Fidelity-Commensurability Tradeoff in Joint Embedding of Disparate Dissimilarities

Sancar Adali[*] Carey E. Priebe[†]

May 13, 2013

Abstract

1 Introduction

We are interested in problems where the data sources are disparate and the inference task requires that the observations from the different data sources can be judged to be similar or dissimilar.

Consider a collection of English Wikipedia articles and French articles on the same topics. A pair of documents in different languages on the same topic are said to be "matched". The "matched" wiki documents are not necessarily direct translations of each other, so we do not restrict "matchedness" to be a well-defined bijection between documents in different languages. However the matched "documents" provide examples of "similar" observations coming from disparate sources, and we assume the training data consist of a collection of "matched" documents.

The inference task we consider is match detection, i.e. deciding whether a new English article and a new French article are on the same topic or not. While a document in one language, say English, can be compared with other documents in English, a French document cannot be represented using the same features, therefore cannot be directly compared to English documents. It is necessary to derive a data representation where the documents from different languages can be compared (are commensurate). We will use a finite-dimensional Euclidean space for this commensurate representation, where standard statistical inference tools can be used.

"Disparate data" means that the observations are from different "conditions", for example, the data might come from different type of sensors. Formally, the original data reside in a heteregenous collection of spaces. In addition, the data might be structured and/or might reside in infinite dimensional spaces. Therefore, it is possible that a feature representation of the data is not available or inference with such a representation is fraught with complications (e.g. feature selection, non-i.i.d. data, infinite-dimensional spaces). This motivates our dissimilarity-centric approach. For

[*] Johns Hopkins University, Department of Applied Mathematics and Statistics, 100 Whitehead Hall, 3400 North Charles Street, Baltimore, MD 21218-2682

[†] Johns Hopkins University, Department of Applied Mathematics and Statistics, 100 Whitehead Hall, 3400 North Charles Street, Baltimore, MD 21218-2682

an excellent resource on the usage of dissimilarities in pattern recognition, we refer the reader to the Pękalska and Duin book [5].

Since we proceed to inference starting from a dissimilarity representation of the data, our methodology may be applicable to any scenario in which multiple dissimilarity measures are available. Some illustrative examples include: pairs of images and their descriptive captions, textual content and hyperlink graph structure of Wikipedia articles, photographs taken under different illumination conditions. In each case, we have an intuitive notion of "matchedness": for photographs taken under different illumination conditions, "matched" means they are of the same person. For a collection of linked Wikipedia articles, the different conditions are the textual content and hyperlink graph structure, "matched" means a text document and a vertex in the graph corresponds to the same Wikipedia article.

To quantify how suitable the commensurate representation is for subsequent inference, two error criteria can be defined: *fidelity*, which refers to how well the available dissimilarities in a condition are preserved and *commensurability*, which refers to how well the dissimilarities between "matched" objects are preserved. These two concepts will be made concrete in a later section 4.

The major question addressed in this paper is whether, in the tradeoff between fidelity and commensurability, there is a "sweet spot": increases in fidelity (or commensurability) do not result in superior performance for the inference task, due to the resulting commensurability (or fidelity) loss.

2 Related Work

There have many efforts toward solving the related problem of "manifold alignment". "Manifold alignment" seeks to find correspondences between disparate datasets in different conditions (which are sometimes referred as "domains") by aligning their underlying manifolds. The setting that is common in the literature is the semi-supervised setting [3], where correspondences between two collections of points are given and the task is to find correspondences between a new set of points in each condition. In contrast, the hypothesis testing task discussed in this paper is to determine whether any given pair of points is "matched" or not. The proposed solutions [2,10,11] follow a common approach: they look for a common commensurate latent space such that the representations (either projections or embeddings) of the observations in this space match.

Wang and Mahedavan [10] suggest an approach that uses embedding followed by Procrustes Analysis to find maps from the embedding spaces to a commensurate space. Given a paired set of points, Procrustes Analysis [8] finds a linear transformation from one set of points to the other that minimizes sum of squared distances between pairs. In the problem considered in [10], the paired set of points are low-dimensional embeddings of kernel matrices. For the embedding step, they chose to use Laplacian Eigenmaps, though their algorithm allows for any appropriate embedding method.

Zhai et al. [11] solves an optimization problem with respect to two projection matrices for the observations in two domains. The energy function that is optimized contains three terms: two *manifold regularization terms* and one *correspondence preserving term*. The *manifold regularization terms* ensure that the local neighborhood of points are preserved in the low-dimensional space, by making use of the reconstruction error for Locally Linear Embedding [7]. The *correspondence preserving term* ensures that "matched" points are mapped to proximate locations in the commensurate space.

Ham and Lee [3] solve the problem in the semi-supervised setting by a similar approach, by optimizing a energy function that has three terms that are analogous to the terms in [11].

3 Problem Description

In the problem setting considered here, n different objects are measured under K different conditions (corresponding to, for example, K different sensors). We assume we begin with dissimilarity measures. These will be represented in matrix form as K $n \times n$ matrices $\{\Delta_k, k = 1, \ldots, K\}$. In addition, for each condition, dissimilarities between a new object and the previous n objects $\{\mathcal{D}_k, k = 1, \ldots, K\}$ are available. Under the null hypothesis, these new dissimilarities represent a *single* new object measured under K different conditions. Under the alternative hypothesis, the dissimilarities $\{\mathcal{D}_k\}$ represent *separate* new objects measured under K different conditions [6].

For the English-French Wikipedia article example in the introduction, the dissimilarities between articles in the same language ($\{\Delta_k\}$) are available. The dissimilarities between the new English article and the other n English articles ($\mathcal{D}_1$) are also available, as well as the dissimilarities between the new French article and the other n French articles ($\mathcal{D}_2$). The null hypothesis is that the new English and French articles are on the same topic, while the alternative hypothesis is that they are on different topics.

In order to derive a data representation where dissimilarities from disparate sources ($\{\mathcal{D}_k\}$) can be compared, the dissimilarities must be embedded in a commensurate metric space where the metric can be used to distinguish between matched and unmatched observations.

To embed multiple dissimilarities $\{\Delta_k\}$ into a commensurate space, an omnibus dissimilarity matrix $M \in \mathbb{R}^{nk \times nk}$ is constructed. Consider, for $K = 2$,

$$M = \begin{bmatrix} \Delta_1 & L \\ L^T & \Delta_2 \end{bmatrix} \tag{1}$$

where L is a matrix of imputed entries to be described later.

Remark For clarity of exposition,we will consider $K = 2$; the generalization to $K > 2$ is straightforward.

We define the commensurate space to be $\mathbb{R}^d$, where the embedding dimension d is pre-specified. The selection of d – model selection – is a task that requires much attention and is beyond the scope of this article. Investigation of the effect of d on testing performance will be pursued in a subsequent paper.

We use multidimensional scaling (MDS) [1] to embed the omnibus matrix in this space, and obtain a configuration of $2n$ embedded points $\{\hat{x}_{ik}; i = 1, \ldots, n; k = 1, 2\}$ (which can be represented as $\hat{X}$, a $2n \times d$ matrix, where each row of the configuration matrix is the coordinate vector of an embedded point). The discrepancy between the interpoint distances of $\{\hat{x}_{ik}\}$ and the given dissimilarities in M is made as small as possible, as measured by an objective function $\sigma(\widetilde{X}; M)$ which will be described later. In matrix form,

$$\hat{X} = \arg\min_{\widetilde{X}} \sigma(\widetilde{X}; M).$$

Remark We will use x_{ik} to denote the (possibly notional) observation for the i^{th} object in the k^{th} condition, $\tilde{x}_{ik}$ to denote an argument of the objective function and $\hat{x}_{ik}$ to denote the arg min of the objective function. The notation for matrices $(X, \widetilde{X}, \hat{X})$ follows the same convention.

Given the omnibus matrix M and the $2n \times d$ embedding configuration matrix $\hat{X}$ in the commensurate space, the out-of-sample extension [9] for MDS will be used to embed the test dissimilarities $\mathcal{D}_1$ and $\mathcal{D}_2$. Once the test similarities are embedded as two points $(\hat{y}_1, \hat{y}_2)$ in the commensurate space, it is possible to compute the test statistic

$$\tau = d\left(\hat{y}_1, \hat{y}_2\right)$$

for the two "objects" represented by $\mathcal{D}_1$ and $\mathcal{D}_2$. For large values of τ, the null hypothesis will be rejected. If dissimilarities between matched objects are smaller than dissimilarities between unmatched objects with large probability, and the embeddings preserve this stochastic ordering, we could reasonably expect the test statistic to yield large power.

4 Fidelity and Commensurability

Regardless of the inference task, to expect reasonable performance from the embedded data in the commensurate space, it is necessary to pay heed to these two error criteria:

- Fidelity describes how well the mapping to commensurate space preserves the original dissimilarities. The *loss of fidelity* can be measured with the within-condition *fidelity error*, given by

  $$\epsilon_{f_{(k)}} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} (d(\widetilde{\boldsymbol{x}}_{ik}, \widetilde{\boldsymbol{x}}_{jk}) - \delta_{ijkk})^2.$$

  Here δ_{ijkk} is the dissimilarity between the i^{th} object and the j^{th} object where both objects are in the k^{th} condition, and $\widetilde{\boldsymbol{x}}_{ik}$ is the embedded representation of the i^{th} object for the k^{th} condition; $d(\cdot, \cdot)$ is the Euclidean distance function.

- Commensurability describes how well the mapping to commensurate space preserves matchedness of matched observations. The *loss of commensurability* can be measured by the between-condition *commensurability error* which is given by

  $$\epsilon_{c_{(k_1,k_2)}} = \frac{1}{n} \sum_{1 \leq i \leq n; k_1 < k_2} (d(\widetilde{\boldsymbol{x}}_{ik_1}, \widetilde{\boldsymbol{x}}_{ik_2}) - \delta_{iik_1k_2})^2$$

  for conditions k_1 and k_2; $\delta_{iik_1k_2}$ is the dissimilarity between the i^{th} object under conditions k_1 and k_2. Although the between-condition dissimilarities of the same object, $\delta_{iik_1k_2}$, are not available, it is reasonable to set these dissimilarities to 0 for all i, k_1, k_2. These dissimilarities correspond to diagonal entries of the submatrix L in the omnibus matrix M in equation (1). Setting these diagonal entries to 0 forces matched observations to be embedded close to each other.

While the above expressions for *fidelity* and *commensurability* errors are specific to the joint embedding of disparate dissimilarities, the concepts of fidelity and commensurability are general enough to be applicable to other dimensionality reduction methods for data from disparate sources.

In addition to fidelity and commensurability, there is the *separability* criteria: dissimilarities between unmatched observations in different conditions should be preserved (so that unmatched pairs are not embedded close together). The error for this criteria can be measured by $\epsilon_{s_{k_1k_2}} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n; k_1 < k_2} (d(\widetilde{\boldsymbol{x}}_{ik_1}, \widetilde{\boldsymbol{x}}_{jk_2}) - \delta_{ijk_1k_2})^2$ for conditions k_1 and k_2.

Let us now show how fidelity and commensurability errors can be made explicit in the objective function. Consider the weighted raw stress criterion ($\sigma_W(\cdot)$) which we choose as the objective function for the embedding of M. The entries of M are $\delta_{ijk_1k_2}$ for the available dissimilarities. As the between-condition dissimilarities, $\delta_{ijk_1k_2}$ for $i \neq j$, are not available in general, the entries corresponding to the unavailable dissimilarities can be imputed as $\delta_{ijk_1k_2} = \frac{\delta_{ijk_1k_1} + \delta_{ijk_2k_2}}{2}$.

$$\sigma_W(\widetilde{X}; M) = \sum_{i \leq j, k_1 \leq k_2} w_{ijk_1k_2} (D_{ijk_1k_2}(\widetilde{X}) - \delta_{ijk_1k_2})^2. \tag{2}$$

Here, ijk_1k_2 subscript of a partitioned matrix refers to the entry in the i^{th} row and j^{th} column of the sub-matrix in k_1^{th} row partition and k_2^{th} column partition, W is the weight matrix, $\widetilde{X}$ is the configuration matrix that is the argument of the stress function, D is the Euclidean distance function of the rows of its matrix argument. *Each of the individual terms in the sum (2) can be ascribed to fidelity, commensurability or separability.*

$$\begin{aligned} \sigma_W(\cdot; M) &= \sum_{i,j,k_1,k_2} \underbrace{w_{ijk_1k_2}(D_{ijk_1k_2}(\cdot) - M_{ijk_1k_2})^2}_{term_{i,j,k_1,k_2}} \\ &= \underbrace{\sum_{i=j,k_1<k_2} term_{i,j,k_1,k_2}}_{Commensurability} + \underbrace{\sum_{i<j,k_1=k_2} term_{i,j,k_1,k_2}}_{Fidelity} + \underbrace{\sum_{i<j,k_1<k_2} term_{i,j,k_1,k_2}}_{Separability} \quad . \end{aligned} \tag{3}$$

Due to the fact that data sources are "disparate", it is not obvious how a dissimilarity between an object in one condition and another object in another condition can be computed or defined in a sensible way. Although these unavailable dissimilarities can be imputed as mentioned, they can also be set to any finite number and ignored in the embedding by setting the associated weights in the raw stress function to be 0 for the weighted raw stress criterion. We choose to do the latter to restrict our attention to the fidelity-commensurability tradeoff.

As mentioned in description of commensurability, we set the between-condition dissimilarities of the same object ($\{M_{iik_1k_2}\}$) to 0. Then the raw stress function can be written as

$$\sigma_W(\widetilde{X};M) \;=\; \underbrace{\sum_{i=j,k_1<k_2} w_{ijk_1k_2}(D_{ijk_1k_2}(\widetilde{X}))^2}_{Commensurability} \;+\; \underbrace{\sum_{i<j,k_1=k_2} w_{ijk_1k_2}(D_{ijk_1k_2}(\widetilde{X})-M_{ijk_1k_2})^2}_{Fidelity} \quad .$$

This motivates the naming of the omnibus embedding approach as Joint Optimization of Fidelity and Commensurability (JOFC).

The weights in the raw stress function allow us to address the question of the "sweet spot" in Fidelity-Commensurability tradeoff. Let $w \in (0,1)$. Setting the weights $(w_{ijk_1k_2})$ for the commensurability and fidelity terms to w and $1-w$, respectively, will allow us to control the relative importance of fidelity and commensurability terms in the objective function.

Let us denote the raw stress function with these simple weights by $\sigma_w(\widetilde{X};M)$. With simple weighting, when $w = 0.5$, all terms in the objective function have the same weights. We will refer to this weighting scheme as *uniform weighting.* Uniform weighting does not necessarily yield the best fidelity-commensurability tradeoff in terms of subsequent inference.

Previous investigations of the JOFC approach [6] did not consider the effect of non-uniform weighting. Our thesis is that using non-uniform weighting in the objective function will allow for superior performance. That is, for a given exploitation task there is an optimal w, denoted w^*, and in general $w^* \neq 0.5$. In particular, we consider hypothesis testing, as in [6], and we let the area under the ROC curve, $AUC(w)$, be our measure of performance for any $w \in [0,1]$. In this case, we show that $AUC(w)$ is continuous, and hence $w^* = \arg\max_{w\in[0,1]} AUC(w)$ exists. We demonstrate the potential practical advantage of our weighted generalization of JOFC via simulations.

5 Definition of w^*

Let us denote the test dissimilarities $(\mathcal{D}_1,\mathcal{D}_2)$ by $(\mathcal{D}_1^{(m)},\mathcal{D}_2^{(m)})$ under "matchedness" hypothesis, and by $(\mathcal{D}_1^{(u)},\mathcal{D}_2^{(u)})$ under the alternative. The out-of-sample embedding of $(\mathcal{D}_1^{(m)},\mathcal{D}_2^{(m)})$ involves the augmentation of the omnibus matrix M, which consists of n "matched" pairs of dissimilarities, with $(\mathcal{D}_1^{(m)},\mathcal{D}_2^{(m)})$. The resulting augmented $(2n+2)\times(2n+2)$ matrix has the form:

$$\Delta^{(m)} = \left[\begin{array}{cccc} \multicolumn{2}{c}{\multirow{2}{*}{M}} & \mathcal{D}_1^{(m)} & \vec{\mathcal{D}}_{NA} \\ & & \vec{\mathcal{D}}_{NA} & \mathcal{D}_2^{(m)} \\ \mathcal{D}_1^{(m)T} & \vec{\mathcal{D}}_{NA}^T & 0 & \mathcal{D}_{NA} \\ \vec{\mathcal{D}}_{NA}^T & \mathcal{D}_2^{(m)T} & \mathcal{D}_{NA} & 0 \end{array}\right]. \tag{4}$$

where the scalar $\mathcal{D}_{NA}$ and the vector of length n $\vec{\mathcal{D}}_{NA}$ represent dissimilarities that are not available. In our JOFC procedure, these unavailable entries in $\Delta^{(m)}$ are either ignored in the embedding optimization or imputed using other dissimilarities that are available. For a simpler notation, let us assume it is the latter case. Also note that $\Delta^{(u)}$ has the same form as $\Delta^{(m)}$ where $\mathcal{D}_k^{(m)}$ is replaced by $\mathcal{D}_k^{(u)}$.

We define the dissimilarity matrices $\{\Delta^{(m)},\Delta^{(u)}\}$ to be two matrix-valued random variables : $\Delta^{(m)} : \Omega \rightarrow \mathbf{M}_{(2n+2)\times(2n+2)}$ and $\Delta^{(u)} : \Omega \rightarrow \mathbf{M}_{(2n+2)\times(2n+2)})$ for the appropriate sample space (Ω).

Remark Suppose the objects in k^{th} condition can be represented as points in a measurable space Ξ_k, and the dissimilarities in k^{th} condition are given by a dissimilarity measure δ_k acting on pairs of points in Ξ_k. Assume $\mathcal{P}_{(m)}$ is the joint probability distribution over "matched" objects, while the joint distribution of "unmatched" objects $\{k = 1, \ldots, K\}$ is $\mathcal{P}_{(u)}$. Assuming the data are i.i.d., under the two hypotheses ("matchedness" and "unmatchedness", respectively), the $n+1$ pairs of objects are governed by the product distributions $\{\mathcal{P}_{(m)}\}^n \times \mathcal{P}_{(m)}$ and $\{\mathcal{P}_{(m)}\}^n \times \mathcal{P}_{(u)}$. The distributions of $\Delta^{(m)}$ and $\Delta^{(u)}$ are the induced probability distributions of these product distributions (induced by the dissimilarity measure δ_k applied to objects in k^{th} condition $\{k = 1, \ldots, K\}$) .

The criterion function for the embedding is $\sigma_W(\widetilde{X}; M)$ which can be written as $f_w(\widetilde{X}, \Delta)$ for the simple weighting scheme with w, and an omnibus dissimilarity matrix Δ. The embedding coordinates for the unmatched pair are $\hat{y}_1^{(u)}, \hat{y}_2^{(u)}$ where

$$\hat{y}_1^{(u)}, \hat{y}_2^{(u)} = \arg\min_{\widetilde{y}_1^{(u)}, \widetilde{y}_2^{(u)}} \left[\min_{\widetilde{\mathcal{T}}} f_w \left(\begin{bmatrix} \widetilde{\mathcal{T}} \\ \widetilde{y}_1^{(u)} \\ \widetilde{y}_2^{(u)} \end{bmatrix}, \Delta^{(u)} \right) \right].$$

A similar expression gives the embedding for the matched pair.

Remark Note that the in-sample embedding of $\widetilde{\mathcal{T}}$ is necessary but irrelevant for the inference task, hence the minimization with respect to $\widetilde{\mathcal{T}}$ is denoted by min instead arg min.

Remark Note also that all of the random variables following the embedding, such as $\hat{y}_1^{(u)}$, are dependent on w; for the sake of simplicity, this will not be shown in the notation.

Assuming $\Delta^{(m)} \to \{\hat{y}_1^{(m)}, \hat{y}_2^{(m)}\}$ and $\Delta^{(u)} \to \{\hat{y}_1^{(u)}, \hat{y}_2^{(u)}\}$ are measurable maps, consider the test statistic τ which equals $d(\hat{y}_1^{(m)}, \hat{y}_2^{(m)})$ under null hypothesis of matchedness and $d(\hat{y}_1^{(u)}, \hat{y}_2^{(u)})$ under alternative. Under the null hypothesis, the distribution of the statistic is governed by the distribution of $\hat{y}_1^{(m)}$ and $\hat{y}_2^{(m)}$, under the alternative it is governed by the distribution of $\hat{y}_1^{(u)}$ and $\hat{y}_2^{(u)}$.

Then, $\beta(w, \alpha) = 1 - F_{d\left(\hat{y}_1^{(u)}, \hat{y}_2^{(u)}\right)}(F^{-1}_{d\left(\hat{y}_1^{(m)}, \hat{y}_2^{(m)}\right)}(1-\alpha))$ where F_Y denotes the cumulative distribution function of Y. The AUC function is defined as:

$$AUC(w) = \int_0^1 \beta(w, \alpha) \, \mathrm{d}\alpha \, .$$

Although we might care about optimal w with respect to $\beta(w, \alpha)$ (with a fixed type I error rate α), it will be more convenient to define w^* in terms of the AUC function.

Finally, define

$$w^* = \arg\max_w AUC(w).$$

Some important questions about w^* are related to the nature of the AUC function. While finding an analytical expression for the value of w^* is intractable, an estimate $\hat{w}^*$ based on estimates of $AUC(w)$ can be computed. For the Gaussian setting described in 6.1 , a Monte Carlo simulation is run in Section 6 to find the estimate of $AUC(w)$ for different w values.

5.1 Continuity of $AUC(\cdot)$

Let $T_0(w) = d(\hat{y}_1^{(m)}, \hat{y}_2^{(m)})$ and $T_a(w) = d(\hat{y}_1^{(u)}, \hat{y}_2^{(u)})$ denote the value of the test statistic under null and alternative distributions for the embedding with the simple weighting w. The area under the curve measure can be written as:

$$AUC(w) = P\left[T_a(w) > T_0(w)\right]$$

where $T_a(\cdot)$ and $T_0(\cdot)$ can be regarded as stochastic processes whose sample paths are functions of w. We prove $AUC(w)$ is continous with respect to w. We start with this lemma from [**?**].

Lemma 1. *Let z be a random variable. The functional $g(z;\gamma) = P\left[z \geq \gamma\right]$ is upper semi-continuous in probability with respect to z . Furthermore, if $P\left[z = \gamma\right] = 0$, $g(z;\gamma)$ is continuous in probability with respect to z.*

Proof. Suppose z_n converges to z in probability. Then by definition, for any $\delta > 0$ and $\epsilon > 0$, $\exists N$ such that for all $n > N$

$$Pr\left[|z_n - z| \geq \delta\right] \leq \epsilon.$$

The functional $g(z;\gamma)$ is non-increasing and left continuous with respect to γ. Therefore, for $\delta > 0$, $g(z_n;\gamma) - g(z;\gamma) \geq g(z_n;\gamma) - g(z;\gamma - \delta)$ and due to the left-continuity, the difference between the two sides of the inequality can be made as small as desired.

$$\begin{aligned}
g(z_n;\gamma) - g(z;\gamma-\delta) &= Pr\left[z_n \geq \gamma\right] - Pr\left[z \geq \gamma - \delta\right] &(5)\\
&\leq Pr\left[\{z_n \geq \gamma\} \backslash \{z \geq \gamma - \delta\}\right] &(6)\\
&\leq Pr\left[\{\{z_n \geq \gamma\} \backslash \{z \geq \gamma - \delta\}\} \cap \{z_n \geq z\}\right] &(7)\\
&= Pr\left[\{z_n - z \geq \delta\}\right] \leq \epsilon &(8)
\end{aligned}$$

Since ϵ and δ are arbitrary, $\limsup_{n\to\infty}(g(z_n;\gamma) - g(z;\gamma)) = 0$ for any $\delta > 0$, i.e. $g(z;\gamma)$ is upper semi-continuous .

By arguments symmetric to (5)-(8) , we can show that

$$g_{\gamma+\delta}(z) - g(z_n;\gamma) \leq \epsilon \qquad (9)$$

In addition , assume that $P\left[z = \gamma\right] = 0$, $g(z;\gamma)$ is also right-continuous with respect to γ. $g(z_n;\gamma) - g(z;\gamma) \leq g(z_n;\gamma) - g_{\gamma+\delta}(z)$ and the difference can be made as small as possible. Along with 9, this means that

$$\liminf_{n\to\infty}(g(z_n;\gamma) - g(z;\gamma)) = 0.$$

Therefore , $\lim_{n\to\infty} g(z_n;\gamma) = g(z;\gamma)$. □

Theorem 1. *Let $T(w,\omega)$ be continuous with respect to w at w_0 for almost all ω and let $T(w_0,\omega)$ be measurable map $\Omega \to \mathbb{R}$, then $F_\gamma(w) = Pr\left[T(w,\omega) > \gamma\right]$ is upper semi-continous with respect to w at w_0. If, in addition, $Pr\left[T(w,\omega) = \gamma\right] = 0$ where $\omega \in \Omega$, and Ω is the sample space), then $F_\gamma(w)$ is continuous at w_0.*

Theorem 2. *Let $T(\cdot)$ be a stochastic process indexed by w in the interval (0,1). Assume the process is continuous in probability (stochastic continuity) at w_0 i.e.*

$$\forall a > 0 \quad \lim_{\delta \to 0} Pr\left[|T(w+\delta) - T(w)| > a\right] \to 0 \quad (*)$$

$\forall w \in (0,1)$. Furthermore, assume that $Pr\left[T(w) = 0\right] = 0$.
Then, for any $w > 0, \epsilon > 0$, there exists δ_ϵ

$$|Pr\left[T(w+\delta_\epsilon) > 0\right] - Pr\left[T(w) > 0\right]| < \epsilon.$$

and
$Pr\left[T(w) > 0\right]$ is continuous at w_0.

Proof. Assume the sequence w_n converges to w_0 with $w_n \in (0,1)$. Consider $\|Pr\left[T(w_n) > 0\right] - Pr\left[T(w_0) > 0\right]\|$, the difference of the probabilities of the two events $T(w) > 0$ and $T(w_n) > 0$. Both of these two events are the union of two disjoint events.

$$Pr\left[T(w_n) > 0\right] = Pr\left[T(w_n) > 0 \cap T(w) \leq 0\right] + Pr\left[T(w_n) > 0 \cap T(w) > 0\right]$$

$$Pr\left[T(w) > 0\right] = (Pr\left[T(w_n) > 0 \cap T(w) > 0\right] + Pr\left[T(w_n) \leq 0 \cap T(w) > 0\right]$$

Removing the common term from both probabilities,

$$\begin{aligned}
\|Pr\left[T(w_n) > 0\right] - Pr\left[T(w) > 0\right]\| &= \|Pr\left[T(w_n) > 0 \cap T(w) \leq 0\right] \\
&- Pr\left[T(w_n) \leq 0 \cap T(w) > 0\right]\| \\
&= Pr\left[T(w_n) \leq 0 | T(w) > 0\right] Pr\left[T(w) > 0\right] \\
&\leq \max(Pr\left[T(w_n) > 0 | T(w) \leq 0\right] Pr\left[T(w) \leq 0\right], \\
&\quad Pr\left[T(w_n) \leq 0 | T(w) > 0\right] Pr\left[T(w) > 0\right]) \\
&\leq \max(Pr\left[T(w_n) > 0 | T(w) \leq 0\right], \\
&\quad Pr\left[T(w_n) \leq 0 | T(w) > 0\right]) \\
&\leq \max(Pr\left[T(w_n) > 0 | T(w) \leq 0\right], \\
&\quad Pr\left[T(w_n) \leq 0 | T(w) > 0\right])
\end{aligned}$$

For any w, choose δ_ϵ such that such that both $Pr\left[T(w+\delta_\epsilon) > 0 | T(w) \leq 0\right] < \epsilon$ and $Pr\left[T(w+\delta_\epsilon) \leq 0 | T(w) > 0\right] < \epsilon$. Such a value of δ_ϵ exists , since the conditional probabilities $Pr\left[T(w_n) > 0 | T(w) \leq 0\right]$ and $Pr\left[T(w_n) \leq 0 | T(w) > 0\right]$ can be made smaller than ϵ due to stochastic continuity. Therefore

$$\|Pr\left[T(w_n) > 0\right] - Pr\left[T(w) > 0\right]\| \leq \max(Pr\left[T(w_n) > 0 | T(w) \leq 0\right], Pr\left[T(w_n) \leq 0 | T(w) > 0\right]) < \epsilon$$

Therefore $Pr\left[T(w) > 0\right]$ is continuous with respect to w.

□

Remark An equivalent theorem (Theorem 2.1 in [4]) states that : if $T(w,\omega)$ is continuous with respect to w almost everywhere ($Pr[\omega : T(w,\omega)$ is discontinuous with respect to $w] = 0$ where $\omega \in \Omega$, and Ω is the sample space), then $F(x) = Pr\left[T(w) > 0\right]$ is continuous.

Corollary 1. *$AUC(w) = P\left[T_a(w) - T_0(w) > 0\right]$ is continuous with respect to w.*

Since $AUC(w)$ is continuous with respect to w in $(0,1)$, a global maximum w^* exists. We do not have closed-form expressions for the null and alternative distributions of the test statistic τ (with w as a parameter), so we cannot provide a rigorous proof of the uniqueness of w^*. However, for various data settings, simulations always resulted in *unimodal* estimates for the AUC function.

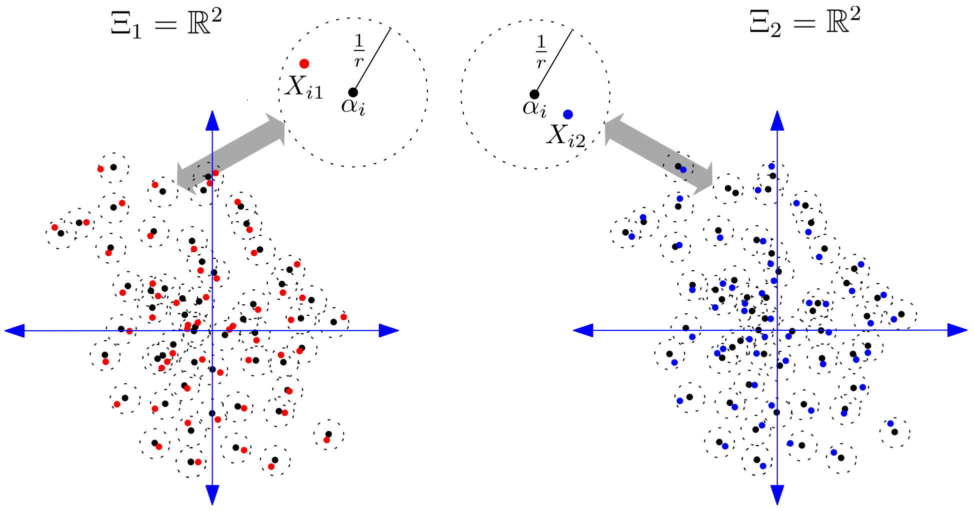

Figure 1: For the Gaussian setting (Section 6.1), the $\boldsymbol{\alpha_i}$, are denoted by black points and the $\boldsymbol{x}_{ik}$ are denoted by red and blue points respectively.

6 Simulation Results

6.1 Gaussian setting

Let n "objects" be represented by $\boldsymbol{\alpha}_i \sim^{iid} \mathcal{N}(\mathbf{0}, I_p)$. Let the $K = 2$ measurements for the i^{th} object under the different conditions be represented $\boldsymbol{x}_{ik} \sim^{iid} \mathcal{N}(\boldsymbol{\alpha_i}, \Sigma)$ represent $K = 2$ matched measurements (each under a different condition). Σ is a positive-definite $p \times p$ matrix whose maximum eigenvalue is $\frac{1}{r}$. See Figure 1.

Dissimilarities (Δ_1 and Δ_2) for the omnibus embedding are the Euclidean distances between the measurements in the same condition.

The parameter r controls the variability between "matched" measurements. If r is large, it is expected that the distance between matched measurements $\boldsymbol{x}_{i1}$ and $\boldsymbol{x}_{i2}$ is stochastically smaller than $\boldsymbol{x}_{i1}$ and $\boldsymbol{x}_{i'2}$ for $i \neq i'$; if r is small, then dissimilarities between pairs of "matched" measurements and "unmatched" are less distinguishable. Smaller r will make the decision problem harder and will lead to higher rate of errors or tests with smaller AUC measure.

6.2 Simulation

We generate the training data of matched sets of measurements (instantiation of $\mathcal{T}$) according to the Gaussian setting. Dissimilarity representations are computed from pairwise Euclidean distances of these measurements. We also generate a set of matched pairs and unmatched pairs of measurements for testing with the same distribution. Following the out-of-sample embedding of the dissimilarities test pairs we compute test statistics for matched and unmatched pairs. This allows us to compute the empirical power at different α (Type I error rate) values and the empirical AUC measure.

The signal and noise dimensions (p and q) were chosen as 5 and 10, respectively. For $nmc = 150$ Monte Carlo replicates, $n = 150$ matched training pairs and $m = 150$ matched and unmatched test pairs (generated according to the Gaussian

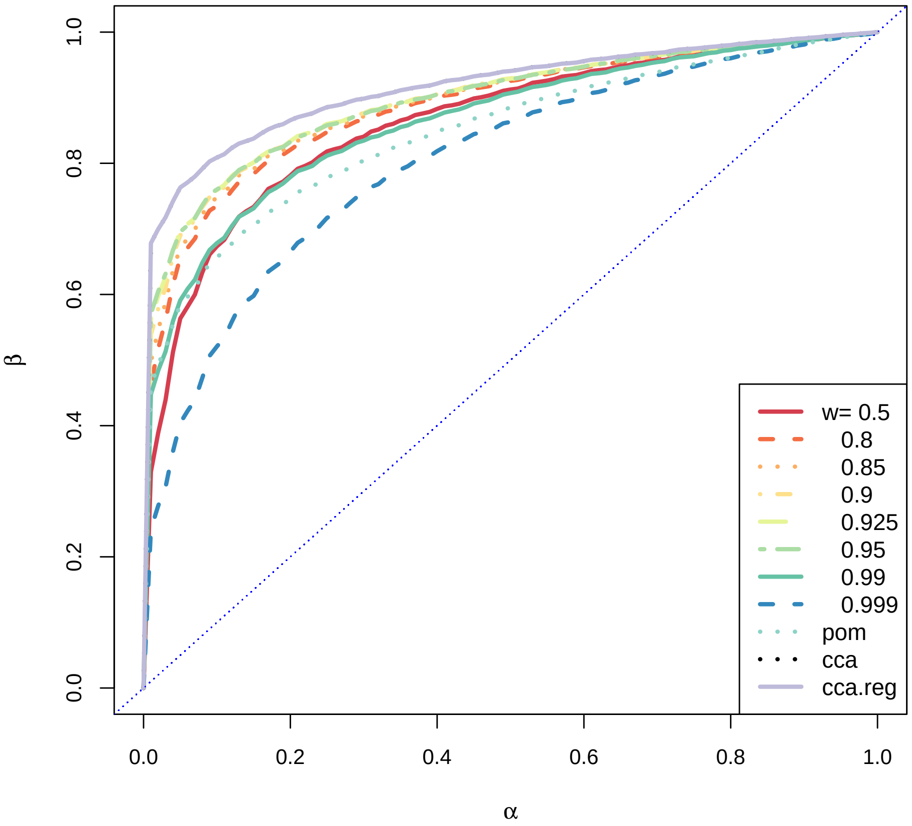

Figure 2: β vs α for different w values

setting) were generated. Using the resulting test statistic values for matched and unmatched test pairs, the AUC measure was computed for different w values along with the average of the power(β) values at different αs. The plot in Figure 2 shows the β-α curves for different values of w. In Figure 3, $\beta(w)$ is plotted against w for fixed values of α. The average AUC measure for these $nmc = 150$ MC replicates are in Table 1.

w	0.1	0.4	0.5	0.8	0.85	0.9	0.925	0.95	0.99	0.999
AUC	0.811	0.822	0.834	0.886	0.896	0.902	0.902	0.898	0.849	0.782

Table 1: average AUC(w) for $nmc = 150$ MC replicates

The test statistic that used a particular w value which had the highest AUC measure is not the same for each MC replicate. The number of MC replicates for which a particular w value led to the highest AUC is shown in the histogram in 4.

Note that the estimate of the optimal w^* has an AUC measure higher than that of w=0.5 (uniform weighting). This finding was confirmed using data generated according to the Gaussian setting with different set of parameters.

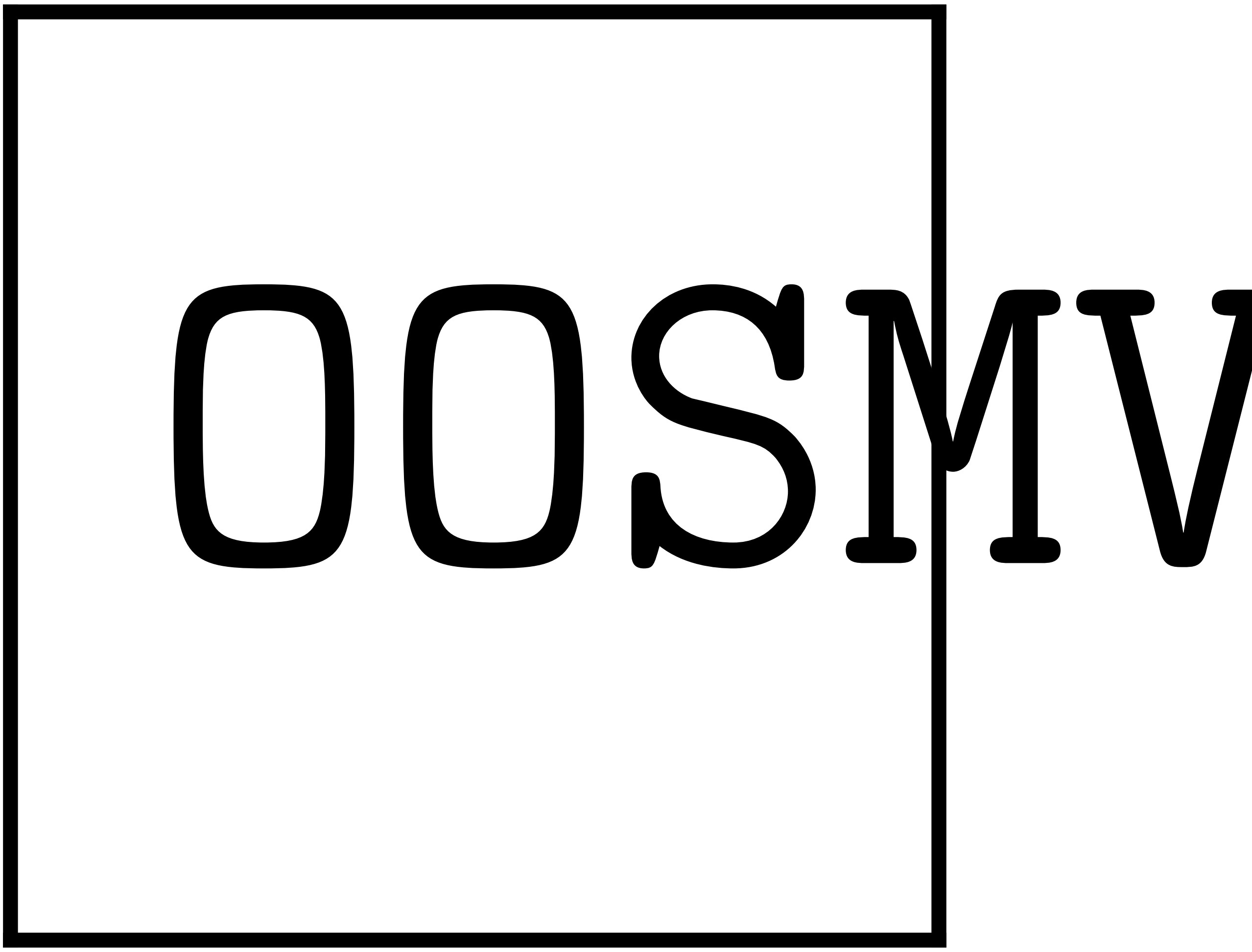

Figure 3: β vs w plot for different α values

7 Conclusion

The tradeoff between Fidelity and Commensurability and the relation to the weighted raw stress criterion for MDS were both investigated with simulations . For hypothesis testing as the exploitation task, the three approaches were compared in terms of testing power. The results indicate that when doing a joint optimization, one should consider an optimal compromise point between Fidelity and Commensurability, which corresponds to an optimal weight w^* of the weighted raw stress criterion in contrast to the uniform weighting for omnibus matrix embedding.

References

[1] I. Borg and P. Groenen. *Modern Multidimensional Scaling. Theory and Applications*. Springer, 1997.

[2] Brent Castle, Michael W. Trosset, and Carey E. Priebe. A nonmetric embedding approach to testing for matched pairs. (TR-11-04), October 2011.

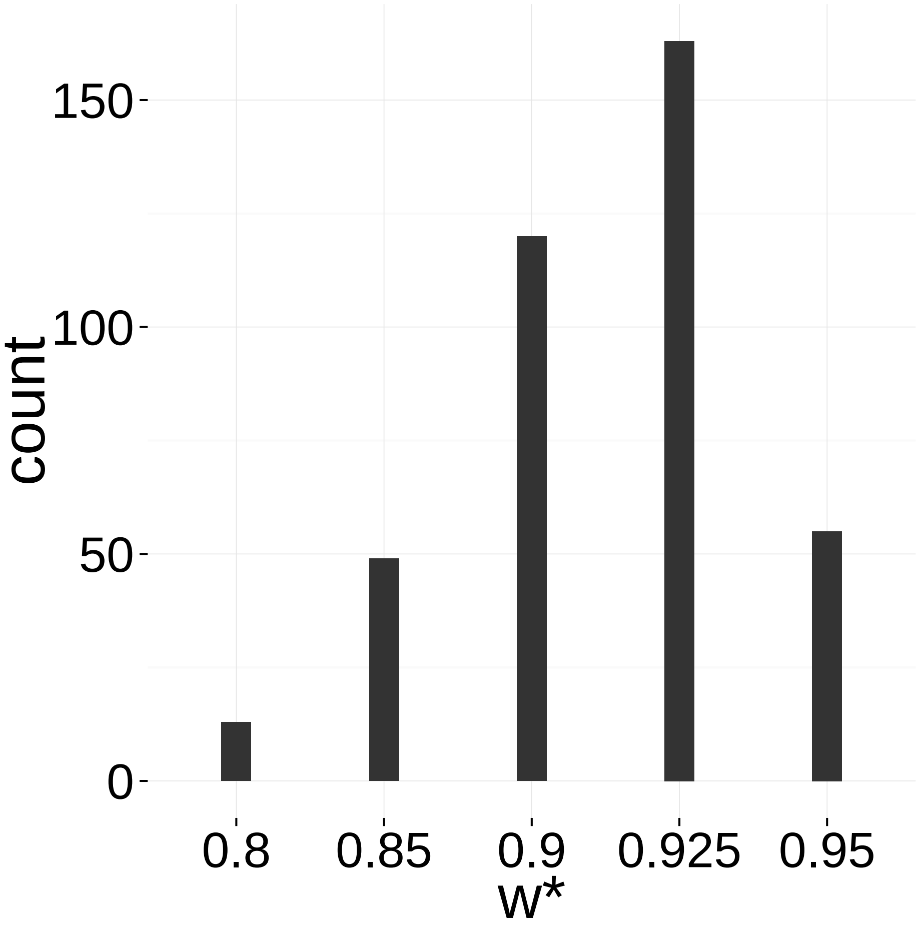

Figure 4: Histogram of $\arg\max_w AUC(w)$ for $nmc = 150$ replicates

[3] Jihun Ham, D Lee, and L. Saul. Semisupervised alignment of manifolds. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence, Z. Ghahramani and R. Cowell, Eds*, volume 10, pages 120–127. Citeseer, 2005.

[4] V.I. Norkin. The analysis and optimization of probability functions. *International Institute for Applied Systems Analysis technical report, Tech. Rep*, 1993.

[5] E. Pekalska and R.P.W. Duin. *The dissimilarity representation for pattern recognition: foundations and applications*. Series in machine perception and artificial intelligence. World Scientific, 2005.

[6] C.E. Priebe, D.J. Marchette, Z. Ma, and S. Adali. Manifold matching: Joint optimization of fidelity and commensurability. *Brazilian Journal of Probability and Statistics*. Submitted for publication.

[7] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

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

[9] Michael W. Trosset and Carey E. Priebe. The out-of-sample problem for classical multidimensional scaling. *Comput. Stat. Data Anal.*, 52:4635–4642, June 2008.

[10] C. Wang and S. Mahadevan. Manifold alignment using Procrustes analysis. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 1120–1127, New York, New York, USA, 2008. ACM Press.

[11] D. Zhai, B. Li, H. Chang, S. Shan, X. Chen, and W. Gao. Manifold alignment via corresponding projections. In *Proceedings of the British Machine Vision Conference*, pages 3.1–3.11. BMVA Press, 2010. doi:10.5244/C.24.3.