Fidelity-Commensurability Tradeoff in Joint Embedding of Disparate Dissimilarities

Sancar Adali* Carey E. Priebe†

August 1, 2011

Abstract

For matched data from disparate sources (objects observed under different conditions), optimality of information fusion must be defined with respect to the inference task at hand. Defining the task as matched/unmatched hypothesis testing for dissimilarity observations, Priebe et al. [4] presents manifold matching using an embedding method based on joint optimization of fidelity (preservation of within-condition dissimilarities between observations of an object) and commensurability (preservation of between-condition dissimilarities between observations) . The tradeoff between fidelity and commensurability is investigated by varying weights in weighted embedding of an omnibus dissimilarity matrix. Optimal (defined with respect to the power of the test) weights for the optimization correspond to an optimal compromise between fidelity and commensurability. Results indicate optimal weights are different than equal weights for commensurability and fidelity and the proposed weighted embedding scheme provides significant improvements in test power.

1 Introduction

It is a challenge to do a tractable analysis on data from disparate sources of data (such as multiple sensors). The multitude of sensors technology and large numbers of sensors both are sources of difficulty and hold promise for efficient inference.

In the problem setting considered, n different objects/instances are measured/judged under K different conditions (corresponding possibly to K different sensors) using dissimilarity measures. These will be represented in matrix form as K $n \times n$ matrices $\{\Delta_k, k = 1, \ldots, K\}$. In addition, dissimilarities between K new measurements/observations and the previous n objects under K conditions are given. The inference task is to test the null hypothesis that "these measurements are from the same object" (matched) against the alternative hypothesis that "they are not from the same object" (unmatched) [4]: In order to get a data representation where dissimilarities from disparate sources can be compared, the dissimilarities must be mapped to a commensurate metric space where the metric can be used to distinguish between "matched" and "unmatched" pairs.

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 space , imputing entries if necessary. Consider, for $K = 2$ [1],

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

*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

[1] Throughout this paper , it will be assumed, the number of conditions, K, is equal to 2 for the simplicity of presentation. The approaches are easily generalizable to $K > 2$ conditions.

where L is a matrix of imputed entries. Using MDS to embed this omnibus matrix into a commensurate space , $2n$ embedded observations $\{\tilde{y}_i^{(k)}; i = 1, \ldots, n; k = 1, 2\}$ are obtained. [2], 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. Out-of-sample extension for MDS will be used throughout this paper [5].

Regardless of the inference task, to expect reasonable performance from the embedded data in the commensurate space, two criteria must be adhered to:

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

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

  where δ_{ijk} is the dissimilarity between i^{th} object and j^{th} object where both objects are in 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).

- 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 *incommensurability error* which is given by

$$\epsilon_{c_{k_1 k_2}} = \frac{1}{n} \sum_{1 \leq i \leq n; k_1 < k_2} (d(\widetilde{\boldsymbol{x}}_{ik_1}, \widetilde{\boldsymbol{x}}_{ik_2}) - \delta_{ijk_1k_2})^2$$

  for conditions k_1 and k_2; $\delta_{ijk_1k_2}$ is the dissimilarity between i^{th} object under k_1^{th} condition and j^{th} object under k_2^{th} condition.

While *infidelity* and *incommensurability* errors as measured by the above expressions is specific to the joint embedding scheme by MDS, the concepts of fidelity and commensurability are general enough to be applicable to other dimensionality reduction methods that can be used on data from disparate sources.

The joint embedding results in the joint optimization of fidelity and commensurability and this relation can be made explicit by the use of the raw stress function in MDS embedding. Denoting $(s,t)^{th}$ entry of M by M_{st} ,

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

is the MDS criterion function , where individual terms in the sum can be ascribed to fidelity and commensurability respectively.

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

[2] It will be assumed the commensurate space $\mathcal{X}$ is $\mathbb{R}^d$ where d is pre-specified. The selection of d – model selection – is a task that requires much attention and is beyond the scope of this article. Discussion of the effect of d on matching performance will be available at a later paper.

The separability error terms will be ignored herein (by setting the associated weights to be 0), due to the fact that $\delta_{k_1k_2}(\boldsymbol{x}_{ik_1}, \boldsymbol{x}_{jk_2})$ is not available [3] and it will be easier to focus on just the fidelity-commensurability tradeoff. Setting $\delta_{ijk_1k_2}$ to 0 [4],

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

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

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 inference task. The weights in raw stress allow us to answer this question relatively easily. Setting w_{ij} to w and $1-w$ for commensurability and fidelity terms respectively will allow us to control the relative importance of fidelity and commensurability terms in the optimization.

Previous investigations of JOFC approach assumed unweighted(equivalently, equally weighted) stress functions for MDS embedding. This assumption corresponds to $w = 0.5$ case in this presentation . The expectation here is that there is a w^* such that $w^* \neq 0.5$ that is optimal for the specific exploitation task (hypothesis testing, which w value has the best power). That is , the default value of $w = 0.5$ is not the "optimal" value and w should be deliberately chosen.

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.

2.1 Gaussian setting

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

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

2.2 Alternative Methodologies

Two alternative methodologies exist that correspond roughly to the extreme ends of the range of w values ($w \in (0,1)$). For the optimization of commensurability with fidelity as secondary priority($w \approx 1$), an alternative method is Canonical Correlational Analysis (CCA) [2], which finds optimally correlated projections of two random vectors. For the optimization of fidelity , Principal Components Analysis (PCA) finds "best"' representation separately for two conditions . To optimize commensurability as secondary priority, one

[3]Due to the fact that data sources are "disparate", it is not obvious how a dissimilarity between an object in one condition and another object in another condition can be computed or defined in a sensible way.

[4]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 (1) are chosen to be all zeroes. Setting these diagonal entries to zero forces matched points to be embedded close to each other.

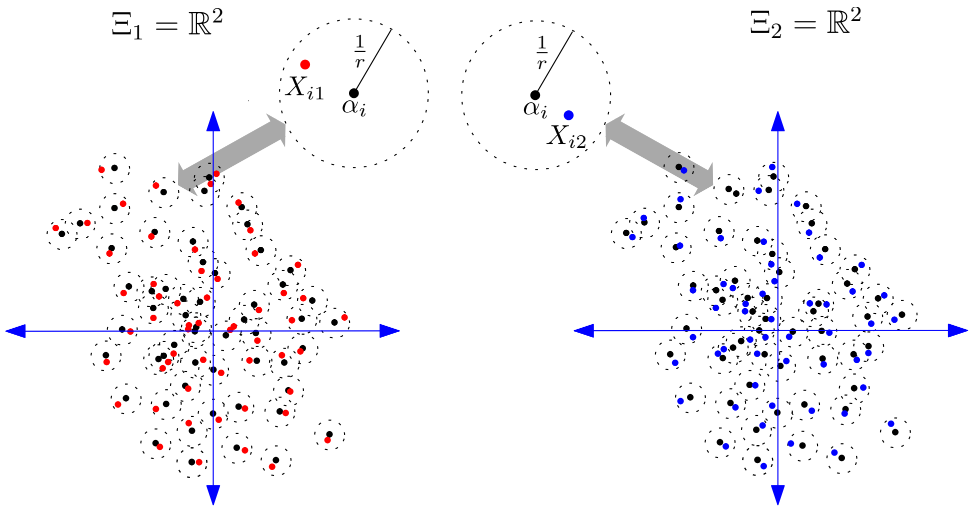


Figure 1: For the Gaussian setting (Section 2.1), the $\boldsymbol{\alpha_i}$, are denoted by black points and the $\boldsymbol{x}_{ik}$ are denoted by red and blue points respectively.

can use the projections computed by PCA to compute a Procrustes transformation that will make the projections commensurate. This $Procrustes \circ MDS$ approach is analagous to $w \approx 0$ case for JOFC.

3 Related Work

There have many efforts toward solving the related problem of "manifold alignment"."Manifold alignment" seeks to find correspondences between observations from different "conditions". The setting that is most similar to ours is the semi-supervised setting [?], where a set of correspondences are given and the task is to find correspondences between a new set of points in each condition. In contrast, the hypothesis testing task discussed in this paper is to determine whether any given pair of points is "matched" or not. The proposed solutions follow a common approach: 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. [1, 6, 7]

4 Definition of w^*

Two dissimilarity matrices to be embedded are $\Delta^{(m)}\left(\begin{bmatrix} \mathcal{T} \\ X_1^{(m)} \\ X_2^{(m)} \end{bmatrix}\right)$ and $\Delta^{(u)}\left(\begin{bmatrix} \mathcal{T} \\ X_1^{(u)} \\ X_2^{(u)} \end{bmatrix}\right)$ as two matrix-valued random variables $\Delta^{(m)} : \Omega \to \mathbf{M}_{(n+2)\times(n+2)}$,$\Delta^{(u)} : \Omega \to \mathbf{M}_{(n+2)\times(n+2)}$ for the appropriate sample space (Ω). $\mathcal{T}$ is i.i.d random sample of $(X_1^{(m)}, X_2^{(m)})$.

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

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

A similar expression gives the embedding for the matched pair, where (u) superscript is replaced by (m). Assuming the necessary conditions for $\hat{X}_1^{(m)}$, $\hat{X}_2^{(m)}$, $\hat{X}_1^{(u)}$, $\hat{X}_2^{(u)}$ to be random vectors hold, consider the test statistic $\tau = d(\hat{X}_1^{\cdot}, \hat{X}_2^{\cdot})$. Under null hypothesis of matchedness, the distribution of the statistic is governed by the distribution of $\hat{X}_1^{(m)}$ and $\hat{X}_2^{(m)}$, under the alternative it is governed by $\hat{X}_1^{(u)}$ and $\hat{X}_2^{(u)}$.

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

Then

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

Note that all random variables such as $\hat{X}_k^{(m)}$ are dependent on w. Finally, define

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

Given specific distributions $\mathbf{F}_{\hat{X}_k^{(m)}}$,$\mathbf{F}_{\hat{X}_k^{(u)}}$, 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 5.

w^* is defined to be the argmin of the power function with respect to w and some important questions about w^* are related to the nature of this function $\beta_\alpha(w)$. While finding an analytical expression for the value of w^* is intractable, an estimate $\hat{w}^*$ based on noisy evaluations of $\beta_\alpha(w^*)$ can be computed. A Monte Carlo simulation is run in Section 5 to find the estimate of $\beta_\alpha(w)$ at various values of w and α .

4.1 Continuity of $\beta(\cdot)$

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

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

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

where $T_a(\cdot)$ and $T_0(\cdot)$ can also be regarded as stochastic processes whose sample paths are continuous functions of w except at a finite number of points in $(0, 1)$.

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

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

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

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

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

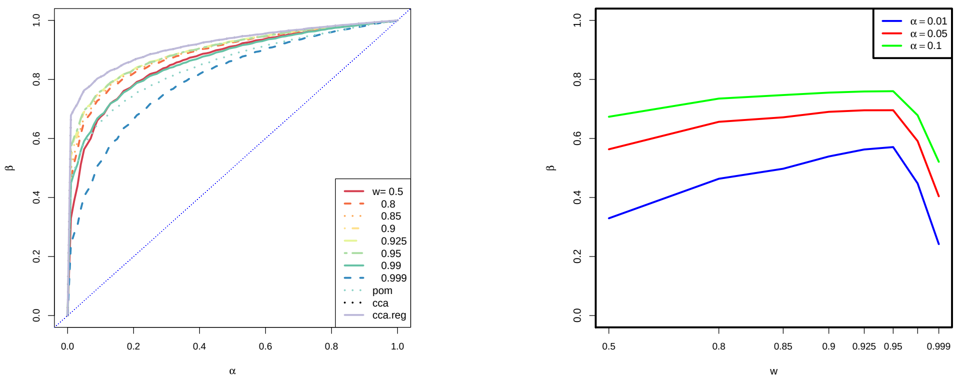
and

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

[5]

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

[5] Theorem 2.1 in [3] states the same theorem : if $T(w, \omega)$ is continuous with respect to w almost everywhere ($Pr[\omega : T(w, \omega)$ is discontinuous with respect to $w] = 0$ where $\omega \in \Omega$, and Ω is the sample space) , then $F(x) = Pr[T(w) > 0]$ is continuous.



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

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

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

5 Simulation Results

Training data of matched sets of measurements (instantiation of $\mathcal{T}$) were generated according to the Gaussian setting. 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 distribution. Following the out-of-sample embedding of the dissimilarities test pairs (computed via by one of the three P∘M, CCA and JOFC approaches) test statistics for matched and unmatched pairs were used to compute power values at different type I error rate α values.

Setting p and q to 5 and 10, respectively, for $n = 150$ matched training pairs and $m = 150$ matched and unmatched test pairs, the average of the power values for $nmc = 150$ Monte Carlo replicates are computed at different αs and are plotted in Figure 2 against α for the Gaussian setting. Qualititatively similar plots for the Dirichlet setting are not included for brevity. The plot in Figure 2 shows that for different values of w, β-α curves vary significantly. The conclusion is that the match detection tests with JOFC embedding using specific w values have better performance than other w values in terms of power. In Figure 2, $\beta(w)$ is plotted against w for fixed values of α. It is interesting that the optimal value of w seems to be in the range of $(0.85, 1)$ for this setting, which suggests a significant emphasis on commensurability might be critical for the match detection task. Simulations with other paramater values resulted in w^* estimates closer to 0.5.

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

Note that for all of the settings, the estimate of the optimal w^* has higher power than w=0.5 (the unweighted case).

6 Conclusion

The tradeoff between Fidelity and Commensurability and the relation to the weighted raw stress criterion for MDS were both investigated with several simulations and experiments on real data. For hypothesis testing as the exploitation task, the three approaches were compared in terms of testing power. The results indicate that 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.

References

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

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

[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] 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.

[6] 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.

[7] 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.