{Y} \in G_{m,n}$. It is shown that $\Phi(\theta_1(\mathcal{X},\mathcal{Y}), \ldots, \theta_m(\mathcal{X},\mathcal{Y}))$ defines a unitarily invariant metric on $G_{m,n}$ for every symmetric gauge function $\Phi$. This provides a wide class of new metrics on $G_{m,n}$. Some related results on perturbation and approximation of subspaces in $G_{m,n}$, as well as the canonical angles between them, are also discussed. Furthermore, the equality cases of the triangle inequalities for several unitarily invariant metrics are analyzed.
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
- textual content of the document
- links in the document to other documents
Consider a subset of English and French Wikipedias that are 1-1 correspondent
We take the (directed) 2-neighborhood of the document “Algebraic Geometry” in the English Wikipedia, with the associated documents in the French Wikipedia ($n = 1382$)
Experimental Results

Experimental results indicate that
joint optimization of fidelity & commensurability
via omnibus embedding approach
is (for this case)
superior to canonical correlation and procrustes·mds
Minimum Reciprocal Rank experimental paired results indicate that **joint optimization of fidelity & commensurability** via omnibus embedding approach is (for this case) superior to **procrustes o mds**.
Exploitation Task: Classification

\[(X, Y, Z) \sim F_{X,Y,Z}\]
\[Y : \mathcal{J} \cup \overline{\mathcal{J}}\]
\[Z : \mathcal{J} \rightarrow \{0, 1\}\]
\[X | Z = 2 : \mathcal{J} \rightarrow \nabla Z\]

- Available training data:
  \[\mathcal{Z}_0 = \{(x_i, y_i \in \mathcal{J}, z_i = 0)\}\]
  \[\mathcal{Z}_e = \{(x_i, y_i \in \overline{\mathcal{J}}, z_i = 0)\}\]
  \[\mathcal{Z}_1 = \{(x_i, y_i \in \mathcal{J}, z_i = 1)\}\]

- Data to be classified:
  \[(x, y \in \mathcal{J}, z = 1)\]

\[g_e \sim x \sim \nabla e \rightarrow \mathcal{J}\]
\[\tilde{g}_e \sim x \sim \nabla \tilde{e} \rightarrow \overline{\mathcal{J}}\]
\[g_0 \sim \mathcal{J}\]
\[
\begin{align*}
(1) & \quad \tilde{g}_1 = g_0 \circ \omega \\
(2) & \quad \text{LDA approach (implicitly) identifies portions } P(q, Z) \text{ for which } \hat{\omega}(x) \text{ identify} \\
(3) & \quad \tilde{g}_1 = \tilde{g}_0 \circ \omega \\
& \quad \text{“known” \ estimate} \\
& \quad \text{unknown \ \ needed}
\end{align*}
\]
Integrated Sensing and Processing

\[ \Delta(\theta) \]

\[ \Xi_1, \Xi_2, \ldots, \Xi_{K-1}, \Xi_K \]

ISP

\[ \rho_1, \rho_2, \ldots, \rho_{K-1}, \rho_K \]

\[ \mathbb{R}^{d_1}, \mathbb{R}^{d_2}, \ldots, \mathbb{R}^{d_{K-1}}, \mathbb{R}^{d_K} \]

MM

\[ W \]

Fidelity

Commensurability

Separability
Fusion and Inference from Multiple and Massive Disparate Data Sources

\[ f: \quad \Xi \rightarrow \quad M = \{ (V_i, x_i, t_i) \} \]

\[ g: \quad M \rightarrow \]

\( f \) extracts “events” \((V_i, x_i, t_i)\) -- who does what when

\( g \) produces time series of attributed graphs \( \{ G_t (M) \} \)

Anomaly Detection in Time Series of Attributed Graphs
“The wealth of your practical experience with sane and interesting problems will give to mathematics a new direction and a new impetus.”

– Leopold Kronecker to Hermann von Helmholtz –