Fidelity-Commensurability Tradeoff in Joint Embedding of Disparate Dissimilarities

Sancar Adali[*] Carey E. Priebe[†]

April 11, 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 [6].

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 take under different illumination conditions. In each case, we have an intuitive notion of "matchedness": for photographs take 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 corresponds to the same Wikipedia article.

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 [2,11,12] follow a common approach: they look for a common commensurate or a latent space, such that the representations (either projections or embeddings) of the observations in this space match.

Wang and Mahedavan [11] 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 [11], 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. [12] solves an optimization problem with respect to two projection matrices for the datasets 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 of Locally Linear Embedding. 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 analagous to the terms in [12].

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 compared to the previous n objects", measured under K different conditions (the dissimilarities are matched). Under the alternative hypothesis, "the dissimilarities $\{\mathcal{D}_k\}$ represent separate *new* objects compared to the the previous n objects" measured under K different conditions (the dissimilarities are unmatched) [7].

For the English-French Wikipedia article example in the introduction, dissimilarities between the new English article and n other English articles ($\mathcal{D}_1$) are available, and likewise for the new French article and other n French articles ($\mathcal{D}_2$) [1]. 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 [2] is constructed. Consider, for $K = 2$,

$$M = \left[\begin{array}{cc} \Delta_1 & L \\ L^T & \Delta_2 \end{array} \right] \tag{1}$$

where L is a matrix of imputed entries.

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). 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 $\sigma(\widetilde{X})$ [3]). In matrix form,

$$\hat{X} = \arg\min_{\tilde{X}} \sigma(\tilde{X}).$$

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, which coordinates of the embedded point. The notation for matrices $(X, \tilde{X}, \hat{X})$ follows the same convention.

[1] in addition to the dissimilarities between articles in the same language ($\{\Delta_k\}$)

[2] a $nk \times nk$ partitioned matrix whose diagonal blocks are given by $\{\Delta_k\}$

[3] $\sigma(\widetilde{X})$ implicitly depends on the omnibus matrix M

Given the omnibus matrix M and the $2n \times d$ embedding configuration matrix $\hat{X}$ in the commensurate space, the out-of-sample extension [10] 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 *infidelity error*, given by

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

  Here δ_{ijk} 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 *incommensurability 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 [4]. Setting these diagonal entries to 0 forces matched observations to be embedded close to each other.

While the above expressions for *infidelity* and *incommensurability* 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

[4] These dissimilarities correspond to diagonal entries of the submatrix L in the omnibus matrix M in equation (1).

preserved (so that unmatched pairs are not embedded close together). The error for this criteria can be measured by $\epsilon_{s_{k_1k_2}} = \frac{1}{\binom{n}{2}} \sum_{1\leq i<j\leq n;k_1<k_2} (d(\widetilde{\boldsymbol{x}}_{ik_1}, \widetilde{\boldsymbol{x}}_{jk_2}) - \delta_{k_1k_2}(\boldsymbol{x}_{ik_1}, \boldsymbol{x}_{jk_2}))^2$ for conditions k_1 and k_2.

Let us now show how infidelity and incommensurability errors appear in the objective function. For the joint MDS embedding of the omnibus matrix M, the objective function ($\sigma(\cdot)$) we consider is the raw stress function:

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

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[5], 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.*

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

The separability error terms will be ignored herein, due to the fact that the between-condition dissimilarities, $\delta_{ijk_1k_2}$ for $i \neq j$, are not available. 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. We restrict our attention to the fidelity-commensurability tradeoff.

Although the between-condition dissimilarities of the same object, $\delta_{iik_1k_2}$, are not available, it is not unreasonable to set these dissimilarities to 0 for all i, k_1, k_2 [6]. Setting these diagonal entries to 0 forces matched observations to be embedded close to each other. Setting $\delta_{iik_1k_2}$ to 0, the raw stress function can be written as

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

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

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.

[5] Each row of the configuration matrix is the coordinate vector of an embedded point.

[6] These dissimilarities correspond to diagonal entries of the submatrix L in the omnibus matrix M (1).

The weights in the raw stress function allow us to address this question relatively easily. 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 [7] did not consider the effect of non-uniform weighting. Our hypothesis 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 [7], 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^*

Under each hypothesis of the hypothesis testing task, we have a separate distribution for the pair of test dissimilarities. Under matchedness, the dissimilarities $(\mathcal{D}_1^{(m)}, \mathcal{D}_2^{(m)})$ come from matched objects in different conditions. Under the alternative, the dissimilarities $(\mathcal{D}_1^{(u)}, \mathcal{D}_2^{(u)})$ represent unmatched objects. The out-of-sample embedding of the test dissimilarities involves the augmentation of the omnibus matrix (M) with the "matched" test dissimilarities which yields a $(2n+2) \times (2n+2)$ matrix:

$$\Delta^{(m)} = \begin{bmatrix} \Delta_1 & L & \mathcal{D}_1^{(m)} & \mathcal{D}_{NA} \\ L^T & \Delta_2 & \mathcal{D}_{NA} & \mathcal{D}_2^{(m)} \\ \mathcal{D}_1^{(m)T} & \mathcal{D}_{NA} & 0 & \mathcal{D}_{NA} \\ NA & \mathcal{D}_2^{(m)T} & \mathcal{D}_{NA} & 0 \end{bmatrix}. \tag{4}$$

Likewise, extending M with "unmatched" test dissimilarities yields $\Delta^{(u)}$. These are two matrix-valued random variables : $\Delta^{(m)} : \Omega \to \mathbf{M}_{(2n+2)\times(2n+2)}$ and $\Delta^{(u)} : \Omega \to \mathbf{M}_{(2n+2)\times(2n+2)})$ for the appropriate sample space (Ω).

Remark $\mathcal{D}_{NA}$ represent dissimilarities that are not available. These entries are either ignored in the embedding optimization or imputed using other dissimilarities that are available.

The criterion function for the embedding is $\sigma_W(\widetilde{X})$ 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 $\mathcal{T}$ is necessary but irrelevant for the inference task[7].

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

Assuming the necessary conditions hold for $\hat{y}_1^{(m)}$, $\hat{y}_2^{(m)}$, $\hat{y}_1^{(u)}$ and $\hat{y}_2^{(u)}$ to be random vectors, 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 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 $\hat{y}_2^{(u)}$.

Denote the cumulative distribution function of Y by F_Y.

Then, $\beta\left(w,\alpha\right) = 1 - F_{d\left(\hat{y}_1^{(u)}, \hat{y}_2^{(u)}\right)}(F^{-1}_{d\left(\hat{y}_1^{(m)}, \hat{y}_2^{(m)}\right)}(1-\alpha))$. Define the AUC function:

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

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

Finally, define

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

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 also be regarded as stochastic processes whose sample paths are continuous functions of w except at a finite number of points in $(0, 1)$.

Theorem 1. [8] *Let $T(\cdot)$ be a stochastic process indexed by w in the interval (0,1). Assume the process is continuous in probability (stochastic continuity) everywhere in the interval i.e.*

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

[7] hence the minimization with respect to $\widetilde{\mathcal{T}}$ is denoted by min instead arg min

[8] An equivalent theorem (Theorem 2.1 in [5]) 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.

$\forall w \in (0,1)$.

Then, for any $w > 0, \epsilon > 0$, *there exists* δ_ϵ

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

and

$Pr\,[T(w) > 0]$ *is continuous with respect to* w.

Corollary 1. $AUC(w) = P\,[T_a(w) - T_0(w) > 0]$ *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 .

5.2 Alternative Methodologies

Two alternative methodologies exist that correspond roughly to the extreme ends of the range of w values.

If we are concerned with only the optimization of commensurability with fidelity as secondary priority ($w \approx 1$), Canonical Correlational Analysis (CCA) [4] –which finds optimally correlated projections of two random vectors– can be used as an alternative method. Since the projections in CCA is computed for vectors in finite-dimensional Euclidean space, the given dissimilarities $(\Delta_1, \ldots, \Delta_k)$ have to be embedded first. CCA is then applied to the embeddings.

For the optimization of fidelity, the projections to the commensurate space can be found for the two conditions separately using Principal Components Analysis (PCA). PCA, like CCA, has to be applied to the embeddings of dissimilarities. Instead of applying PCA to embeddings to get a low-dimensional representation, it is possible to embed the dissimilarities directly in the low-dimensional space. The equivalence of PCA and Classical Multidimensional Scaling [9] under certain conditions suggests that this embedding approach is the right analog for PCA. To optimize commensurability as secondary priority, one can then compute a Procrustes transformation between the two configurations to make them as commensuratte as possible. This $Procrustes \circ MDS$ approach which we denote by $P \circ M$ is analogous to $w \approx 0$ case for JOFC.

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 $\boldsymbol{x}_{i1}$

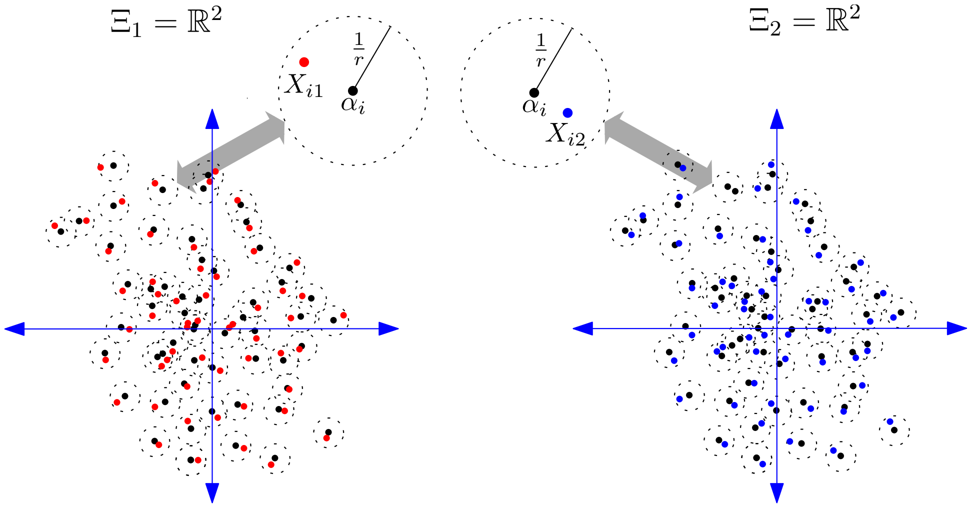

Figure 1: For the Gaussian setting (Section 6.1), the $\boldsymbol{\alpha_i}$, are denoted by black points and the $\boldsymbol{x}_{ik}$ are denoted by red and blue points respectively.

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 (computed via by one of the three P∘M, CCA and JOFC approaches), 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 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

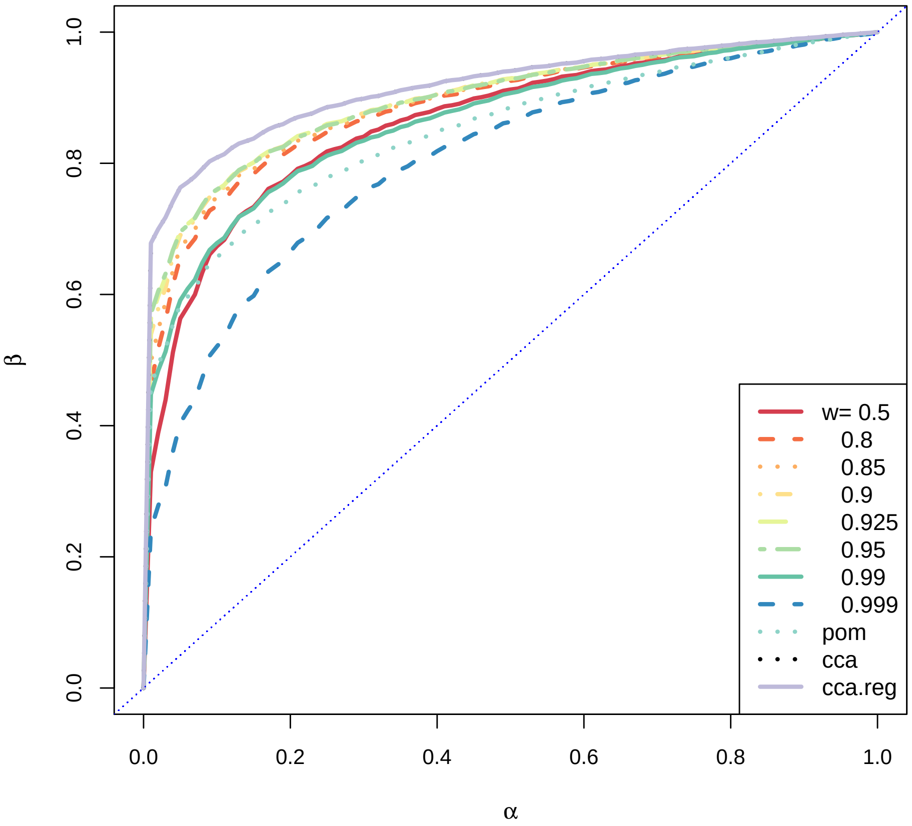

Figure 2: β vs α for different w values

The histogram in 4 shows for how many MC replicates a w value had the highest AUC measure.

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.

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.

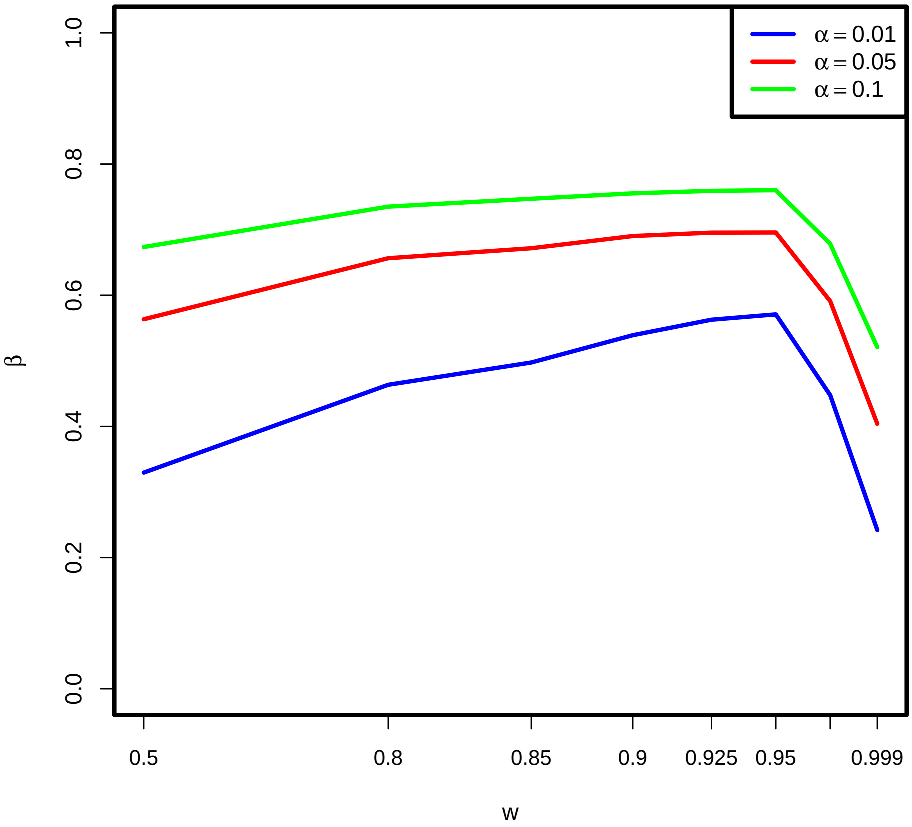

Figure 3: β vs w plot for different α values

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.

[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] David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16:2639–2664, December 2004.

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

[6] 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.

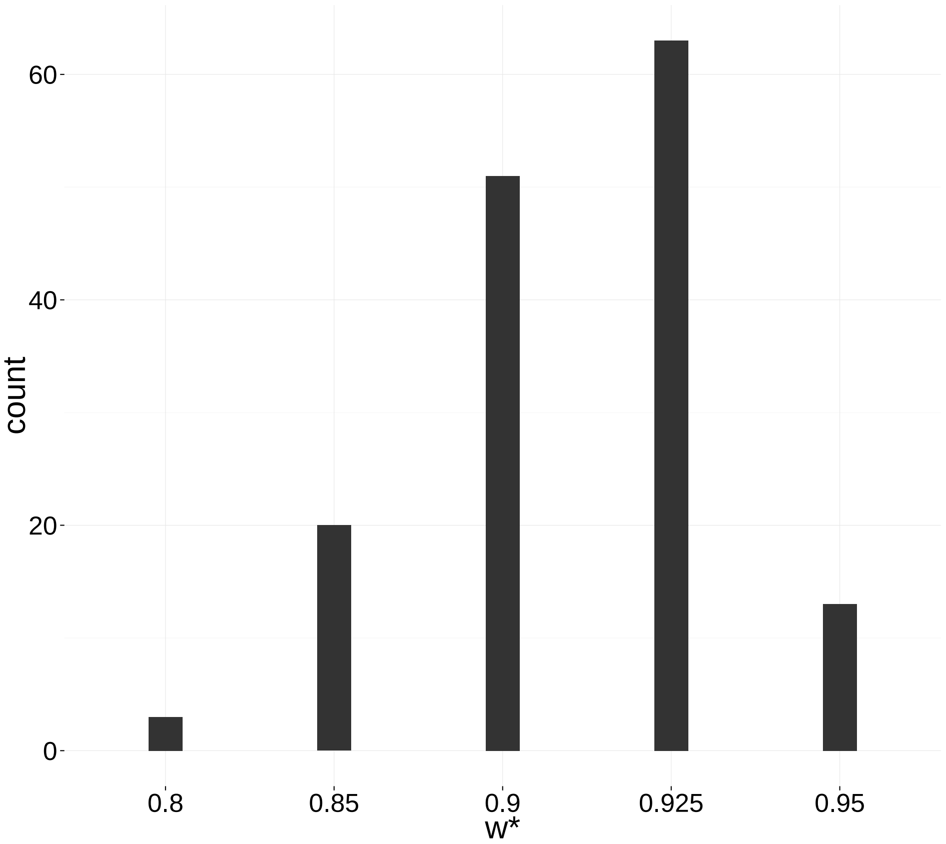

Figure 4: Histogram of $\arg\max_w AUC(w)$ for $nmc = 150$ replicates

[7] 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.

[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] W. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17:401–419, 1952.

[10] 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.

[11] 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.

[12] 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.