

Fusion and Inference from Multiple Data Sources

Priebe, Carey E. ¹

E-mail: cep@jhu.edu

Ma, Zhiliang ¹

Marchette, David J. ²

Hohman, Elizabeth ²

Coppersmith, Glen ³

¹ *Johns Hopkins University, Applied Mathematics & Statistics*

3400 North Charles Street, Baltimore, MD 21218, USA

² *Naval Surface Warfare Center, Code Q21, Dahlgren, VA 22448, USA*

³ *Johns Hopkins University, Human Language Technology Center of Excellence, USA*

Abstract: Given K matched feature vectors $\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,K}$ for each of n objects, with $\mathbf{x}_{i,k} \in \Xi_k$, and given additional feature vectors $\{\mathbf{y}_k\}_{k=1}^K$, we consider testing $H_0 : \{\mathbf{y}_k\}_{k=1}^K$ are matched feature vectors representing a single object measured under K conditions versus H_A : they do not represent a single object. We develop an approach to this problem which uses only the interpoint dissimilarities for each condition separately. We impute the dissimilarities between matched measurements of different conditions to obtain one omnibus dissimilarity matrix, which is then embedded into Euclidean space. Out-of-sample embedding is used to embed the new measurements $\{\mathbf{y}_k\}_{k=1}^K$ into this same space, and we determine whether a match is present by examining the distance between the corresponding embeddings. We illustrate our methodology on English and French documents collected from Wikipedia, demonstrating superior performance compared to that obtained via standard Procrustes analysis.

Key words: fusion, inference, dissimilarity, embedding

1. Problem

Consider two Wikipedias, English and French. Two collections of $n = 1382$ matched Wikipedia documents are given and denoted by $\mathbf{E} = \{\mathbf{x}_{i,1}\}_{i=1}^n$, $\mathbf{F} = \{\mathbf{x}_{i,2}\}_{i=1}^n$, and $\mathbf{x}_{i,1} \sim \mathbf{x}_{i,2}$, meaning that the English document $\mathbf{x}_{i,1}$ and the French document $\mathbf{x}_{i,2}$ are matched. Our goal is to determine whether a match is present between two new documents \mathbf{y}_1 and \mathbf{y}_2 . That is, we consider testing: $H_0 : \mathbf{y}_1 \sim \mathbf{y}_2$ versus $H_A : \mathbf{y}_1 \not\sim \mathbf{y}_2$. We assume the two new documents represent a matched pair under H_0 . This allows us to control the probability of missing a true match. This is practical when computer algorithms are used to eliminate easily rejected pairs and the remaining possibly matched pairs will be manually examined. We consider two types of dissimilarity matrices, denoted generically as D_E and D_F : (1) G_E and G_F , developed from the graph structures; (2) T_E and T_F , obtained from the textual contents.

2. Approach

The basic idea is to (i) embed the two sets of matched documents, \mathbf{E} and \mathbf{F} , into the *same* space Ξ ($\Xi = \mathbb{R}^d$ in most cases, but with the possibility of being an uncommon space); (ii) embed the two new documents \mathbf{y}_1 and \mathbf{y}_2 , referred to as the out-of-sample documents, into the space Ξ ; (iii) determine whether a match is present by examining the distance between the embeddings of \mathbf{y}_1 and \mathbf{y}_2 , with a large distance being evidence against H_0 . The key to this approach is the embedding: how shall we determine the space Ξ and how shall we embed the two new documents into Ξ ? A traditional method is via Procrustes analysis (Sibson 1978), which we refer to as the P -approach.

Our approach is to impute W , the dissimilarities between \mathbf{E} and \mathbf{F} , by the average of D_E and D_F to obtain one omnibus dissimilarity matrix M , and then embed M into the space Ξ . For any two additional documents \mathbf{y}_1 and \mathbf{y}_2 , let u_1 and v_2 denote the dissimilarities (vector) between \mathbf{y}_1 and

$$\begin{array}{c}
M \\
y_1 \\
y_2
\end{array}
=
\begin{array}{c}
\left[\begin{array}{cc}
D_E & W \\
W^T & D_F
\end{array} \right]
\begin{array}{c}
\begin{array}{c}
n \times n \\
n \times n
\end{array} \\
\begin{array}{c}
n \times 1 \\
n \times 1
\end{array} \\
\begin{array}{c}
n \times 1 \\
n \times 1
\end{array} \\
\begin{array}{c}
1 \times 1 \\
1 \times 1
\end{array}
\end{array}
\begin{array}{c}
\begin{array}{c}
u_1 \\
u_2
\end{array} \\
\begin{array}{c}
v_1 \\
v_2
\end{array} \\
\begin{array}{c}
a \\
a
\end{array} \\
\begin{array}{c}
0 \\
0
\end{array}
\end{array}
\end{array}$$

Figure 1: We impute W by $(D_E + D_F)/2$ to construct M , u_2 and v_1 by $(u_1 + v_2)/2$.

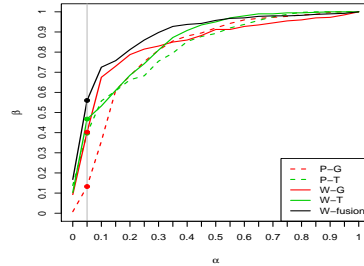


Figure 2: The ROC curve depicts that W -approach is generally superior to P -approach; T is generally superior to G; Fusion is generally superior to either G or T alone.

E , y_2 and F , respectively. Under the null hypothesis that y_1 and y_2 are matched, we impute the dissimilarities between y_1 and F (denoted by v_1), and dissimilarities between y_2 and E (denoted by u_2) by averaging, in the same way that we obtain W . That is, $v_1 = u_2 = (u_1 + v_2)/2$. Out-of-sample embedding (Trosset and Priebe, 2008) is used to embed $(u_1^T, v_1^T)^T$ and $(u_2^T, v_2^T)^T$ into Ξ . Notice that this approach embeds D_E and D_F simultaneously into Ξ . We refer this approach as the W -approach. Figure 1 depicts the construction of the omnibus dissimilarity matrix M .

We consider one additional step, to combine the data of textual content and graph structure. Ideally both sources of data contain complementary information so that their fusion leads to superior power in testing compared to either textual content data or graph structure data alone. We achieve the fusion by combining the embeddings obtained in the W -approach via the Cartesian product (Ma, Cardinal-Stakenas, Park, and Priebe, 2009). Distances in the embedding product space are then computed and examined to test the presence of a match.

3. Results

We randomly select two pairs of matched documents from E and F , then apply the approaches introduced in Section 2 to obtain the distances between the two matched pairs (denoted by d_0), and the distances between the two non-matched pairs (denoted by d_A). We use Classical Multidimensional Scaling (CMDS) (Torgerson 1952) in the embedding and specify $\Xi = \mathbb{R}^d$ ($d = 6$ is determined by examining the “scree” plot of the full embeddings). We use ranks of the distances d_A based on 200 Monte Carlo simulations to estimate the powers for different levels of α . That is, for each $\alpha \in [0, 1]$, the critical value c_α is defined as the (100α) th percentile of d_0 , and the corresponding power is the percentage of distances in d_A that are larger than the critical value c_α . The power at level α is our performance in determining that a non-match is in fact a non-match. The β against α ROC curves are shown in Figure 2. For example, at $\alpha = 0.05$ (missing 5% of the true matches), we obtain a power of $\hat{\beta}_{W-fusion} = 0.560$ (correctly eliminating 56% of the false matches) via W -fusion. This is a statistical significant improvement over the results obtained sans fusion ($\hat{\beta}_{P-G} = 0.135$, $\hat{\beta}_{P-T} = 0.379$, $\hat{\beta}_{W-G} = 0.403$, $\hat{\beta}_{W-T} = 0.468$. See Figure 2).

REFERENCES

Ma, Z., Cardinal-Stakenas, A., Park, Y., and Priebe, C.E. (2009). Combining Dissimilarity Representations in Embedding Product Space. *submitted for publication*.

Sibson, R. (1978). Studies in the Robustness of Multidimensional Scaling: Procrustes Statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(2):234–238.

Torgerson, W. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419.

Trosset, M.W. and Priebe, C.E. (2008). The out-of-sample problem for classical multidimensional scaling. *Computational Statistics & Data Analysis*, 52(10):4635–4642.