

A Nonmetric Embedding Approach to Testing for Matched Pairs

Brent Castle^a Michael W. Trosset^b Carey E. Priebe^c

October 9, 2011

Technical Report 11-04
Department of Statistics
Indiana University
Bloomington, IN

^aSchool of Informatics & Computing, Indiana University. Email: bscastle@cs.indiana.edu

^bDepartment of Statistics, Indiana University. Email: mtrosset@indiana.edu

^cDepartment of Applied Mathematics & Statistics, Johns Hopkins University. Email: cep@jhu.edu

A Nonmetric Embedding Approach to Testing for Matched Pairs

Brent Castle*

Michael W. Trosset[†]

Carey E. Priebe[‡]

Abstract

We consider the problem of matched pair hypothesis testing, i.e., the problem of determining whether or not a pair of disparate feature vectors correspond to a common object. Our approach relies on measures of pairwise dissimilarity in the respective feature spaces. We use three-way nonmetric multidimensional scaling to embed a training set of matched pairs in a low-dimensional Euclidean space, then embed test pairs by a corresponding out-of-sample technique. The Euclidean distance between the points obtained by embedding a test pair is used as a statistic to test whether or not the pair is matched. We demonstrate our methodology on simulated data and a collection of Wikipedia articles.

Key Words: three-way multidimensional scaling; multi-view learning; multiple modalities; heterogeneous data; data fusion

1 Introduction

Data mining often necessitates the fusion of disparate sources of data. For example, a captioned image has both text and an image that describe the same object, and images may be compared by measures of color, texture, etc. Often, domain experts have constructed specialized measures of proximity for each data source, but the measures are not on commensurate scales. Thus, a fundamental problem in data fusion is the problem of constructing a suitable representation of the data in which inference can proceed.

In machine learning, recent approaches to combining disparate data have assumed the existence of pairwise similarities, interpreted as inner products (kernels). From a set of base kernels, multiple kernel learning (MKL) constructs a single kernel, typically for use in a support vector machine. See [8] and [15] for key references and [11] for a rare example of a nonmetric ap-

proach. MKL is often used for multi-view learning, i.e., learning about objects that are observed from multiple views. MKL was used in [18] to combine information from multiple base features for the purpose of classifying images. A Bayesian approach to MKL for multi-view learning was proposed in [2]. MKL has also been used to combine kernels for use across multiple tasks [5] and for boosting [19].

We consider the problem of matched pair hypothesis testing, i.e., the problem of determining whether or not a pair of disparate feature vectors correspond to a common object. Instead of working with pairwise similarities and positive semidefinite kernels, we work with pairwise dissimilarities and Euclidean distances. Instead of learning a convex combination of multiple kernels for the purpose of classification, we embed multiple dissimilarities in a common Euclidean representation for the purpose of hypothesis testing. Our emphasis on dissimilarities resembles [12]. Methods for constructing configurations of points from pairwise dissimilarities is the province of multidimensional scaling (MDS). See [1] for an introduction to MDS.

Priebe et al. [14] proposed an embedding approach for matched pair hypothesis testing. Given two sets of pairwise dissimilarities for a training set of matched pairs, their JOFC approach constructs two distinct configurations in the same Euclidean space in such a way that elements of a matched pair tend to be near each other. A test pair is then embedded, each element with respect to the corresponding configuration. The Euclidean distance between the locations of the embedded test pair is then used as a test statistic. We propose an alternative approach that uses three-way MDS to construct a single configuration of points from two sets of pairwise dissimilarities. To facilitate the fusion of disparate measures of dissimilarity, our approach is nonmetric, i.e., only the ranks of the pairwise dissimilarities are used to construct the common representation of the matched pairs and to embed the elements of test pairs.

Section 2 introduces notation and describes the JOFC methodology in [14]. We modify JOFC in Section 3, deriving a three-way nonmetric MDS approach to matched pair hypothesis testing. Section 4 summarizes

*School of Informatics & Computing and Department of Statistics, Indiana University, 309 N Park Ave, Bloomington, IN 47408. Email: bscastle@cs.indiana.edu

[†]Department of Statistics and School of Informatics & Computing, Indiana University. Email: mtrosset@indiana.edu

[‡]Department of Applied Mathematics & Statistics, Johns Hopkins University. Email: cep@jhu.edu

the algorithms that implement our approach. Section 5 provides an illustrative example. A reader unfamiliar with embedding methods might benefit from examining the figures in this section before studying the details of the formulations. Section 6 reports results from several numerical experiments and Section 7 concludes.

2 Background

Let $\omega_1, \omega_2, \dots, \omega_N \in \Omega$ be independent and identically distributed objects drawn from an abstract probability space. The objects may or may not be directly observable, but the feature maps $\pi_k : \Omega \mapsto \Xi_k$ represent the objects in observable feature spaces. Feature spaces are often Euclidean, but we allow more general representations, e.g., graphs.

We restrict attention to the case of $K = 2$ feature maps. If $x_{i1} = \pi_1(\omega_i) \in \Xi_1$ and $x_{i2} = \pi_2(\omega_i) \in \Xi_2$ for some $\omega_i \in \Omega$, then we say that (x_{i1}, x_{i2}) is a *matched pair* and we write $x_{i1} \sim x_{i2}$. If $x_{i1} = \pi_1(\omega_i)$ and $x_{j2} = \pi_2(\omega_j)$ for some $\omega_i \neq \omega_j$, then we say that (x_{i1}, x_{j2}) is a *mismatched pair* and we write $x_{i1} \not\sim x_{j2}$. Unless the feature maps are injective, a pair may be both matched and mismatched.

We assume that each feature space is equipped with a measure of dissimilarity between pairs of points. Formally, a measure of dissimilarity is a function $\delta_k : \Xi_k \times \Xi_k \mapsto \mathbb{R}$ that is symmetric ($\delta_k(x_{ik}, x_{jk}) = \delta_k(x_{jk}, x_{ik})$), nonnegative, and zero if its arguments are identical. Informally, the dissimilarity of $\pi_k(\omega_i)$ and $\pi_k(\omega_j)$ should be small if ω_i and ω_j are similar, large if they are dissimilar. In many applications, domain experts have developed specialized measures of dissimilarity, e.g., for comparing the color or texture of images. Absent such measures, one often relies on standard measures of distance, e.g., Euclidean distance between feature vectors or shortest path distance between nodes of a graph.

2.1 Matched Pair Hypothesis Testing Consider two probability models for generating pairs in $\Xi_1 \times \Xi_2$: either (H_0) draw a single object, ω_i , then compute $\pi_1(\omega_i)$ and $\pi_2(\omega_i)$; or (H_A) draw two objects, ω_i and ω_j , independently, then compute $\pi_1(\omega_i)$ and $\pi_2(\omega_j)$. We assume that a *training set* of matched pairs,

$$\mathcal{S}_N = \{(x_{11}, x_{12}), \dots, (x_{N1}, x_{N2})\},$$

was generated by H_0 . If it is not known whether (y_1, y_2) was generated by H_0 or H_A , then (y_1, y_2) is a *test pair*. Given a training set and a test pair, we propose a test of the null hypothesis that the test pair was generated by H_0 against the alternative hypothesis that the test pair was generated by H_A . We write these simple hypotheses as $H_0 : y_1 \sim y_2$ and $H_A : y_1 \not\sim y_2$. This formulation of *matched pair hypothesis testing* is a special case (pairs

instead of K -tuples) of a more general decision problem described in [14].

2.2 Joint Optimization of Fidelity and Commensurability (JOFC) Priebe et al. [14] described an approach to matched pair hypothesis testing that they termed Omnibus Embedding. For each dissimilarity function δ_k , they used MDS to represent the training objects in a p -dimensional Euclidean space. The two representations were aligned in a manner described below. Each y_k was then embedded in representation k and the Euclidean distance between the embedded points was used as a test statistic.

More precisely, let

$$\Delta_k = [\delta_k(x_{ik}, x_{jk})]_{ij}$$

and let $\tilde{x}_{1k}, \tilde{x}_{2k}, \dots, \tilde{x}_{Nk} \in \mathbb{R}^p$ denote the corresponding configuration of points constructed by minimizing the raw stress criterion,

$$\begin{aligned} \sigma_r(\tilde{X}_k; \Delta_k) &= \sum_{i < j} [\delta_k(x_{ik}, x_{jk}) - d(\tilde{x}_{ik}, \tilde{x}_{jk})]^2 \\ &= \frac{1}{2} \left\| \Delta_k - D(\tilde{X}_k) \right\|_F^2, \end{aligned}$$

where d denotes Euclidean distance in \mathbb{R}^p , \tilde{X}_k is the $N \times k$ *configuration matrix* that contains \tilde{x}_{ik}^t in row i , $D(\tilde{X}_k) = [d(\tilde{x}_{ik}, \tilde{x}_{jk})]_{ij}$ is the corresponding matrix of pairwise Euclidean distances, and $\|\cdot\|_F$ denotes the Frobenius norm. Priebe et al. [14] termed $\sigma_r(\tilde{X}_k; \Delta_k)$ the *fidelity* of embedding Δ_k . Aligning \tilde{X}_1 and \tilde{X}_2 requires minimizing a measure of *commensurability*, e.g.,

$$\eta(\tilde{X}_1, \tilde{X}_2) = \left\| \tilde{X}_1 - \tilde{X}_2 \right\|_F^2.$$

One might simply compose a Procrustes analysis with MDS ($p \circ m$), first embedding to minimize fidelity, then rotating one configuration to improve commensurability. In contrast, Priebe et al. [14] proposed jointly optimizing fidelity and commensurability (JOFC). Their general framework for JOFC allows one to choose various measures of commensurability and to vary the tradeoff between fidelity and commensurability. One of the key insights in [14] is that jointly optimizing fidelity and commensurability leads to a more powerful test than does optimizing each in turn, e.g., by $p \circ m$.

We present a variation of JOFC that minimizes

$$\begin{aligned} (2.1) \quad \sigma_j(\tilde{X}_1, \tilde{X}_2) &= \sigma_r(\tilde{X}_1; \Delta_1) + \sigma_r(\tilde{X}_2; \Delta_2) + \eta(\tilde{X}_1, \tilde{X}_2) \\ &= \frac{1}{2} \sum_{k=1,2} \left\| \Delta_k - D(\tilde{X}_k) \right\|_F^2 + \left\| \tilde{X}_1 - \tilde{X}_2 \right\|_F^2, \end{aligned}$$

subject to $\tilde{X}_1, \tilde{X}_2 \in \mathbb{R}^{N \times p}$, p fixed. Writing $\tilde{X}_1 = \tilde{X}_1 Q$ for an orthogonal matrix Q , we see that $p \circ m$ first minimizes $\sigma_r(\tilde{X}_1; \Delta_1) + \sigma_r(\tilde{X}_2; \Delta_2)$, then minimizes $\|\tilde{X}_1 - \tilde{X}_2\|_F^2 = \|\tilde{X}_1 Q - \tilde{X}_2\|_F^2$ with respect to Q . In general, $p \circ m$ does not minimize σ_j . Figure 1 displays an embedding of 20 matched pairs by JOFC. One configuration is represented by circles, the other by crosses, and the pairs are connected by line segments.

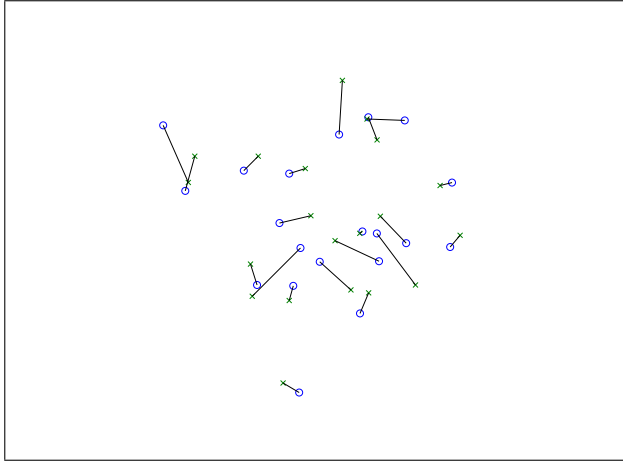


Figure 1: JOFC embedding of 20 matched pairs.

Now suppose that (y_1, y_2) have dissimilarity vectors

$$\bar{\delta}_k = (\delta_k(y_k, x_{1k}), \dots, \delta_k(y_k, x_{Nk}))^t.$$

Each y_k is embedded in relation to \tilde{X}_k by minimizing

$$\sigma_{j-o}(\tilde{y}_k) = \frac{1}{2} \left\| \begin{bmatrix} \Delta_k & \bar{\delta}_k \\ \bar{\delta}_k^t & 0 \end{bmatrix} - D \left(\begin{bmatrix} \tilde{X}_k \\ \tilde{y}_k^t \end{bmatrix} \right) \right\|_F^2,$$

the raw stress criterion for an augmented dissimilarity matrix and configuration in which only one point is free to vary. This is an example of *out-of-sample* embedding. Figure 2 illustrates the out-of-sample embedding of three matched pairs, whereas Figure 3 illustrates the out-of-sample embedding of three mismatched pairs. In both cases, the reference configuration is the configuration displayed in Figure 1.

If both fidelity and commensurability are good, then the Euclidean distance $d(\tilde{y}_1, \tilde{y}_2)$ should be small for matched pairs and larger for mismatched pairs. This distance is the JOFC test statistic. Its null distribution can be estimated from the training set by cross-validation, then used to determine critical values for achievable levels of significance. For a given critical value, power can be estimated from the empirical distribution of $d(\tilde{y}_1, \tilde{y}_2)$ for mismatched pairs.

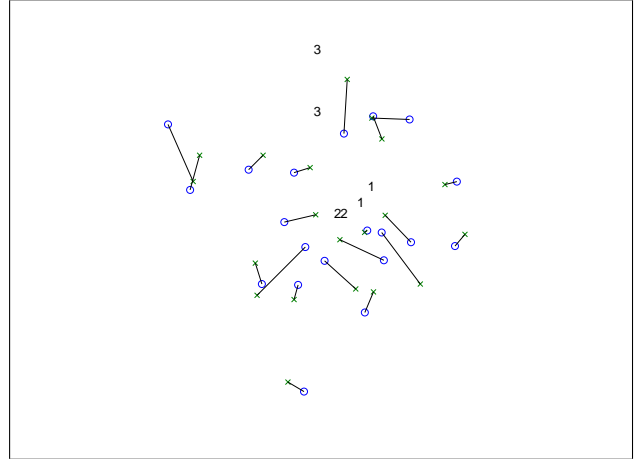


Figure 2: Three matched pairs have been added to the configuration in Figure 1.

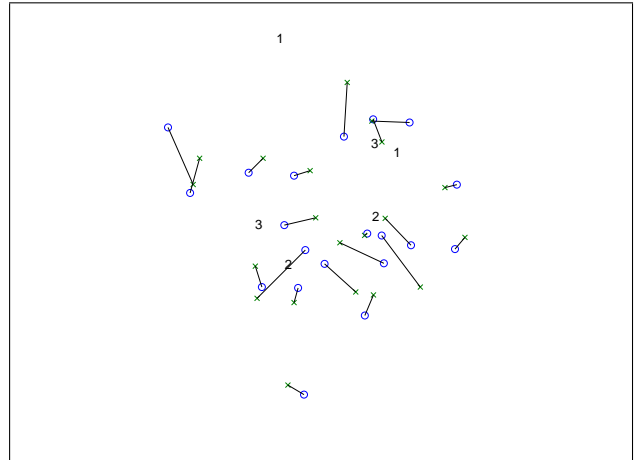


Figure 3: Three mismatched pairs have been added to the configuration in Figure 1.

3 Three-Way Nonmetric MDS (3WNM)

Inspired by a comment at the end of [14], that it might be of interest to compare three-way MDS to JOFC, we now describe an alternative approach to matched pair hypothesis testing. In contrast to JOFC, three-way MDS constructs a single configuration from both Δ_k . Because the δ_k may measure dissimilarity in different ways, e.g., Euclidean distance versus shortest path distance, we only use information about how the entries of each Δ_k are ordered, i.e., we allow monotonic transformations of each δ_k . Thus, our approach to embedding is an example of three-way nonmetric MDS. For each a test pair embedded as $\tilde{y}_1, \tilde{y}_2 \in \mathbb{R}^p$, we use $d(\tilde{y}_1, \tilde{y}_2)$ as a test statistic, as described in Section 2.2.

We describe 3WNM for the special case of matched pairs ($K = 2$), but note that our formulation naturally extends to $K > 2$ representations. The test statistic is then the square root of the sum of the squared distances between the K out-of-sample embedded points.

3.1 Embedding the Training Pairs Fix p , the embedding dimension, and let \tilde{X} denote the $N \times p$ configuration matrix constructed from the training pairs. In three-way MDS, the common representation constructed from multiple dissimilarity matrices is often called the group space. Because \tilde{X} models both Δ_1 and Δ_2 , a natural extension of the raw stress criterion is

$$\sigma_r(\tilde{X}; \Delta_1) + \sigma_r(\tilde{X}; \Delta_2) = \frac{1}{2} \sum_{k=1,2} \left\| \Delta_k - D(\tilde{X}) \right\|_F^2,$$

the identity model for three-way MDS [3]. More generally, one might weight the summands. Notice that the identity model is just (2.1) with $\tilde{X}_1 = \tilde{X}_2$.

Nonmetric MDS constructs configurations using only the ranks of the dissimilarities. Following Kruskal [7], we replace each Δ_k with $\mathcal{M}(\Delta_k)$, the cone of all dissimilarity matrices whose entries have the same order as the entries of Δ_k , i.e.,

$$\mathcal{M}(\Delta_k) = \left\{ \tilde{\Delta}_k : \begin{array}{l} \tilde{\delta}_k(x_{ik}, x_{jk}) \leq \tilde{\delta}_k(x_{mk}, x_{nk}) \text{ if } \\ \delta_k(x_{ik}, x_{jk}) \leq \delta_k(x_{mk}, x_{nk}) \end{array} \right\}$$

Equivalently, $\mathcal{M}(\Delta_k)$ is the cone of all dissimilarity matrices that can be obtained from Δ_k by a nondecreasing transformation of its entries. Note that $[0] \in \mathcal{M}(\Delta_k)$. We will refer to the elements of $\mathcal{M}(\Delta_k)$ as surrogate dissimilarity matrices.

Conceptually, we would like to find a triple $(\tilde{X}, \tilde{\Delta}_1, \tilde{\Delta}_2)$ that minimizes

$$\sigma_n(\tilde{X}, \tilde{\Delta}_1, \tilde{\Delta}_2) = \frac{1}{2} \sum_{k=1,2} \left\| \tilde{\Delta}_k - D(\tilde{X}) \right\|_F^2$$

subject to $\tilde{\Delta}_k \in \mathcal{M}(\Delta_k)$. Unfortunately, embedding each point at the same location (resulting in $D(\tilde{X}) = [0]$) and choosing $\tilde{\Delta}_k = [0]$ results in a degenerate global solution. To preclude such solutions, we are obliged to modify the problem. Traditionally, degenerate solutions have been precluded by adopting a scale-invariant objective function. Instead, following Trosset's [16] observation that degenerate solutions can be precluded by imposing explicit constraints, we require the surrogate dissimilarity matrices to lie in

$$\mathcal{N} = \left\{ \tilde{\Delta}_k : \left\| \tilde{\Delta}_k \right\|_F^2 \geq 1 \right\}.$$

Our optimization problem is then

$$(3.2) \quad \begin{array}{ll} \text{minimize} & \sigma_n(\tilde{X}, \tilde{\Delta}_1, \tilde{\Delta}_2) \\ \text{subject to} & \tilde{\Delta}_k \in \mathcal{M}(\Delta_k) \cap \mathcal{N}. \end{array}$$

Thus, 3WNM constructs a single \tilde{X} that attempts to respect the orderings of both Δ_1 and Δ_2 .

3.2 Embedding the Test Pairs After embedding the training pairs in \mathbb{R}^p , we embed the test pairs in relation to the training pairs. Given a pair of test objects (y_1, y_2) , let

$$\bar{\delta}_k = (\delta_k(y_k, x_{1k}), \dots, \delta_k(y_k, x_{Nk}))^t \text{ for } k = 1, 2$$

denote the vector of δ_k -dissimilarities between y_k and each of the objects in the training set. Our out-of-sample embedding of y_1 and y_2 resembles the technique described in Section 2.2; however, we embed with respect to a single configuration and allow monotonic transformations of $\bar{\delta}_k$. Hence, we obtain $\tilde{y}_1, \tilde{y}_2 \in \mathbb{R}^p$ by minimizing

$$\sigma_{n-o}(\tilde{y}_k, \bar{\delta}_k) = \frac{1}{2} \left\| \begin{bmatrix} \tilde{\Delta}_k & \bar{\delta}_k \\ \bar{\delta}_k^t & 0 \end{bmatrix} - D \left(\begin{bmatrix} \tilde{X} \\ \tilde{y}_k \end{bmatrix} \right) \right\|_F^2$$

subject to the constraint that

$$(3.3) \quad \begin{bmatrix} \tilde{\Delta}_k & \bar{\delta}_k \\ \bar{\delta}_k^t & 0 \end{bmatrix} \in \mathcal{M} \left(\begin{bmatrix} \Delta_k & \bar{\delta}_k \\ \bar{\delta}_k^t & 0 \end{bmatrix} \right).$$

Because $\tilde{\Delta}_k$ is fixed, it is not necessary to impose an additional nondegeneracy constraint. Notice that (3.3) reduces to simple bound constraints on the components of $\bar{\delta}_k$, the bounds determined by entries in $\tilde{\Delta}_k$.

4 Algorithms

We now describe algorithms for finding good solutions of (3.2) and its out-of-sample extension. Our algorithms rely on a general search strategy, sometimes called variable alternation, in which the objective function is successively minimized with respect to distinct blocks of decision variables. The classic example of minimization by variable alternation is coordinate descent, in which the blocks consist of the individual decision variables. Variable alternation produces a nondecreasing sequence of objective function values. Under mild conditions, the sequence of iterates converges to a connected set of stationary points.

The optimization problems formulated in Section 3 have a natural block structure. For (3.2), the blocks are \tilde{X} , $\tilde{\Delta}_1$, and $\tilde{\Delta}_2$. For the out-of-sample problem, the blocks are \tilde{y}_k and $\bar{\delta}_k$.

4.1 Three-Way Nonmetric MDS (3WNM) Algorithm 4.1 describes our variable alternation strategy for solving (3.2). Many MDS algorithms have this general structure [1]. The alternating subproblems are solved by Algorithms 4.2 and 4.3. The loop may be terminated by imposing a maximum number of iterations ($\ell \leq L$) and/or by monitoring how much an iteration decreases the value of the objective function. If N is large, then one might choose L small. In practice, much of the progress is made in the first several iterations.

ALGORITHM 4.1. 3WNM

0. Set $\tilde{\Delta}_k = \Delta_k / \min(1, \|\Delta_k\|)$ and initialize \tilde{X} .
1. Compute $\sigma_n^{[0]}$ and set $\ell = 0$.
2. Do until termination:
 - (a) Increment ℓ .
 - (b) Fix $\tilde{\Delta}_1$ and $\tilde{\Delta}_2$ and minimize σ_n with respect to \tilde{X} . This step uses Algorithm 4.2.
 - (c) Fix \tilde{X} and minimize σ_n with respect to $\tilde{\Delta}_1$ and $\tilde{\Delta}_2$, subject to $\tilde{\Delta}_k \in \mathcal{M}(\Delta_k) \cap \mathcal{N}$. This step uses Algorithm 4.3.
 - (d) Compute $\sigma_n^{[\ell]}$.

Step 2b in Algorithm 4.1 fixes the surrogate dissimilarity matrices $\tilde{\Delta}_1$ and $\tilde{\Delta}_2$ and minimizes σ_n with respect to \tilde{X} . This problem is the standard identity model for three-way MDS [3]. Borg and Groenen [1] simplified this problem by writing

$$\begin{aligned} \sigma_n(\tilde{X}; \tilde{\Delta}_1, \tilde{\Delta}_2) &= \frac{1}{2} \sum_{k=1,2} \left\| \tilde{\Delta}_k - D(\tilde{X}) \right\|_F^2 \\ &= \left\| \bar{\Delta} - D(\tilde{X}) \right\|_F^2 + \frac{1}{2} \left\| \bar{\Delta} - \tilde{\Delta}_1 \right\|_F^2 + \frac{1}{2} \left\| \bar{\Delta} - \tilde{\Delta}_2 \right\|_F^2, \end{aligned}$$

where $\bar{\Delta} = (\tilde{\Delta}_1 + \tilde{\Delta}_2)/2$. Only the first term in the final expression depends on \tilde{X} , and this term is just the raw stress criterion with the average surrogate dissimilarities. Hence, Step 2b requires only an algorithm that minimizes the raw stress criterion for a single fixed dissimilarity matrix.

The raw stress criterion is often minimized by iterative majorization, specifically by repeated application of a fixed point mapping known as the Guttman transform. If, as here, the summands in the raw stress criterion are equally weighted, then computing the Guttman transform is not expensive. Let $\bar{\delta}_{ij}$ denote entry ij in $\bar{\Delta}$ and let $d_{ij}(\tilde{X})$ denote $\|\tilde{x}_i - \tilde{x}_j\|$. Let $B(\tilde{X})$ denote the $N \times N$ matrix with off-diagonal entries $b_{ij} = -\bar{\delta}_{ij}/d_{ij}(\tilde{X})$ and

diagonal entries $b_{ii} = -\sum_{j:i \neq j} b_{ij}$. Then the Guttman transform of \tilde{X} is

$$\Gamma(\tilde{X}) = B(\tilde{X})\tilde{X}/N.$$

See [1] for the derivation of Γ .

The complete algorithm for step 2b is described in Algorithm 4.2. Again, the loop may be terminated by imposing a maximum number of iterations and/or by monitoring how much an iteration decreases the value of the raw stress criterion. In our experience, 5–10 iterations usually produce a good \tilde{X} .

ALGORITHM 4.2. Update \tilde{X}

1. Compute $\bar{\Delta} = (\tilde{\Delta}_1 + \tilde{\Delta}_2)/2$.
2. Compute $\sigma_n^{[0]}$ and set $\ell' = 0$.
3. Do until termination:
 - (a) Increment ℓ' .
 - (b) $\tilde{X} \leftarrow \Gamma(\tilde{X})$
 - (c) Compute $\sigma_n^{[\ell']}$.

Step 2c in Algorithm 4.1 updates the surrogate dissimilarity matrices. For a fixed \tilde{X} , the objective function σ_n is separable and it suffices to update each $\tilde{\Delta}_k$ by projecting $D(\tilde{X})$ into $\mathcal{M}(\Delta_k) \cap \mathcal{N}$. This projection is accomplished by first projecting $D(\tilde{X})$ into $\mathcal{M}(\Delta_k)$, then (if necessary) rescaling the solution. Projection into order constraints is called isotonic regression, for which various algorithms exist. For the case of a complete ordering, Grotzinger and Witzgall [4] describe an algorithm that is linear in the number of variables, hence $O(N^2)$ in our application. Algorithms for general partial orderings are quadratic in the number of variables, hence $O(N^4)$ in our application. If $\mathcal{M}(\Delta_k)$ only defines a partial ordering (because there are ties in the entries of Δ_k), then we circumvent the expense of partial ordering by imposing a complete ordering. If $\delta_{ij} = \delta_{kl}$ and $d_{ij}(\tilde{X}) < d_{kl}(\tilde{X})$, then we impose $\tilde{\delta}_{ij} \leq \tilde{\delta}_{kl}$. In the unlikely event that $\delta_{ij} = \delta_{kl}$ and $d_{ij}(\tilde{X}) = d_{kl}(\tilde{X})$, then we randomly impose either $\tilde{\delta}_{ij} \leq \tilde{\delta}_{kl}$ or $\tilde{\delta}_{ij} \geq \tilde{\delta}_{kl}$. Admittedly, our approach is *ad hoc*. If we impose additional order constraints, we may not actually compute the projection of $D(\tilde{X})$ into $\mathcal{M}(\Delta_k)$. Furthermore, how we modify $\mathcal{M}(\Delta_k)$ may vary from iteration to iteration of Algorithm 4.1. Nevertheless, our approach appears to work well in practice.

ALGORITHM 4.3. Update $\tilde{\Delta}_k$

1. Compute $D(\tilde{X})$.

2. Update $\tilde{\Delta}_k$ by projecting $D(\tilde{X})$ into $\mathcal{M}(\Delta_k)$, suitably modified to define a complete ordering, using the algorithm described in [4].
3. If $\|\tilde{\Delta}_k\|_F^2 < 1$, then set $\tilde{\Delta}_k = \tilde{\Delta}_k / \|\tilde{\Delta}_k\|$.

The primary computational burden of 3WNM lies in updating $\tilde{\Delta}_1$ and $\tilde{\Delta}_2$. Note that these updates can be performed in parallel. Furthermore, as described in [6], isotonic regression can itself be parallelized efficiently.

4.2 Out-of-sample Embedding Trosset and Priebe [17] proposed an algorithm for out-of-sample embedding with respect to the error criterion used in classical MDS. They also observed that out-of-sample embedding has a natural formulation with respect to the raw stress criterion. Ma and Priebe [10] derived an iterative majorization algorithm for this case. Here we describe an algorithm for out-of-sample embedding in which the raw stress criterion is minimized subject to order constraints on a set of surrogate dissimilarities.

The algorithm for embedding a test pair resembles the algorithm for embedding the set of training pairs. Again we use variable alternation and terminate the loop by imposing a maximum number of iterations and/or by monitoring how much an iteration decreases the value of the objective function. The following algorithm is repeated for $k = 1, 2$. Because the greatest expense lies in computing the distances between \tilde{y}_k and each \tilde{x}_{ik} , computation is $O(N)$.

ALGORITHM 4.4. Out-of-sample Embedding

0. Compute $\bar{\delta}_k = (\delta_k(y_k, x_{1k}), \dots, \delta_k(y_k, x_{Nk}))^t$.
1. Initialize \tilde{y}_k , e.g., by placing \tilde{y}_k at the origin or at the \tilde{x}_{ik} for which $\bar{\delta}_k(y_k, x_{1k})$ is smallest.
2. Compute $\sigma_{n-o}^{[0]}$ and set $\ell = 0$.
3. Do until termination:
 - (a) Increment ℓ .
 - (b) Fix $\tilde{\delta}_k$ and minimize σ_{n-o} with respect to \tilde{y}_k . This step uses Algorithm 4.5.
 - (c) Fix \tilde{y}_k and minimize σ_{n-o} with respect to $\tilde{\delta}_k$. This step uses Algorithm 4.6.
 - (d) Compute $\sigma_{n-o}^{[\ell]}$.

Algorithm 4.5 is a variant of Algorithm 4.2. Instead of varying every point in the configuration, N points are fixed and one point is varied. Let $b(\tilde{y}_k)$ have entries $\tilde{\delta}_{k,i}/d(\tilde{y}_k, \tilde{x}_{ik})$, where $\tilde{\delta}_{k,i}$ is entry i in $\tilde{\delta}_k$, and let

$$\gamma(\tilde{y}_k) = \frac{1}{N} \sum_{i=1}^N [(1 - b_i) \tilde{x}_i + b_i \tilde{y}_k].$$

Then out-of-sample embedding is performed as follows:

ALGORITHM 4.5. Update \tilde{y}_k

1. Compute $\sigma_{n-o}^{[0]}$ and set $\ell' = 0$.
2. Do until termination:
 - (a) Increment ℓ' .
 - (b) Compute $d(\tilde{y}_k, \tilde{x}_1), \dots, d(\tilde{y}_k, \tilde{x}_N)$.
 - (c) $\tilde{y}_k \leftarrow \gamma(\tilde{y}_k)$
 - (d) Compute $\sigma_{n-o}^{[\ell']}$.

Algorithm 4.6 is a variant of Algorithm 4.3:

ALGORITHM 4.6. Update $\tilde{\delta}_k$

1. Compute $d((\tilde{X}) = (d(\tilde{y}_k, \tilde{x}_1), \dots, d(\tilde{y}_k, \tilde{x}_N))^t$.
2. Project $d(\tilde{X})$ into the rectangle defined by (3.3) and denote the projection $\tilde{\delta}_k$.

5 Illustrative Example

We illustrate 3WNM with a small version of an experiment performed in [14]. The original $N = 20$ objects are drawn from a multivariate normal distribution on $\Omega = \mathbb{R}^3$,

$$\omega_1, \omega_2, \dots, \omega_N \sim \text{Normal}(0, I_3).$$

The feature spaces are $\Xi_1 = \Xi_2 = \mathbb{R}^6$ and the observed feature vectors are

$$x_{ik} = \pi_k(\omega_i) = ((1 - a)s_{ik}^t, a\epsilon_{ik}^t)^t,$$

where

$$s_{ik} \sim \text{Normal}(\omega_i, I_3/30) \quad \text{and} \quad \epsilon_{ik} \sim \text{Normal}(0, I_3).$$

The s_{ik} represent observations of the original ω_{ik} that have been corrupted by measurement error, while the ϵ_{ik} represent pure noise in extraneous dimensions. The constant a controls the relative magnitudes of the s_{ik} and the ϵ_{ik} , hence the extent to which the measured dissimilarities in \mathbb{R}^3 will be affected by noise in the extraneous dimensions. We used $a = 0.4$, resulting in a fairly challenging inferential task.

The dissimilarity measure δ_k was Euclidean distance in Ξ_k . We formed Δ_1 and Δ_2 , then used 3WNM to embed the (x_{i1}, x_{i2}) in \mathbb{R}^p . The problem of choosing p is a problem of model selection; we chose $p = 2$ to facilitate visualization. The resulting points, $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{20}$, are displayed in Figure 4.

Figures 5 and 6 each display the \tilde{x}_i and three additional pairs (identified by plotting symbols 1, 2, and

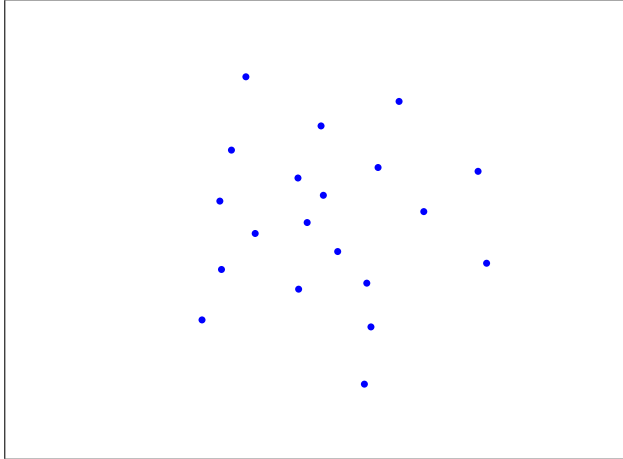


Figure 4: Illustrative embedding of $N = 20$ matched pairs in \mathbb{R}^2 by three-way nonmetric MDS.

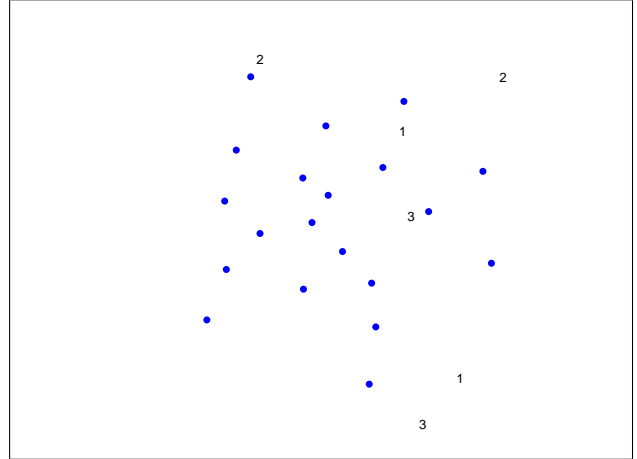


Figure 6: Three mismatched pairs have been added to the configuration in Figure 4.

3) embedded out-of-sample. These pairs are matched in the case of Figure 5, mismatched in the case of Figure 6. The intrapair distances of the matched pairs are much smaller than the intrapair distances of the mismatched pairs. This observation is the basis for our approach to matched pair hypothesis testing: we reject $H_0 : y_1 \sim y_2$ if and only if $d(\tilde{y}_1, \tilde{y}_2)$ is sufficiently large.

null hypothesis of matched pairs.

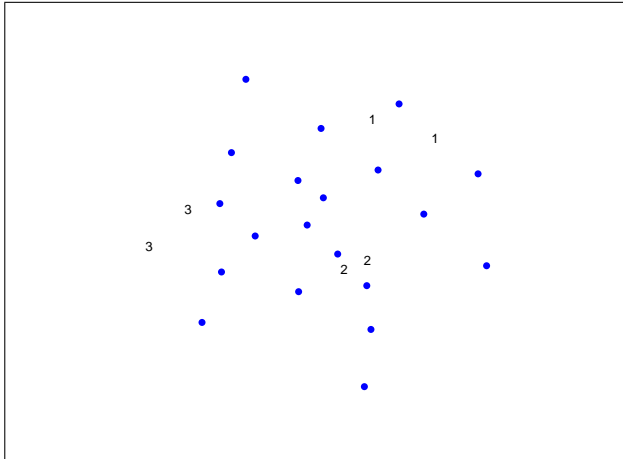


Figure 5: Three matched pairs have been added to the configuration in Figure 4.

We generated 1000 matched and 1000 mismatched (y_1, y_2) and computed $d(\tilde{y}_1, \tilde{y}_2) = \|\tilde{y}_1 - \tilde{y}_2\|$ for each pair. Figure 7 displays the empirical cumulative distribution functions (CDFs) of $d(\tilde{y}_1, \tilde{y}_2)$ and Figure 8 displays corresponding kernel density estimates. Evidently, $d(\tilde{y}_1, \tilde{y}_2)$ is stochastically larger under the alternative hypothesis of mismatched pairs than under the

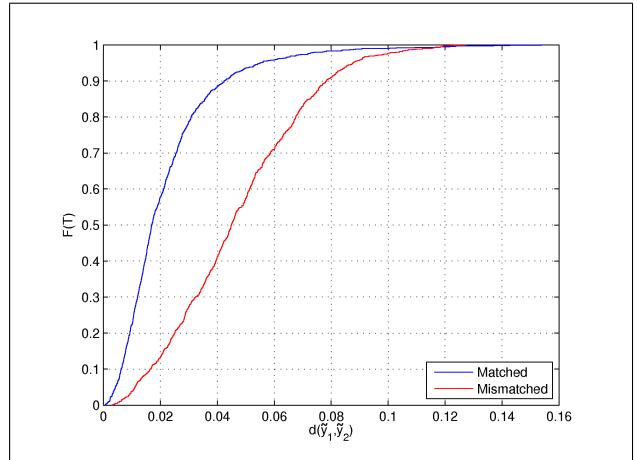


Figure 7: Empirical CDF of the test statistic under the matched and mismatched conditions.

6 Numerical Experiments

We performed three numerical experiments designed to explore the robustness of 3WNM to changes in δ_1 and δ_2 . Our experiments extend experiments reported in [14], in each case comparing the performance of 3WNM to JOFC. Sections 6.1 and 6.2 describe simulation experiments in which $\Xi_1 = \Xi_2$. Section 6.3 describes an example of disparate feature representations.

6.1 Gaussian Simulation Here we extend the illustrative example. We chose δ_1 to measure Euclidean distance and considered three choices of δ_2 : Euclidean

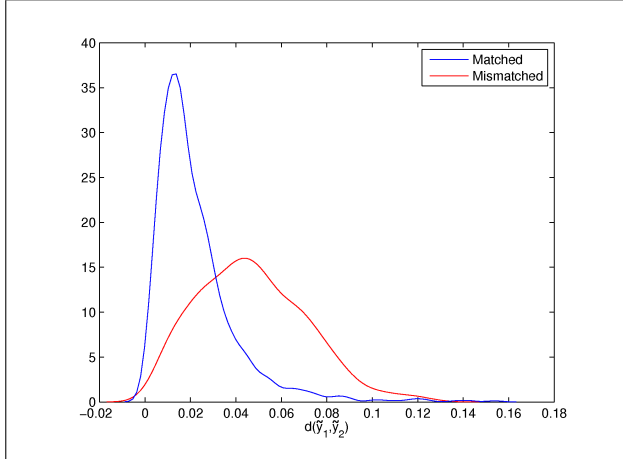


Figure 8: Kernel density estimates of the test statistic under the matched and mismatched conditions.

distance, squared Euclidean distance, and “city block” (L^1) distance. For each choice of δ_2 we performed 100 replications of the following experiment.

First, using the probability model described in Section 5, we generated $N = 20$ training pairs, (x_{i1}, x_{i2}) , and 200 test pairs, (y_{i1}, y_{i2}) . Of the 200 test pairs, 100 were matched and 100 were mismatched. Second, we constructed representations of the data in \mathbb{R}^2 using the 3WNM and JOFC procedures. Third, for each representation we estimated the tradeoff between significance level (α) and power (β) by (a) using the empirical distribution of $d(\tilde{y}_1, \tilde{y}_2)$ for the matched pairs to determine a critical value c_α for each achievable α , then (b) estimating the power of the level- α test by the proportion of mismatched pairs for which $d(\tilde{y}_1, \tilde{y}_2) > c_\alpha$.

Figures 9, 10, and 11 plot (α, β) , averaged over 100 replications, for the three choices of δ_2 . Because 3WNM is nonmetric and squaring is a monotonic transformation, any differences between 3WNM’s (α, β) curves in Figures 9 and 10 is due to chance variation. In contrast, Euclidean and city block distances are not monotonically related, resulting in an intrinsically different (α, β) curve. The 3WNM curves are slightly above (more powerful) the JOFC curves, and slightly more so for $\delta_2 \neq \delta_1$.

6.2 Dirichlet Simulation Priebe et al. [14] used Dirichlet distributions to model the pairing of documents written in different languages. Accordingly, we repeated the experiments described in Section 6.1 using Dirichlet distributions instead of normal distributions and drawing $\omega_i \sim \text{Dirichlet}(1, 1, 1)$, $s_{ik} \sim \text{Dirichlet}(30\omega_i + 1)$, and $\epsilon_{ik} \sim \text{Dirichlet}(1, 1, 1)$. Figures 12, 13, and 14 display (α, β) curves for 3WNM and

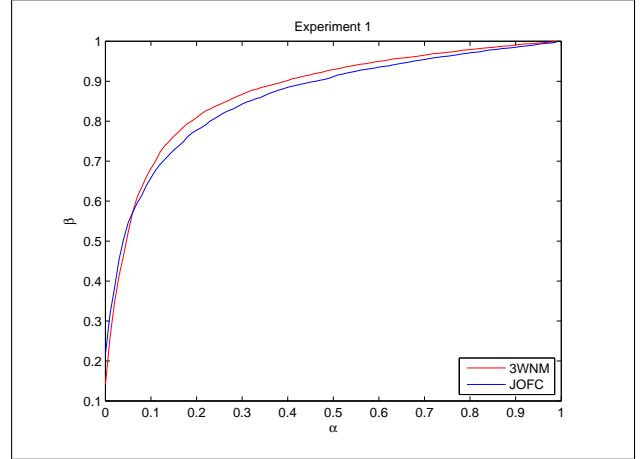


Figure 9: (Gaussian Setting) δ_1 and δ_2 are both Euclidean distance.

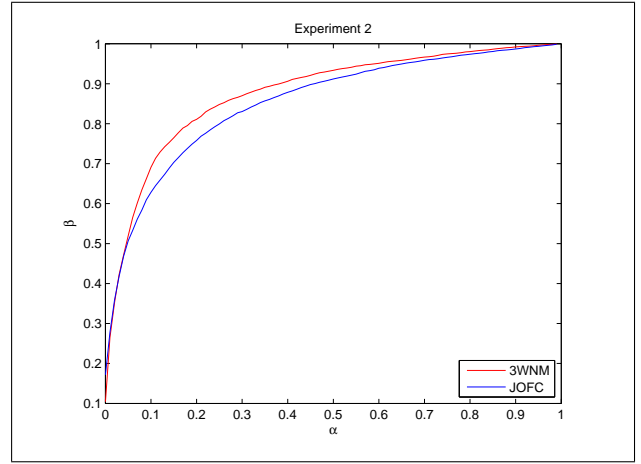


Figure 10: (Gaussian Setting) δ_1 is Euclidean distance and δ_2 is squared Euclidean distance.

JOFC for the same three choices of δ_2 , with patterns comparable to Figures 9, 10, and 11.

6.3 Wikipedia Priebe et al. [14] compared English and French Wikipedia articles related to algebraic topology. Here we compare two disparate representations of the English articles. The objects are the $N = 1382$ articles in the directed two neighborhood of the “Algebraic Topology” node of the English Wikipedia link graph. We represented each article as (1) a node in the link graph of the extracted articles, and (2) a point in a vector space model of article content. The corresponding dissimilarity measures were (1) the shortest path distance in the link graph, and (2) the cosine dissimilarity of the discounted mutual information [14, 9, 13]. Here

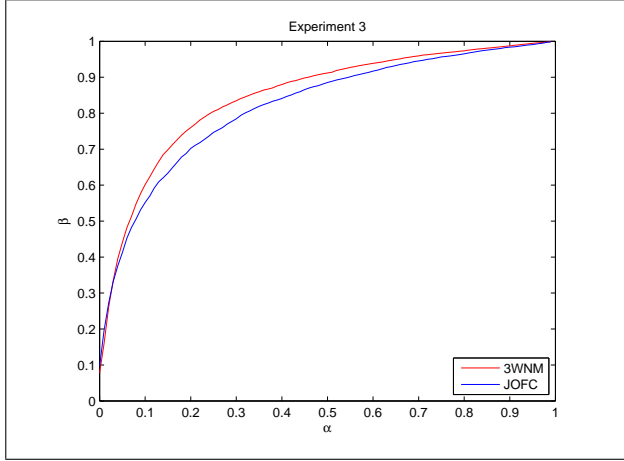


Figure 11: (Gaussian Setting) δ_1 is Euclidean distance and δ_2 is city block distance.

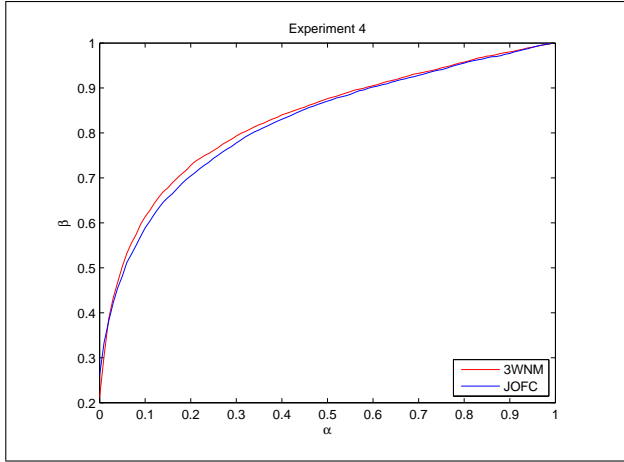


Figure 12: (Dirichlet Setting) δ_1 and δ_2 are both Euclidean distance.

δ_1 and δ_2 are completely disparate measures of dissimilarity, a circumstance for which 3WNM was designed but JOFC was not. Note that the problem of determining whether or not an article's content matches its location in the link graph mimics the problem of identifying spam websites on the world wide web.

Because we lacked a probability model from which to generate additional matched and unmatched pairs, we used five-fold cross-validation to estimate the relationship between α and β . We partitioned the $N = 1382$ articles into F_0, \dots, F_4 , then repeated the following for $i = 0$ to 4:

1. Let F_i denote the training set.
2. Let $F_{(i+1) \bmod 5}$ denote a set of matched test pairs.

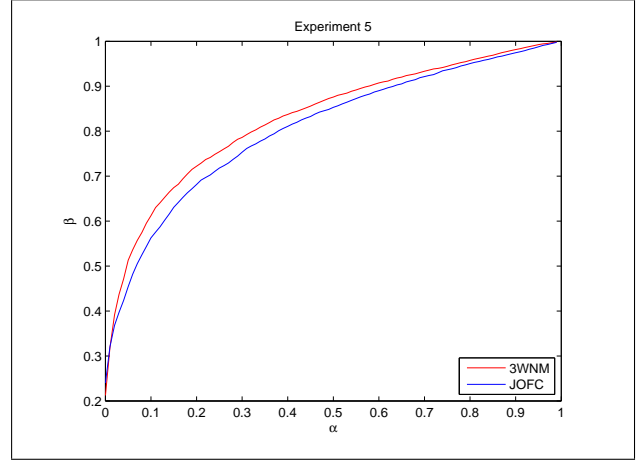


Figure 13: (Dirichlet Setting) δ_1 is Euclidean distance and δ_2 is squared Euclidean distance.

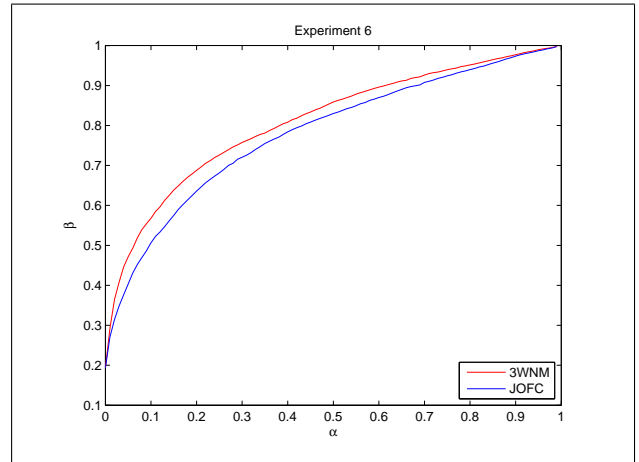


Figure 14: (Dirichlet Setting) δ_1 is Euclidean distance and δ_2 is city block distance.

3. Let respective pairs in $F_{(i+2) \bmod 5}$ and $F_{(i+3) \bmod 5}$ be used as mismatched test pairs.

Both 3WNM and JOFC embedded in \mathbb{R}^{10} . The (α, β) curves were estimated by averaging β over the five testing scenarios.

Figure 15 shows the estimated (α, β) curves for 3WNM and JOFC. The JOFC approach was not designed for disparate measures of dissimilarity and displays almost no power. The 3WNM approach, which ignores metric information and only considers rank information, displays considerably greater power.

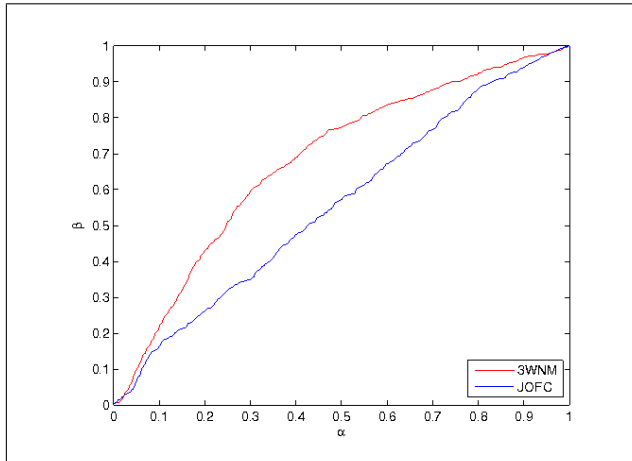


Figure 15: Wikipedia Experiment

7 Discussion

Priebe et al. [14] proposed a novel approach to the problem of matched pair hypothesis testing. We have modified their ideas for use with disparate dissimilarity measures. Nonmetric MDS allows monotonic transformations of the input dissimilarities, thereby using the ranks of the dissimilarities but not their numeric values.

Although we might have used any of several nonmetric embedding methodologies, we adopted a three-way approach because of the relation between the so-called identity model for three-way MDS and the JOFC formulation proposed in [14]. Other formulations may also have merit. Our numerical experiments suggest that using three-way nonmetric MDS approach provides good results both when the dissimilarity measures are similar and when they are disparate.

8 Acknowledgements

This research was supported by the Office of Naval Research, the Air Force Office of Scientific Research, and by a National Security Science & Engineering Faculty Fellowship awarded to the third author.

References

- [1] I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*, Second Edition. Springer, New York, 2005.
- [2] M. Christoudias, R. Urtasun, and T. Darrell. *Bayesian Localized Multiple Kernel Learning*. Technical Report UCB/EECS-2009-96, Department of Electrical Engineering & Computer Sciences, University of California, Berkeley, 2009.
- [3] J. Commandeur and W. Heiser. *Mathematical Derivations in the Proximity Scaling (PROXSCAL) of Sym-*

- metric Data Matrices*. Research Report RR-93-04, Department of Data Theory, Leiden University, 1993.
- [4] S. Grotzinger and C. Witzgall. Projections onto order simplexes. *Applied Mathematics and Optimization*, 12:247–270, 1984.
- [5] P. Jawanpuria and J. Saketha Nath. Multi-task multiple kernel learning. In *Proceedings of the 11th SIAM International Conference on Data Mining*, 1:828–838. SIAM, Philadelphia, PA, 2011.
- [6] A. Kearsley, R. A. Tapia, and M. W. Trosset. An approach to parallelizing isotonic regression. In *Applied Mathematics and Parallel Computation: Festschrift for Klaus Ritter*, edited by H. Fischer, B. Riedmüller, and S. Schäffler, pages 141–147. Physica-Verlag, Heidelberg, 1996.
- [7] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.
- [8] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.
- [9] D. Lin and P. Pantel. Concept discovery from text. In *Proceedings of the 19th International Conference on Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, 2002.
- [10] Z. Ma and C. E. Priebe. *Out-of-sample Embedding using Iterative Majorization*. Unpublished manuscript, 2010.
- [11] B. McFee and G. Lanckriet. Learning multi-modal similarity. *Journal of Machine Learning Research*, 12:491–523, 2011.
- [12] E. Pekalska and R. P. W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations And Applications*. Series in Machine Perception and Artificial Intelligence, Volume 64. World Scientific Publishing, River Edge, NJ, 2005.
- [13] P. Pantel and D. Lin. Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619. ACM, New York, 2002.
- [14] C. E. Priebe, D. J. Marchette, Z. Ma, and S. Adali. *Manifold Matching: Joint Optimization of Fidelity and Commensurability*. Submitted for publication, 2010.
- [15] S. Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.
- [16] M. W. Trosset. A new formulation of the nonmetric strain problem in multidimensional scaling. *Journal of Classification*, 15:15–35, 1998.
- [17] M. W. Trosset and C. E. Priebe. The out-of-sample problem for classical multidimensional scaling. *Computational Statistics and Data Analysis*, 52:4635–4642, 2008.
- [18] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Proceedings of the 11th IEEE International Conference on Computer Vision*. IEEE, New York, 2007.

- [19] H. Xia and S. C. H. Hoi. MKBoost: A framework of multiple kernel boosting. In *Proceedings of the 11th SIAM International Conference on Data Mining*, 1:199–210. SIAM, Philadelphia, PA, 2011.