Dimension Reduction With Prior Information
for Knowledge Discovery

Anh Tuan Bui

Abstract—This paper addresses the problem of mapping high-dimensional data to a low-dimensional space, in the presence of other known features. This problem is ubiquitous in science and engineering as there are often controllable/measurable features in most applications. To solve this problem, this paper proposes a broad class of methods, which is referred to as conditional multidimensional scaling (MDS). An algorithm for optimizing the objective function of conditional MDS is also developed. The convergence of this algorithm is proven under mild assumptions. Conditional MDS is illustrated with kinship terms, facial expressions, textile fabrics, car-brand perception, and cylinder machining examples. These examples demonstrate the advantages of conditional MDS over conventional dimension reduction in improving the estimation quality of the reduced-dimension space and simplifying visualization and knowledge discovery tasks. Computer codes for this work are available in the open-source cml R package.

Index Terms—Distance scaling, ISOMAP, multidimensional scaling, Sammon mapping, SMACOF.

I. INTRODUCTION

D

imension reduction is necessary for scientists to discover hidden features from high-dimensional data. Specifically, dimension reduction methods meaningfully map high-dimensional observations to some low-dimensional space. Ideally, the reduced dimension should be as small as the intrinsic dimension of the data [1]. To achieve this goal, principal component analysis (PCA) and factor analysis are widely used for learning a linear map. For nonlinear cases, kernel PCA [2] and manifold learning [3] are among popular choices. Autoencoder is another nonlinear dimension reduction method that has been successfully applied in many studies (see, e.g., [4, [5]).

Most of dimension reduction methods require input data to be in the vector form. However, data are sometimes in dissimilarity/similarity forms (e.g., similarity rating or intercorrelation). In many applications, practitioners may also want to define suitable pairwise dissimilarity measures for their data, such as dissimilarity indices for community ecology data [6], dissimilarity for image data of random heterogeneous materials [7], dissimilarity between unstructured point cloud data [8], the cosine dissimilarity commonly used for text data [9], and distance measures for statistical distributions [10]. This paper develops a dimension reduction method that directly takes pairwise dissimilarities as the input.

Existing dissimilarity-based dimension reduction methods (see Section II.A for a brief review) largely cannot incorporate existing knowledge of known features into the dimension reduction process. For instance, Section IV.E discusses a cylinder machining example, in which the cylinders are machined with different combinations of lathe-turning parameters. Indeed, there are often such controllable and/or measurable features in most science and engineering applications. This limitation generally leads to two potential problems. First, failing to utilize all available information generally leads to poorer estimation of the reduced-dimension space. Second, the known features may mask other features in visualization (due to the influence of dimension arrangement [11], dimension order [12], or visualization distortion [13]); this can hamper visualization and knowledge discovery tasks.

In light of this, this paper proposes a novel dissimilarity-based dimension reduction framework, called conditional MDS. Additionally, an optimization algorithm is developed to solve the learning problem of conditional MDS. This algorithm is shown to converge under mild assumptions. The main contribution of this paper is that the proposed conditional MDS methodology and theory endow MDS analysis with the ability to incorporate available knowledge of the known features. This leads to the following advantages:

- Improving the estimation quality of the reduced-dimension space by utilizing more fully available data.
- Simplifying visualization and knowledge discovery tasks in two ways. First, marginalizing the known features out of the reduced-dimension space helps avoid the aforementioned masking problem. Second, the discovered features in previous analyses can be used as the known features in subsequent analyses. This knowledge discovery process is more straightforward because instead of having to recognize all unknown (including unanticipated) features at once, practitioners can now focus on one or a few features at a time.

The organization of this paper is as follows. Section II reviews related works, and Section III presents the proposed method. Section IV shows the advantages of the proposed method over existing dissimilarity-based dimension reduction methods via five examples. Discussions on a number of issues of the proposed method are in Section V. Finally, Section VI concludes the paper.
II. RELATED WORKS

A. Dissimilarity-Based Dimension Reduction

Dissimilarity-based dimension reduction methods directly take pairwise dissimilarities as the input. Specifically, given a $N \times N$ pairwise dissimilarity matrix $\Delta = [\delta_{ij}]_{i,j=1,...,N}$ of $N$ objects, their goal is to find low-dimensional embeddings $\{z_1, z_2, \ldots, z_N\}$ of these $N$ objects.

Classical MDS [14], Sammon mapping [15], and curvilinear component analysis [16] are among the earliest dissimilarity-based dimension reduction techniques. These methods can be considered as instances of a broad class of metric MDS techniques [17]. Given $\Delta = [\delta_{ij}]_{i,j=1,...,N}$, metric MDS finds $\{z_1, z_2, \ldots, z_N\}$ via minimizing the stress function:

$$
\min_{z_1, z_2, \ldots, z_N} \sum_{i<j} w_{ij} \left( ||z_i - z_j|| - \delta_{ij} \right)^2,
$$

(1)

where $w_{ij}$ is a nonnegative weight, $||z_i - z_j||$ is the Euclidean distance between $z_i$ and $z_j$, and $\delta_{ij}$ is some non-negative monotonic transformation of $\delta_{ij}$ ($i, j = 1, 2, \ldots, N$). The weights are only zero when the corresponding dissimilarities are missing.

Different variants of metric MDS can be derived by changing the weighting and transformation schemes. For instance, classical MDS sets $w_{ij} = 1 \forall i, j$; Sammon mapping sets $w_{ij} = 1/(\delta_{ij} \sum_{j<i} \delta_{ij}) \forall i, j$; and curvilinear component analysis sets the $w_{ij}$’s as some decreasing function of $||z_i - z_j||$. All these three methods do not require transformation of the $\delta_{ij}$’s (i.e., $\delta_{ij} = \delta_{ij} \forall i, j$). Classical MDS assumes that $\delta_{ij} = ||z_i - z_j|| \forall i, j$, in which case its solution can be found via an eigendecomposition of the “doubly centered” matrix of $\Delta$.

This assumption is often violated in practice. Therefore, metric MDS generally relaxes this assumption and finds the solution via iterative optimization. SMACOF [18] is the most popular algorithm for solving the optimization problem of metric MDS via majorization [17].

Metric MDS can be viewed as a class of manifold learning algorithms, which aim to find an implicit manifold representation of the data via the embeddings $z_i$’s. Sammon mapping and curvilinear component analysis can learn nonlinear manifolds they emphasize the local behavior through the weights $w_{ij}$’s. Classical MDS can only handle linear manifolds because it does not use any weighting or transformation scheme.

More recently, some popular dissimilarity-based dimension reduction approaches include isometric mapping (ISOMAP) [19], t-distributed stochastic neighbor embedding (t-SNE) [20], and uniform manifold approximation and projection (UMAP) [21]. ISOMAP extends classical MDS to handle nonlinear manifolds by transforming the distances $\delta_{ij}$’s to “geodesic” distances. The geodesic distance is defined as the shortest distance between points along the manifold. It can be estimated by forming a neighborhood graph and minimizing the sum of the Euclidean distances on this graph. ISOMAP obtains the embeddings $z_i$’s via applying classical MDS to the geodesic distances.

Unlike MDS-based approaches, which focus on preserving dissimilarities, t-SNE aims to conserve probabilities. In particular, t-SNE requires that the conditional probability of picking a point as a neighbor of a given point in the embedding space is similar to that in the data space. To achieve this, t-SNE minimizes the Kullback-Leibler divergence of the probability distribution in the embedding space from that in the data space.

UMAP is another approach that diverges from dissimilarity preservation. Particularly, UMAP tries to maintain topological representations of the embeddings and the data. To construct these representations, UMAP uses fuzzy simplicial set representations of local manifold approximations. It finds the solution via minimizing the cross-entropy of the two fuzzy sets.

Numerous variants of these methods have been proposed in the literature, such as local MDS [22], spectral MDS [23], robust kernel ISOMAP [24], upgraded landmark ISOMAP [25], dynamic t-SNE [26], opt-SNE [27], progressive UMAP [28], and density-preserving UMAP [29]. However, these methods are unable to incorporate available knowledge of the manifolds. In many situations, we already know a subset of the manifold features. Recall the cylinder machining example in Section IV.E, if a conventional dimension reduction method is applied in this example, the lathe-turning parameters used to machine the cylinders will be rediscovered [8]. This is unnecessary and may hinder the discovery of other unknown features. Moreover, failing to utilize the available information of these parameters could adversely affect the learning accuracy of the reduced-dimension space. To address these issues, this paper develops the conditional MDS framework to marginalize the effects of the known features to learn solely the other unknown features.

B. Supervised Dimension Reduction

A common usage of dimension reduction methods is for mitigating the curse of dimensionality, to build a better predictive model for some outcome based on the reduced-dimension data. In this setting, dimension reduction can be viewed as a preprocessing step. However, by disregarding the outcome, one could expect that this solution is suboptimal.

To address this problem, dimension reduction methods have been extended to consider the outcome. Many such extensions are based on the manifold learning techniques presented in Section II.A. For example, supervised ISOMAP (S-ISOMAP) [30] transforms the given dissimilarities such that the transformed dissimilarities of intra-class observations are less than that of inter-class observations. Then, ISOMAP is applied to the transformed dissimilarities. Marginal ISOMAP (M-ISOMAP) [31] uses class information to construct constrained neighborhood graphs, in which edges exist if and only if observations are in the same class. This helps avoid several hyper-parameters required by the dissimilarity transformation of S-ISOMAP. Multi-Manifold Discriminant ISOMAP (MMD-ISOMAP) [32] extends M-ISOMAP to the multi-manifold setting (i.e., observations in different classes lie on different manifolds). To this end, MMD-ISOMAP maximizes the inter-class distances in the reduced-dimension space. Semi-supervised local multi-manifold ISOMAP [33] further extends MMD-ISOMAP to utilize both labeled and unlabeled data, via maximizing...
inter-manifold distances and minimizing intra-manifold distances. Note that these extensions can be readily adopted for MDS techniques. Also, there have been extensions to the supervised setting for t-SNE [34, 35], and UMAP [36]. In addition to supervised dissimilarity-based manifold learning, other supervised dimension reduction methods based on PCA and non-negative matrix factorization can be found in the review paper of Chao et al. [37].

However, supervised dimension reduction solves a fundamentally different problem from the one in this paper. Essentially, supervised dimension reduction assumes that there exists a function that maps the features in the reduced-dimension space to the outcome. If we treat the known features as the outcome, we are assuming that such functional relationships exist between the known and unknown features. However, in the context of this paper, the known features do not necessarily depend on the unknown features. For example, in the kinship terms example in Section IV.A, Gender (say, a known feature) is not a function of Generation (say, an unknown feature).

Likewise, the problem and methods proposed in this paper are not the same as those in the literature of “sufficient dimension reduction”. Sufficient dimension reduction approaches (see, e.g., Cook [38] and the references therein) seek for a sufficient statistic $z$, as a function of the given data in a reduced-dimension space, such that the conditional distribution of the outcome given the data is the same as the conditional distribution of the outcome given $z$. Again, this entails a functional relationship between the outcome and $z$.

“Interactive dimension reduction” methods [39], [40] are somewhat related to the proposed conditional MDS framework in the sense that these approaches also incorporate domain knowledge into the dimension reduction process. Nevertheless, a key difference between these methods and conditional MDS is that these methods require practitioners to interact actively with the dimension reduction process. For example, the practitioners adjust features weights, change model parameters, or manipulate outputs, so as to produce comprehensible dimension reduction results. In contrast, there is no direct interaction between conditional MDS and practitioners during the dimension reduction process.

III. CONDITIONAL MULTIDIMENSIONAL SCALING

This section presents the proposed conditional MDS framework and its optimization algorithm. Let $\{u_i = [u_{i1}, u_{i2}, \ldots, u_{ip}]^T \in \mathbb{R}^p: i = 1, 2, \ldots, N\}$ and $\{v_i = [v_{i1}, v_{i2}, \ldots, v_{iq}]^T \in \mathbb{R}^q: i = 1, 2, \ldots, N\}$ denote the values of $p$ unknown features and $q$ known features, respectively, of $N$ objects. Note that $u$ and $v$ constitute all the underlying features that govern the variation in the observed data. Because the intrinsic dimension of the observed data is at most $N - 1$, it is required that $N > p + q$. To make the unknown features $u$ and the known features $v$ compatible in a single manifold coordinate system, an affine transformation $B^Tv$, where $B \in \mathbb{R}^{p \times q}$, is needed. Moreover, denote $U = [u_1, u_2, \ldots, u_N] \in \mathbb{R}^{N \times p}$ and $V = [v_1, v_2, \ldots, v_N] \in \mathbb{R}^{N \times q}$. Thus, $[U, VB] \in \mathbb{R}^{N \times (p + q)}$ contain all the manifold feature values of the $N$ objects.

The objective function of conditional MDS is referred to as the “conditional stress”:

$$\sigma(U, B) = \sum_{i<j} w_{ij} (\tilde{d}_{ij} - d_{ij}(U, B))^2,$$

where the $w_{ij}$’s and $\tilde{d}_{ij}$’s are as defined in (1), and

$$d_{ij}(U, B) = \sqrt{\|u_i - u_j\|^2 + \|B^T(v_i - v_j)\|^2}$$

is the Euclidean distance between the $i$th and $j$th objects in the aforementioned manifold coordinate system for all $i, j$. Conditional MDS minimizes the conditional stress over $U$ and $B$ to reduce the discrepancy between each pair of the $\tilde{d}_{ij}$’s and the $d_{ij}(U, B)$’s. In a sense, this preserves the original information (in the form of dissimilarities) in the reduced-dimension space (in the form of Euclidean distance). Like metric MDS, various variants of the conditional MDS framework can be derived by using different weighting and dissimilarity transformation schemes for the conditional stress function.

In optimization language, the learning task of conditional MDS is solving

$$U^*, B^* = \arg\min_{U \in \mathbb{R}^{N \times p}, B \in \mathbb{R}^{p \times q}} \sigma(U, B).$$

The SMACOF algorithm [18], which solves the learning problem of metric MDS, does not consider the existence of the known features $v$. It attempts to learn all the known and unknown features altogether. It is inapplicable to solve the learning problem (4) of conditional MDS, which utilizes the information of the known features. Hence, this paper develops a “conditional SMACOF” algorithm to solve (4), based on the majorization technique in SMACOF [18]. In brief, this majorization technique constructs an easy-to-minimize majorizing function for an objective function. In conditional SMACOF, the majorizing function for the conditional stress is quadratic in $U$ and $B$. By minimizing this majorizing function, the conditional stress function is also minimized. The update formulas (7) and (8) for $U$ and $B$, respectively, in Theorem 1 indeed correspond to a stationary point of this majorizing function. Under Assumption 1, these updates yield a monotonically decreasing sequence of conditional stress values. The proof of Theorem 1 is based on the majorization technique and Lemma 1.

**Theorem 1:** Define $H = [h_{ij}]_{i,j=1}^N \in \mathbb{R}^{N \times N}$ with

$$h_{ij} = \begin{cases} -w_{ij} & \text{if } i \neq j \\ \sum_{k=1, k \neq i}^N w_{ik} & \text{if } i = j \end{cases},$$

and let $H^+$ be the Moore-Penrose inverses of $H$. Furthermore, denote $C(U, B) = [c_{ij}]_{i,j=1}^N \in \mathbb{R}^{N \times N}$ with

$$c_{ij} = \begin{cases} -w_{ij} \tilde{d}_{ij} & \text{if } d_{ij}(U, B) \neq 0 \\ 0 & \text{if } d_{ij}(U, B) = 0 \end{cases} \quad \text{for } i \neq j,$$

and $c_{ii} = \sum_{j=1}^N c_{ij}$.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Additionally, let $U^{[l]}$ and $B^{[l]}$ denote the values of $U$ and $B$, respectively, at the $l$th iteration. If $U^{[l]}$ and $B^{[l]}$ are updated by

$$U^{[l]} = H^T C \left( U^{[l-1]}, B^{[l-1]} \right) U^{[l-1]}$$

$$B^{[l]} = (V^T H V)^{-1} V^T C \left( U^{[l-1]}, B^{[l-1]} \right) V B^{[l-1]}$$

then under Assumption 1, $\sigma(U^{[l]}, B^{[l]}) \leq \sigma(U^{[l-1]}, B^{[l-1]})$, and the equality occurs when $U^{[l]} = U^{[l-1]}$ and $B^{[l]} = B^{[l-1]}$.

Assumption 1:

a) The dissimilarities are symmetric, and thereby, so are the weights $w_{ij}$ ($i \neq j = 1, 2, \ldots, N$).

b) The weight matrix $[w_{ij}]_{N \times N}$ is irreducible, i.e., it does not contain groups within the $N$ objects for which the intergroup weights are always 0.

c) $V$ contains $q$ linearly independent difference vectors $v_j - v_k$ ($1 \leq j < k \leq N$).

Lemma 1: $V^T H V$ is positive definite under Assumption 1.

In practice, it is recommended that we stop the iterative updates (7) and (8) when the reduction in the conditional stress is less than some threshold. Note that the conditional stress between consecutive iterations, and the max number of iterations $l_{\max}$.

Algorithm 1 summarizes the main steps of the conditional MDS algorithm, based on conditional SMACOF. The required inputs include the given $N \times N$ dissimilarity matrix $\Delta = [\hat{d}_{ij}]_{i,j=1, \ldots, N}$, the known feature values $V$, the weights $w_{ij}$s, the threshold $\gamma$ for min reduction in the normalized conditional stress between consecutive iterations, and the max number of iterations $l_{\\text{max}}$.

Step 1 of Algorithm 1 initializes and pre-computes necessary quantities for the iterations in Step 2. $B$ can be initialized by an identity matrix, and $U$ can be initialized randomly. In Step 2, $U$ and $B$ are iteratively updated based on (7) and (8), respectively, until the number of iterations reaches $l_{\max}$ or the reduction of the normalized conditional stress $\sigma(U, B)$ from the previous iteration is not greater than $\gamma$. Finally, Step 3 returns the optimal $U^{*}, B^{*}$, and normalized conditional stress. Note that practitioners may want to run the algorithms a few times to find the initial configuration that yields the smallest normalized conditional stress. The convergence of Algorithm 1 is guaranteed by Corollary 1. All the proofs can be found in the supplementary materials of this paper.

Corollary 1: Under Assumption 1, $U$ and $B$ will converge to a critical point of the conditional stress function if they are iteratively updated by (7) and (8) as in Algorithm 1 (with $l_{\max} = \infty$ and $\gamma = 0$), respectively.

Algorithm 1: Conditional MDS.

Inputs: $\Delta$, $V$, $w_{ij}$ ($\forall i, j$), $\gamma$, and $l_{\max}$

Step 1:

a) Initialize $U^{[0]}$ and $B^{[0]}$

b) Pre-compute $H$ and $(V^T H V)^{-1} V^T$

c) Compute the initial normalized conditional stress $\sigma_n^{[0]}(U^{[0]}, B^{[0]})$ using (9).

d) Set $l = 0$

Step 2: While ($l < l_{\max}$ and $(\sigma_n^{[l]} - \sigma_n^{[l-1]} > \gamma$) do:

a) $l = l + 1$

b) Update $U^{[l]}$ and $B^{[l]}$ by (7) and (8), respectively.

c) Compute the current normalized conditional stress $\sigma_n^{[l]}(U^{[l]}, B^{[l]})$ using (9).

Step 3: Set $U^{*} = U^{[l]}$, $B^{*} = B^{[l]}$, $\sigma_n(U^{*}, B^{*}) = \sigma_n(U^{[l]}, B^{[l]})$

Outputs: $U^{*}$, $B^{*}$, and $\sigma_n(U^{*}, B^{*})$
Fig. 1. Scatterplot of the 2D coordinates of 14 kinship terms found by: (a) metric MDS, (b) conditional MDS using Gender as the known feature, and (c) conditional MDS, using Gender and Kinship Degree as the known features. In Fig. 1(a), up to some rotation, a feature is apparently Gender, but the other is unclear as it is possibly a combination of several features. In Fig. 1(b), conditional MDS marginalizes Gender out of the dimension reduction result, revealing a change in the kinship degree level along the dashed arrow. In Fig. 1(c), conditional MDS marginalizes both Gender and Kinship Degree out of the dimension reduction result, exposing a generation transition along the dashed curved arrow.

the underlying function that maps $\mathbf{u}$ and $\mathbf{v}$ to the manifold embedded in the data space, say $f(\mathbf{u}, \mathbf{v}) = f(\mathbf{u}, g(\mathbf{u}))$.

Remark 5: Note in (3) that $f$ accounts for the joint effect of the known features $\mathbf{v}$ on the pairwise distances, in compliance with the unknown features $\mathbf{u}$. Therefore, by learning $f$, we can separate the joint effect of the known features on the pairwise distances, from that of the unknown features. This is how conditional MDS can marginalize the known features in the dimension reduction result.

IV. EXAMPLES

This section demonstrates the advantages of conditional MDS via a number of examples. Specifically, Section IV.A illustrates how conditional MDS can simplify visualization and knowledge discovery tasks via a kinship terms example. To demonstrate the benefit of incorporating available knowledge of the known features for improving estimation accuracy of the reduced-dimension space, Sections IV.B, IV.C, IV.D, and IV.E compare conditional MDS with four dissimilarity-based dimension reduction methods (metric MDS, ISOMAP, t-SNE, and UMAP) in four examples (facial expressions, textile fabrics, simulated car-brand perception, and cylinder machining). As mentioned in the introduction, we assume that the input data are pairwise dissimilarities. As such, dimension reduction methods that cannot take pairwise dissimilarities as the input (e.g., PCA and its variants) are not considered here.

A. Kinship Terms

This section demonstrates the advantages of conditional MDS in simplifying visualization and knowledge discovery tasks. As mentioned in Section I, known features can mask unknown features in visualization. To illustrate this problem, Fig. 1(a) plots the two-dimensional (2D) coordinates found by applying metric multidimensional scaling (MDS) to the pairwise dissimilarities of 14 kinship terms in the study of Rosenberg and Kim [41]. Up to some rotation, a feature can be inferred to be Gender (see Borg et al. [42] for a full treatment of MDS result interpretation). Nevertheless, the interpretation for the other feature in Fig. 1(a) is not obvious because Gender may mask this feature.

Conditional MDS addresses this problem by marginalizing the known features out of the dimension reduction results, to expose unknown features. To see this, Fig. 1(b) shows a similar plot to Fig. 1(a), but for the dimension reduction result of conditional MDS, using Gender as the known feature for the kinship terms. Interestingly, we can see in Fig. 1(b) seven pairs of highly similar kinship terms: Sister/Brother (i.e., Sibling), Mother/Father (i.e., Parent), Daughter/Son (i.e., Child), Grandmother/Grandfather (i.e., Grandparent), Granddaughter/Grandson (i.e., Grandchild), Niece/Nephew (i.e., Nibling), and Aunt/Uncle (i.e., Pibling). With Gender marginalized out of the reduced-dimension space, Fig. 1(b) reveals a change in the kinship degree level along the dashed arrow. In this example, conditional MDS helps identify a new “Kinship Degree” feature.

Moreover, conditional MDS allows practitioners to add the discovered features in previous analyses to the known feature set in subsequent analyses, repeatedly. For example, Fig. 1(c) shows a similar plot to Fig. 1(b), but for the dimension reduction result of conditional MDS using Gender and Kinship Degree as the known features. With both Gender and Kinship Degree marginalized out of the reduced-dimension space, Fig. 1(c) exposes a generation transition (Grandparent → Parent/Pibling → Sibling → Child/Nibling → Grandchild) along the dashed curved arrow. Thus, conditional MDS helps identify additionally a new “Generation” feature.

Remark 6: Conditional MDS is not limited for use only with 2D scatterplots as in Fig. 1. One can also use conditional MDS in conjunction with other higher-dimensional visualization techniques (e.g., RadViz plots [43] or star coordinates [44]) to discover several features at once. For illustration, Fig. S1 in the supplementary materials of this paper shows a RadViz plot of the three metric MDS coordinates learned for the 14 kinship terms. This plot reveals the Gender and Kinship Degree features, but not Generation. Nevertheless, a further round of
conditional MDS analysis using Gender and Kinship Degree as the known features can expose Generation as in Fig. 1(c). For ease of comprehension for general readers and practitioners, this paper uses 2D scatterplots.

**B. Facial Expressions**

In this section, conditional MDS is tested with the facial expressions data in the study of Abelson and Sermat [45]. In this work, the authors obtained pairwise dissimilarities for 13 facial expressions (by asking 30 women to give nine-point dissimilarity ratings of each pair of 13 photographs). They showed that the estimates by Engen et al. [46] of the three Schlosberg scales (Pleasant-Unpleasant (PU), Attention-Rejection (AR), and Tension-Sleep (TS)) for these 13 facial expressions accounted for 75% of the variation in the dissimilarities. Note that the absolute correlations between PU-AR, PU-TS, and AR-TS computed from the data by Engen et al. [46] are 0.18, 0.15, and 0.75, respectively. To validate the performance of conditional MDS, one or two of these Schlosberg scales (PU, AR, and TS) is used as the known feature(s). We treat the other scale(s) as unknown for evaluation purposes. In total, this example considers six different scenarios of the known feature sets.

Conditional MDS is compared with metric MDS based on the average canonical correlation (ACC), as follows. For conditional MDS, we first combine the known feature(s) with the learned feature(s) to have three features in total for each scenario. Then, we compute their ACC with the unknown feature(s). The rationale for this calculation is that the reduced-dimension space of conditional MDS can be viewed to consist of the p learned features and the q known features. For metric MDS, we compute the ACC between the three metric MDS dimensions and all three Schlosberg scales, because metric MDS treats all these scales as unknown. Table I reports the ACCs of metric MDS (Column 1) and of conditional MDS for six different scenarios of the known feature set (Columns 2–7). It can be seen from Table I that conditional MDS significantly improves the estimation accuracy of the reduced-dimension space in all scenarios.

| Metric          | Conditional MDS |
|-----------------|-----------------|
| MDS             | [PU] [AR] [TS]  |
| [PU, AR]        | 0.71 0.85 0.94  |
| [PU, AR, TS]    | 0.91 0.89 0.91  |

To gain more insight of conditional MDS, Table II reports the absolute correlations between the first conditional MDS dimension and TS, this dimension also exhibits somewhat high correlation with AR (0.72). This suggests that this dimension is related to both TS and AR, which is due to their inherent correlation. The moderate correlations of the first dimension of conditional MDS with the unknown features when using {PU, AR} or {PU, TS} as the known feature set are also likely due to the inherent strong correlation between AR and TS.

| Known feature set | PU   | AR   | TS   |
|-------------------|------|------|------|
| [PU]              | 0.18 | 0.72 | 0.91 |
| [AR]              | 0.93 | 0.02 | 0.31 |
| [TS]              | 0.92 | 0.14 | 0.15 |
| [PU, AR]          | 0.07 | 0.04 | 0.38 |
| [PU, TS]          | 0.03 | 0.52 | 0.13 |
| [AR, TS]          | 0.93 | 0.07 | 0.16 |

The bold numbers indicate the correlations with the unknown feature(s).

**C. Textile Fabrics**

This section tests conditional MDS on textile fabric images from the textile2 dataset [47]. This dataset contains images taken on non-overlapped areas of two fabric samples. These two fabric samples have slightly different textures. We randomly select 15 images from each sample. For evaluation purposes, we randomly apply the following three changes to these 30 images. The first is a 90-degree rotation of the images. The other two are digital contractions of the images in the horizontal and vertical directions by up to 50%. We use the Bernoulli distribution to determine if an image is rotated. And we use the uniform distribution for the amounts of contractions. The resulting images are then cropped to be of size $250 \times 250$ pixels. Thus, four underlying features govern the differences among the images: texture, rotation, and vertical/horizontal contractions.

We compute the pairwise dissimilarities of these 30 images by two different measures. The first is the Kullback-Leibler (KL) dissimilarity measure for surface images with random textures of Bui and Apley [7]. The second is the Euclidean distance (ED) when we treat the collection of pixel intensities of each image as a vector of $250 \times 250 = 62500$ dimensions. We apply conditional MDS to these dissimilarities, using texture and rotation as the known features. We set the number of unknown features $p = 2$, which is found via the scree-plot-based approach discussed in Section V.V (see Fig. S2 in the supplementary materials of this paper). We compare conditional MDS with metric MDS in the same manner as in Section IV.B, but additionally along with ISOMAP, t-SNE, and UMAP.

For conditional MDS, we compute the ACC of the four underlying features with the combined set of the two known and two learned features by conditional MDS. For the metric MDS, ISOMAP, t-SNE, and UMAP, we compute the ACCs between
the four underlying features and the four learned features by these methods. Table III shows the ACCs of all methods for both KL dissimilarity and ED. It can be seen from Table III that when using the KL dissimilarity measure, conditional MDS almost produces perfect canonical correlations between its learned four-dimensional space and the four underlying features. This performance of all other dimension reduction methods are much poorer because they do not utilize the available information of the known features. Furthermore, the performances of all methods are poor when using ED to calculate the dissimilarities. This is due to the random nature of textile fabric textures, which ED cannot capture completely. The stark contrast between ED and KL here highlights the need for dissimilarity-based dimension reduction, so that suitable dissimilarity measures can be used.

To show more insights of conditional MDS, Table IV reports the correlation of the two learned dimensions by conditional MDS from the KL dissimilarities with the four underlying features. Again, the learned dimensions are strongly correlated with at least one unknown feature (horizontal/vertical contraction) and weakly correlated with the known features (texture and rotation).

### Table III

#### Average Canonical Correlations Between the Unknown Features and the Four-Dimensional Space Learned for the Fabric Images by Conditional MDS and Four Dimension Reduction Methods

| Metric MDS | ISOMAP | t-SNE | UMAP | Conditional MDS |
|------------|--------|-------|------|-----------------|
| KL         | 0.66   | 0.65  | 0.65 | 0.58            | 0.99 |
| ED         | 0.24   | 0.26  | 0.27 | 0.30            | 0.61 |

#### Correlations Between the Two Learned Dimensions by Conditional MDS with the Four Underlying Features in the Textile Fabrics Example

| Horizontal contraction | Vertical contraction | Rotation | Texture |
|------------------------|----------------------|----------|---------|
| 1st dim.               | 0.43                 | 0.87     | 0.05    | 0.14   |
| 2nd dim.               | 0.86                 | 0.46     | 0.03    | 0.08   |

#### Car Features and Their Weights

| Quality | Safety | Value | Perf | Eco | Design | Tech |
|---------|--------|-------|------|-----|--------|------|
| 90      | 88     | 83    | 82   | 81  | 70     | 68   |

### Table IV

D. Car-Brand Perception

This section evaluates conditional MDS with a simulated car-brand perception example. Suppose that the seven features in Table V contribute to the dissimilarities between car brands. The Uniform(0, 1) distribution is used to generate the values of these features for \( N = 30 \) car brands. Then, the pairwise dissimilarities between car brands are calculated as the weighted Euclidean distances of the feature vectors (using the weights in Table V divided by their sum), plus normal noises with zero means and standard deviations equal to 20% of the weighted Euclidean distances. The features and weights are taken from the 2014 Car-Brand Perception Survey of Consumer Reports [48].

Out of the seven features in Table V, we consider three scenarios of the known feature set: Scenario #1 - \{Quality, Safety, Value, Perf\}, Scenario #2 - \{Quality, Safety, Value, Perf, Eco\}, and Scenario #3 - \{Quality, Safety, Value, Perf, Eco, Design\}. We also assume that efforts have been made to estimate the consumers’ perception of the values of the known features for the \( N \) car brands. To simulate the estimates, we add noises to the generated values of the known features for the \( N \) car brands. These noises follow normal distributions with zero means and standard deviations equal to 5% of the generated values of the corresponding known features.

As with the example in Section IV.C, conditional MDS is compared with metric MDS, ISOMAP, t-SNE, and UMAP in terms of the ACCs in this example. However, the ACCs here are calculated between the reduced-dimension space of all methods (including conditional MDS) and the true values of all seven features in Table V (i.e., before adding noises). Table VI reports the medians (and the interquartile ranges in the parentheses) of the ACCs over 100 Monte Carlo replicates for all methods. Table VI shows that conditional MDS improves the learning accuracy of the reduced-dimension space. The more known features are utilized, the better the learning accuracy is.

To provide more insights of conditional MDS, Table VII reports the medians of the absolute correlations of the first conditional MDS dimension with each feature over the 100 Monte Carlo replicates (along with the interquartile ranges in

### Table V

| Quality | Safety | Value | Perf | Eco | Design | Tech |
|---------|--------|-------|------|-----|--------|------|
| 90      | 88     | 83    | 82   | 81  | 70     | 68   |

| Scenario # | Quality | Safety | Value | Perf | Eco | Design | Tech |
|------------|---------|--------|-------|------|-----|--------|------|
| 1          | .11 (.05, .18) | .12 (.05, .17) | .12 (.05, .20) | .10 (.06, .15) | .10 (.06, .16) | .13 (.07, .24) | .12 (.06, .15) | .11 (.06, .19) | .13 (.06, .20) | .10 (.06, .16) | .11 (.06, .18) | .11 (.05, .21) | .91 (.79, .96) | .10 (.04, .16) | .11 (.04, .21) | .27 (.16, .49) | .82 (.32, .93) | .14 (.05, .24) | .28 (.12, .51) | .74 (.41, .94) | .98 (.96, .98) |

The bold numbers indicate correlations with unknown features.
the parentheses). Table VII suggests similar observations with the previous examples. Particularly, the median absolute correlations are weak for the known features and generally strong for at least an unknown feature in all three scenarios. The weak median absolute correlations of the first conditional MDS dimension with the known features in Table VII clearly indicate that conditional MDS effectively marginalizes the effects of the known features.

The results in Table VII also demonstrate the advantage of conditional MDS in simplifying visualization and knowledge discovery tasks. Suppose a practitioner begins with the four known features in Scenario #1 and discover the Eco feature via visualizing the learned conditional MDS dimensions. The practitioner can then repeat this process with these four known features + Eco (i.e., Scenario #2) and discover the Design feature. Repeating this process one more time with the four original known features + Eco + Design (i.e., Scenario #3), the practitioner can discover the Tech feature. This process is illustrated in Fig. S3 in the supplementary materials of this paper. Note that without conditional MDS, the practitioner would need to discover Eco, Design, and Tech together with rediscovering the other four known features at once; this is a much more challenging task.

Because the features in Table V contribute separately to the pairwise dissimilarities, the ground truth of B for each scenario in this example is a diagonal matrix. The diagonal values of this matrix are the weights of the corresponding known features. However, we assume this is unknown and have used the update (8) instead of (10) for B in this example. To demonstrate that conditional MDS can still learn well B, Fig. 2 shows the mean squared error (MSE) \(10^{-5}\) of the estimate of each element in B over 100 Monte Carlo replicates for the car-brand perception example, for three scenarios of the known feature set.

### Table VII

| Scenario | Parameter | MSE (10^-5) |
|----------|-----------|-------------|
| #1       | Turning speed | 0.03        |
|          | Cutting depth | 0.54        |
| #2       | Turning speed | 0.41        |
|          | Cutting depth | 0.00        |
| #3       | Turning speed | 0.02        |
|          | Cutting depth | 0.35        |

The second parameter is cutting depth, with values in \{0.4, 0.8, 1.2\} mm. There are 10 cylinders in each combination.

Suppose that the Euclidean distances between the cylinders are given as their pairwise dissimilarities. Conditional MDS is applied to these dissimilarities, using the two lathe-turning parameters as the known features. The goal is to identify potential unknown features that cause surface variation among the cylinders in this cylinder machining process. Using again the scree-plot-based approach in Section V.B, \(p = 3\) is selected as the number of unknown features (see Fig. S4 in the supplementary materials of this paper for the scree plot of this example).

Table VIII shows the absolute correlations between the lathe-turning parameters and the three conditional MDS dimensions. If the lathe-turning parameters are marginalized completely, we could expect that these correlations are close to zero. However, some of them are moderate. This suggests that the cylinders might not be machined exactly at the set values of the two lathe-turning parameters. Hence, conditional MDS can help the process controller realize that there are potentially unknown factors (e.g., voltage fluctuation or tool wear) causing variation in the two lathe-turning parameters of this machining process.

Recall that the optimal \(p\) value is identified as 3. Therefore, there is another potential unknown feature, other than the two related to the variations in the two lathe-turning parameters. To investigate this, Fig. 3 plots \(e_{i1}^*\) versus \(e_{i2}^*\), noting the top views of some cylinders superimposed at the corresponding values of \(e_{i1}^*\) and \(e_{i2}^*\). Note that these top views are magnified to show
the small radial deviations from the cylinder norm. In general, the top cross-sections of the cylinders are wider than the bottom cross-sections (probably because the cylinders were fixed at the bottom while being machined).

It can be seen from Fig. 3 that the top cross-sections of the cylinders are ellipsoidal. In addition, the angles of their orientations tend to increase in magnitude from the bottom-left to top-right corners. To see this change more clearly, the estimated orientation angles are shown in the middle of the top views. Each estimate is the median of the rotation angles of the 10 top cross-sections of each cylinder. The absolute values of these angle estimates agree with our above observation. Therefore, conditional MDS can help the process controller identify the third unknown feature, which corresponds to rotation of the cylinders around their axis.

V. DISCUSSIONS

A. Validity of Assumptions

This section discusses the validity of the assumptions in Assumption 1, which is required for the convergence of conditional SMACOF. First, Assumption 1(a) virtually always holds because even if the dissimilarities \( \delta_{ij} \) \( (i, j = 1, 2, \ldots, N) \) are not symmetric (which is uncommon in practice), we can make them symmetric by replacing \( \delta_{ij} \) and \( \delta_{ji} \) with \( \delta_{ij} + \delta_{ji} \) \( (i, j = 1, 2, \ldots, N) \). And for symmetric dissimilarities, there are no reasons for using \( w_{ij} \neq w_{ji} \) \( (i, j = 1, 2, \ldots, N) \). Second, Assumption 1(b) causes no loss of generality for conditional SMACOF. If this assumption is violated, we can apply conditional SMACOF to each group. By definition, the weight matrix of each group is irreducible [18], which satisfies Assumption 1(b). Third, Assumption 1(c) is violated if \( V \) contains less than \( q \) linearly independent difference vectors \( v_j - v_k \) \( (1 \leq j < k \leq N) \). This situation is rather rare in practice because the number of the difference vectors grows quadratically with \( N \). Reducing \( q \) is an option to address this situation should it occurs. Alternatively, we can perturb \( V \) slightly such that it contains \( q \) linearly independent difference vectors.

C. Choice of \( p \)

A natural question is how to choose the smallest number of unknown features \( p \) without losing information in the data. The answer is similar to the case of conventional dimension reduction methods, for which the common solution is based on a scree plot of some measure of goodness-of-fit of the dimension reduction result. For example, the eigenvalues are often used in classical MDS, and the stress is commonly used in metric MDS for this purpose. For conditional MDS, we can choose \( p \) at the elbow of the scree plot of the conditional stress versus \( p \).

Fig. 4 shows such scree plots for the car-brand perception example in Section IV.D, using seven different sets of known features: \{Quality\} \((q = 1)\), \{Quality, Safety\} \((q = 2)\), \{Quality, Safety, Value\} \((q = 3)\), \{Quality, Safety, Value, Perf\} \((q = 4)\), \{Quality, Safety, Value, Perf, Eco\} \((q = 5)\), \{Quality, Safety, Value, Perf, Eco, Design\} \((q = 6)\), and \{Quality, Safety, Value, Perf, Eco, Design\} \((q = 7)\). Note that the solutions for \( \delta \) and set \( \hat{b}_m \) \( \leq \) at the elbow of \( N \) are used here, to focus entirely on the estimation quality of the diagonal elements \( b_m \)'s of \( B \).

Denote \( \hat{b}_m \) as the estimate of \( b_m \), and \( s_m \) as the sample standard deviation of the \( m \)th column of \( V \) \((m = 1, 2, \ldots, q)\). Let \( \mathbb{R} \) be the index set of the relevant features and \(|\mathbb{R}|\) be its cardinality. We evaluate the results of conditional MDS in this example based on the following two measures. The first measure is the MSE of the estimates of the \( b_m \)'s corresponding to the relevant features:

\[
\sum_{m \in \mathbb{R}} (\hat{b}_m - b_m)^2 / |\mathbb{R}|
\]

This measure tells us how accurately conditional MDS learns the effects of the relevant known features. The second measure is

\[
r = \frac{\sum_{m \in \mathbb{R}} |b_m| s_m / (q - |\mathbb{R}|)}{\sum_{m \in \mathbb{R}} |b_m| s_m / (q - |\mathbb{R}|)} \]

which can be viewed as a “signal-to-noise” ratio. This measure is the average estimate of the \( b_m \)'s of the relevant features over that

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
of the irrelevant features. The larger $r$ is, the better the accuracy of conditional MDS is. Note that $s_m$ appears in this measure to account for the differences in the feature scales.

Fig. 5 plots these two measures against the number of irrelevant features. We can see that the MSE measure slightly increases as the number of irrelevant features increases up to around 30, but seems to stabilize afterwards. More importantly, the MSEs are very small, even when the number of irrelevant features is close to the number of car brands $N = 100$. This shows that conditional MDS manages to learn well the effects of the relevant known features. We can also see that the “signal-to-noise” ratio $r$ keeps increasing as the number of irrelevant features increases. Because the $b_m$’s of the relevant features remain close to the ground truths (due to the small MSEs), the $b_m$’s of the irrelevant features approach 0 as their number increases.

D. Interpretation With B

This section discusses two interpretations of $B$ (see Section S2 in the supplementary materials of this paper for an illustrative example). The first involves the columns of $B$, where we assume that the known features are standardized. Recall that the known features are linearly combined with each column of $B$, creating affine-transformed features. Thus, we can use the values of each column of $B$ to determine how the known features interrelate in contributing to the pairwise distances in (3). The second involves the rows of $B$, where we do not require standardized known features, but we assume that the dissimilarity $\delta_{ij}$ between the $i$th and the $j$th objects equals to $d_{ij}(U, B)$ plus some noise. From (3), we see that an increase of one unit in the difference in the $m$th known feature between the $i$th and the $j$th objects will contribute an increase to the squared dissimilarity $\delta_{ij}^2$ by the sum of square of the elements in the $m$th row of $B$. Thus, the importance of a known feature to the dissimilarities between objects is reflected in the magnitudes of the elements in its corresponding row in $B$. In the special case when $B$ is strictly diagonal, the sum of square of the $m$th row of $B$ reduces to $b_m^2$. We encounter this situation in the example in Section VC, where the relevant known features have large $b_m^2$’s, whereas the irrelevant features have $b_m^2$’s close to 0.

E. Normalized Conditional Stress

This section illustrates the behavior of the normalized conditional stress in the examples used in this paper. Particularly, Fig. 6 plots the normalized conditional stress (in log scale) against the first 100 iterations of conditional MDS for each example. Obviously, the normalized conditional stress reduces significantly in the first few iterations (especially the first one) in all the examples. Then, it decreases (as expected from Theorem 1) gradually. In general, the normalized conditional stresses reach small values relatively quickly. These observations show that conditional SMACOF can be stopped quite early, at least in these examples. Indeed, this early stopping strategy is also widely used for the conventional SMACOF algorithm.

F. Computational Complexity

This section discusses the computational complexity of conditional SMACOF. The space complexity of conditional SMACOF is $O(N^2)$, and this applies to any algorithm that deals with dissimilarity matrices. Following, we discuss the time complexity of conditional SMACOF and compare it with SMACOF, ISOMAP, t-SNE, and UMAP.

Table S1 in the supplementary materials of this paper provides the time complexities of the main calculations in conditional SMACOF. If we assume that $N >> p + q$, the time complexity of conditional SMACOF can be derived from this table to be: (i) $O(N^2)$ when the $w_{ij}$’s are 1, and (ii) $O(N^{2.373})$ otherwise.

Note that these time complexities are the same with those of the conventional SMACOF algorithm. The time complexities of ISOMAP and t-SNE are both $O(N^2)$ [19], [20], which are the same with that of conditional SMACOF when all the weights $w_{ij}$’s are 1. The time complexity of UMAP largely depends on its nearest neighbor search step [21], which usually have much better empirical performance than $O(N^2)$ if done approximately.

We now compare the computational times of conditional SMACOF (for both weighting schemes of the $w_{ij}$’s) with those of SMACOF, ISOMAP, t-SNE, and UMAP. Here we reuse the car-brand perception example in Section IVD, but vary $N$ to study its impact on the computational time. We use {Quality, Safety, Value, Perf} as the known feature set for conditional SMACOF. If we assume that $N >> p + q$, the time complexity of conditional SMACOF is $O(N^2)$, and this applies to any algorithm that deals with dissimilarity matrices. Following, we discuss the time complexity of conditional SMACOF and compare it with SMACOF, ISOMAP, t-SNE, and UMAP.

Table S1 in the supplementary materials of this paper provides the time complexities of the main calculations in conditional SMACOF. If we assume that $N >> p + q$, the time complexity of conditional SMACOF can be derived from this table to be: (i) $O(N^2)$ when the $w_{ij}$’s are 1, and (ii) $O(N^{2.373})$ otherwise.

Note that these time complexities are the same with those of the conventional SMACOF algorithm. The time complexities of ISOMAP and t-SNE are both $O(N^2)$ [19], [20], which are the same with that of conditional SMACOF when all the weights $w_{ij}$’s are 1. The time complexity of UMAP largely depends on its nearest neighbor search step [21], which usually have much better empirical performance than $O(N^2)$ if done approximately.

We now compare the computational times of conditional SMACOF (for both weighting schemes of the $w_{ij}$’s) with those of SMACOF, ISOMAP, t-SNE, and UMAP. Here we reuse the car-brand perception example in Section IVD, but vary $N$ to study its impact on the computational time. We use {Quality, Safety, Value, Perf} as the known feature set for conditional SMACOF. If we assume that $N >> p + q$, the time complexity of conditional SMACOF is $O(N^2)$, and this applies to any algorithm that deals with dissimilarity matrices. Following, we discuss the time complexity of conditional SMACOF and compare it with SMACOF, ISOMAP, t-SNE, and UMAP.
Conditional MDS is illustrated with several examples, which demonstrate its advantages over conventional dimension reduction methods, due to its capability to utilize available knowledge of known features. This capability helps improve the estimation quality of the reduced-dimension space. Furthermore, this capability helps conditional MDS enhance knowledge discovery processes in two ways. First, conditional MDS exposes only unknown, unanticipated features in the reduced-dimension space; this helps avoid the aforementioned masking problems. Second, conditional MDS enables a repeated, simpler knowledge discovery process, in which discovered features in previous analyses can be used as the known features in subsequent analyses.

Conditional MDS is not without limitations. First, conditional MDS does not directly solve the challenges in interpreting the unknown features. Rather, it helps practitioners approach this task as accurately and easily as possible, by utilizing more fully existing knowledge. Methods that improve the interpretation of dimension reduction results [50] can be explored to address this limitation. Second, the time complexity of conditional MDS does not scale well with $N$. As there is a connection between SMACOF and gradient based optimization [18], stochastic gradient based optimization with an $O(N)$ time complexity may be a potential direction to address the scalability problem. These issues are left for future research.

Another potential research avenue is extending conditional MDS to other application contexts. For instance, a challenge in precipitation forecasting is that data come from nonhomogeneous and overlapped sources, such as satellites, radar, and other weather measurements. If the non-vector data can be transformed into pairwise dissimilarities, the learned conditional MDS features (using the vector data as the known features) can be used as inputs for conventional predictive models. Another example is handling missing data in, for instance, high-dimensional and sparse matrices [51] and dynamically weighted directed network [52]. Data in such cases can be treated as pairwise dissimilarities. Assuming that information of known features is available, conditional MDS can incorporate this knowledge to estimate coordinate representations for the objects with missing dissimilarities. The coordinates can in turn be used to estimate these missing values.

ACKNOWLEDGMENTS

The author thanks the editor and the anonymous associate editor and referees for helping improve this article.

REFERENCES

[1] L. Van Der Maaten, E. O. Postma, and H. J. van den Herik, “Dimensionality reduction: A comparative review,” J. Mach. Learn. Res., vol. 10, no. 66/71, 2009, Art. no. 13.
[2] B. Schölkopf, A. Smola, E. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” Neural Comput., vol. 10, pp. 1299–1319, 1998.
[3] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” Science, vol. 290, no. 5500, pp. 2323–2326, 2000.
[4] M. Cui, L. Li, M. Zhou, J. Li, A. Abusorrah, and K. Sedraoui, “A bi-population cooperative optimization algorithm assisted by an autoencoder for medium-scale expensive problems,” IEEE/CAA J. Autom., vol. 9, no. 11, pp. 1952–1966, Nov. 2022.
