Fidelity-Commensurability tradeoff in Joint Embedding of Disparate Dissimilarities

Sancar Adali[*] Carey E. Priebe[†]

August 1, 2011

Abstract

For matched data from disparate sources (objects observed under different conditions), optimality of information fusion must be defined with respect to the inference task at hand. Defining the task as matched/unmatched hypothesis testing for dissimilarity observations, Priebe et al. [6] presents manifold matching using an embedding method based on joint optimization of fidelity (preservation of within-condition dissimilarities between observations of an object) and commensurability (preservation of between-condition dissimilarities between observations) . The tradeoff between fidelity and commensurability is investigated by varying weights in weighted embedding of an omnibus dissimilarity matrix. Optimal (defined with respect to the power of the test) weights for the optimization correspond to an optimal compromise between fidelity and commensurability. Results indicate optimal weights are different than equal weights for commensurability and fidelity and the proposed weighted embedding scheme provides significant improvements in test power.

1 Introduction

It is a challenge to do a tractable analysis on data from disparate sources of data (such as multiple sensors). The multitude of sensors technology and large numbers of sensors both are sources of difficulty and hold promise for efficient inference. There are quite a few real life cases, where the data is acquired or available exclusively in dissimilarity representation instead of feature representation [1,5,7]. Multidimensional scaling is the processing step used to find the feature representation equivalent of this kind of data, so that statistial machine learning methods can be applied to the data. A specific variant of MDS will be used to get an embedding in Euclidean Space.

In this paper, hypothesis testing is the exploitation task being considered, and the "optimal" embedding refers to that leads to the test with the highest power. Consider the weighted raw-stress function:

$$\sigma_W(X) = \sum_{1\leq s\leq n; 1\leq t\leq n} w_{st}(d_{st}(X) - \delta_{st})^2 \tag{1}$$

for an $n\times p$ configuration matrix (n points in p dimensions) X where $d_{st}(X)$ is the Euclidean distance between s^{th} and t^{th} rows of X and w_{st} is the weight for st^{th} squared difference. $n \times n$ matrix representation of the weights and Euclidean distance will be denoted by W and $D(X)$, respectively. This criterion function which will be minimized for embedding configurations is appropriate for the purpose of finding a tradeoff between two different criteria (namely the preservation of fidelity and commensurability).

[*] 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

The dissimilarity-representation version of a hypothesis testing problem is stated as follows:
n different objects/instances are measured/judged under K different conditions with (possibly notional) measurements x_{ik} indexed by object and condition. Each of the measurements x_{ik} lies in the corresponding space Ξ_k.

$$\begin{array}{ccccc} & \Xi_1 & \cdots & \Xi_K \\ \textit{Object } 1 & \boldsymbol{x}_{11} & \sim \cdots \sim & \boldsymbol{x}_{1K} \\ \vdots & \vdots & \vdots & \vdots \\ \textit{Object } n & \boldsymbol{x}_{n1} & \sim \cdots \sim & \boldsymbol{x}_{nK} \end{array}$$

For each pair of measurements x_{ik}, x_{jk} in the same space, the dissimilarity value $\delta_{ijk} = \delta\{x_{ik}, x_{jk}\}$, is based on the dissimilarity measure on the space Ξ_k. The dissimilarities are assumed to be non-negative and symmetric, and 0 for $\delta\{x_{ik}, x_{ik}\}$. These dissimilarities are exploited to carry out the following hypothesis testing task:

Given dissimilarities between K new measurements/observations ($\boldsymbol{y}_k; k = 1, \ldots, K$) and the previous n objects under K conditions, test the null hypothesis that "these measurements are from the same object" against the alternative hypothesis that "they are not from the same object" [6]:

$$H_0 : \boldsymbol{y}_1 \sim \boldsymbol{y}_2 \sim \cdots \sim \boldsymbol{y}_K \text{ versus } H_A : \exists i, j, 1 \leq i < j \leq K : \boldsymbol{y}_i \nsim \boldsymbol{y}_j$$

The null hypothesis can be restated as the case where the dissimilarities are "matched" and the alternative as the case where they are not "matched".

Dissimilarities are in the form of $n \times n$ dissimilarity matrices $\{\Delta_k; k = 1, \ldots, K\}$ with entries $\{\delta_{ijk}; i = 1, \ldots, n;\ j = 1, \ldots, n\}$ and a vector (of length nK) of dissimilarities $\boldsymbol{\Delta}^{new} = \{\delta^{new}_{ik}; i = 1, \ldots, n;\ k = 1, \ldots, K\}$ where δ^{new}_{ik} is the dissimilarity between x_{ik} and y_k

Since dissimilarities are measured between pairs of objects under the same condition, they will be the entries in separate dissimilarity matrices , each matrix consisting of dissimilarities between pairs of measurements for a separate condition. Due to the fact that data sources are "disparate", it is not immediately obvious how a dissimilarity between an object in one condition and another object in another condition can be computed, or even defined in a sensible way. In general, these between-condition between-object similarities are not available.

Throughout this paper , it will be assumed, the number of conditions, K, is equal to 2 for the simplicity of presentation.

2 "Matched" and "Conditions" in data

"Conditions" and "matched" refer to concepts dependent on the context of the problem. Conditions could be different modalities of data, e.g., one condition could be an image of an object, while the other condition could be a text description of the object. "Matched", in general, means observations of the same object, or realizations of a common concept.

3 Two models for generating data

In this section, two data models are proposed that illustrate the idea of matchedness.

3.1 Gaussian setting

Let $\Xi_1 = \mathbb{R}^p$ and $\Xi_2 = \mathbb{R}^p$. Let $\boldsymbol{\alpha}_i \sim^{iid} MVNormal(\mathbf{0}, I_p)$ represent n "objects". Let $\boldsymbol{x}_{ik} \sim^{iid} MVNormal(\boldsymbol{\alpha_i}, \Sigma)$ represent $K = 2$ matched measurements (each under a different condition). Σ is a positive-definite $p \times p$ matrix such that $\max(\Lambda(\Sigma)) = \frac{1}{r}$ where $\Sigma = U\Lambda(\Sigma)U'$ is the eigenvalue decomposition of Σ. See Figure 1.

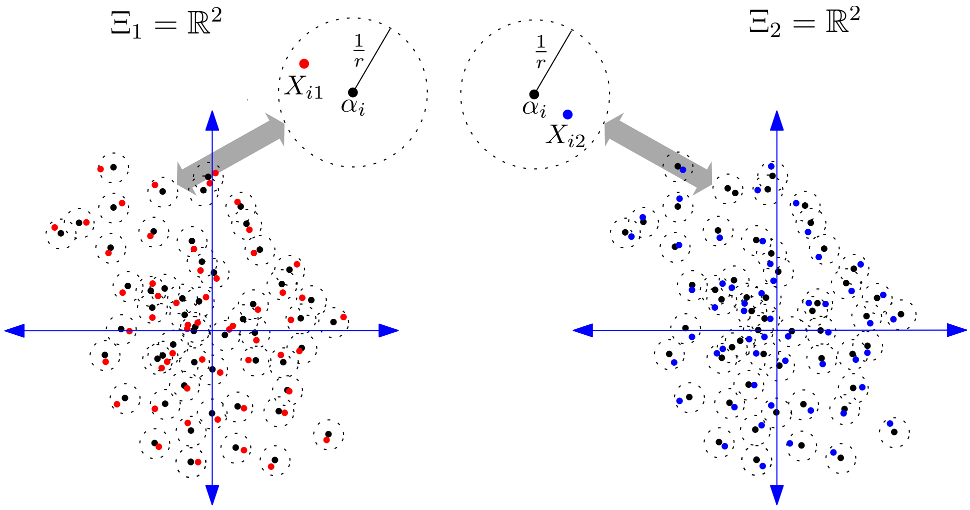


Figure 1: For the Gaussian setting (Section 3.1), the $\boldsymbol{\alpha_i}$, are denoted by black points and the $\boldsymbol{x}_{ik}$ are denoted by red and blue points respectively.

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 $\boldsymbol{x}_{i2}$ to be stochastically smaller than $\boldsymbol{x}_{i1}$ and $\boldsymbol{x}_{i'2}$ for $i \neq i'$; if r is small, then "matched" is not informative in terms of similarity of measurements. Smaller r will make the decision problem harder and will lead to higher rate of errors or tests with smaller power for fixed type I error rate α.

3.2 Dirichlet setting

Let $S^p = \{\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^{(p+1)}, \sum_{l=1}^{p+1} x_l = 1\}$ be the standard p-simplex in $\mathbb{R}^{p+1}$. Let $\Xi_1 = S^p$ and $\Xi_2 = S^p$. Denote a vector of ones by $\mathbf{1}_{p+1} \in \mathbb{R}^{(p+1)}$. Let $\boldsymbol{\alpha}_i \sim^{iid} Dirichlet(\mathbf{1}_{p+1})$ represent n "objects" and let $\boldsymbol{x}_{ik} \sim^{iid} Dirichlet(r\boldsymbol{\alpha}_i + \mathbf{1}_{p+1})$ represent K measurements.

The parameter r again controls the variability between "matched" measurements.

3.3 Noise

Measurements $\boldsymbol{x}_{ik}$ carry the signal that is relevant to the exploitation task. Noise dimensions can be introduced to the measurements by concatenating a q-dimensional error vector whose magnitude is controlled by the parameter $c, c \in (0,1)$. The noisy measurements will be represented by the random vectors

$$\breve{X}_{ik} = [(1-c)\boldsymbol{x}_{ik} \;\; cE_{ik}] \tag{2}$$

where $E_{ik} \sim^{iid} Dirichlet(\mathbf{1}_{(q+1)})$ for the Dirichlet setting and $E_{ik} \sim^{iid} MVNormal(\mathbf{0}, (1 + \frac{1}{r})I_{q+1})$ for the Gaussian setting. $\breve{X}_{ik}$ will be used instead of $\boldsymbol{x}_{ik}$ for computing dissimilarities in the "noisy" version of the problem. These noisy measurements allow the comparison of different methods applied to the problem with respect to their robustness.

4 Manifold Matching

The JOFC approach can be summarized as identifying embeddings of multiple disparate data sources into the same low-dimensional space where joint inference can be pursued.

It will be assumed the commensurate space $\mathcal{X}$ is $\mathbb{R}^d$ where 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. Discussion of the effect of d on matching performance will be available at a later paper.

To embed dissimilarities $\{\Delta_k, k = 1, \ldots, K\}$ from different conditions into a commensurate space in one step, an omnibus dissimilarity matrix M can be embedded in the low-dimensional Euclidean spaceinto one omnibus dissimilarity matrix M, imputing entries if necessary. Consider, for $K = 2$,

$$M = \begin{bmatrix} \Delta_1 & L \\ L^T & \Delta_2 \end{bmatrix} \tag{3}$$

where L is a matrix of imputed entries. One way to impute L is to set it to $\frac{\Delta_1+\Delta_2}{2}$. Another choice for imputation is introduced in 5: the diagonal of L is set to 0, the rest of the entries are NA and are ignored in the optimization of MDS criterion. Using MDS to embed this omnibus matrix into a space $\mathcal{X}$, $2n$ embedded observations $\{\tilde{y}_i^{(k)}; i = 1, \ldots, n; k = 1, 2\}$ are obtained in a single space, with distances between the different observations consistent with the given dissimilarities. Now that the observations are commensurate, it is possible to compute the test statistic

$$\tau = d\left(\tilde{y}_i^{(1)}, \tilde{y}_j^{(2)}\right)$$

for i^{th} and j^{th} observations under different conditions. For “large” values of τ, the null hypothesis will be rejected. This approach will be referred to as the Joint Optimization of Fidelity and Commensurability (JOFC) approach, for reasons that will be explained in Section 5. Out-of-sample extension for MDS will be used throughout this paper [8].

5 Fidelity and Commensurability constraints for Manifold Matching

Unless

- the dissimilarity matrix is the Euclidean distance matrix of the original observations, and,
- the embedding dimension is greater or equal to the dimension of the original observations,

MDS with raw stress will not result in a perfect reconstruction of the original observations. Note that the objective of the embedding is not *perfect* reconstruction, but the best embedding for the exploitation task which is to test whether two sets of dissimilarities are “matched”. What is considered a good “commensurate” representation will be dependent on how well the information in original dissimilarities that is relevant to the the match detection task is preserved. The following two criteria embody the two kind of information that is relevant to this task.

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

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

  where $\boldsymbol{x}_{ik}$ is the original observation of the i^{th} object for the k^{th} condition and $\widetilde{\boldsymbol{x}}_{ik}$ is the embedded configuration of the i^{th} object for the k^{th} condition; $d(\cdot, \cdot)$ is the Euclidean distance function (for the embedding space) and $\delta_k(\cdot, \cdot)$ is the dissimilarity function defined for objects in the k^{th} condition.

- Commensurability is 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* is given by

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

for conditions k_1 and k_2; $\delta_{k_1k_2}(\cdot,\cdot)$ is the (notional) dissimilarity function between measurements in k_1^{th} and k_2^{th} conditions. Note that ,in general, there are K within-condition fidelity error terms and $\frac{K\times(K-1)}{2}$ between-condition commensurability error terms.

Although the between-condition dissimilarities of the same object, $\delta_{k_1k_2}(\boldsymbol{x}_{ik_1},\boldsymbol{x}_{ik_2})$, are not available, it is not unreasonable in this setting to set $\delta_{k_1k_2}(\boldsymbol{x}_{ik_1},\boldsymbol{x}_{ik_2}) = 0$ for all i, k_1, k_2. So diagonal entries of L in equation (3) are chosen to be all zeroes. Setting these diagonal entries to zero forces matched points to be embedded close to each other.

Then, the commensurability error term becomes

$$\epsilon_{c_{k_1k_2}} = \frac{1}{n}\sum_{1\le i\le n;k_1<k_2}(d(\widetilde{\boldsymbol{x}}_{ik_1},\widetilde{\boldsymbol{x}}_{ik_2})))^2$$

There is also between-condition *separability error* given by

$$\epsilon_{s_{k_1k_2}} = \frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n;k_1<k_2}(d(\widetilde{\boldsymbol{x}}_{ik_1},\widetilde{\boldsymbol{x}}_{jk_2}) - \delta_{k_1k_2}(\boldsymbol{x}_{ik_1},\boldsymbol{x}_{jk_2}))^2.$$

This error will be ignored herein, due to the fact that $\delta_{k_1k_2}(\boldsymbol{x}_{ik_1},\boldsymbol{x}_{jk_2})$ is not available.

In the raw stress version of MDS, the individual terms can be separated according to whether they are contributing to fidelity or commensurability error.

Consider the weighted raw stress criterion $\sigma_W(\cdot)$ with a weighting matrix W, given in equation (1).

$$\begin{aligned}\sigma_W(\cdot) = & \sum_{i,j,k_1,k_2} w_{ijk_1k_2}(d_{ijk_1k_2}(\cdot) - M_{ijk_1k_2})^2 \\ = & \underbrace{\sum_{i=j,k_1<k_2} w_{ijk_1k_2}(d_{ijk_1k_2}(\cdot) - M_{ijk_1k_2})^2}_{Commensurability} + \underbrace{\sum_{i<j,k_1=k_2} w_{ijk_1k_2}(d_{ijk_1k_2}(\cdot) - M_{ijk_1k_2})^2}_{Fidelity} \\ + & \underbrace{\sum_{i<j,k_1<k_2} w_{ijk_1k_2}(d_{ijk_1k_2}(\cdot) - M_{ijk_1k_2})^2}_{Separability} \quad . \end{aligned} \tag{4}$$

Since $\delta_{k_1k_2}(\boldsymbol{x}_{ik_1},\boldsymbol{x}_{ik_2})$ are set to 0, the corresponding entries of M in the commensurability terms will be 0.

Since the separability error is ignored, the weights for separability terms are chosen to be 0. This also means off-diagonal elements of L in equation (3) can be ignored. When separability terms are removed from equation (4), the resulting equation is a sum of fidelity and commensurability error terms:

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

This motivates referring to the omnibus embedding approach as Joint Optimization of Fidelity and Commensurabilty (JOFC).

5.1 Alternative Methodologies

Two alternative methodologies exist that correspond roughly to the extreme ends of the range of w values For the optimization of commensurability with fidelity as secondary priority($w \approx 1$), an alternative method is Canonical Correlational Analysis (CCA) [3], which aims to find linear subspaces of the Euclidean space such that the projection of data points to those subspaces results in vectors that are maximally correlated. For the optimization of fidelity , one can use Principal Components Analysis (PCA), which aims to find linear subspaces such that projection of data points to those subspaces results in observation vectors that represent the original data as best as possible. To optimize commensurability as secondary priority, one can use the projections computed by PCA to compute a Procrustes transformation that will make the projections commensurate. This $Procrustes \circ MDS$ approach is analagous to $w \approx 0$ case for JOFC. Details about these methodologies along with the exact nature of relationship to extreme ends of JOFC will be discussed in another paper.

6 Related Work

There have many efforts toward solving the related problem of "manifold alignment"."Manifold alignment" seeks to find correspondences between observations from different "conditions". The setting that is most similar to ours is the semi-supervised setting [?], where a set of correspondences 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 follow a common approach in that they look for a common commensurate or a latent space, such that the representations (possibly projections or embeddings) of the observations in the commensurate space match. [2, 9, 10]

7 Fidelity and Commensurability Tradeoff

The major question addressed in this work is whether in the tradeoff between preservation of fidelity and preservation of commensurability , there is an optimal point for the match detection task. The weights in raw stress allow us to answer this question relatively easily. Since in equation (4), each term indexed with i, j is either a fidelity or a commensurability term, setting w_{ij} to w and $1-w$ for commensurability and fidelity terms respectively will allow us to control the importance of fidelity and commensurability terms in the optimization by varying w.

$$\begin{aligned}\sigma_W(X) &= f_w(D(X), M)\\ &= \underbrace{\sum_{i=j,k_1\neq k_2} w(d_{ijk_1k_2}(X))^2}_{Commensurability} \quad + \quad \underbrace{\sum_{i<j,k_1=k_2} (1-w)(d_{ijk_1k_2}(X)-M_{ijk_1k_2})^2}_{Fidelity}\\ &= (w)\,(n)\,\epsilon_{c_{k_1=1,k_2=2}} \quad + \quad (1-w)\binom{n}{2}(\epsilon_{f_{k=1}}+\epsilon_{f_{k=2}})\end{aligned}$$

The expectation here is that there is a w^* that is optimal for the specific exploitation task (has the best power in hypothesis testing). In fact, exploratory simulations presented in this paper confirm the power of the tests varies with varying w and indicate the range where the optimal w^* lies.

8 Definition of w^*

Two dissimilarity matrices are defined to be $\Delta^{(m)}\left(\begin{bmatrix}\mathcal{T}\\ X_1^{(m)}\\ X_2^{(m)}\end{bmatrix}\right)$ and $\Delta^{(u)}\left(\begin{bmatrix}\mathcal{T}\\ X_1^{(u)}\\ X_2^{(u)}\end{bmatrix}\right)$ as two matrix-valued random variables $\Delta^{(m)}:\Omega\rightarrow\mathbf{M}_{n\times n}$,$\Delta^{(u)}:\Omega\rightarrow\mathbf{M}_{n\times n}$ for the appropriate sample space (Ω). These dissimilarity matrices are result of dissimilarities between an i.i.d sample of matched measurements augmented with a matched and unmatched pair of measurements respectively.

The criterion function for the embedding is $\sigma_W(\cdot) = f_w(D(\cdot), \Delta)$. All of the random variables following the embedding is dependent on w, for the sake of simplicity, it will not be shown in the notation. The embedding for the unmatched pair $\hat{X}_1^{(u)}, \hat{X}_2^{(u)}$ is

$$\hat{X}_1^{(u)}, \hat{X}_2^{(u)} = \underset{\acute{X}_1^{(u)}, \acute{X}_2^{(u)}}{\arg\min} \left[\min_{\acute{\mathbf{T}}} f_w \left(D \left(\begin{bmatrix} \acute{\mathbf{T}} \\ \acute{X}_1^{(u)} \\ \acute{X}_2^{(u)} \end{bmatrix} \right), \Delta^{(u)} \right) \right]$$

where there is an implicit dependence on $\mathbf{T}$, because $\Delta^{(u)}$ depends on $\mathbf{T}$. A similar expression gives the embedding for the matched pair $\hat{X}_1^{(m)}, \ \hat{X}_2^{(m)} = \arg\min_{\acute{X}_1^{(m)}, \acute{X}_2^{(m)}} \left[\min_{\acute{\mathbf{T}}} f_w \left(D \left(\begin{bmatrix} \acute{\mathbf{T}} \\ \acute{X}_1^{(m)} \\ \acute{X}_2^{(m)} \end{bmatrix} \right), \Delta^{(u)} \right) \right]$.
Assuming these minima exist and are unique, the mappings $\hat{X}_1^{(m)} : \omega \to \mathbf{R}^{d'}$, $\hat{X}_2^{(m)} : \omega \to \mathbf{R}^{d'}$, $\hat{X}_1^{(u)} : \omega \to \mathbf{R}^{d'}$, $\hat{X}_2^{(u)} : \omega \to \mathbf{R}^{d'}$ are measurable maps, $\hat{X}_1^{(m)}$, $\hat{X}_2^{(m)}$, $\hat{X}_1^{(u)}$, $\hat{X}_2^{(u)}$ are random vectors.

Consider the test statistic $\tau = d(\hat{X}_1^{\cdot}, \hat{X}_2^{\cdot})$. Under null hypothesis of matchedness, the distribution of the statistic is governed by the distribution of $\hat{X}_1^{(m)}$ and $\hat{X}_2^{(m)}$, under the alternative it is governed by $\hat{X}_1^{(u)}$ and $\hat{X}_2^{(u)}$.

Denote by F_Y the cumulative distribution function of Y where Y can be any function of $\hat{X}_k^{(m)}$ or $\hat{X}_k^{(u)}$. for $k = \{1, 2\}$

Then

$$\beta_\alpha(w) = 1 - F_{d\left(\hat{X}_1^{(u)}, \hat{X}_2^{(u)}\right)}(F^{-1}_{d\left(\hat{X}_1^{(m)}, \hat{X}_2^{(m)}\right)}(1 - \alpha)).$$

Note that all random variables are dependent on w. Finally, define

$$w^* = \arg\max_w \beta_\alpha(w).$$

Even for given $\mathbf{F}_u, \mathbf{F}_m$, w^* must be defined with respect to the value of allowable type I error rate α. For two different α values, it is quite possible that $\beta_{\alpha_1}(w_1) > \beta_{\alpha_1}(w_2)$ and $\beta_{\alpha_2}(w_1) < \beta_{\alpha_2}(w_2)$. This can be observed in results in Section 9. Investigation of some properties of w^* is included in section 9 . w^* is defined to be the argmin of the power function with respect to w and some important questions about w^* are related to the nature of this function $\beta_\alpha(w)$. While finding an analytical expression for the value of w^* is intractable, an estimate $\hat{w}^*$ based on noisy evaluations of $\beta_\alpha(w^*)$ can be computed. A Monte Carlo simulation is run in Section 9 to find the estimate of $\beta_\alpha(w)$ at various values of w and α .

8.1 Continuity of $\beta(\cdot)$

Let $T_0(w)$ and $T_a(w)$ denote the value of the test statistic under null and alternative distributions for the embedding with w . Consider $\beta_\alpha(\cdot)$ as a function of w, which can be written as $P[T_a(\cdot) > c_\alpha(\cdot)]$ where $c_\alpha(\cdot)$ is the critical value for level α. Instead of $\beta_\alpha(\cdot)$

the area under the curve measure will be shown to be continuous as a surrogate:

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

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. *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.*

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

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

Then,

for any $w > 0, \epsilon > 0$*, there exist* 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 with respect to* w*.*

Proof. . For any δ for which $\delta + w \in (0,1)$, consider the difference of the probabilities of the two events $T(w) > 0$ and $T(w+\delta) > 0$ for any w.

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

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 $Pr\left[T(w+\delta) > 0 | T(w) \leq 0\right]$ and $Pr\left[T(w+\delta) \leq 0 | T(w) > 0\right]$ can be made smaller than ϵ due to stochastic continuity. Therefore

$$2\|Pr\left[T(w+\delta) > 0\right] - Pr\left[T(w) > 0\right]\| \leq \max(Pr\left[T(w+\delta) > 0 | T(w) \leq 0\right], Pr\left[T(w+\delta) \leq 0 | T(w) > 0\right]) < \epsilon$$

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

□

Theorem 2.1 in [4] states the same theorem : 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*.*

Proof. $Pr\left[T(w) > 0\right]$ as a function of w is continuous with respect to w. Let $T(w) = T_a(w) - T_0(w)$. □

9 Simulation Results

To compare the different approaches, training data of matched sets of measurements were generated according to the Dirichlet and Gaussian settings. Dissimilarity representations were computed from pairwise Euclidean distances of these measurements. A set of matched pairs and unmatched pairs of measurements were also generated for testing with the same distributions. Following the out-of-sample embedding of the dissimilarities test pairs (computed via by one of the three P∘M, CCA and JOFC approaches) test statistics for matched and unmatched pairs were used to compute power values at a set of fixed type I error rate α values.

Additionally, to take robustness of methods into consideration, "noisy" measurements were created from the original measurements by concatenating randomly generated independent noise vectors (subsection 3.3). The magnitude of noise is controlled by the parameter c in equation (2)). The original setting, with $c = 0$, will be referred as the "noiseless case". If the magnitude of noise is small enough, and the embedding dimension is not larger than signal dimension, the embeddings provided by PCA and MDS will not be affected significantly. However if the number of noisy dimensions (controlled by the parameter q in the distribution of E_{ik} as defined in equation (2)) is large enough, embeddings via CCA will be affected due to spurious correlation between noisy dimensions.

Given the setting (”Gaussian”,”Dirichlet”), the steps for each Monte Carlo replicate are as follows:

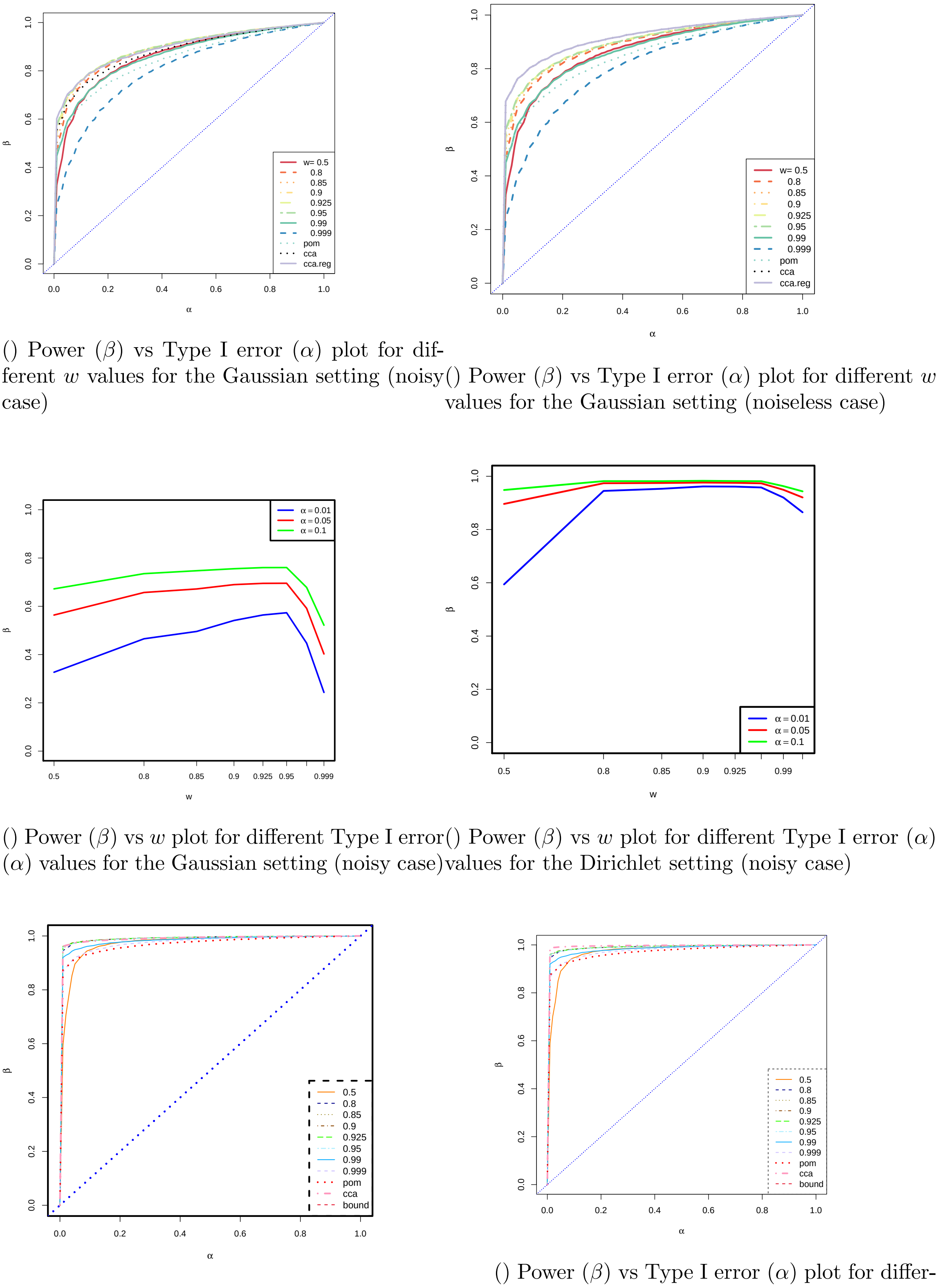
- A training set ($\mathbf{T}_{mc}$) which consists of n pairs of matched measurements is generated. If $c = 0$, the “noiseless” data setting is being simulated and measurements are p-dimensional vectors, otherwise the “noisy” setting is being used to generate data and measurement vectors are $(p+q)$-dimensional. $\mathbf{T}_{mc} = \begin{array}{ccc} X_{11} & \ldots & X_{1K} \\ \cdots & \cdots & \cdots \\ X_{n1} & \ldots & X_{nK} \end{array}$ where each X_{ik} is a random vector of dimension $(p + q \times I(c > 0))$ and the conditional distribution $X_{i.}|\boldsymbol{\alpha}_i$ is specified as an appropriate Multivariate Normal or Dirichlet distribution.
- Dissimilarities are computed and embedded in Euclidean space via MDS (followed by a transformation from $\mathbf{R}^d$ to $\mathbf{R}^d$ and projection into $\mathbf{R}^d$, respectively for P∘M and CCA) . The final embeddings lie in $\mathbb{R}^d$. Denote this in-sample embedding configuration as $\hat{\mathbf{T}}$. Note that if the JOFC method is being used, embedding is carried out with the weighted raw stress function $\sigma_W(\cdot) = f_w(D(\cdot), M)$ in equation (**??**) with a common weight w for commensurability terms and another common weight $1 - w$ for fidelity terms, otherwise unweighted raw stress function $(\sigma(\cdot))$ is used as a criterion function for embedding.
- m pairs of matched measurements are generated which are treated as out-of-sample, and
  - compute the dissimilarities between these out-of-sample points and the points in $\mathbf{T}_{mc}$,
  - embed the OOS dissimilarities as pairs of embedded points via the OOS extension: $(\tilde{y}_1^{(1)}, \tilde{y}_1^{(2)}), \ldots, (\tilde{y}_m^{(1)}, \tilde{y}_m^{(2)})$,
  - compute the test statistic τ for each pair.

  The values of the statistic τ are used for computing the empirical cumulative distribution function under null hypothesis.
- Identical steps for m pairs of unmatched measurements result in the empirical cumulative distribution function of τ under alternative hypothesis.
- For any fixed α value, a critical value for the test statistic and the corresponding power is computed.

Setting p and q to 5 and 10, respectively, for $n = 150$ and $m = 150$, the average of the power values for $nmc = 150$ Monte Carlo replicates are computed at different αs and are plotted in Figure 2 against α for the Gaussian setting. Qualititatively similar plots for the Dirichlet setting are not included for brevity. The plot in Figure 2 shows that for different values of w, β-α curves vary significantly. The conclusion is that the match detection tests with JOFC embedding using specific w values have better performance than other w values in terms of power. In Figure 2, $\beta(w)$ is plotted against w for fixed values of α. It is interesting that the optimal value of w seems to be in the range of $(0.85, 1)$ for all settings, which suggests a significant emphasis on commensurability might be critical for the match detection task.

Note that in Figure 2 for $\alpha = 0.05$, $\beta_{\alpha=0.05}(w = 0.99) \geq \beta_{\alpha=0.05}(w = 0.5)$. However, for $\alpha = 0.3$, $\beta_{\alpha=0.3}(w = 0.99) \leq \beta_{\alpha=0.3}(w = 0.5)$. This justifies our comment that w^* must be defined with respect to α.

Note that for all of the settings, the estimate of the optimal w^* has higher power than w=0.5 (the unweighted case). To test the statistical significance of this observation, the null hypothesis that $H_0 : \beta_\alpha(\hat{w}^*) \leq \beta_\alpha(w = 0.5)$ is tested against the alternative $H_A = \beta_\alpha(\hat{w}^*) > \beta_\alpha(w = 0.5)$. The least favorable null hypothesis is that $H_0 : \beta_\alpha(\hat{w}^*) = \beta_\alpha(w = 0.5)$. Using



() Power (β) vs Type I error (α) plot for different w values for the Gaussian setting (noisy case)

() Power (β) vs Type I error (α) plot for different w values for the Gaussian setting (noiseless case)

() Power (β) vs w plot for different Type I error (α) values for the Gaussian setting (noisy case)

() Power (β) vs w plot for different Type I error (α) values for the Dirichlet setting (noisy case)

() Power (β) vs Type I error (α) plot for different w values for the Dirichlet setting (noisy case)

() Power (β) vs Type I error (α) plot for different w values for the Dirichlet setting (noiseless case)

Figure 2: ROC curves and β vs w plots for simulation experiments

previous notation, the test statistic will be denoted by $T_a(w)$ under the alternative hypothesis and $T_0(w)$ under the null hypothesis.

McNemar's test will be used to compare the two predictors (referred to as C_1 and C_2 with w=0.5 and w=w^* at a fixed α value.

For the noisy version of the Gaussian setting at allowable type I error 0.05 for the two tests, when comparing the null hypothesis that $H_0 : \beta_\alpha(\hat{w}^*) = \beta_\alpha(w = 0.5)$ against the alternative $H_A = \beta_\alpha(\hat{w}^*) > \beta_\alpha(w = 0.5)$, the p-value is $p < 1.09E - 24$ which indicates the power using estimate of optimal w^* is significantly greater than the power when using $w = 0.5$.

10 Conclusion

The tradeoff between Fidelity and Commensurability and the relation to the weighted raw stress criterion for MDS were both investigated with several simulations and experiments on real data. For hypothesis testing as the exploitation task, the three approaches were compared in terms of testing power. The results indicate that the joint optimization (JOFC) approach is superior to CCA and P∘M, and is also robust to spurious correlations CCA suffers from. Also 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 unweighted raw stress for omnibus matrix embedding. The JOFC approach is quite versatile and can be applied to many problems where data of multiple modalities have to be made commensurate. JOFC approach was also applied to test for matches between Wiki articles and pairs of vertices in random graph data. Performance of JOFC in these simulations and experiments shows that it it is an appropriate method for these settings.

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] 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.

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

[8] 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.

[9] 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.

[10] 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.