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

Sancar Adali[*] Carey E. Priebe[*]

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. [4] 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) . We investigate the tradeoff between fidelity and commensurability 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 our weighted embedding scheme provides significant improvements in test power.

Key Words: Multidimensional Scaling, Manifold Matching, Disparate Sources, Fidelity, Commensurability, Information Fusion, Canonical Correlation, Procrustes Matching

1. Problem Setting

The problem setting is one where the data are available in a dissimilarity representation, possibly because the observations themselves are not available, or too complex to be used in inference tasks. If our inference task is one where we can not use the dissimilarities directly, Multidimensional Scaling (MDS) [6, 1, 3] can be used for embedding the observations in the Euclidean space with a chosen dimension d such that the distances between the observations are as close as possible (in various senses) to the original dissimilarities. Different criterion functions can be used to measure how close the distances are to the given dissimilarities, leading to different embedded configurations. Given $n \times n$ dissimilarity matrix $\Delta = \{\delta_{st}; 1 \leq s \leq n; 1 \leq t \leq n\}$, one possible function is the weighted raw-stress:

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

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. We will refer to the $n \times n$ matrix representation of the weights and Euclidean distance as W and $D(X)$, respectively.

We wish to study the dissimilarity-representation version of the following hypothesis testing problem: Suppose we have n different objects/entities that are measured under K different conditions with measurement vectors x_{ik} indexed by object and condition. Each of the measurements x_{ik} lies in the corresponding space Ξ_k.

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

Given K new measurements/observations, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k, \ldots, \boldsymbol{y}_K$, $\boldsymbol{y}_k \in \Xi_k$, we wish to test the null hypothesis that "these observations are from the same object" versus the alternative hypothesis that

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

"they are not from the same object" [4]:

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

We can restate the null hypothesis as the case where the observations are "matched" and the alternative as the case where they are not "matched".

As it turns out, all of the observations are in a dissimilarity representation; that is, instead of $\{\boldsymbol{x}_{ik}; i = 1, \ldots, n;\ k = 1, \ldots, K\}$, and $\{\boldsymbol{y}_k; k = 1, \ldots, K\}$, we are given $n \times n$ 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. We again wish to test whether these dissimilarities arose from measurements for which the null hypothesis is true – whether the measurements are "matched" or not. Since dissimilarities are measured between pairs of objects under the same condition, we have separate dissimilarity matrices consisting of dissimilarities between pairs of measurements for each 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.

2. Manifold Matching

This hypothesis testing task is a specific case of a joint inference task from disparate data sources. Our approach requires the measurements in different conditions to be commensurate. As a solution, we will propose "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. We can formalize this approach by considering maps $\rho_k, k = 1, \ldots, K$ from measurement spaces Ξ_k to a low-dimensional commensurate space $\mathcal{X}$. The learning problem involves estimating these maps from a training data of matched measurements, though the maps might not necessarily be explicit, as in the following case.

Suppose we are able to combine the given dissimilarity matrices $\{\Delta_k, k = 1, \ldots, K\}$ into one omnibus dissimilarity matrix M, imputing entries if necessary. Consider, for $K = 2$,

$$M = \begin{bmatrix} \Delta_1 & L \\ L^T & \Delta_2 \end{bmatrix} \tag{2}$$

where L is a matrix of imputed entries. Using MDS to embed this omnibus matrix into a space $\mathcal{X}$, we obtain $2n$ embedded observations $\{\tilde{y}_i^{(k)}; i = 1, \ldots, n; k = 1, 2\}$ in a single space, with distances between the different observations consistent with the given dissimilarities. Now that the observations are commensurate, we can compute a 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, we will reject the null hypothesis. We will refer to this approach as the Joint Optimization of Fidelity and Commensurability (JOFC) approach, for reasons that will be explained in Section 6. In this approach, the mappings $\{\rho_k, ; k = 1, \ldots, K\}$ are not explicitly defined.

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

Given dissimilarities between K new test observations and the previous nK training observations (referred to as out-of-sample (OOS)), it is necessary to embed them along with the original (in-sample) $nK \times nK$ dissimilarities between training observations. In many applications, it is practical to pre-embed the in-sample dissimilarities and extend the embedding with out-of-sample dissimilarities, rather than re-embed the augmented dissimilarity matrix which includes both in-sample and out-of-sample dissimilarities. This out-of-sample extension is used in various approaches in this paper. Out-of-sample embedding can be done one observation at a time, or jointly if the dissimilarities among multiple test observations are also available.

We will assume 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.

For the remainder of this paper, for simplicity we consider the $K = 2$ case, although the generalization to $K > 2$ conditions is straightforward.

3. "Matched" and "Conditions" in data

What we mean by "conditions" and "matched" is 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.
- We might be talking about objects in a single space with multiple dissimilarities, where dissimilarities are measured for different purposes, or judged by different people.

4. Two models for generating data

Here we propose two data models that illustrate our idea of matchedness.

4.1 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 1.

The parameter r controls the variability between "matched" measurements. If r is large, we expect the distance between matched measurements X_{i1} and X_{i2} to be stochastically smaller than $X_{i'1}$ 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 a given value of allowable type I error rate α.

4.2 Dirichlet setting

Let $S^p = \{\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^{(p+1)}, \sum_{l=1}^{p+1} x_l = 1, x_l > 0 \quad \forall \quad l = 1, \ldots, p+1\}$ 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 2.

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

4.3 Noise

Measurements X_{ik} carry the signal that is relevant to our 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

$$\breve{X}_{ik} = [(1-c)X_{ik} \;\; cE_{ik}] \tag{3}$$

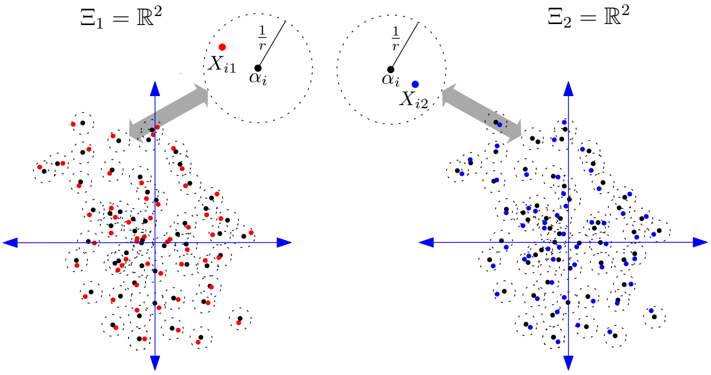

Figure 1: For the Gaussian setting (Section 4.1), the α_i are denoted by black points and the X_{ik} are denoted by red and blue points respectively.

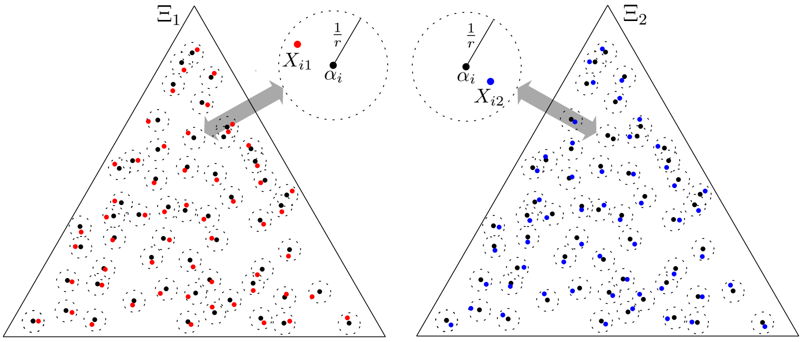

Figure 2: For the Dirichlet setting (Section 4.2), the α_i are denoted by black points and the X_{ik} are denoted by red and blue points respectively.

:

where $E_{ik} \sim^{iid} Dirichlet(\mathbf{1}_{(q+1)})$ for the Dirichlet setting and $E_{ik} \sim^{iid} MVNormal(\mathbf{0}, (1+\frac{1}{r})I_q)$ for the Gaussian setting. $\check{X}_{ik}$ 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.

5. 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, our hypothesis testing task 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 [7] 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 [5], 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 [7], 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. [8] finds two projection matrices to minimize three terms in an energy function similar to our JOFC approach (see Section 2). One of the terms is the *correspondence preserving term* which is the sum of the squared distances between corresponding points and is analogous to our commensurability error term. 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 our 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.

6. 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 we are not neccessarily interested in perfect reconstruction, but the best embedding for our exploitation task which is to test whether two sets of dissimilarities are "matched". Two concepts will help us argue how the manifold matching should proceed in order to optimize the matching for the exploitation task we are to carry out.

- Fidelity is how well the mapping to commensurate space preserves the original dissimilarities. Our within-condition *fidelity error* is given by

$$\epsilon_{f_k} = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} (d(\widetilde{\boldsymbol{x}}_{ik}, \widetilde{\boldsymbol{x}}_{jk}) - \delta_k(\boldsymbol{x}_{ik}, \boldsymbol{x}_{jk}))^2 \qquad (4)$$

  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. Between-condition *commensurability error* is given by

$$\epsilon_{c_{k_1 k_2}} = \frac{1}{n} \sum_{1 \le i \le n; k_1 < k_2} (d(\widetilde{\boldsymbol{x}}_{ik_1}, \widetilde{\boldsymbol{x}}_{ik_2}) - \delta_{k_1 k_2}(\boldsymbol{x}_{ik_1}, \boldsymbol{x}_{ik_2}))^2 \qquad (5)$$

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.

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 we choose diagonal entries of L in equation (2) to be all zeroes. Setting these diagonal entries to zero forces matched points to be embedded close to each other. We ignore the possibility that this choice for between-condition dissimilarities might not be optimal, in order to concentrate on the main problem.

Then, the commensurability error 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. 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 we are considering is a partitioned matrix consisting of matrices from different conditions ($k = 1, 2$), so we index the entries of the matrix 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,

$$\begin{aligned}\sigma_W(\cdot) &= \sum_{i,j,k_1,k_2} w_{ijk_1k_2}(d_{ijk_1k_2}(\cdot) - M_{ijk_1k_2})^2 \\ &= \underbrace{\sum_{i=j,k_1<k_2} w_{ijk_1k_2}(d_{ijk_1k_2}(\cdot) - M_{ijk_1k_2})^2}_{Commensurability} + \underbrace{\sum_{i<j,k_1=k_2} w_{ijk_1k_2}(d_{ijk_1k_2}(\cdot) - M_{ijk_1k_2})^2}_{Fidelity} \\ &+ \underbrace{\sum_{i<j,k_1<k_2} w_{ijk_1k_2}(d_{ijk_1k_2}(\cdot) - M_{ijk_1k_2})^2}_{Separability} \quad . \end{aligned} \tag{6}$$

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

Since we choose to ignore the separability error, we choose the weights for separability terms to be 0. This also means off-diagonal elements of L in equation (2) can be ignored. When separability terms are removed from equation (6), 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 our omnibus embedding approach as Joint Optimization of Fidelity and Commensurabilty (JOFC).

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

$$\sigma(X) = \sum_{1\leq s\leq n;1\leq t\leq n} (d_{st}(X) - \delta_{st})^2 \tag{7}$$

7. Alternative Methodologies

For the optimization of commensurability with fidelity as secondary priority, one can use Canonical Correlational Analysis (CCA) [2], which aims to find linear subspaces of the space the data resides in such that the projection of data to those subspaces results in observation 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 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, we can 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 [6] under certain conditions suggests that this approach is the right analog of Procrustes ∘ PCA 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.

7.1 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.

Procrustes Analysis [5] 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.,

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

where $\|\cdot\|_F$ is the Frobenius norm on matrices. The transformation ρ represented by $\mathbf{Q}^*$ makes the separate MDS embeddings commensurate. 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. We will refer to this approach as P∘M.

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 we allow a larger class of mappings to be considered, we would have a smaller model bias for the mapping function, but we would be paying for it in the form of larger variability. By only considering the class of linear transformations, we are able to learn $\mathbf{Q}^*$ with our limited dataset.

7.2 Canonical Correlational Analysis on Multidimensional Scaling Embeddings

Again MDS is used to compute embedding configurations, X_1 and X_2. We want to embed into the highest dimensional space possible (e.g., $\mathbb{R}^{d'}$ where $d' = p + q$ for our Gaussian and Dirichlet settings) to preserve as many of the signal dimensions as possible (at the risk of possibly including some noise dimensions). CCA [2], 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∘M, new dissimilarities are out-of-sample embedded and mapped to a commensurate space by maps provided by CCA. We can now compute the test statistic and reject the null hypothesis for "large" values of the test statistic τ as in Section 7.1.

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

Suppose we let $w_{ijk_1k_2} = w$ for commensurability terms and $w_{ijk_1k_2} = 1 - w$ for fidelity terms in equation (6). For the resulting weight matrix W, define

$$f_w(D(\cdot), M) = \sigma_W(\cdot) \tag{8}$$

where M is the omnibus matrix obtained from a given pair of dissimilarity matrices, Δ_1 and Δ_2, as in equation (2). 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∘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,

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

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

7.4 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 $U_1^TX_1^TX_1U_1 = I_d$ and $U_2^TX_2^TX_2U_2 = I_d$ (I_d is the $d \times d$ identity matrix).*

8. Fidelity and Commensurability Tradeoff

The major question addressed in this work is whether preservation of fidelity or commensurability is more essential for our hypothesis testing task. The weights in raw stress allow us to answer this question relatively easily. Since in equation (6), 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.

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

Our expectation is that there is a w^* that is optimal for the specific exploitation task (has the best power in hypothesis testing). In fact, our exploratory simulations 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

- we are given a sample of $\mathcal{T}$ ($\mathbf{T}_{mc}$) and a sample of X_m and X_u $\left(\boldsymbol{x_m} = \begin{bmatrix} \boldsymbol{x}_{1m} \\ \boldsymbol{x}_{2m} \end{bmatrix}, \boldsymbol{x_u} = \begin{bmatrix} \boldsymbol{x}_{1u} \\ \boldsymbol{x}_{2u} \end{bmatrix}\right)$ and we compute Euclidean distances 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
- we are given values of dissimilarity matrix-valued function D of the sample of X_m, X_u and $\mathbf{T}_{mc}$:

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

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

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, we define the disssimilarity matrices $\Delta_m \left(\left[\begin{array}{c} \mathcal{T} \\ X_{1m} \\ X_{2m} \end{array} \right] \right)$ and $\Delta_u \left(\left[\begin{array}{c} \mathcal{T} \\ X_{1u} \\ X_{2u} \end{array} \right] \right)$ as two matrix-valued random variables.

The criterion function for the embedding is $\sigma_W(\cdot) = f_w(D(\cdot), \Delta)$. The embedding for the unmatched pair $\hat{X}_{1u}, \hat{X}_{2u}$ is

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

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 $\hat{X}_{1m},\ \hat{X}_{2m}$. Define F_Y as the cumulative distribution function of Y where Y can be any function of $\hat{X}_m$ or $\hat{X}_u$.

Then

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

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.

One of the important questions to be explored is the uniqueness 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 generated from pairwise distances of measurements. A set of matched pairs of measurements and unmatched pairs of measurements were also generated for testing. The test statistics (computed via P∘M, CCA and JOFC approaches) 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 4.3). This setting will be referred to as the “noisy case”. The original setting, with $c = 0$, will be referred as the “noiseless case”. If the magnitude of noise (controlled by the parameter c in equation (3)) is small enough, PCA and MDS will not be affected significantly, but if the number of noisy dimensions (controlled by the parameter q in the distribution of E_{ik} as defined in equation (3)) is large enough, 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 $c = 0$, we are in the “noiseless” setting and measurements are p-dimensional vectors, otherwise we are in the “noisy” setting 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 (8) 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.
- We generate m pairs of matched measurements which we treat 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 the 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.

For $n = 150$ and $m = 150$, the average of the power values for $nmc = 150$ Monte Carlo replicates are computed at different values of α and are plotted in Figure 3 against α for the Gaussian setting. Qualititatively similar plots for the Dirichlet setting are not included herein for brevity. The plot in Figure 3 shows that for different values of w, we have varying power curves, and some w values outperform others in terms of power. In Figure 4, $\beta(w)$ 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 commensurability might be more critical for our hypothesis testing task.

Note that in Figure 3 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, we wish to test the null hypothesis that $H_0 : \beta_\alpha(\hat{w}^*) \leq \beta_\alpha(w = 0.5)$ against the alternative $H_A = \beta_\alpha(\hat{w}^*) > \beta_\alpha(w = 0.5)$. The least favorable null hypothesis is that $H_0 : \beta_\alpha(\hat{w}^*) = \beta_\alpha(w = 0.5)$.

For a fixed α value, one can compute two critical values using the test statistic values for the two w values that are being compared. Using these critical values, we can determine the decision made by each test for each pair of embedded points, $(\tilde{y}_i^{(1)}, \tilde{y}_i^{(2)}), \quad i = 1, \ldots, m$. To compare the two statistical tests with different w values, 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 c_1 for the first statistical test and c_2 for the second statistical test. If $c_1 = True$ and $c_2 = False$ 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, $e_{c_1 c_2}^{(l)}$ is equal to the number of instances at which the true hypothesis were identified correctly ($c_1 = True$) or incorrectly ($c_1 = False$) by the first test, and correctly ($c_2 = True$) or incorrectly ($c_2 = False$) by the second test in that MC replicate.

Under the null hypothesis, $Pr[(c_1 c_2) = (TF)] = Pr[(c_1 c_2) = (FT)]$, 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 $H_A = \beta_\alpha(\hat{w}^*) > \beta_\alpha(w = 0.5)$, we find $p < 1.09E{-}24$ which indicates the power using estimate of optimal w^* is significantly greater than the power when using $w = 0.5$.

11. Conclusion

We have investigated the tradeoff between Fidelity and Commensurability and the relation to the weighted raw stress criterion for MDS. 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.

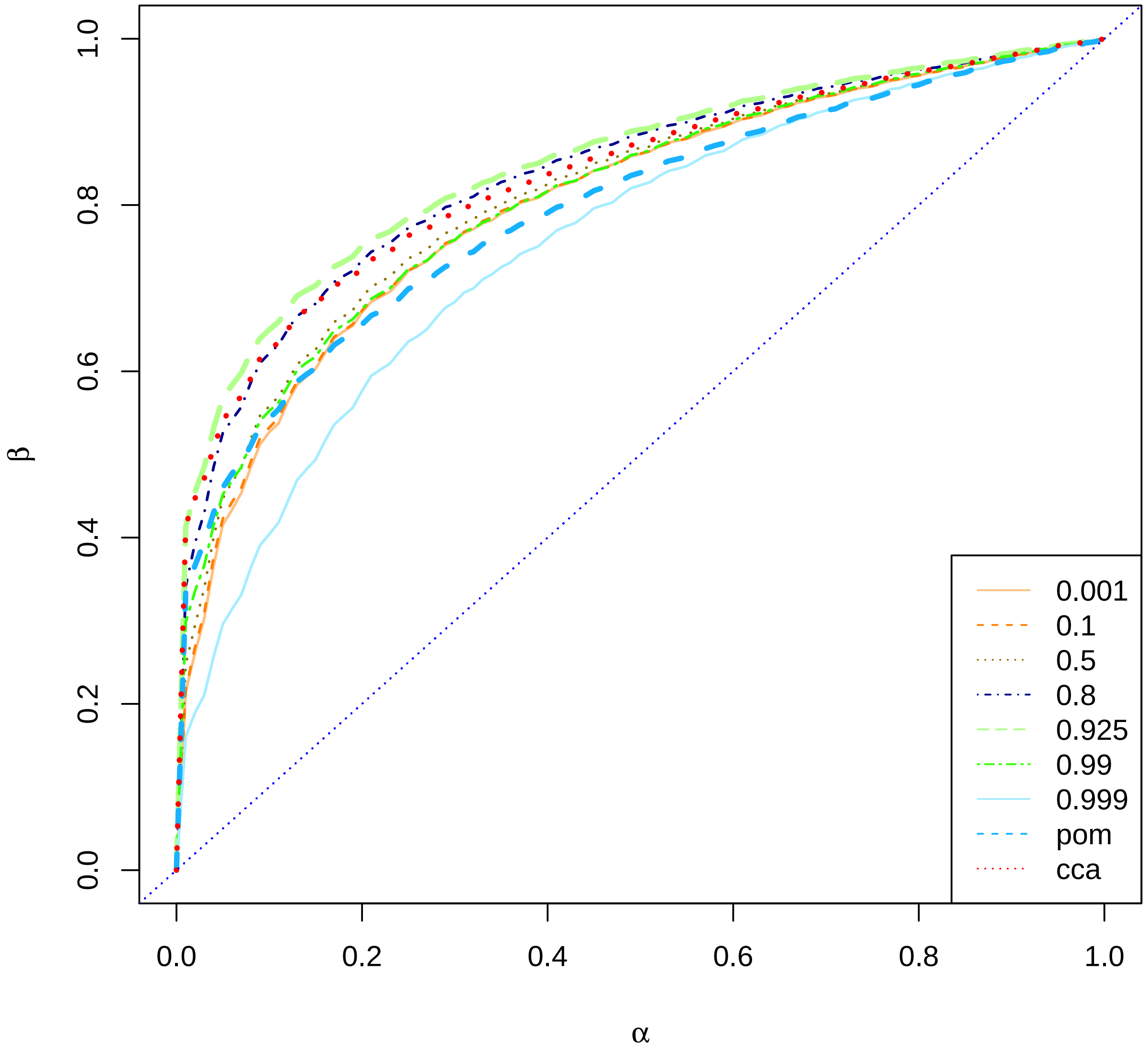

Figure 3: Power (β) vs Type I error (α) plot for different w values for the Gaussian setting (noisy case)

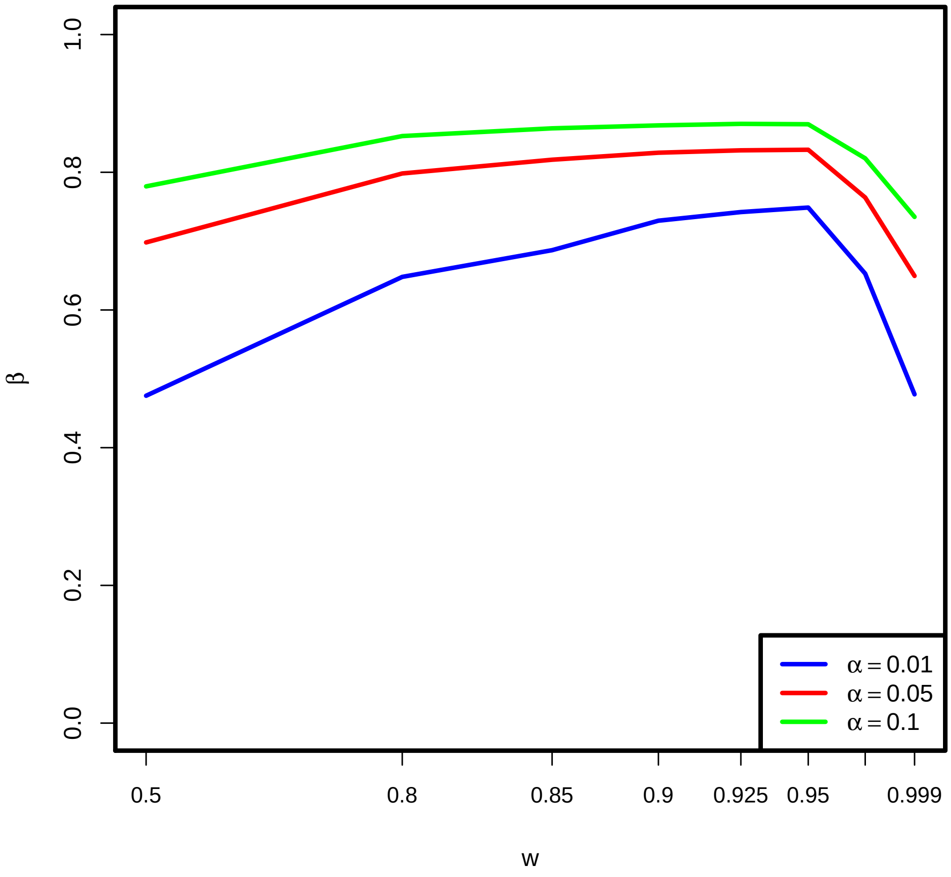

Figure 4: Power (β) vs w plot for fixed Type I error (α) values for the Gaussian setting (noisy case)

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.