Optimal Weighting for Joint Optimization of Fidelity and Commensurability in Tests of Matchedness

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, the forthcoming Manifold Matching paper by Priebe et al. [12] presents 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. The two extremes of this tradeoff are commensurability optimization prioritized over fidelity optimization and vice versa. 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. The typical multiple sensor setting is visualized in 1.

It is assumed some objects lie in some "object" space Ξ, and each sensor has another "view" of the objects. The measurements recorded by the i^{th} sensor lie in some "measurement space" Ξ_i. The usual approach in pattern recognition is to use feature extractors on the spaces to get a feature representation in Euclidean space and use classical pattern recognition tools to carry out the exploitation task. The alternative approach is to acquire dissimilarities between the group of objects, and use the dissimilarities to either find an embedding in a low-dimensional Euclidean space where classic statistical tools are available for inference or use dissimilarity-based versions of pattern recognition tools [11]. The embedding approach in a low-dimensional Euclidean space will be used where the small number of embedding dimension allow us to avoid "curse of dimensionality". Also, the embeddings of dissimilarities from different conditions will be required to be "commensurate" so that sensor measurements can be compared or jointly used in inference. This is accomplished by maps $\rho_k, k = 1, \ldots, K$ from measurement spaces Ξ_k to a low-dimensional commensurate space $\mathcal{X}$ visualized in 1. Learning these maps from data is an important portion of the proposed approach.

*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

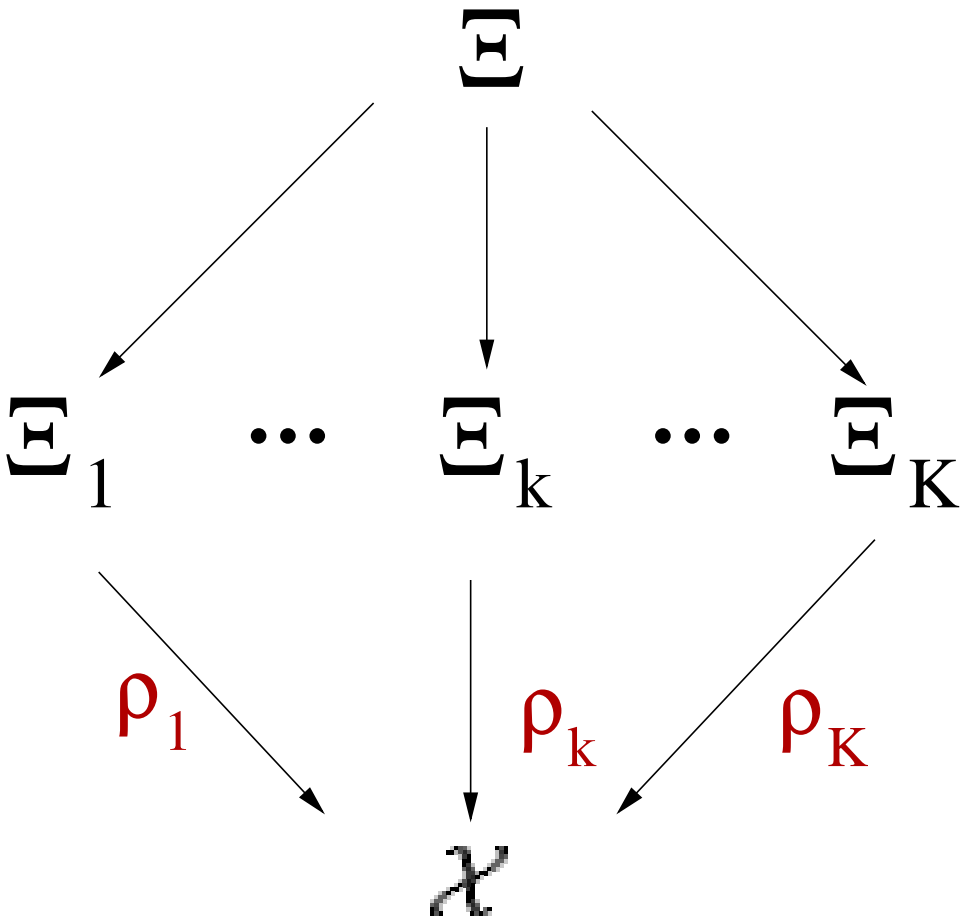


Figure 1: Multiple Sensor setting

There are quite a few real life cases, where the data is acquired or available exclusively in dissimilarity representation instead of feature representation [1, 11, 14]. 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. Multidimensional scaling computes a configuration of points that has interpoint distances as close as possible to the dissimilarities in dissimilarity representation. Different criterion functions can be used to measure how close the distances are to the given dissimilarities, the optimization of which leads to different embedded configurations. Among these, configurations that are as close as possible to (in some sense) optimal for the exploitation task that is of interest are preferred. 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:

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

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).

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.

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

To each pair of measurements x_{ik}, x_{jk} in the same space, a dissimilarity value $\delta_{ijk} = \delta\{x_{ik}, x_{jk}\}$, can be assigned, possibly dependent 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 do inference on 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" [12]:

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

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 $\boldsymbol{\Delta}^{new} = \{\delta_{ik}^{new}; i = 1, \ldots, n;\ k = 1, \ldots, K\}$ where δ_{ik}^{new} 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 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. However, the proposed JOFC approach is generalizable to $K > 2$ and simulation results for a data setting where $K = 3$ will be presented.

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. Some specific examples include:

- If the objects are wiki documents, a condition could be the textual content of the wiki document and another condition could be the wiki hyperlink graph. "Matched" could mean two wiki articles "are on the same topic".
- The condition of a text document can be the language it is in and "matched" could mean two documents "are about the same topic" or translations of each other.
- For photos, "conditions" are different acquisition conditions and "matched" photos mean they are "of the same person". Acquisition conditions could be
  - indoor lighting vs outdoor lighting
  - two cameras of different quality
  - passport photos and airport surveillance photos.
- objects in a single space with multiple dissimilarities each of which is considered a "separate" condition, where dissimilarities may be judged by different people, or measured using separate sensors.

3 Two models for generating data

Here 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 $X_{ik} \sim^{iid} MVNormal(\boldsymbol{\alpha_i}, \Sigma)$ 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 2.

The parameter r controls the variability between "matched" measurements. If r is large, it is expected that the distance between matched measurements X_{i1} and X_{i2} to be stochastically smaller than X_{i1} and $X_{i'2}$ 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 $S^p = \{\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^{(p+1)}, \sum_{l=1}^{p} x_l = 1\}$ 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 $X_{ik} \sim^{iid} Dirichlet(r\boldsymbol{\alpha}_i + \mathbf{1}_{p+1})$ represent K measurements. See Figure 3.

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

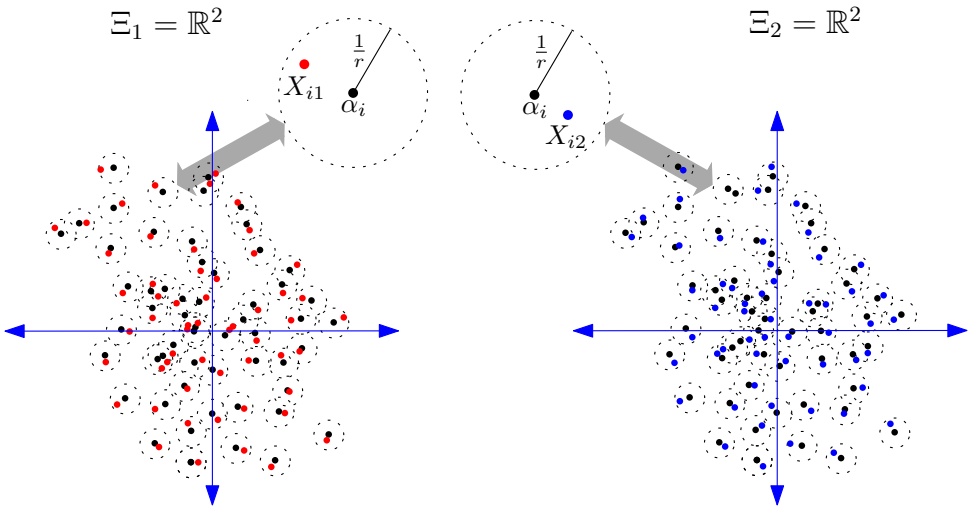


Figure 2: For the Gaussian setting (Section 3.1), the α_i are denoted by black points and the X_{ik} are denoted by red and blue points respectively.

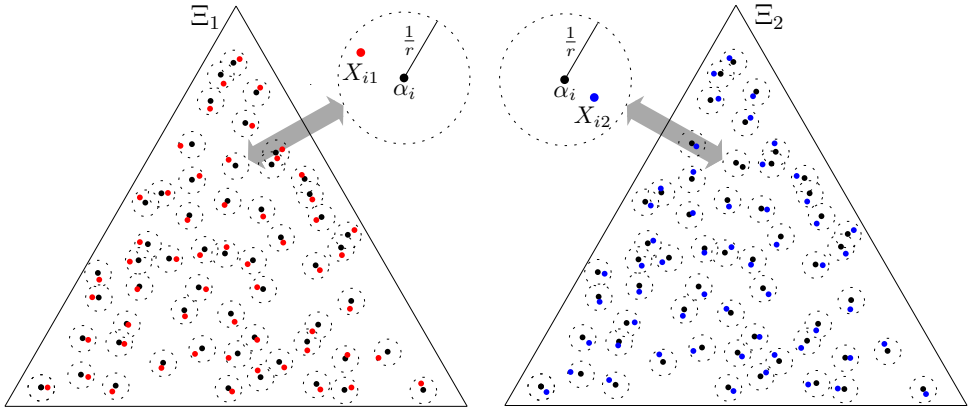


Figure 3: For the Dirichlet setting (Section 3.2), the α_i are denoted by black points and the X_{ik} are denoted by red and blue points respectively.

3.3 Noise

Measurements 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. The noisy measurements will be represented by the random vectors

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

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 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 "manifold matching", which is defined as simultaneous "manifold learning" and "manifold alignment" – identifying embeddings of multiple disparate data sources into the same low-dimensional space where joint inference can be pursued. The inference task either requires the fusion of data from disparate sources(as in the test of matchedness) or gathers performance gains from the fusion. The assumption is that the data in each measurement space lie approximately in a low dimensional manifold. An effort is made to match the low-dimensional manifolds so that matched measurements from different conditions are as close as possible to each other. The embeddings from the dissimilarities in the measurement spaces into the commensurate space are implicit mapping from the measurement space(Ξ_k) to the commensurate space($\mathcal{X}$ in 1). The learning problem involves estimating these maps(whether implicit as in the dissimilarity representation of multiple sensor measurements or explicit in the feature representation setting) from a training data of matched measurements.

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.

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. In this approach, the mappings $\{\rho_k, ; k = 1, \ldots, K\}$ in 1 are not explicitly defined.

In any exploitation task that necessitates such an matching of manifolds or where the matching is expected to improve performance, the omnibus embedding approach can be used to embed the observations in a single space where they are commensurate.

4.1 Out-of-sample Extension

Manifold learning algorithms computes embeddings of training data on low-dimensional manifolds, however they may not always give a map as output to be used to compute embeddings of new points. Applying manifold learning algorithms to the combination of training and test data will give us such a map, but this would need to be repeated for any new test point, and embeddings of training dissimilarities would be partially dependent on test data. The solution is to extend the embeddings of training dissimilarities by out-of-sample embedding the test dissimilarities. In the hypothesis testing task, the dissimilarities between a new pair of test observations and the previous nK training observations are available as test data. As embedding the original (in-sample) $nK \times nK$ dissimilarities doesn't result in a explicit map for embedding test data, out-of-sample embedding will be used to embed the test dissimilarities. Out-of-sample extension for MDS will be used throughout this paper [16].

Out-of-sample embedding can be done one observation at a time, or jointly if the dissimilarities among multiple test observations are also available.

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*

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

  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*

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

  for conditions k_1 and k_2; $\delta_{k_1 k_2}(\cdot, \cdot)$ 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. (It is possible that this choice for between-condition dissimilarities might not be optimal. This issue will be ignored in order to focus on the main problem.)

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. Although it is possible to impute these dissimilarities, the optimal imputation is an open question and ignoring these terms provides for investigation of simpler, still open questions.

Note that the omnibus embedding approach tries to jointly optimize fidelity and commensurability , by minimization of some measure of discrepancy between the given dissimilarities (all of which are between-condition or within-condition dissimilarities) and the distances of the embedded configuration. This is most obvious in the raw stress version of MDS, since 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). The omnibus matrix M is a partitioned matrix consisting of matrices from different conditions ($k = 1, 2$), the entries of the matrix will be indexed by 4-tuple i, j, k_1, k_2 which refers to the entry in the i^{th} row and j^{th} column of the submatrix in the k_1^{th} row partition and k_2^{th} column partition. For example, the entry $M_{2n,n}$ will have the indices $\{i, j, k_1, k_2\} = \{n, n, 2, 1\}$ in the new indexing scheme. $D(\cdot)$ and W, which are the same size as M, follow the same 4-tuple indexing. Then,

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

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:

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

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

Note that for purpose of minimization, setting all weights $w_{ijk_1k_2}$ equal is equivalent to the unweighted raw stress $\sigma(X)$:

$$\sigma(X) = \sum_{1 \leq s \leq n; 1 \leq t \leq n} (d_{st}(X) - \delta_{st})^2 \tag{5}$$

6 Alternative Methodologies

For the optimization of commensurability with fidelity as secondary priority, an alternative method is Canonical Correlational Analysis (CCA) [5], 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. CCA finds a basis for these subspaces iteratively: For each new component in the basis, CCA finds the pair of directions that maximizes correlation with the constraint that the projections along the new directions are uncorrelated with projections along previous components. The latter constraint results in additional preservation of fidelity for each new direction. 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. Since the data is originally in a dissimilarity representation, it is possible to directly embed in the low-dimensional space and use Procrustes Analysis to find a mapping between the two separate embeddings. The equivalence of PCA and Classical Multidimensional Scaling [14] under certain conditions suggests that this approach is the right analog of Procrustes $\circ$ PCA to utilize when in a dissimilarity setting.

The omnibus embedding approach is expected to be more powerful for the exploitation task than either of the sequential optimizations, since the exploitation task (testing matchedness) requires both optimization of fidelity and commensurability.

6.1 Procrustes Analysis on Multidimensional Scaling Embeddings

Since separate condition dissimilarities are available, a straightforward approach is to embed each conditional dissimilarity matrix, Δ_1 and Δ_2, separately in d-dimensional Euclidean space (call these embedded configurations X_1 and X_2, respectively) and then find a mapping function $\rho : \mathbb{R}^d \to \mathbb{R}^d$ that maps each point in X_2 to approximately its corresponding point in X_1. This approach can be seen as a specific example of the general setting where the commensurate space is d-dimensional Euclidean space and ρ_1 in 1 is the identity map.

Estimation of ρ is carried out using Procrustes Analysis on training data. Procrustes Analysis [13] finds a orthonormal matrix $\mathbf{Q}^*$ that minimizes the sum of squared distances between the target configuration X_1 and the configuration X_2 transformed by $\mathbf{Q}^*$, i.e.,

$$\mathbf{Q}^* = \underset{Q^TQ=I}{\arg\min} \|X_1 - X_2Q\|_F$$

where $\|\cdot\|_F$ is the Frobenius norm on matrices. The map ρ estimated by the linear map $\mathbf{Q}^*$ makes the separate MDS embeddings as commensurate as possible. Once such a mapping is computed, one can out-of-sample embed new dissimilarities for each condition (separately) and use $\mathbf{Q}^*$ to make the embeddings commensurate. One can then compute the test statistic τ (the distance between commensurate embeddings) for the hypothesis testing problem. This approach will be referred to as P$\circ$M - Procrustes $\circ$MDS.

Note that the Procrustes transformation $\mathbf{Q}^*$ is limited to a linear transformation consisting of rotation and reflection and possibly also scaling components. The optimal mapping might very well be non-linear. If a larger class of mappings is considered, this would result

in a smaller model bias but also larger variability for the mapping function. By only considering the class of linear transformations, it is possible to learn $\mathbf{Q}^*$ with the limited sample size.

6.2 Canonical Correlational Analysis on Multidimensional Scaling Embeddings

Again MDS is used to compute embedding configurations, X_1 and X_2. It is desirable to embed into the highest dimensional space possible ($\mathbb{R}^{d'}$ where $d' = p + q$ for the Gaussian and Dirichlet settings) to preserve as many of the signal dimensions as possible (at the risk of possibly including some noise dimensions). CCA [5], then, yields two mappings $\mathcal{U}_1$ and $\mathcal{U}_2$ that map these embeddings in $\mathbb{R}^{d'}$ to the low-dimensional commensurate space ($\mathbb{R}^d$).

Canonical Correlational Analysis

Let X and Y be two s-dimensional random vectors. If one wants to find the pair of linear projection operators $U_1 : \mathbb{R}^s \to \mathbb{R}$, $U_2 : \mathbb{R}^s \to \mathbb{R}$ that maximize correlation between the projections of X and Y, CCA finds the solution as stated in the optimization problem

$$\hat{u}_1, \hat{u}_2 = \arg \max_{u_1 \in \mathbb{R}^s, u_2 \in \mathbb{R}^s} \frac{E[u_1^T X Y^T u_2]}{E[u_1^T X X^T u_1] E[u_2^T Y Y^T u_2]}$$

with the constraints $E[u_1^T X X^T u_1] = 1, E[u_2^T Y Y^T u_2] = 1$ for uniqueness. The constraints simplify the optimization function to

$$\arg \max_{u_1 \in \mathbb{R}^s, u_2 \in \mathbb{R}^s} E[u_1^T X Y^T u_2].$$

If the projections are to a pair of d-dimensional linear subspaces, the additional pairs of projection vectors can be computed sequentially, with the constraints that the projections along the new directions are uncorrelated with projections along previous directions. That is, i^{th} pair of directions that maximize correlation is computed by

$$\hat{u}_{1(i)}, \hat{u}_{2(i)} = \arg \max_{u_{1(i)}, u_{2(i)} \in \mathbb{R}^s} E[u_{1(i)}^T X Y^T u_{2(i)}].$$

subject to constraints $E[u_{1(i)}^T X X^T u_{1(i)}] = 1$, $E[u_{2(i)}^T Y Y^T u_{2(i)}] = 1$, $E[u_{1(i)}^T X X^T u_{1(j)}] = 0$, $E[u_{2(i)}^T Y Y^T u_{2(j)}] = 0 \; \forall \quad j = 1, \ldots, i-1$. For sample CCA, $E[XX^T]$,$E[YY^T]$ and $E[XY^T]$ are replaced with their sample estimates. The direction vectors $\hat{u}_{1(i)}, \hat{u}_{2(i)}, i = 1, \ldots, d$ form the rows of projection matrices which represent the mappings $\mathcal{U}_1$ and $\mathcal{U}_2$.

Note that s, the dimension of X and Y, is the embedding dimension d' in the CCA approach.

As in P$\circ$M, new dissimilarities are out-of-sample embedded and mapped to a commensurate space by maps provided by CCA. The test statistic can now be computed and the null hypothesis is rejected for "large" values of the test statistic τ as in Section 6.1.

6.3 Relation of $P \circ M$ and Joint Optimization of Fidelity and Commensurability

Suppose the weights are chosen to be $w_{ijk_1k_2} = w$ for commensurability terms and $w_{ijk_1k_2} = 1 - w$ for fidelity terms in equation (4). For the resulting weight matrix W, define

$$f_w(D(\cdot), M) = \sigma_W(\cdot) \tag{6}$$

where M is the omnibus matrix obtained from a given pair of dissimilarity matrices, Δ_1 and Δ_2, as in equation (3). As w goes to 0, the configuration embedded by JOFC converges to a configuration equivalent to (up to rotation and reflection) the configuration embedded by P$\circ$M.

Theorem 1. *Define* $\sigma(\cdot) = \sigma_{W=\mathbf{1}}(\cdot)$ *(unweighted raw stress) where* $\mathbf{1}$ *is a matrix of 1's. Let* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *be the corresponding* $n \times p$ *configuration matrices with column means of* $\mathbf{0}$ *(obtained from separately embedding* Δ_1 *and* Δ_2 *by minimizing the raw stress* $\sigma(\cdot)$ *). Let* $\mathbf{Q} = \arg\min_{\mathbf{P^T P = P P^T = I}} ||\mathbf{X}_1 - \mathbf{X}_2\mathbf{P}||^2$ *,* $\tilde{\mathbf{X}}_2 = \mathbf{X}_2\mathbf{Q}$*, and let* $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \tilde{\mathbf{X}}_2 \end{bmatrix}$.

For $w > 0$*, let* $\mathbf{Y}_w = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix}$ *be a* $2n \times p$ *configuration matrix obtained by minimization of* $f(\mathcal{Y}, M) = (1-w)\left(\sigma(\mathcal{Y}_1) + \sigma(\mathcal{Y}_2)\right) + w||\mathcal{Y}_1 - \mathcal{Y}_2||^2$ *with respect to* $\mathcal{Y} = \begin{bmatrix} \mathcal{Y}_1 \\ \mathcal{Y}_2 \end{bmatrix}$ *with the constraint that* $\mathcal{Y}_1$ *and* $\mathcal{Y}_2$ *are two* $n \times p$ *configuration matrices having column means of* $\mathbf{0}$*. Then,*

$$lim_{w \to 0} \mathbf{Y}_w = \mathbf{XR}$$

for a $p \times p$ *orthonormal matrix* $\mathbf{R}$*. (*$\mathbf{R}$ *is a transformation matrix with a rotation and possibly a reflection component.)*

6.4 Relation of CCA and Commensurability

Theorem 2. *Let* $\mathcal{U}$ *be the set of all orthogonal d-frames (ordered set of d linearly independent vectors) of* $R^{d'}$*. Let* X_1 *and* X_2 *be two* $n \times d'$ *(configuration) matrices that are perfectly "matched" (there exists a transformation matrix* $\mathbf{Q}$ *such that* $\|X_1\mathbf{Q} - X_2\| = 0$*). If commensurability is defined as in equation* (5)*, where the embedded configurations are* $\tilde{X}_1 = X_1U_1$ *and* $\tilde{X}_2 = X_2U_2$ *for some* $U_1 \in \mathcal{U}$ *and* $U_2 \in \mathcal{U}$*, and the original dissimilarities are* $D(X_1)$ *and* $D(X_2)$*, CCA on* X_1 *and* X_2 *gives* $\mathbf{U}_1 \in \mathcal{U}$ *and* $\mathbf{U}_2 \in \mathcal{U}$*, the two elements of* $\mathcal{U}$ *that maximize commensurability, subject to* $U_1^T X_1^T X_1 U_1 = I_d$ *and* $U_2^T X_2^T X_2 U_2 = I_d$ *(*I_d *is the* $d \times d$ *identity matrix).*

7 Related Work

There have many efforts toward solving "manifold alignment", which is a related problem. "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.

Wang and Mahedavan [17] suggest an approach that uses embedding followed by Procrustes Analysis to find a map to a commensurate space. Given a paired set of points, Procrustes Analysis [13], finds a transformation from one set of points to another in the same space that minimizes sum of squared distances, subject to some constraints on the transformation. In the case mentioned in [17], the paired set of points are corresponding low-dimensional embeddings of kernel matrices. For the embedding step, they made the choice of using Laplacian Eigenmaps, though their algorithm allows for any appropriate embedding method.

Zhai et al. [18] finds two projection matrices to minimize three terms in an energy function similar to the JOFC approach (see Section 4). One of the terms is the *correspondence preserving term* which is the sum of the squared distances between corresponding points and is analogous to the commensurability error term in JOFC. The other two terms are *manifold regularization terms* and consist of the reconstruction error for a Locally Linear Embedding of the projected points. These terms, analogous to fidelity, make sure the projections in the lower dimension retain the structure of the original points. For fidelity error terms in the JOFC setting, this is done by preserving dissimilarities. For manifold regularization terms, this is done by preserving the local neighborhood of points, such that close points are

not mapped apart. Ham and Lee solve the problem in semi-supervised setting by a similar approach, by minimizing a cost function of three terms, two terms for fidelity of embedding, one term of commensurability.

Another view to look at the data from different sources is to consider disparate data as different views to be reconciled. According to this view, for observations of n objects under K conditions, n points are embedded instead of nK points. Choi et al. [3] use the Markov random walk interpretation of multiple kernel matrices to combine into one kernel matrix. Many other "Multiple Kernel Learning" methods exist in the literature [8–10].

Another approach is Three-way Multidimensional scaling [1, 2]. This approach assumes the different "conditions" of the data are linear transformations of a single configuration and aims to find this single configuration. In the JOFC approach, one would first embed the in-sample dissimilarities via the three-way MDS, which would give as the linear transformations that map from group configuration to individual configurations under each condition. This is followed by out-of-embedding the OOS dissimilarities, and use the inverse of the transformation matrices to find the out-of-sample embeddings with respect to the group configuration. Since the out-of-sample embeddings are commensurate, the test statistic can be computed as the distance between the OOS embeddings.

8 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.

9 Definition of w^*

Let $\mathbf{F}_m$ be the joint distribution of $X_m = \begin{bmatrix} X_{1m} \\ X_{2m} \end{bmatrix}$ and $\mathbf{F}_u$ be the joint distribution of $X_u = \begin{bmatrix} X_{1u} \\ X_{2u} \end{bmatrix}$ where X_{1m}, X_{2m} are the random vectors of dimension d' for the matched observation pair and X_{1u}, X_{2u} are the random vectors of dimension d' for the unmatched data pair. The constraint on $\mathbf{F}_m$ is that correlation matrix of X_{1m}, X_{2m} is non-zero, while the constraint on $\mathbf{F}_u$ is that correlation matrix of X_{1u}, X_{2u} is zero.

Let $\mathcal{T}$ denote the random variable for a data matrix $(2n \times d')$ for an i.i.d. sample of $\begin{bmatrix} X_{1m} \\ X_{2m} \end{bmatrix}$ and let $\mathbf{T}_{mc}$ denote realization of $\mathcal{T}$ for any Monte Carlo replicate.

For the exploitation task at hand, it is assumed that either

- a sample of $\mathcal{T}$ ($\mathbf{T}_{mc}$) and a sample of X_m and X_u $\left(\boldsymbol{x_m} = \begin{bmatrix} \boldsymbol{x}_{1m} \\ \boldsymbol{x}_{2m} \end{bmatrix}, \boldsymbol{x_u} = \begin{bmatrix} \boldsymbol{x}_{1u} \\ \boldsymbol{x}_{2u} \end{bmatrix}\right)$ are given and Euclidean distances are computed between $\boldsymbol{x}_{.m}$ and the rows in $\mathbf{T}_{mc}$ and Euclidean distances between $\boldsymbol{x}_{.u}$ and the rows in $\mathbf{T}_{mc}$ to form dissimilarity matrices Δ_m and Δ_u, or
- values of dissimilarity matrix-valued function D of the sample of X_m, X_u and $\mathbf{T}_{mc}$:

$$\boldsymbol{\Delta_m} = D\left(\begin{bmatrix} \mathbf{T}_{mc} \\ \boldsymbol{x}_{1m} \\ \boldsymbol{x}_{2m} \end{bmatrix}\right)$$

$$\boldsymbol{\Delta_u} = D\left(\begin{bmatrix} \mathbf{T}_{mc} \\ \boldsymbol{x}_{1u} \\ \boldsymbol{x}_{2u} \end{bmatrix}\right)$$

  are available, where the $(s,t)^{th}$ entry of $D(\cdot)$ ($d_{st}(\cdot)$ in equation (1)) is the Euclidean distance between the s^{th} and t^{th} rows of its argument.

Either way, the dissimilarity matrices are defined to be $\Delta_m\left(\begin{bmatrix} \mathcal{T} \\ X_{1m} \\ X_{2m} \end{bmatrix}\right)$ and $\Delta_u\left(\begin{bmatrix} \mathcal{T} \\ X_{1u} \\ X_{2u} \end{bmatrix}\right)$ as two matrix-valued random variables $\Delta_m : \Omega \to \mathbf{M}_{n\times n}$,$\Delta_u : \Omega \to \mathbf{M}_{n\times n}$ for the sample space (Ω).

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 $\hat{X}_{1u}, \hat{X}_{2u}$ is

$$\hat{X}_{1u}, \hat{X}_{2u} = \underset{\acute{X}_{1u}, \acute{X}_{2u}}{\arg\min}\left[\min_{\acute{\mathbf{T}}} f_w\left(D\left(\begin{bmatrix} \acute{\mathbf{T}} \\ \acute{X}_{1u} \\ \acute{X}_{2u} \end{bmatrix}\right), \Delta_u\right)\right]$$

where there is an implicit dependence on $\mathbf{T}$, because Δ_u depends on $\mathbf{T}$. A similar expression gives the embedding for the matched pair $\hat{X}_{1m}$, $\hat{X}_{2m} = \arg\min_{\acute{X}_{1m}, \acute{X}_{2m}}\left[\min_{\acute{\mathbf{T}}} f_w\left(D\left(\begin{bmatrix} \acute{\mathbf{T}} \\ \acute{X}_{1m} \\ \acute{X}_{2m} \end{bmatrix}\right), \Delta_u\right)\right]$.
Assuming these minima exist and are unique, the mappings $\hat{X}_{1m} : \omega \to \mathbf{R}^{d'}$, $\hat{X}_{2m} : \omega \to \mathbf{R}^{d'}$, $\hat{X}_{1u} : \omega \to \mathbf{R}^{d'}$, $\hat{X}_{2u} : \omega \to \mathbf{R}^{d'}$ are measurable maps, $\hat{X}_{1m}$, $\hat{X}_{2m}$, $\hat{X}_{1u}$, $\hat{X}_{2u}$ are random vectors.

Consider the test statistic $\tau = d(\hat{X}_{1.}, \hat{X}_{2.})$. Under null hypothesis of matchedness, the distribution of the statistic is governed by the distribution of $\hat{X}_{.m}$, under the alternative it is governed by $\hat{X}_{.u}$.

Denote by F_Y the cumulative distribution function of Y where Y can be any function of $\hat{X}_m$ or $\hat{X}_u$.

Then

$$\beta_\alpha(w) = 1 - F_{d(\hat{X}_{1u}, \hat{X}_{2u})}(F^{-1}_{d(\hat{X}_{1m}, \hat{X}_{2m})}(1-\alpha)).$$

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

$$w^* = argmax_w \beta_\alpha(w).$$

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 10. Investigation of some properties of w^* is included in section 10 . 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 $\beta_\alpha(w)$. Note that in a general setting, finding the exact value of w^* is intractable. The estimate $\hat{w}^*$ will be based on noisy evaluations of $\beta(w^*)$. A Monte Carlo simulation is run in Section 10 to find the estimate of this function at various values of w.

9.1 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 $(0,1)$. By assuming stochastic continuity of $T_a(w)$ and $T_0(w)$ in the interval $(0,1)$, ie assuming the probability of a jump discontinuity of $T_0(w)$ at a particular w_0 is 0, the continuity of $AUC(w)$ will be proven.

Theorem 3. *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

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

and
$Pr[T(w) > 0]$ *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.

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

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

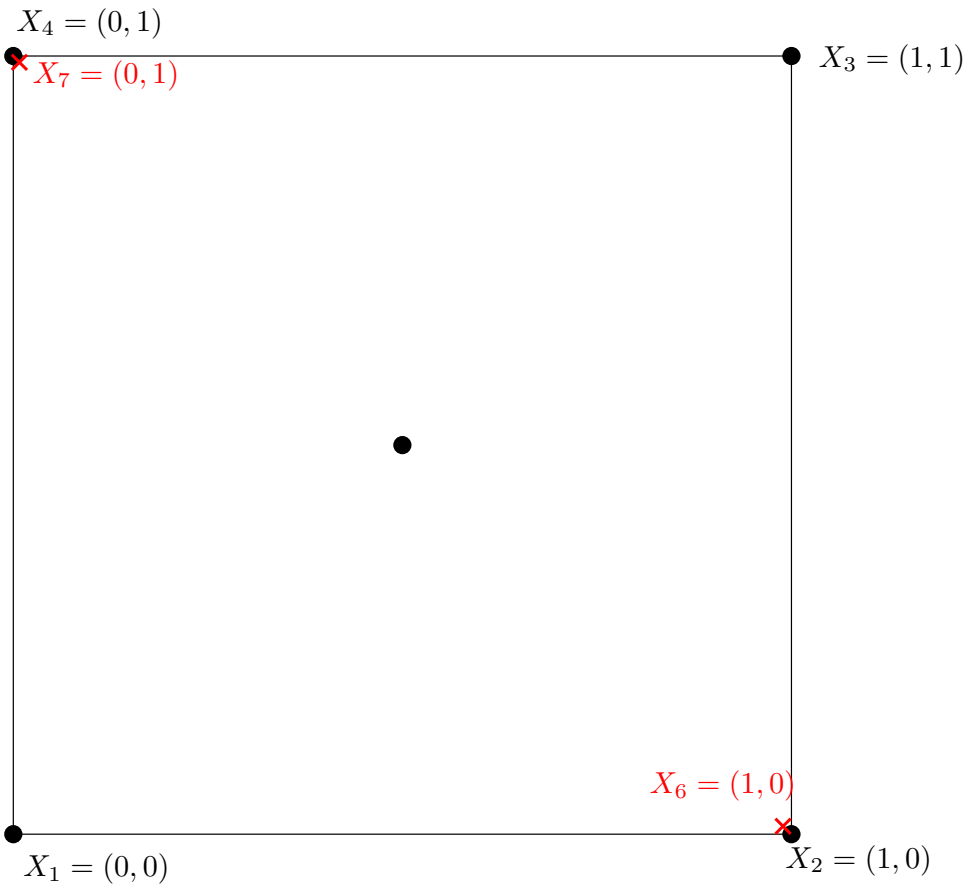
$$\|Pr[T(w+\delta) > 0] - Pr[T(w) > 0]\| \le \max(Pr[T(w+\delta) > 0 | T(w) \le 0], Pr[T(w+\delta) \le 0 | T(w) > 0]) < \epsilon$$

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

□

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

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



The main assumption used in the proof is the stochastic continuity of $T_0(w)$ and $T_a(w)$. In the constructed example that is in the detour to follow, a certain symmetry in the realized configuration is assumed, which results in a discontinuity at a particular point of w. Any change in this symmetry will give a discontinuity at a different w. When the probability distribution of the data that leads to the dissimilarities is continuous, it can be argued that the probability of a jump discontinuity at a particular w has measure zero which is equivalent to stochastic continuity.

9.1.1 A short detour : Discontinuity in weighted raw stress OOS configurations

Note that it is possible to have multiple local minima in the embedding step(see example in [15]). It is possible to construct an example where w controls which of the local minima is the global minimum among the configurations of $\hat{X}_.$.

Consider five in-sample points in $\mathbb{R}^2$ with locations $X_1 = (0, 0)$, $X_2 = (1, 0)$, $X_3 = (1, 1)$ and $X_4 = (1, 0)$, $X_5 = (.5, .5)$ and two out-of-sample points with coordinates $X_6 = (1, 0)$ and $X_7 = (0, 1)$. Suppose X_6 is matched with X_2 and X_7 is matched with X_4. Denote the Euclidean distance matrix by D. Suppose, due to noise, or due to dissimilarities not being Euclidean distances, the dissimilarity matrix is

$$D'_{ij} = \begin{cases} D_i j - 1.4 & \text{if } (i, j) \in \{(4, 6), (6, 4), (2, 7), (7, 2)\} \\ D_i j & \text{otherwise} \end{cases} .$$

Qualitatively, the three points X_1, X_7 and X_3 form a barrier which the OOS points need to cross to reach their matched counterparts.

Based on the initial configuration, the embedding coordinates of $\hat{X}_6$ might be closer to X_4 compared to X_2. This is due to a local minimum in the configuration space. If $\hat{X}_6$ starts

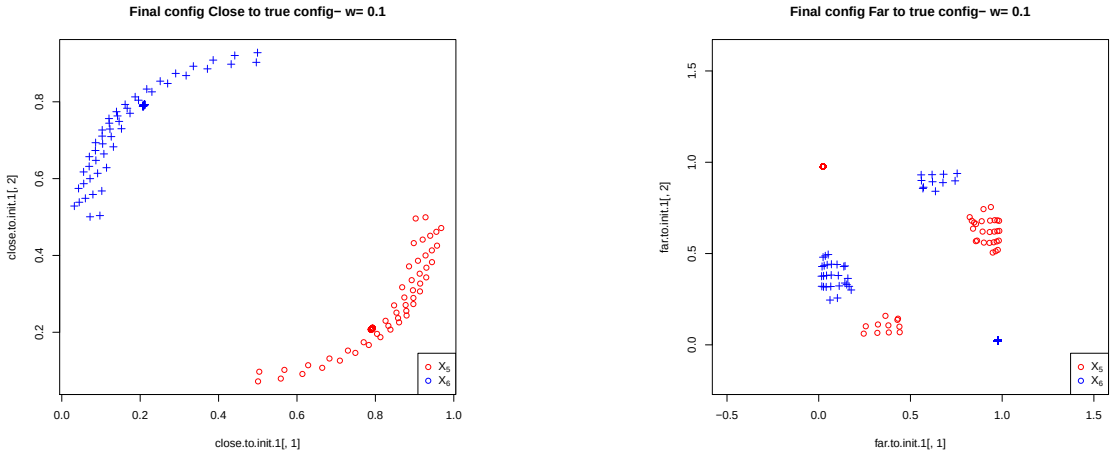


Figure 4: Final configurations for for different $w = 0.1$

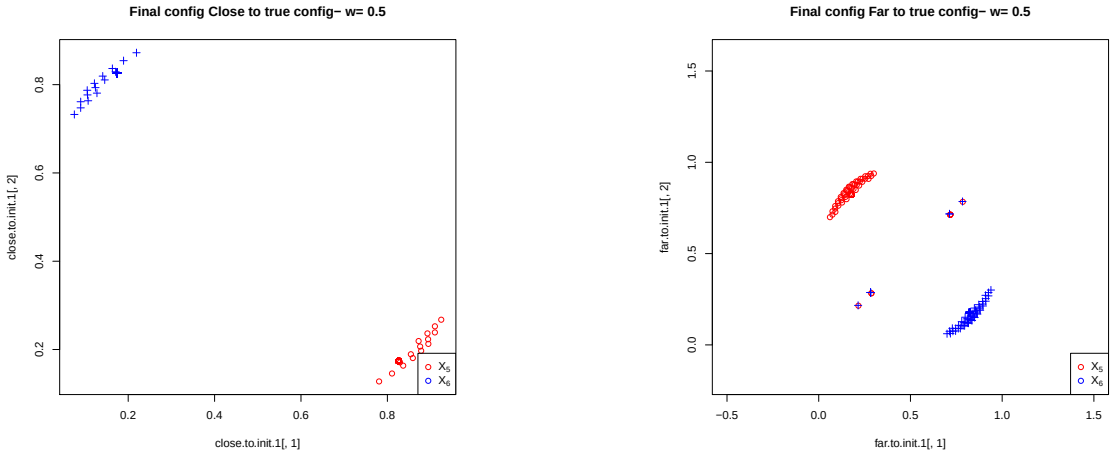


Figure 5: Final configurations for for different $w = 0.5$

from the initial configuration where it's located on the X_4 side of the y=x line at the start of optimization, it might have to cross paths with the embedding of X_1, X_3, X_7 with which it has a nonzero dissimilarity. Based on value of w, it might be easier to get out of this local minimum. In fact, depending on w can this local minimum can be a global minimum. That is, the configuration where X_6 stays on the side of X_4 instead of X_2 might have a lower stress than the configuration where X_6 is near X_2. The following plots shows the local minimum $\hat{X}_6$ ends up in, depending on initial configuration. Depending on w value, one of these minima has a lower weighted stress than the other.

w value	0.1	0.45	0.5	0.55	0.99
Local min for real config.	2.80	1.77	1.62	1.47	0.04
Alternative local min	0.39	1.53	1.65	1.74	100.00

Other than such carefully constructed examples, it is expected that w will have no effect on the ordering of the local minima according to their stress values. Therefore, the argmin among these local minima is independent of w. The minimum configuration is then a continuous function of w. By the continuity of the distance function with respect to configurations, the test statistic is continuous with respect to w. One can conclude $\beta(w)$ is

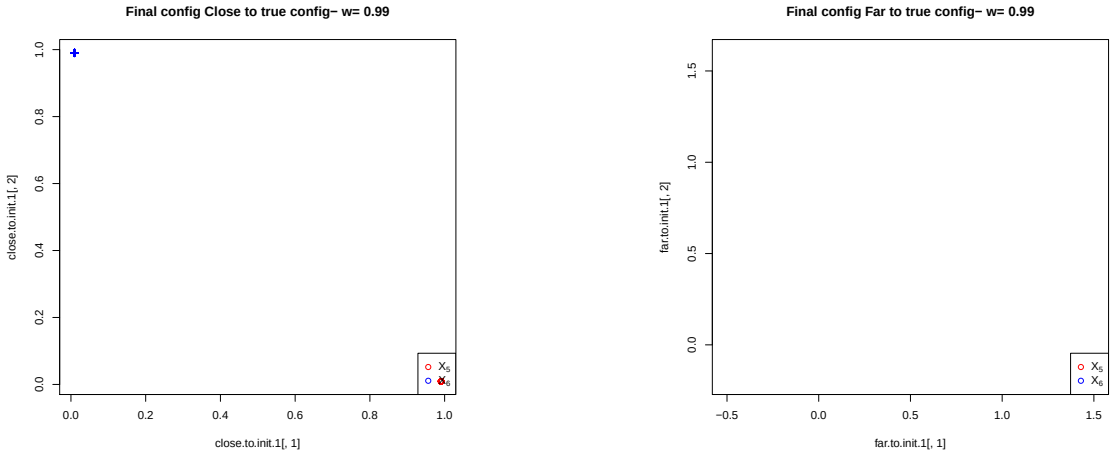


Figure 6: Final configurations for for different $w = 0.99$

a continuous function of w.

10 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). This setting will be referred to as the "noisy case". 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.
- 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 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 (6) 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.

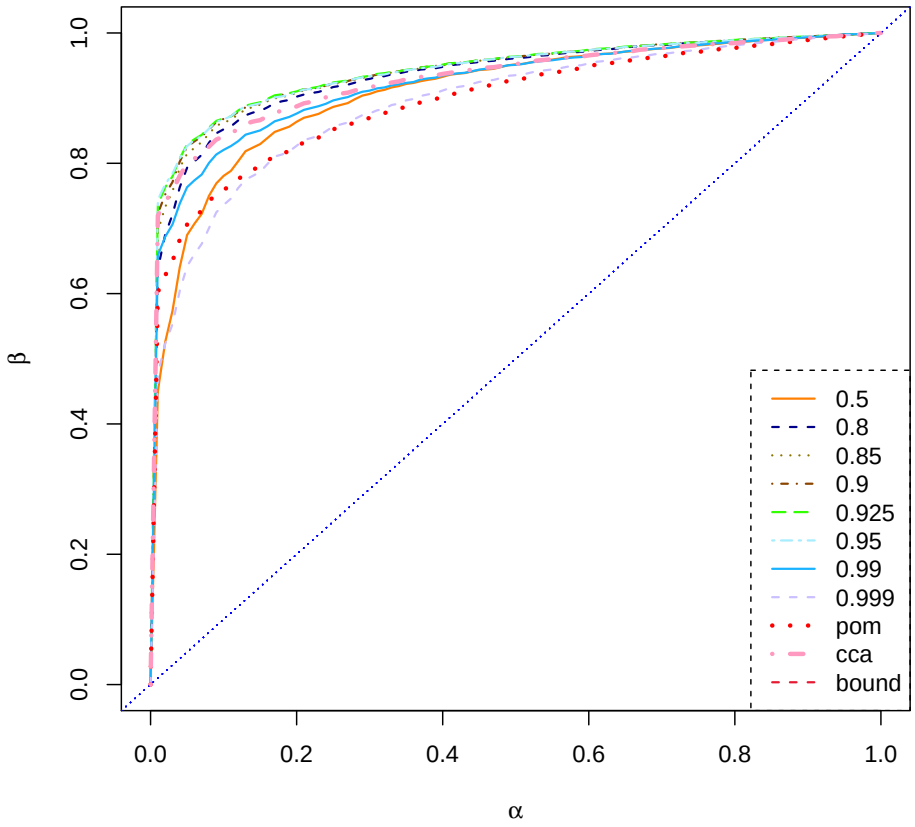


Figure 7: Power (β) vs Type I error (α) plot for different w values for the Gaussian setting (noisy case)

- 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 12 against α for the Gaussian setting. Qualititatively similar plots for the Dirichlet setting are not included for brevity. The plot in Figure 12 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

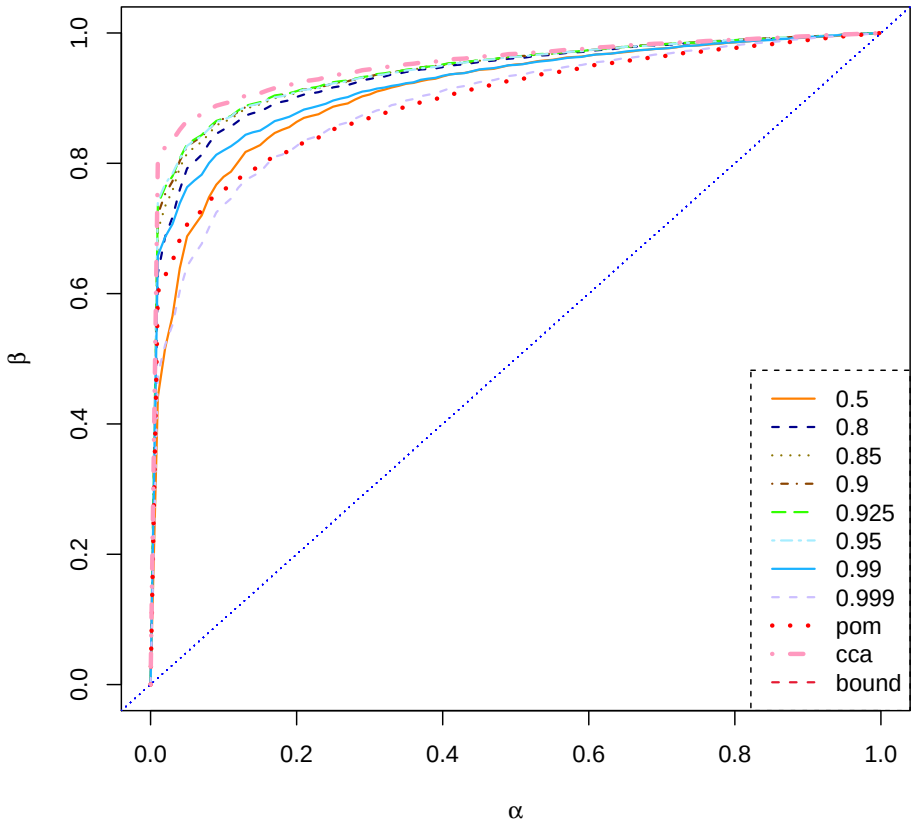


Figure 8: Power (β) vs Type I error (α) plot for different w values for the Gaussian setting (noiseless case)

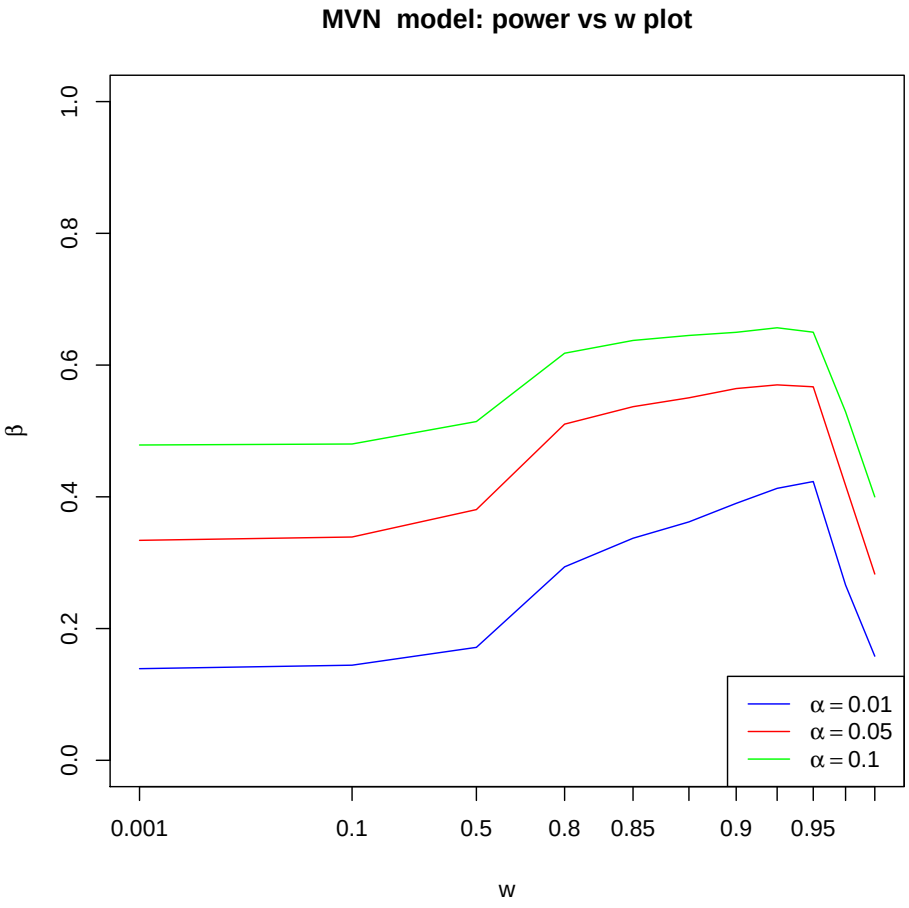


Figure 9: Power (β) vs w plot for different Type I error (α) values for the Gaussian setting (noisy case)

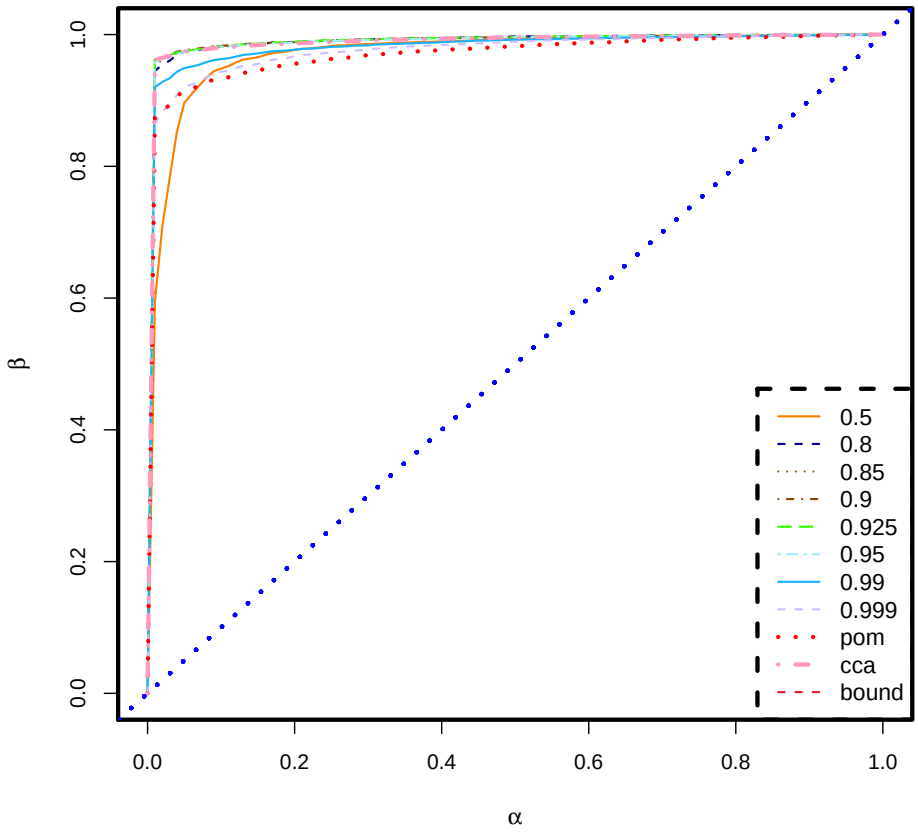


Figure 10: Power (β) vs Type I error (α) plot for different w values for the Dirichlet setting (noisy case)

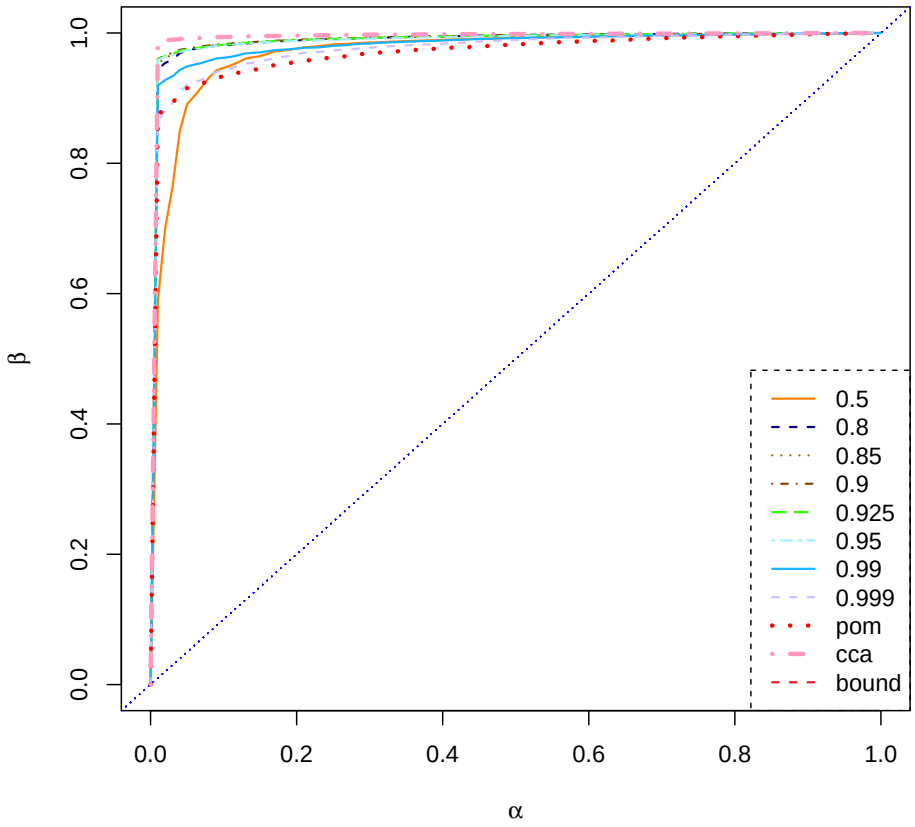


Figure 11: Power (β) vs Type I error (α) plot for different w values for the Dirichlet setting (noiseless case)

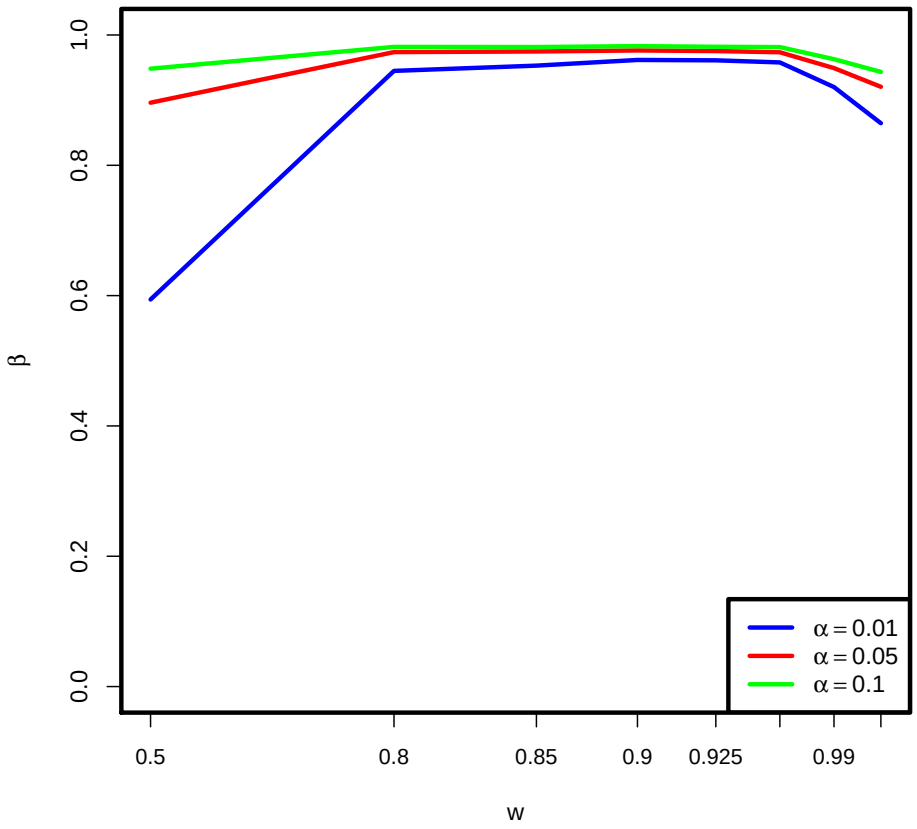


Figure 12: Power (β) vs w plot for different Type I error (α) values for the Gaussian setting (noisy case)

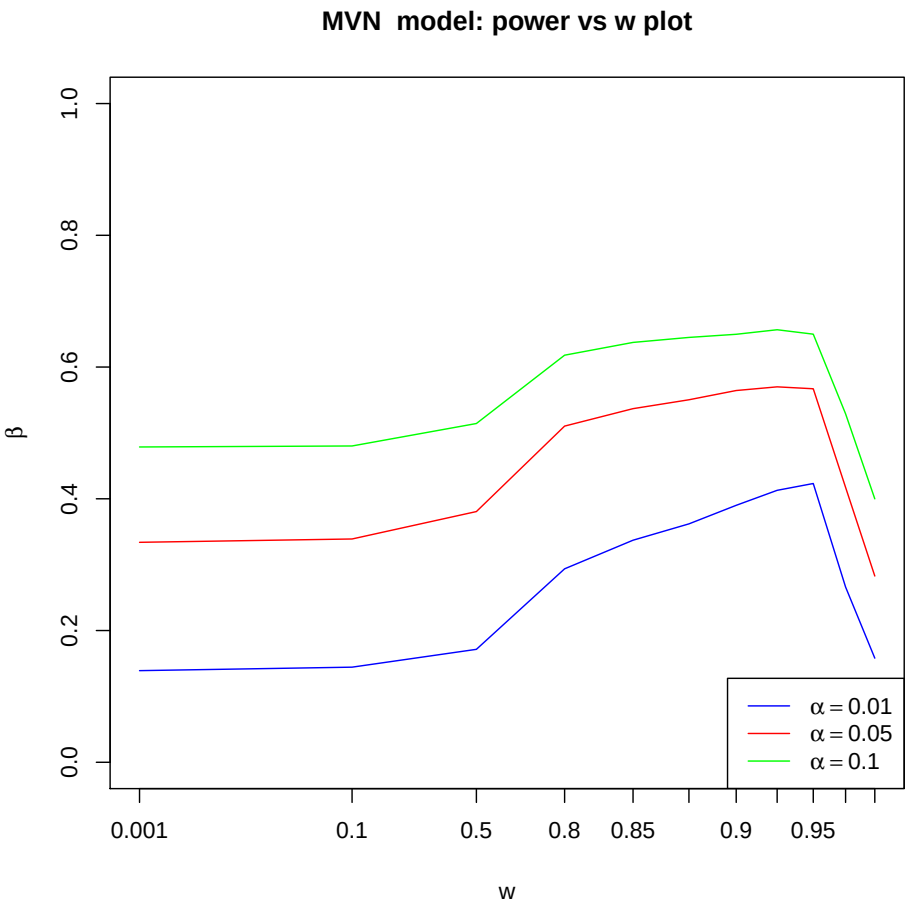


Figure 13: Power (β) vs w plot for fixed Type I error (α) values for the Gaussian setting (noisy case)

in terms of power. In Figure 13, $\beta(w)$ 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 13 for $\alpha = 0.05$, $\beta_{\alpha=0.05}(w = 0.99) \geq \beta_{\alpha=0.05}(w = 0.5)$. However, for $\alpha = 0.2$, $\beta_{\alpha=0.2}(w = 0.99) \leq \beta_{\alpha=0.2}(w = 0.5)$. 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 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 a fixed α value, one can compute two critical values $c(0.5) = max_l\{P(T_0(0.5) > c) < \alpha\}$, $c(w^*) = max_l\{P(T_0(w_2) > c) < \alpha\}$. The values of the decision function that uses these critical values, for each pair of embedded points (indexed by i, are $(\tilde{y}_i^{(1)}, \tilde{y}_i^{(2)})$, $i = 1, \ldots, m$. To compare the two statistical tests with $w = 0.5$ and w=w^* , one can prepare a 2×2 contingency-table of correct decisions and incorrect decisions made by each statistical test

(or equivalently true and false classifications made by two classifiers). Denote decision outcome as g_1 for the first statistical test and g_2 for the second statistical test. If $g_1 = True$ and $g_2 = False$ for an instance, the first test made the correct decision and the second test made the incorrect decision with regard to the null and alternative hypotheses. Consider the contingency table for a Monte Carlo replicate given by

$$G^{(l)} = \begin{array}{|c|c|} \hline e_{FF}^{(l)} & e_{TF}^{(l)} \\ \hline e_{FT}^{(l)} & e_{TT}^{(l)} \\ \hline \end{array}$$

where l is the index of the MC replicate, $e_{g_1 g_2}^{(l)}$ is equal to the number of instances at which the true hypothesis were identified correctly ($g_1 = True$) or incorrectly ($g_1 = False$) by the first test, and correctly ($g_2 = True$) or incorrectly ($g_2 = False$) by the second test in that MC replicate.

Under the null hypothesis that the two predictors have the same power at α , $Pr[(g_1 g_2) = (TF)] = Pr[(g_1 g_2) = (FT)]$, so $\sum_l I\{e_{TF}^{(l)} > e_{FT}^{(l)}\}$ will be distributed according to the binomial distribution, $\mathcal{B}(nmc, 0.5)$. ($I\{\cdot\}$ is the indicator function.)

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 $w = 0.5$.

11 The general case of more than two conditions

As it was mentioned, all of the approaches are generalizable to $K > 2$ conditions, though an ambiguity need to be resolved. The alternative hypothesis could be defined as the event that at least one of the K new dissimilarities are pairwise unmatched ($H_{A1} : \exists i, j, 1 \leq i < j \leq K : \boldsymbol{y}_i \nsim \boldsymbol{y}_j$) or it could be defined as the case that absolutely none of the K dissimilarities are pairwise matched ($H_{A2} : \forall i, j, 1 \leq i < j \leq K : \boldsymbol{y}_i \nsim \boldsymbol{y}_j$) .

For P∘M approach, one can use Procrustes analysis generalized to more than two configurations. Generalized Procrustes Analysis [4] is a well established method for finding a collection of linear transformations that minimizes the mismatch between the transformed configurations and a "mean" shape computed from these configurations.

For CCA approach, there are multiple generalizations available as the correlation between more than two configurations can be defined in multiple ways [6]. Let $X_1, \ldots, X_K$ be random vectors. Consider the first set of canonical variates to be computed, $Z_1^1, \ldots, Z_K^1$. Denote the correlation matrix of $Z_1^1, \ldots, Z_K^1$ by $\Phi^{(1)}$. The following three criteria are proposed in [6].

- SUMCOR. Maximize the sum of the elements of $\Phi^{(1)}$: $\mathbf{1}'(\Phi^{(1)})\mathbf{1}'$
- MAXVAR. Maximize the largest eigenvalue of $\Phi^{(1)}$: $\lambda_1^{(1)}$
- MINVAR. Minimize the smallest eigenvalue of $\Phi^{(1)}$: $\lambda_1^{(m)}$

One can think of all of these criteria as different norms on the correlation matrix. An interesting question that will not be addressed here is whether any of these generalizations is more appropriate for H_{A1} or H_{A2}.

To test whether JOFC and generalized CCA approaches are appropriate for this setting as well, The simulations in the previous section10 were repeated with K-condition data , generated by a multivariate normal model with $K = 3$ conditions. The alternative was chosen as H_{A1} and SUMCOR criterion was chosen as the generalization of CCA. The "Noisy" case will be investigated, i.e. q-dimensional noise vectors of magnitude c were added to the matched measurements, and "signal" vectors were multiplied by $1 - c$ The ROC curves for these simulations are shown in 14.

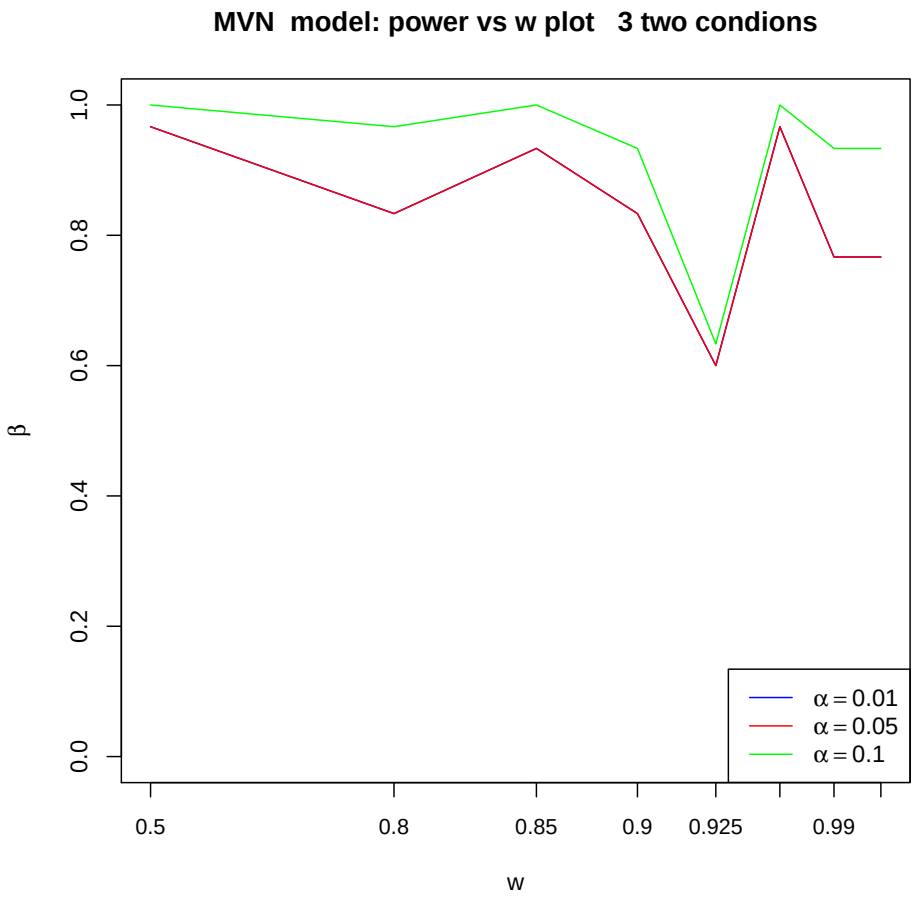
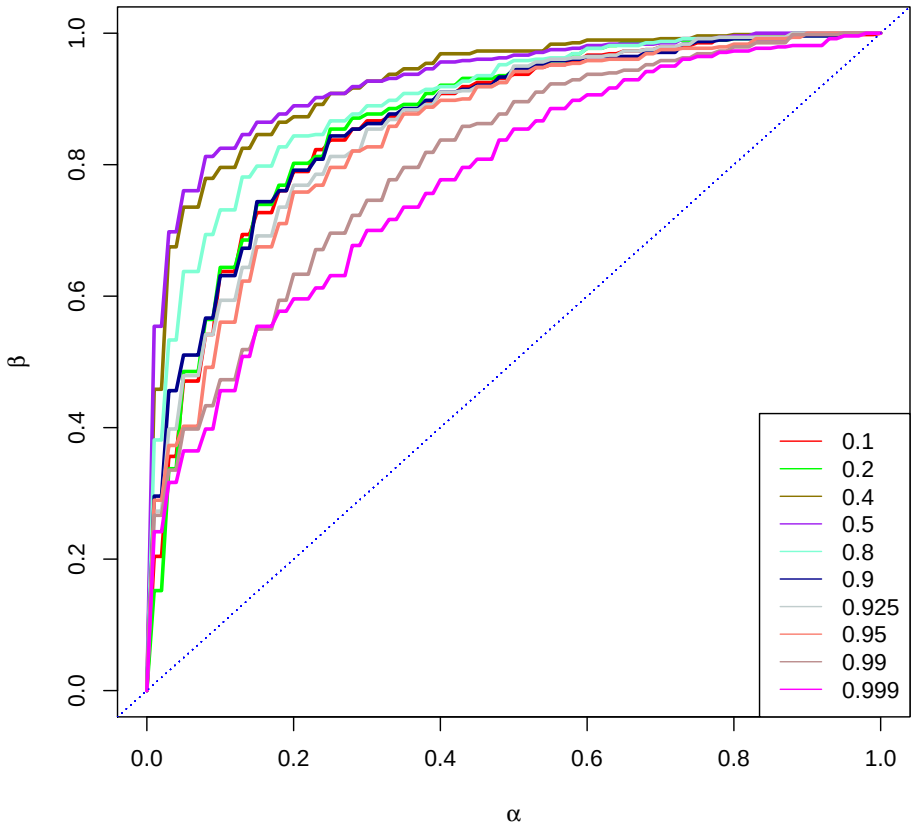


Figure 14: Power (β) vs Type I error (α) plot for different w values for the Gaussian setting (noisy case)

12 Experiments on Wiki Data

To test the JOFC approach with real data, a collection of articles are collected from the English Wikipedia, consisting of the directed 2-neighborhood of the document "Algebraic Geometry". This collection of 1382 articles and the correspondence of each article in French Wikipedia is our real-life dataset. It is possible to utilize both textual content of the documents and the hyperlink graph structure. The textual content of the documents is summarized by the bag-of-words model. Dissimilarities between documents in the same language are computed by the Lin-Pantel discounted mutual information [**?**] and cosine dissimilarity $k(x_{ik}; x_{jk}) = 1 - (x_{ik}x_{jk})/(\|x_{ik}\|_2\|x_{ik}\|_2)$. The dissimilarities based on the hyperlink graph of the collection of the articles are for each pair of vertices i and j, the number of vertices one must travel to go from i to j. Further details about this dataset is available in [**?**] Only dissimilarities based on the textual content will be considered in this example.

The exploitation task is still testing for matchedness of vertices between different conditions, in this case wiki articles that are on the same topic in different languages. For hypothesis testing, randomly held out four documents - one matched pair and one unmatched pair - are used to compute empirical type I error α and estimate of power based on the critical value computed from the distribution of the test statistic for the remaining 1380 matched pairs. The test statistic is computed using one of the three approached mentioned cca, $p \circ m$, and $jofc$. The two sets of held-out matched pairs are embedded as $\tilde{y}_1$ and $\tilde{y}_2$, via out-of-sample embedding, to estimate the null distribution of the test statistic $T = d(\tilde{y}_1; \tilde{y}_2)$. This allows us to estimate critical values for any specified Type I error level. Then the two sets of heldout unmatched pairs are embedded as $\tilde{y'}_1$ and $\tilde{y'}_2$, via out-of-sample embedding. $T' = d(\tilde{y'}_1; \tilde{y'}_2)$ will give us an empirical distribution of the test statistic under the alternative hypothesis. And the distribution under null hypothesis and under alternative hypothesis can be used to estimate power. Target dimensionality d is determined by the Zhu and Ghodsi automatic dimensionality selection method [19], resulting in d = 6 for this data set.



13 Model Selection

For the simulations presented up to now, the embedding dimension d was set to 2. This was a convenient choice which allowed us to investigate various aspects of JOFC and competing approaches. However, more care is required in selection of this parameter, since it plays such a big role in performance in general learning settings. Specifically, the effect of this parameter on Fidelity and Commensurability should be elucidated. First consider the distribution of the test statistic for different values of the embedding dimension in the Gaussian setting.

?? shows the effect of d on Fidelity and Commensurability errors.

The following plot of ROC curves [**?**] shows the effect of d parameter on the performance of different methods for the Gaussian setting.

This plot shows the histogram of the estimates of w^* (each estimate comes from a single MC replicate) for different d values.

14 Experiments on Graph Data

Another application of the JOFC approach to the vertex matching problem in multiple graphs. These simulations will be for the same semi-supervised setting as mentioned in 7, where matchings between some vertices in different graphs are known and the task is to infer the correspondences between the remaining collection of vertices in the graphs.

Suppose $A, B \in \mathcal{R}^{(r+s)\times(r+s)}$ are adjacency matrices for graphs partitioned as (r rows

then s rows, r columns then s columns)

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \qquad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

To simplify suppose that $A_{11} = B_{11}$, ie the first r vertices of A's graph correspond respectively to the first r vertices of B's graph, and we wish to complete the isomorphism by determining the correspondences between the pairs of s vertices. That is, we seek a permutation matrix $P \in \{0,1\}^{s\times s}$ such that $A = (I_{r\times r} \oplus P)B(I_{r\times r} \oplus P)^T$, ie

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} I_{r\times r} & 0_{r\times s} \\ 0_{s\times r} & P \end{bmatrix} = \begin{bmatrix} I_{r\times r} & 0_{r\times s} \\ 0_{s\times r} & P \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

Using omnibus embedding, it is possible to embed the vertices of two graphs in a commensurate space. Therefore, the JOFC approach can be used here for determining the pairwise distances between the vertices of A and B. The next step is to use the pairwise distances to find the optimal 1-1 matchings by the Hungarian algorithm [7]. The Hungarian algorithm finds an optimal matching between two sets of vertices such that the total cost which is the sum of the pairwise distances of matched nodes is minimized.

One useful property of dissimilarity representation is that the structure of data is irrelevant once an appropriate dissimilarity function for the data is available. There are many distances that can be defined between vertices in graphs. We assume that an appropriate distance measure is available to us. In our experiments we will use three different dissimilarities between vertices in a graph:

- the shortest path on the unweighted graph whose adjacency matrix is available
- the shortest path on a weighted version of the graph whose adjacency matrix is available
- diffusion distance between vertices on the (unweighted) graph.
- DICE distance [**?**, **?**, **?**] "????"

We will omit the results for weighted graph dissimilarities, since they seem to have the same performance as the weighted dissimilarities.

Note that these dissimilarities can only be defined between vertices of the same graph. We impute the inter-condition dissimilarities as described before3.

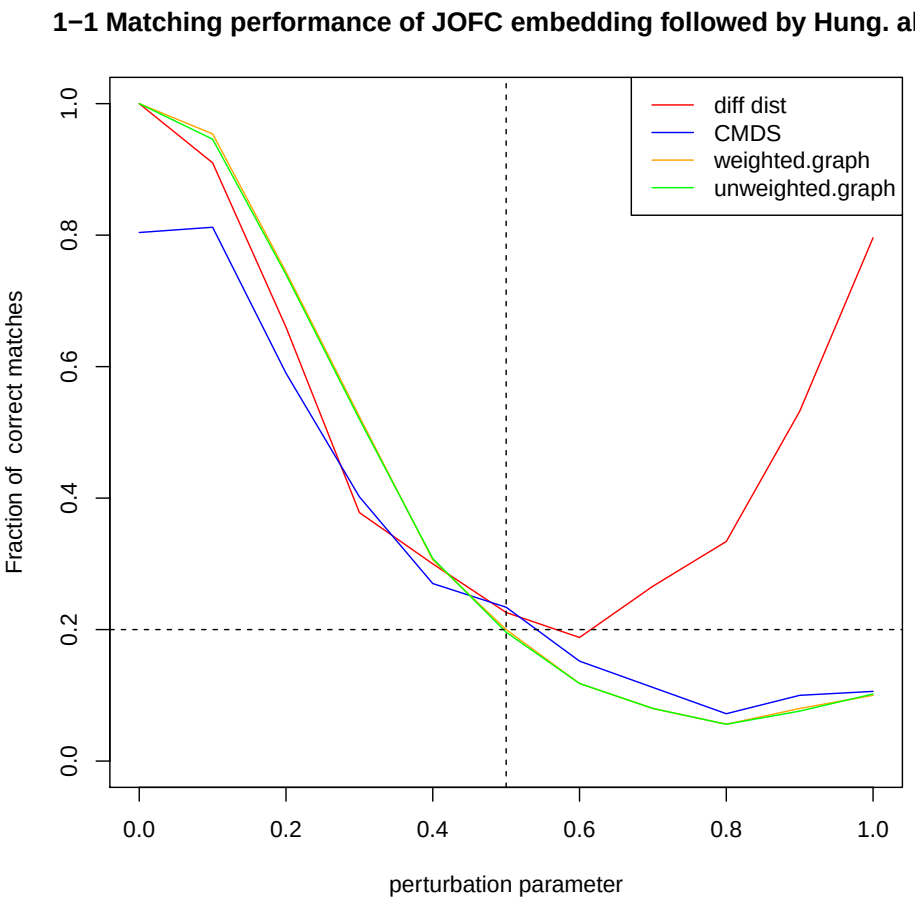
To test JOFC approach, consider the following simulation: A is the adjacency matrix of an Erdos-Renyi graph, that is $[A]_{ij} \sim Binomial(p)$ where $[A]_{ij}$ is ij-th entry of the adjacency matrix A. and the adjacency matrix B is a entry-wise bit-flipped version of the adjacency matrix of A, that is In the following simulation, A is the adjacency matrix of an Erdos-Renyi graph, that is $[A]_{ij} \sim Binomial(p)$ where $[A]_{ij}$ is ij-th entry of the adjacency matrix A. and the adjacency matrix B is a entry-wise bit-flipped version of the adjacency matrix of A, that is $[B]_{ij} \,|\, [A]_{ij} = 0 \sim Binomial(p_{10})$ $[B]_{ij} \,|\, [A]_{ij} = 1 \sim Binomial(p_{11})$. Suppose $p_{10} = p_{11} = p$.

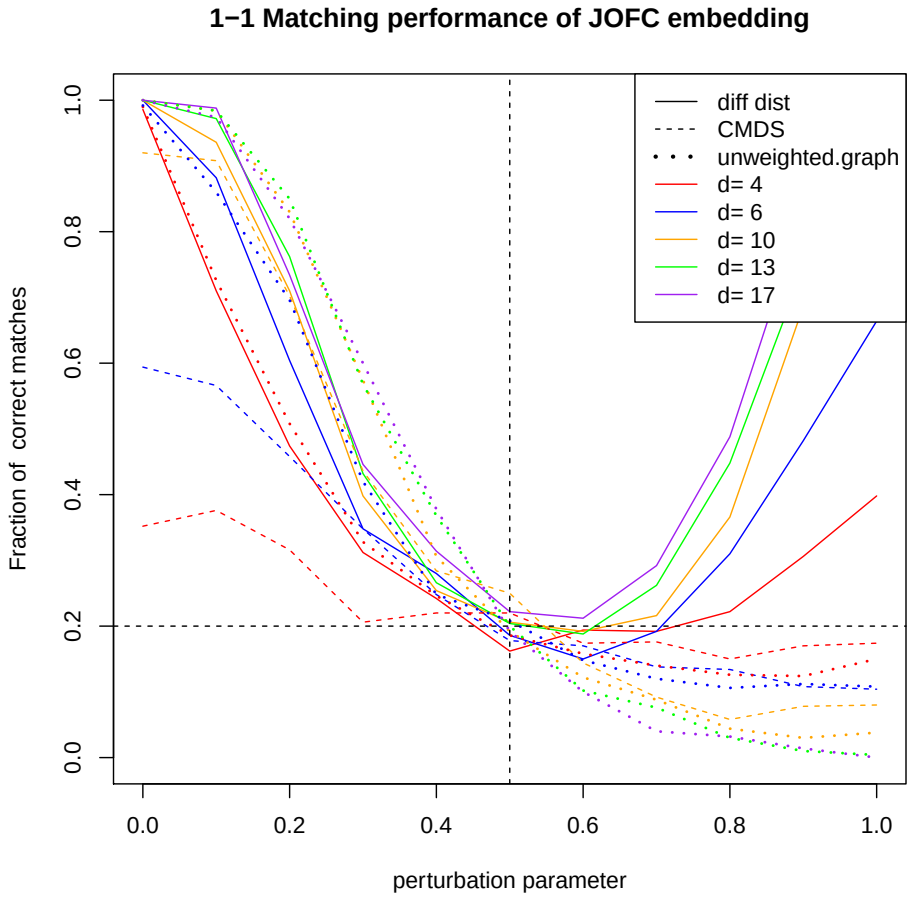
The probability of flipping an entry of the adjacency matrix is the perturbation parameter p_{pert} which is the variable on the x-axis. The performance measure is the proportion of true matches to the number of matches. Note that under chance, the expected number of true matches is 1, as shown with the dashed line. In the simulation, $r = 20$ and $s = 5$. p_{pert} varies from 0 to 1 in increments of 0.1.

In the plot above, JOFC approach applied to dissimilarities based on weighted and unweighted graphs is compared with classical MDS on dissimilarities of weighted graphs.

Note that JOFC for unweighted and weighted graphs have better performance compared to CMDS. As the perturbation parameter gets larger, the performance degrades until it is indistinguishable from random chance at $pert = 0.5$.

Another feature of the plot is the U-shape of the curve for diffusion-distance based dissimilarities. This invariancy with respect to complement of the graph should be investigated further.


1−1 Matching performance of JOFC embedding followed by Hung. algo.
diff dist
CMDS
weighted.graph
unweighted.graph
Fraction of correct matches
0.0
0.2
0.4
0.6
0.8
1.0
perturbation parameter



An interesting trend in the graph is that shortest-path based dissimilarities are an improvement over diffusion-path dissimilarities for perturbation parameter less than 0.5 , but as perturbation parameter increases past 0.5, fraction of correct matches for diffusion distance based dissimilarity recovers, while for other dissimilarities the fraction continues to fall.

The dissimilarity type that has the best improvement in performance is JOFC with shortest path distances in weighted graphs(unweighted graphs have similar performance)

This graph shows the effect of the weight parameter of stress w on the probability of true matches.

This graph shows the effect of the diffusion distance parameter T on the probability of true matches for dissimilarities based on diffusion distance.

15 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. Two alternative approaches, P∘M and CCA, were presented as extremes of the tradeoff between Fidelity and Commensurability. 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] H. Choi, S. Choi, and Y. Choe. Manifold integration with markov random walks. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 1*, pages 424–429. AAAI Press, 2008.

[4] J. Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975. 10.1007/BF02291478.

[5] 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.

[6] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):pp. 433–451, 1971.

[7] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52(1):7–21, 2005.

[8] G.R.G. Lanckriet, N. Cristianini, Peter Bartlett, L.E. Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

[9] Y.Y. Lin, T.L. Liu, and C.S. Fuh. Dimensionality reduction for data in multiple feature representations. *Advances in Neural Information Processing Systems*, 21:961–968, 2009.

[10] B. McFee and G.R.G Lanckriet. Learning multi-modal similarity. *The Journal of Machine Learning Research*, 12:491–523, February 2011.

[11] 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.

[12] 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.

[13] 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.

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

[15] Michael W. Trosset and Rudolf Mathar. On the existence of nonglobal minimizers of the stress criterion for metric multidimensional scaling. 1997.

[16] 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.

[17] 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.

[18] 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.

[19] Mu Zhu and Ali Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930, 2006.