BIOT

Bibal, Adrien; Marion, Rebecca; von Sachs, Rainer; Frénay, Benoît

Published in:
Neurocomputing

DOI:
10.1016/j.neucom.2021.04.088

Publication date:
2021

Document Version
Peer reviewed version

BIOT: Explaining Multidimensional Nonlinear MDS Embeddings using the Best Interpretable Orthogonal Transformation

Adrien Bibal[a,*], Rebecca Marion[b,*], Rainer von Sachs[b], Benoît Frénay[a]

[a] *PReCISE, NADI, Faculty of Computer Science, University of Namur, Rue Grandgagnage 21, B-5000 Namur, Belgium*
[b] *ISBA, LIDAM, Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium*

Abstract

Dimensionality reduction (DR) is a popular approach to data exploration in which instances in a given dataset are mapped to a lower-dimensional representation or "embedding." For nonlinear dimensionality reduction (NLDR), the dimensions of the embedding may be difficult to understand. In such cases, it may be useful to learn how the different dimensions relate to a set of external features (i.e., relevant features that were not used for the DR). A variety of methods (e.g., PROFIT and BIR) use external features to explain embeddings generated by NLDR methods with rotation-invariant objective functions, such as multidimensional scaling (MDS). However, these methods are restricted to two-dimensional embeddings. In this paper, we propose BIOT, which makes it possible to explain an MDS embedding with any number of dimensions without requiring visualization.

Keywords: Multidimensional Scaling, Explainability, Lasso, Orthogonal Transformations

1. Introduction

Interpretability and explainability are hot topics in machine learning. Interpretability refers to the intrinsic capacity of a model to be understandable for a user [1, 2], and the problem of explainability arises for non-interpretable (i.e. black-box) models [3]. Indeed, when machine learning models are black boxes, techniques that are external to the model must be used to provide explanations.

While most of the machine learning literature on interpretability and explainability is framed for a supervised learning context, the need for such concepts also exists in unsupervised learning. For instance, in clustering (or cluster analysis), users may want to understand the meaning behind the clusters found. Similarly, users that perform dimensionality reduction (DR) on their data may be interested in understanding the meaning of the reduced dimensions.

DR is often used when the high-dimensionality of the original dataset makes it difficult to perform data exploration and/or makes data analysis victim to the curse of dimensionality [4, 5], among other problems. However, some of the most effective DR techniques (i.e. UMAP [**?**], *t*-SNE [6], MDS [7], etc.) are nonlinear, which makes the embeddings they generate difficult to interpret. One solution to this problem is to use a set of additional features to explain the dimensions of the low-dimensional embedding.

For example, in psychology, nonlinear dimensionality reduction is commonly applied to datasets containing pairwise comparisons between objects (e.g., the perceived (dis)similarity between pairs of social groups [8]). Additional interpretable features are then used to determine the meaning of the embedding dimensions [8]. In sensometrics, it is also common to study the relationship between embedding dimensions and an external feature set. For instance, some studies seek to identify sensory attributes in one dataset (e.g., flavor, smell of products) that could be used to explain embeddings of consumer preferences in a second dataset (e.g., product appreciation scores) [9].

The aforementioned examples depend on the assumption that the embedding dimensions are them-

*Corresponding authors. Both authors contributed equally.
Email addresses: adrien.bibal@unamur.be (Adrien Bibal), rebecca.marion@uclouvain.be (Rebecca Marion), rainer.vonsachs@uclouvain.be (Rainer von Sachs), benoit.frenay@unamur.be (Benoît Frénay)

selves meaningful. This is not necessarily the case for neighborhood-preserving methods such as t-SNE or UMAP, which do not preserve small distances in the same way as large distances, generating embedding dimensions that can be spatially misleading. However, methods that seek to preserve pairwise distances between instances, like multidimensional scaling (MDS) [7], are good candidates for this explanation approach.

Metric and non-metric MDS [7] are very popular nonlinear dimensionality reduction (NLDR) methods [10] in this category, especially in fields like psychology and ecology, and they are well-developed in the literature. Explanation techniques, such as property fitting (PROFIT), exist to explain MDS embeddings by regressing external features onto the embedding dimensions [11]. PROFIT has several shortcomings [12], but these limitations can be overcome by regressing the embedding dimensions onto the external features using sparse regression techniques such as the Lasso. However, for NLDR methods with objective functions invariant to rotation, such as MDS, this approach requires the optimization of the embedding orientation. Indeed, all rotations of an MDS embedding are equivalent for MDS, but can result in very different regression models in terms of sparsity, interpretability and error.

Best interpretable rotation (BIR) is a state-of-the-art method for solving this problem [13, 12], but it (i) involves exhaustively exploring all possible rotation angles and (ii) is restricted to explanations of two-dimensional (2D) embeddings. In this paper, we propose best interpretable orthogonal transformation (BIOT), a new method that tackles these two issues. First, the objective function for BIOT can be optimized without performing an exhaustive exploration of all possible rotation angles. Second, BIOT makes it possible to easily explain embeddings with more than two dimensions. This second feature of BIOT lifts the requirement of having two dimensions to explore the data, which makes, e.g., 5D and 6D embeddings now useful. Thanks to this, embeddings that have a lower DR loss, and are thus more faithful to the original high-dimensional data, can be studied. Moreover, we show that the performance of BIOT is better than BIR and other state-of-the-art techniques.

This paper is structured as follows. Section 2 motivates the need for explaining NLDR embeddings and highlights the potential explainability of MDS. Section 3.1 introduces the notations used in this paper. The problem tackled in this paper is formally stated in Section 3.2. BIOT, the method proposed to solve this problem, even for embeddings with more than two dimensions, is introduced in Section 3.3. Section 4 presents how regressing embedding dimensions onto external features can be performed using state-of-the-art techniques. A numerical evaluation of the proposed method and state-of-the-art methods is presented in Section 5. In order to clearly highlight the usefulness of BIOT, a case study demonstrates the application of BIOT to explain MDS embeddings in Section 6. Finally, Section 7 concludes the paper.

2. Motivation

The nonlinear dimensionality reduction (NLDR) methods used today produce embeddings that are not always understandable. To compensate for this lack of understandability, or interpretability, NLDR embeddings are often restricted to two or three dimensions so that the data can be explored and analyzed visually. Furthermore, some methods are not even designed to produce higher-dimensional embeddings. For example, Barnes-Hut, the widely used approximation for accelerating the optimization of t-distributed stochastic neighbor embedding (t-SNE) [6], is technically restricted to produce embeddings with three or fewer dimensions (because it uses quadtree for two-dimensional embeddings and octtree for three-dimensional embeddings) [**?**].

One problem with using visualization to analyze NLDR embeddings is that it inherently limits the amount of information from the original dataset that can be represented in the embedding. Moreover, the relative positions of instances in the visualization are not always easy to explain (e.g., why some instances are close together or far apart). This is especially true for neighborhood-preserving NLDR methods (such as t-SNE [6] and uniform manifold approximation and projection (UMAP) [**?**]). These techniques can provide interesting visual results, but are not completely faithful to the original space, as large distances in the original space are less well preserved than small distances [**?**]. As a result, the axes of the visualization (i.e. the embedding dimensions) have no particular meaning.

In contrast, methods that attempt to preserve pairwise distances (e.g., multidimensional scaling (MDS) [7]) are able to generate more spatially meaningful embedding dimensions. As a result, the embedding dimensions can be used as features for characterizing the instances. Moreover, if the meaning of these dimensions is identified, the data can be explored without necessarily resorting to visualization: similarities and dissimilarities between instances can be explained by the embedding dimensions that characterize them.

In this paper, we are interested in the problem of exploring high dimensional datasets using NLDR embeddings with more than two or three dimensions. In particular, we focus on embeddings generated by MDS, a popular distance-preserving method in the literature. The next section describes this problem in detail.

3. Proposed Method

3.1. Notations

Matrices are indicated with bold, upper-case letters (e.g., $\mathbf{X}$) and vectors are indicated using bold, lower-case letters with dot notation, where $\mathbf{x}_{\bullet,j}$ is the j-th column vector in $\mathbf{X}$ and $\mathbf{x}_{i,\bullet}$ is the i-th row vector. Scalar elements from a matrix or vector are indicated using lower-case letters (e.g., x_{ij}). Instances are indexed with the letter $i \in \{1, ..., n\}$, external features with the letter $j \in \{1, ..., d\}$ and embedding dimensions with the letter $k \in \{1, ..., m\}$.

3.2. Problem Definition and Background

Metric and non-metric multidimensional scaling (MDS) [7] are nonlinear dimensionality reduction (NLDR) [10] techniques that are widely used in academia (e.g., in psychology), as well as in industry. Given an $n \times n$ (dis)similarity matrix $\mathbf{Q}$, where n is the number of instances, MDS produces an $n \times m$ embedding $\mathbf{X}$ for a chosen number of dimensions m.

In its most classical form, the objective of MDS is to minimize the stress, a measure of reconstruction error. This means maximizing the match between the dissimilarities of instances in the high-dimensional (HD) space and the pairwise distances in the low-dimensional (LD) space. For instance, Kruskal's stress is defined as

$$\text{Stress} = \sqrt{\frac{\sum_{ii'} \left(d_{ii'}^{\text{HD}} - d_{ii'}^{\text{LD}}\right)^2}{\sum_{ii'} d_{ii'}^{\text{HD}^2}}}, \tag{1}$$

where $d_{ii'}^{\text{HD}}$ (resp. $d_{ii'}^{\text{LD}}$) is the dissimilarity (resp. distance) between the i-th and i'-th instances in HD (resp. LD).

The embedding $\mathbf{X}$ obtained when minimizing the stress is usually used to visually explore the data when $m = 2$. This latter case is called visualization through NLDR [10]. In either case, it is often important to understand the meaning of the MDS dimensions in order to draw conclusions about the data.

One approach for explaining MDS embeddings consists of using an $n \times d$ matrix $\mathbf{F}$ of external features (i.e. features that were not used to produce $\mathbf{Q}$, and therefore not involved in the NLDR process). These external features also allow users to test whether they can explain the embedding with features that were not used to produce it. One popular technique for explaining MDS embeddings with external features is to regress each external feature $\mathbf{f}_{\bullet,j}$ in $\mathbf{F}$ onto the embedding $\mathbf{X}$:

$$\mathbf{f}_{\bullet,j} = \mathbf{X}\mathbf{w} + \mathbf{e}, \tag{2}$$

where $\mathbf{w}$ is a vector of regression weights and $\mathbf{e}$ is an error vector [7]. Property fitting (PROFIT) is based on this idea of fitting external features (called properties) to the induced embedding [11].

Two main issues arise from classical approaches like PROFIT [12]. First, rather than using combinations of external features to explain the embedding, external features are used one by one, thereby providing less insight about the dimensions. Second, the solution requires that the embedding $\mathbf{X}$ be visualized. Indeed, the goal of PROFIT is to show trends in an NLDR visualization. However, one may be interested in explaining an NLDR embedding with more than two dimensions.

One approach to solving the first issue is (i) to reverse the regression direction in order to explain each dimension of the embedding $\mathbf{X}$ on the basis of a linear combination of the external features $\mathbf{F}$, and (ii) to apply a sparsity penalty to the regression weights $\mathbf{W}$ so that each dimension of $\mathbf{X}$ is explained by as few features in $\mathbf{F}$ as possible [13, 12]:

$$\mathbf{X} = \mathbf{F}\mathbf{W} + \mathbf{E}, \tag{3}$$

where $\mathbf{W}$ is sparse. However, the authors in [13, 12] demonstrate that the arbitrary orientation of an MDS embedding is often not the best for balancing model error with sparsity. They show that it is necessary to simultaneously optimize both the sparse weight matrix $\mathbf{W}$ and a rotation matrix $\mathbf{R}$ that controls the orientation of the embedding. The model of interest becomes

$$\mathbf{X}\mathbf{R} = \mathbf{F}\mathbf{W} + \mathbf{E}, \tag{4}$$

where $\mathbf{W}$ is constrained to be sparse. In other words, one must find the rotation $\mathbf{R}$ leading to the sparse regression model that best explains the rotated MDS embedding $\mathbf{X}\mathbf{R}$.

In principle, any transformation matrix $\mathbf{R}$ that preserves all meaningful structure from the original embedding could be used in this framework. Orthogonal transformations, which preserve all pairwise Euclidean distances between instances, are thus good candidates. In this paper, we are interested in the problem of finding the best orthogonal transformation of MDS embeddings of any number of dimensions such that they can

be explained with sparse linear models based on external features. The next section introduces our proposed method, Best Interpretable Orthogonal Transformation (BIOT), for solving this problem.

3.3. BIOT, the Proposed Method

The overall objective of Best Interpretable Orthogonal Transformation (BIOT) is to explain the dimensions of an embedding $\mathbf{X}$ ($n \times m$) using a matrix of external features $\mathbf{F}$ ($n \times d$). BIOT does this by finding an orthogonal $m \times m$ matrix $\mathbf{R}$ such that the transformed embedding can be explained by a sparse weight matrix $\mathbf{W}$ ($d \times m$). Given a hyperparameter $\lambda > 0$, the optimization problem for BIOT is

$$\arg\min_{\mathbf{R},\mathbf{W}} \frac{1}{2n}\|\mathbf{XR} - \mathbf{FW}\|_F^2 + \lambda \sum_{k=1}^{m} \|\mathbf{w}_{\bullet,k}\|_1 \qquad (5)$$

s.t. $\mathbf{R}$ is an un-truncated orthogonal matrix, i.e. $\mathbf{RR}^\top = \mathbf{R}^\top\mathbf{R} = \mathbf{I}_m$.

The orthogonality constraint for $\mathbf{R}$ ensures that the transformed embedding $\mathbf{XR}$ retains all meaningful structure from the original embedding: pairwise euclidean distances and the dimensionality of the embedding are preserved. The Lasso penalty on the columns of $\mathbf{W}$ (i.e. $\sum_{k=1}^{m} \|\mathbf{w}_{\bullet,k}\|_1$) encourages the selection of fewer features per embedding dimension. As a result, the transformed dimensions can be explained by potentially distinct sets of features. The best $\mathbf{R}$ for this problem is the orthogonal transformation that results in the model with the best balance between model error and sparsity, as controlled by the hyperparameter λ.

3.3.1. Optimizing $\mathbf{W}$ for Fixed $\mathbf{R}$

Given a fixed embedding orientation $\mathbf{R}$, the optimization of the weights $\mathbf{W}$ is a Lasso problem. For a particular embedding dimension k, the optimal weight vector is

$$\arg\min_{\mathbf{w}_{\bullet,k}} \frac{1}{2n}\|\mathbf{X}\mathbf{r}_{\bullet,k} - \mathbf{F}\mathbf{w}_{\bullet,k}\|_2^2 + \lambda\|\mathbf{w}_{\bullet,k}\|_1. \qquad (6)$$

Following cyclic coordinate descent optimization [14], all values of $\mathbf{w}_{\bullet,k}$ are fixed, except a certain value w_{jk} at each iteration. The problem to solve can therefore be rewritten as

$$\arg\min_{w_{jk}} \frac{1}{2n}\|\mathbf{e}_{-jk} - \mathbf{f}_{\bullet,j}w_{jk}\|_2^2 + \lambda\|\mathbf{w}_{-jk}\|_1 + \lambda|w_{jk}|, \qquad (7)$$

where $\mathbf{e}_{-jk} = \mathbf{X}\mathbf{r}_k - \mathbf{F}_{-j}\mathbf{w}_{-jk}$, $\mathbf{F}_{-j}$ is $\mathbf{F}$ without its j-th column $\mathbf{f}_{\bullet,j}$ and $\mathbf{w}_{-jk}$ is the weight vector $\mathbf{w}_{\bullet,k}$ without its j-th value w_{jk}. The optimal w_{jk} can be calculated using soft thresholding [14]:

$$w_{jk} = \frac{\text{sign}(\mathbf{f}_{\bullet,j}^\top\mathbf{e}_{-jk})(|\mathbf{f}_{\bullet,j}^\top\mathbf{e}_{-jk}| - n\lambda)_+}{\mathbf{f}_{\bullet,j}^\top\mathbf{f}_{\bullet,j}}. \qquad (8)$$

3.3.2. Optimizing $\mathbf{R}$ for Fixed $\mathbf{W}$

When $\mathbf{W}$ is found, the next step is to adjust the orientation of the embedding. Since $\mathbf{RR}^\top = \mathbf{I}_m$, for fixed $\mathbf{W}$, Eq. (5) can be rewritten as

$$\arg\min_{\mathbf{R}} \frac{1}{2n}\|\mathbf{X} - \mathbf{FWR}^\top\|_F^2 + \lambda \sum_{k=1}^{m} \|\mathbf{w}_{\bullet,k}\|_1 \qquad (9)$$
$$\text{s.t. } \mathbf{R} \text{ is an un-truncated orthogonal matrix,}$$

because

$$\begin{aligned}
&\|\mathbf{XR} - \mathbf{FW}\|_F^2 \\
&= \mathbf{tr}((\mathbf{XR} - \mathbf{FW})^\top(\mathbf{XR} - \mathbf{FW})) \\
&= \mathbf{tr}(\mathbf{R}^\top\mathbf{X}^\top\mathbf{XR} - \mathbf{R}^\top\mathbf{X}^\top\mathbf{FW} - \mathbf{W}^\top\mathbf{F}^\top\mathbf{XR} + \mathbf{W}^\top\mathbf{F}^\top\mathbf{FW}) \\
&= \mathbf{tr}(\mathbf{X}^\top\mathbf{X} - \mathbf{X}^\top\mathbf{FWR}^\top - \mathbf{RW}^\top\mathbf{F}^\top\mathbf{X} + \mathbf{RW}^\top\mathbf{F}^\top\mathbf{FWR}^\top) \\
&= \mathbf{tr}((\mathbf{X} - \mathbf{FWR}^\top)^\top(\mathbf{X} - \mathbf{FWR}^\top)) \\
&= \|\mathbf{X} - \mathbf{FWR}^\top\|_F^2,
\end{aligned} \qquad (10)$$

thanks to the cyclic property of the trace and the fact that $\mathbf{R}$ is an un-truncated orthogonal matrix.

Finding the optimal matrix $\mathbf{R}$ is an orthogonal Procrustes problem [15]. Indeed, for a fixed $\mathbf{W}$, Eq. (9) can be rewritten as

$$\arg\min_{\mathbf{T}} \|\mathbf{A} - \mathbf{BT}\|_F^2 \text{ s.t. } \mathbf{TT}^\top = \mathbf{T}^\top\mathbf{T} = \mathbf{I}_m, \qquad (11)$$

where $\mathbf{A} = \mathbf{X}/\sqrt{2n}$, $\mathbf{B} = \mathbf{FW}/\sqrt{2n}$ and $\mathbf{T} = \mathbf{R}^\top$. The matrix $\mathbf{T}$ that minimizes Eq. (11) can then be found by decomposing the matrix $\mathbf{C} = \mathbf{B}^\top\mathbf{A} = \frac{1}{2n}(\mathbf{FW})^\top\mathbf{X}$ using SVD, such that $\mathbf{C} = \mathbf{U\Sigma V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ contain the left- and right-singular vectors of $\mathbf{C}$ and $\mathbf{T} = \mathbf{UV}^\top$ [16]. The transformation matrix $\mathbf{R}$ optimizing Eq. (9) is thus $\mathbf{R} = \mathbf{T}^\top = \mathbf{VU}^\top$.

3.3.3. Optimization Algorithm

Algorithm 1, inspired by [17], presents BIOT. It is composed of two repeated steps: 1) optimizing $\mathbf{W}$ given an embedding transformation $\mathbf{R}$ and 2) optimizing $\mathbf{R}$ given regression weights $\mathbf{W}$. These steps are repeated until the change of $\mathbf{W}$ from one iteration to another is lower than a predefined threshold[1].

[1] The implementation of BIOT in R can be found at https://github.com/rebeccamarion/BIOT.

Algorithm 1: BIOT algorithm, inspired by [17].

Data: MDS embedding $\mathbf{X}$ and feature matrix $\mathbf{F}$
Result: Explanation of $\mathbf{X}$ with sparse weights $\mathbf{W}$
$\mathbf{R} = \mathbf{I}_m$;
$\mathbf{X} = \mathbf{XR}$;
$\mathbf{W}$ is obtained by solving Eq. (6) for each k of $\mathbf{X}$;
while $\mathbf{W}$ *changes* **do**
  `// Optimizing R`
  $\mathbf{R}$ is obtained by solving Eq. (9);
  $\mathbf{X} = \mathbf{XR}$;
  `// Optimizing W`
  for *each dimension k of* $\mathbf{X}$ **do**
    $\mathbf{w}_{\bullet,k}$ is obtained by solving Eq. (6);
return $\mathbf{W}$ *and* $\mathbf{R}$

3.3.4. Selecting the Hyperparameters λ and m

BIOT requires the selection of two hyperparameters: the λ used for the Lasso penalty, which represents the relative importance of sparsity with respect to error, and the number m of embedding dimensions to analyze. The first hyperparameter, λ, is common to all Lasso problems and can be set according to the same strategies. For instance, the λ leading to the smallest validation mean squared error (validation MSE) can be considered. Alternatively, the "one-standard error" rule [5] may be used, whereby the largest λ within one-standard deviation of the minimum validation MSE is chosen. This corresponds to a sparser model than for the minimum validation MSE model, without resulting in a significantly different level of error.

While sparsity helps avoid issues like overfitting, it is mainly used, in this work, as a means to obtain interpretable regression models (i.e. models with a reasonable number of non-zero weights). Therefore, while the above heuristics can be used to select λ, the final choice remains with the user. In practical settings, it may be interesting to increase the sparsity of regression models, and thus their interpretability, even at the cost of increasing their validation MSE. For the evaluation of BIOT in this paper (Section 5), however, one of the methods presented in the previous paragraph is used to maintain objectivity.

The number m of embedding dimensions is more similar to the hyperparameters used in unsupervised learning. In clustering, for instance, different numbers of clusters must be tested and analyzed, given the knowledge of experts, to see which choice makes sense. Similarly, for BIOT, different numbers of dimensions can be tested to observe how the analysis changes. Oftentimes, increasing m results in explanations with more and more nuance, as seen in the example provided in Section 6.

In addition to increasing the granularity of the explanation, increasing the number of dimensions reduces the information loss in the embedding, making it more faithful to the original dataset. It also makes the explanation of each individual dimension easier, as less information must be explained by the external features. However, these advantages come at the cost of increasing the cognitive load for the user: understanding 10 dimensions simultaneously may be difficult, even if each dimension is explained by only two or three features. Therefore, some balance must be found between cognitive ease and the level of nuance and faithfulness. Gradually increasing the number of dimensions provides a practical means of evaluating when the number of dimensions m becomes too high for cognitive processing.

The next section presents methods that can be seen as competitors to BIOT.

4. Related Work

Best interpretable rotation (BIR) finds rotations of 2D MDS embeddings that can be explained by sparse multiple regression models [13, 12]. The authors of BIR demonstrated that both the model error and number of non-zero weights of Lasso multiple regression models depend on the rotation of the response matrix $\mathbf{X}$. In order to find a rotation balancing model error with interpretability, they proposed finding the best rotation angle θ^* as follows:

$$\theta^* = \arg\min_{\theta} \frac{1}{2n} \|\mathbf{XR}^\theta - \mathbf{FW}^\theta\|_F^2 + \lambda \sum_{k=1}^{m} \|\mathbf{w}^\theta_{\bullet,k}\|_0, \quad (12)$$

where $m = 2$, $\mathbf{R}^\theta$ is the 2D rotation matrix for a given angle θ and $\mathbf{w}^\theta_{\bullet,k}$ is the Lasso solution explaining the k^{th} dimension of $\mathbf{X}$ rotated by $\mathbf{R}^\theta$. While the matrix of weights $\mathbf{W}^\theta$ is the solution to a regression problem with an ℓ_1-norm penalty, BIR's objective function is minimized with respect to a scalar θ, making it feasible to impose an ℓ_0-norm penalty.

Looking for such a θ^* results in better solutions than other potential competitors from the literature [12], but BIR suffers from two important issues. First, θ is optimized by performing an exhaustive search. In practice, an optimization method for non-convex objective functions is used, such as simulated annealing. The solution for this kind of optimization depends on how long users accept to wait for a solution, as a time-stopping threshold is provided as input. Second, BIR can only find a rotation matrix for 2D MDS embeddings.

BIOT addresses both of these weaknesses. It relaxes BIR's constraint that $\mathbf{R}$ be a rotation matrix, allowing $\mathbf{R}$ to be any type of orthogonal matrix (which includes rotation and reflection matrices as special cases). This makes it possible to apply the method to higher-dimensional embeddings, while preserving the meaningful structure in the transformed embedding. BIOT also relaxes the ℓ_0 norm in BIR's objective function to an ℓ_1 norm applied to the columns of $\mathbf{W}$. This makes the objective function bi-convex, and the solution can be found using alternating optimization instead of an exhaustive search.

For MDS embeddings with two or more dimensions m, sparse reduced rank regression (SRRR) [17] could potentially be used to regress transformed embedding dimensions on external features. SRRR was originally introduced as a method for predicting an untransformed response matrix using a weight matrix $\mathbf{C} = \mathbf{W}\mathbf{R}^\top$ of fixed rank r. The original problem presented in [17] is to find $\mathbf{R}$ ($m \times r$) and $\mathbf{W}$ ($d \times r$) by solving

$$\begin{aligned} \arg\min_{\mathbf{R},\mathbf{W}} \frac{1}{2n}\|\mathbf{X} - \mathbf{F}\mathbf{W}\mathbf{R}^\top\|_F^2 + \lambda \sum_{j=1}^{d} \|\mathbf{w}_{j,\bullet}\|_2 \\ \text{s.t. } \mathbf{R}^\top\mathbf{R} = \mathbf{I}_r \text{ and } \operatorname{rank}(\mathbf{W}\mathbf{R}^\top) = r, \end{aligned} \tag{13}$$

where $\mathbf{w}_{j,\bullet}$ is the j^{th} row of $\mathbf{W}$, $\lambda > 0$ and $r \in \{1, ..., \min(d, m)\}$. The second term in Eq. (13) is a Group-Lasso penalty that forces the elements of $\mathbf{w}_{j,\bullet}$ to be either all zero or non-zero [18]. As λ increases, more rows of $\mathbf{W}$ are set to zero, meaning that fewer features are used to explain the response matrix.

The objective function in Eq. (13) can be reformulated to show that the matrix $\mathbf{W}$ in SRRR contains the regression weights for predicting a transformed response matrix $\mathbf{XR}$. Indeed, thanks to the rotational invariance of the Frobenius norm, the first term in Eq. (13) can be rewritten as follows:

$$\frac{1}{2n}\|\mathbf{X} - \mathbf{F}\mathbf{W}\mathbf{R}^\top\|_F^2 = \frac{1}{2n}\|\mathbf{X}\mathbf{R} - \mathbf{F}\mathbf{W}\|_F^2. \tag{14}$$

For the current application, the meaningful structure from the original embedding $\mathbf{X}$ must be preserved. Therefore, SRRR is only applicable when its hyperparameter r (the rank of $\mathbf{W}\mathbf{R}^\top$ and the number of columns in $\mathbf{R}$) is fixed to $r = m$, the number of embedding dimensions. The setting $r = m$ is the only one that ensures that $\mathbf{R}$ is an orthogonal matrix and that the transformed embedding $\mathbf{XR}$ retains the same number of dimensions as the original embedding $\mathbf{X}$.

Despite its potential relevance for the problem at hand, the sparsity constraints in SRRR are less well adapted than the constraints in BIOT. Indeed, for SRRR, the same set of features would be selected for each transformed embedding dimension, making it difficult to attribute a distinct meaning to each dimension. In contrast, BIOT makes it possible to select potentially distinct sets of features for each embedding dimension, providing greater model interpretability.

Other methods in the literature address either the problem of finding an orthogonal transformation or finding a sparse multiple regression model, but not both. Sparse multi-task regression methods (e.g., multi-task Lasso [19], adaptive multi-task Lasso [20], robust feature selection [21] and joint rank and row selection [22]) find a sparse weight matrix but do not transform the response matrix in any way. Latent variable methods, such as eigenvector partial least squares regression (eigen PLS-R) [23, 24], find an orthogonal transformation of a response matrix that improves the prediction of subsequent multiple regression models, but the models are entirely non-sparse. Sparse latent variable approaches such as sparse canonical correlation analysis (SCCA) [25, 26, 27, 28] and sparse partial least squares regression (SPLS-R) [29] estimate sparse regression models, but the transformation of the response matrix is not orthogonal.

The next section evaluates BIOT by comparing its performance with state-of-the-art methods.

5. Evaluation of BIOT

This section compares BIOT with competitors from the literature for MDS embeddings of two or more dimensions.

5.1. Evaluation Datasets

Three real-world datasets are drawn from the field of ecology: the Doubs river fish communities dataset (Doubs) [30], the Oribatid mites dataset (Mite) [31, 32] and the hunting spider dataset (Spider) [33]. Each dataset is made up of two distinct feature sets. The first feature set $\mathbf{Q}$ contains abundances of p different species (of fish, mites and spiders, respectively) measured at n different sampling sites. The second feature set $\mathbf{F}$, in each dataset, corresponds to d features measured at the n sites, such as Cartesian coordinates, water pH and altitude. For each dataset, ordinal MDS is applied to the first feature set $\mathbf{Q}$ in order to produce several embeddings with a number of dimensions m ranging from 2 to $\min(p, d) - 1$.

Table 1: Characteristics of the evaluation datasets

dataset	instances	features		
		Q	F	total
Doubs	30	27	13	40
Mite	70	35	16	51
Spider	28	12	15	27
Stereotypes	80	80	31	111

The fourth dataset used in our evaluation comes from an experiment in psychology about stereotypes (Stereotypes) [8]. In this dataset, the first feature set **Q** contains similarity comparisons made by participants between n social groups (e.g., students, homeless and athletes). The second feature set **F** contains features that encode stereotypes about these social groups (e.g., degree of smartness, trustworthiness and sincerity). The first feature set **Q** ($n \times n$) is used to generate several MDS embeddings, as for the other datasets.

For all datasets, the second feature set **F** (normalized) is used to explain the mean-centered embeddings produced by the MDS of feature set **Q**. Table 1 summarizes the characteristics of the datasets used in the experiments.

5.2. Experimental Protocol

Four methods are compared in this study: BIOT, BIR (for 2D embeddings only), SRRR and eigen PLS with Lasso regression (ePLS+Lasso). For ePLS+Lasso, we add a Lasso step to ordinary eigen PLS in order to benchmark BIOT and to make the results comparable. Eigen PLS is used to estimate a transformation matrix **R**, then Lasso regression is performed based on the transformed embedding. A range of 20 values for λ ($[0.0001, 3.5]/\sqrt{d}$ in logarithmic scale) was chosen such that each method produces solutions ranging from entirely sparse to entirely non-sparse. For SRRR, the rank r of the matrix $\mathbf{WR}^\top$ is fixed to the number of embedding dimensions, as explained in Section 4.

For each method, embedding and value of λ, 10-fold cross-validation is performed to evaluate the average validation mean squared error (MSE), where

$$\text{MSE} = \frac{1}{m}\sum_{k=1}^{m}\frac{1}{p}\sum_{i=1}^{p}(\mathbf{x}_{i,\bullet}^\top\hat{\mathbf{r}}_{\bullet,k} - \mathbf{f}_{i,\bullet}^\top\hat{\mathbf{w}}_{\bullet,k})^2, \qquad (15)$$

with m being the total number of embedding dimensions, p being the total number of instances for which a prediction is made, $\hat{\mathbf{r}}_{\bullet,k}$ is column k from the estimated orthogonal matrix $\hat{\mathbf{R}}$ and $\hat{\mathbf{w}}_{\bullet,k}$ is column k from the estimated weight matrix $\hat{\mathbf{W}}$. Note that the instances i in the formula were not used when estimating $\hat{\mathbf{R}}$ and $\hat{\mathbf{W}}$. The average validation MSEs for each method and embedding are plotted with respect to the average number of non-zero weights in **W** per dimension, where each point represents a value of λ. The minimum of each plotted curve is the minimum validation MSE.

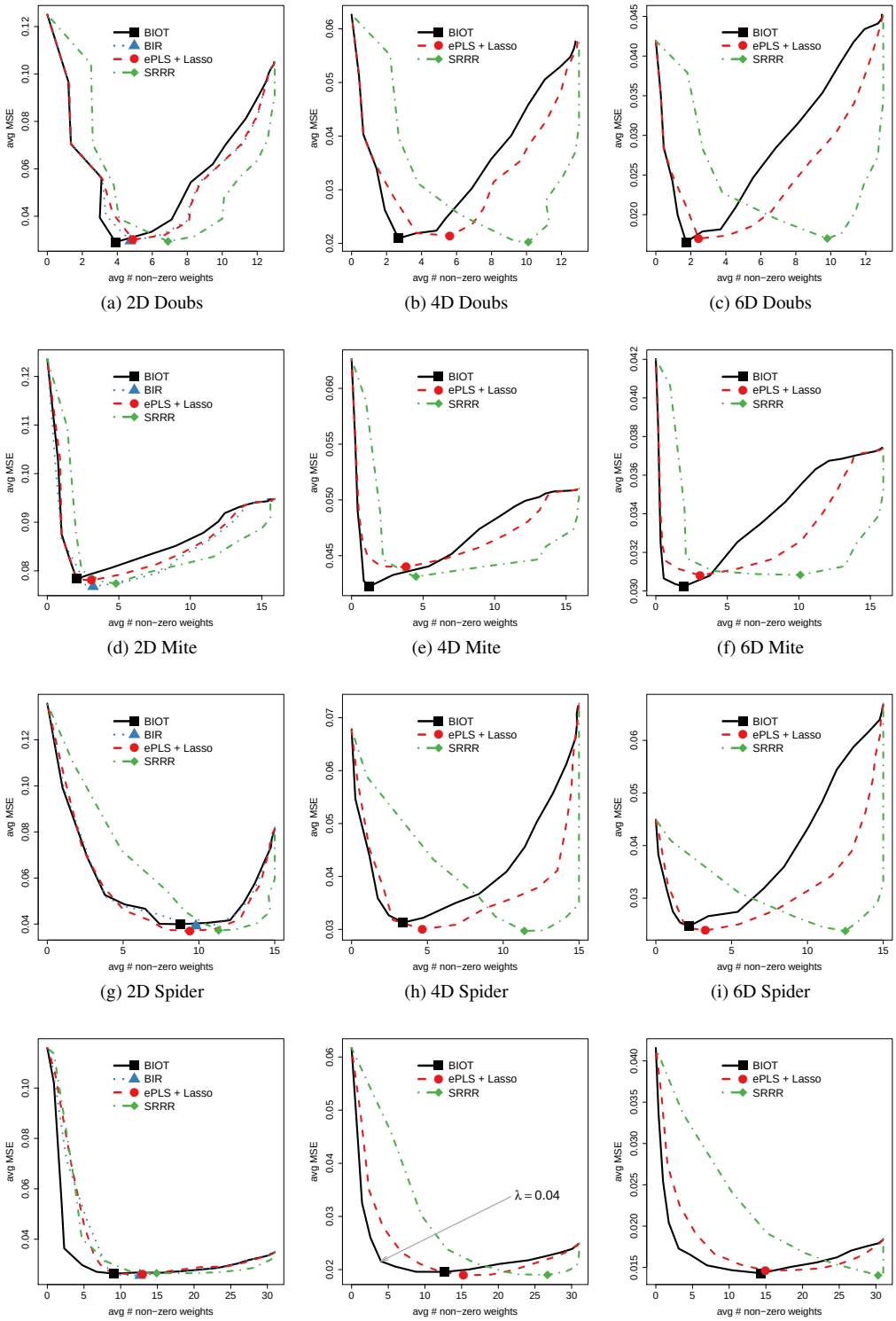
In order to statistically analyze the results obtained for a given dataset and number of dimensions m, the performance of the methods is compared using nested k-fold cross-validation. Each dataset is first split into $k = 10$ outer folds. For each iteration ℓ of an outer loop, the instances in folds $1, ..., \ell-1, \ell+1, ..., 10$ are split into $k = 10$ inner folds. The instances in each fold are characterized by both the (external) features and the response variables, i.e. the MDS coordinates to predict. k-fold cross-validation is performed using the inner folds in order to calculate the average validation MSE for each method and λ. For each method, λ is chosen as the value with the smallest average validation MSE, then the matrices $\hat{\mathbf{R}}$ and $\hat{\mathbf{W}}$ are estimated based on all instances in the inner folds and used to predict the instances in outer fold ℓ. The average test MSE is calculated as the average out-of-sample prediction error (MSE) in the outer loop.

5.3. Results and Discussion

In this experiment, BIOT and the competing methods are applied to embeddings of different numbers of dimensions. The curves for 2D, 4D and 6D embeddings are presented in Fig. 1. For 2D embeddings (Figs. 1a, 1d, 1g and 1j), BIOT finds solutions that are generally as sparse or sparser than those of the other methods, for a similar MSE. Indeed, if a horizontal line is drawn in the graphs (representing a fixed MSE value), BIOT is almost always to the left of the other curves.

Similar trends can be observed for embeddings of more than two dimensions, as shown in the second and third columns of Fig. 1. Note that BIR is not present in these plots, as it can only be applied to 2D embeddings. Interestingly, the difference between the curves is accentuated as the number of embedding dimensions increases. This can be observed as a shifting pattern in the 4D and 6D embeddings of Stereotypes in Fig. 1k and Fig. 1l, compared to a similar but less clear pattern in the 2D embedding of Stereotypes in Fig. 1j. This observation is important, as BIOT is designed for use on higher-dimensional embeddings, where reconstruction error (like the stress) is lower.

The comparison of all methods applied to all embeddings is shown in Table 2. The last embeddings of Stereotypes ($m > 13$) are omitted, as they have stress levels equivalent to the stress for $m = 13$. The results



(a) 2D Doubs (b) 4D Doubs (c) 6D Doubs

(d) 2D Mite (e) 4D Mite (f) 6D Mite

(g) 2D Spider (h) 4D Spider (i) 6D Spider

(j) 2D Stereotypes (k) 4D Stereotypes – The arrow indicates the λ used in the case study in Section 6. (l) 6D Stereotypes

Figure 1: Performance of BIOT, BIR, ePLS+Lasso and SRRR for several λ values. The average validation MSE is plotted against the average number of non-zero weights per dimension. The three columns represent 2D, 4D and 6D embeddings, and the four rows represent the datasets Doubs, Mite, Spider and Stereotypes. The minimum validation MSE is highlighted for each method with a colored symbol.

for each method are shown as a pair of values (average number of non-zero weights per dimension, average test MSE). In order to report an error value that was not used for selecting λ, nested 10-fold cross-validation is used (see Section 5.2 for more details). On each line, results with the highest sparsity (resp. lowest MSE) are highlighted in bold (resp. italics), as well as all other results that are not significantly different according to a pairwise Wilcoxon signed-rank test ($\alpha = 0.05$). Any results not shown in bold or italics are significantly worse than the best results across the different folds.

As seen in Table 2, the best MSE is generally not significantly different for all methods, but the average number of non-zero weights often is. Most of the time, BIOT provides solutions with a lower number of features per dimension, while having a test MSE similar to its competitors. The average number of features used to explain a dimension generally decreases as the number of dimensions m increases. This can be explained by the fact that each new embedding dimension adds less information than previous ones (the stress decreases less). Therefore, fewer and fewer features are needed to explain each additional dimension.

In the next section, several embeddings of Stereotypes are analyzed using BIOT to demonstrate the interpretation of an MDS embedding with more than two dimensions.

Table 2: Results for four datasets Doubs (Do), Mite (Mi), Spider (Sp) and Stereotypes (St). Each result is a pair (average number of non-zero weights, average test MSE) corresponding to the average across the 10 outer folds of the nested 10-fold cross-validation.

	m	stress	BIR	BIOT	ePLS	SRRR
Do	2	0.070	**6.7**, *0.096*	**6.3**, *0.092*	**7.0**, *0.094*	7.0, *0.098*
	3	0.038		**4.8**, *0.060*	**5.4**, *0.061*	5.4, *0.063*
	4	0.026		**4.8**, *0.046*	**5.3**, 0.056	5.3, *0.049*
	5	0.018		**3.9**, *0.041*	**3.4**, *0.047*	3.4, *0.040*
	6	0.013		**2.9**, *0.037*	**2.9**, *0.036*	2.9, *0.038*
	7	0.012		**1.8**, *0.034*	1.8, *0.032*	1.8, *0.037*
	8	0.008		**1.5**, *0.029*	1.5, *0.030*	1.5, *0.033*
	9	0.006		**1.3**, *0.026*	1.3, *0.028*	1.3, *0.031*
	10	0.005		**1.0**, *0.022*	1.0, 0.026	1.0, 0.030
	11	0.004		**1.1**, *0.023*	1.1, *0.024*	1.1, 0.028
	12	0.003		**1.0**, *0.022*	1.0, *0.022*	1.0, 0.026
Mi	2	0.144	1.0, *0.156*	**3.0**, *0.161*	**3.4**, *0.160*	3.4, *0.159*
	3	0.112		**1.7**, *0.104*	1.7, *0.107*	1.7, *0.105*
	4	0.091		**1.2**, *0.084*	1.2, 0.090	1.2, *0.088*
	5	0.077		**1.2**, *0.072*	1.2, 0.075	1.2, *0.073*
	6	0.065		**1.8**, *0.061*	1.8, *0.063*	1.8, 0.063
	7	0.057		**1.5**, *0.053*	1.5, *0.054*	1.5, 0.055
	8	0.049		**1.5**, *0.048*	1.5, *0.048*	1.5, *0.049*
	9	0.044		**1.2**, *0.043*	1.2, *0.044*	1.2, *0.044*
	10	0.040		**1.3**, *0.040*	1.3, *0.040*	1.3, *0.039*
	11	0.036		**1.0**, *0.036*	1.0, *0.037*	1.0, *0.036*
	12	0.032		**1.1**, *0.033*	1.1, *0.034*	1.1, *0.033*
	13	0.029		**0.9**, *0.031*	0.9, *0.031*	0.9, *0.031*
Sp	2	0.089	0.9, *0.085*	**8.5**, *0.081*	8.5, *0.077*	8.5, *0.075*
	3	0.055		**4.0**, *0.071*	4.0, *0.069*	4.0, *0.080*
	4	0.037		**4.3**, *0.065*	4.3, *0.062*	4.3, *0.063*
	5	0.025		**3.3**, *0.060*	**3.6**, *0.055*	3.6, *0.059*
	6	0.019		**2.4**, *0.053*	2.4, *0.048*	2.4, *0.048*
	7	0.016		**1.9**, *0.044*	1.9, *0.043*	1.9, *0.042*
	8	0.012		**1.6**, *0.039*	1.6, *0.039*	1.6, *0.040*
	9	0.007		**1.5**, *0.036*	1.5, *0.036*	1.5, *0.036*
	10	0.004		**1.4**, *0.033*	1.4, *0.033*	1.4, *0.033*
	11	0.001		**1.4**, *0.031*	1.4, *0.030*	1.4, *0.030*
St	2	0.291	**11.3**, *0.053*	**10.8**, 0.058	**10.4**, *0.054*	10.4, *0.060*
	3	0.207		**12.7**, 0.036	**11.7**, *0.033*	11.7, *0.033*
	4	0.169		**9.7**, *0.040*	9.7, *0.039*	9.7, *0.039*
	5	0.146		**10.2**, 0.035	10.2, *0.034*	10.2, *0.032*
	6	0.134		**9.7**, 0.030	9.7, *0.030*	9.7, *0.029*
	7	0.127		**6.5**, *0.028*	6.5, 0.028	6.5, *0.027*
	8	0.122		**7.2**, *0.025*	7.2, *0.025*	7.2, *0.025*
	9	0.120		**8.0**, *0.022*	8.0, *0.023*	8.0, *0.022*
	10	0.118		**7.0**, 0.022	7.0, 0.022	7.0, *0.021*
	11	0.116		**6.4**, *0.020*	6.4, *0.020*	6.4, *0.020*
	12	0.116		**6.5**, *0.018*	6.5, 0.019	6.5, *0.018*
	13	0.115		**5.0**, 0.018	5.0, 0.018	5.0, *0.017*

6. Case Study: Applying BIOT to Stereotypes

The Stereotypes dataset was collected in order to study how people (in the US) implicitly assign stereotypes to social groups. In a first experiment, participants ranked the similarity between social groups, such as celebrities, students and criminals (feature set **Q**). In a second experiment, participants scored these social groups with respect to stereotypes, such as wealthy, altruistic and skillful (external features **F**). The goal was then to see how these stereotypes could explain perceived similarities between social groups.

Let us consider that a researcher in psychology decides to use Lasso regression models to explain the dimensions of three MDS embeddings of the Stereotypes dataset **Q** (embeddings with $m = 3$, $m = 4$ and $m = 5$ dimensions). In this scenario, the researcher initially assumes that transforming the embeddings is not necessary (i.e. $\mathbf{R} = \mathbf{I}_m$). For Lasso, it is common practice to choose a λ resulting in the sparsest, most interpretable model possible while maintaining a low MSE. In order to remain objective, λ is chosen as follows. For λ with the smallest average validation MSE, a 95% confidence interval is calculated. Then, the largest λ value with an average validation MSE within this confidence interval is selected. The regression weights for this approach are shown in Table 3. It can be seen that the number of external features used to explain each of the embedding dimensions is large, and that the same external feature is sometimes used to explain several dimensions at once (e.g., for $m = 3$, wealthy is used with a large coefficient to explain the first and the third dimensions).

In order to generate results that are more interpretable, the researcher then applies BIOT (i.e. **R** is optimized). The same approach is used to select the value of the hyperparameter λ (the chosen value is highlighted

Table 3: Lasso weights for three untransformed embeddings of the Stereotypes dataset (embedding dimensions in rows, Lasso weights for the original, untransformed dimensions in parentheses). The most important features for each dimension are in bold.

$m = 3$	$m = 4$	$m = 5$
wealthy (0.14)	**wealthy** (0.14)	**wealthy** (0.13)
conservative (0.09)	conservative (0.08)	conservative (0.08)
conventional (0.07)	conventional (0.1)	conventional (0.06)
safety (0.07)	safety (0.06)	safety (0.05)
not smart (-0.03)	not smart (-0.01)	not smart (-0.03)
uniformity (0.02)	uniformity (0.01)	uniformity (0.02)
traditional (0.02)		traditional (0.02)
confident (0.01)	confident (0.01)	powerful (0.02)
loyalty (0.01)	loyalty (0.01)	communal (0.02)
competitive (-0.08)	competitive (-0.03)	competitive (-0.06)
masculine (-0.03)	masculine (-0.06)	masculine (-0.04)
communal (0.02)	communal (0.01)	
religious (0.02)	religious (0.01)	religious (0.04)
typical (-0.02)	typical (-0.01)	typical (-0.03)
egoistic (-0.1)	**egoistic** (-0.11)	**egoistic** (-0.09)
not smart (-0.1)	not smart (-0.06)	**not smart** (-0.1)
		uniformity (-0.01)
religious (-0.2)	**religious** (-0.2)	**religious** (-0.17)
wealthy (0.13)	wealthy (0.1)	wealthy (0.12)
familiarity (-0.06)	familiarity (-0.04)	familiarity (-0.05)
masculine (-0.05)	masculine (-0.05)	masculine (-0.05)
traditional (-0.02)	traditional (-0.02)	traditional (-0.05)
conventional (-0.01)		promotion (0.01)
intolerant (-0.01)		
loyalty (-0.01)		
	comfort (0.06)	**comfort** (0.1)
	friendly (-0.06)	friendly (-0.03)
	conventional (-0.04)	conventional (-0.05)
	loyalty (-0.02)	loyalty (-0.04)
		cold (0.02)
		untrustworthy (0.02)
		powerful (0.02)
		typical (-0.02)
		competitive (-0.01)
		threatening (0.01)
		change (0.06)
		communal (0.06)
		competitive (0.04)
		masculine (0.02)
		conservative (-0.01)
		friendly (0.01)
		powerful (0.01)
		promotion (0.01)

in Fig. 1k) and regression weights are estimated for each embedding (see Table 4).

In contrast to the initial approach (Lasso without embedding transformation), BIOT selects fewer external features per dimension, and each dimension is explained by a distinct set of features. For the first embedding ($m = 3$, first column of Table 4), BIOT explains the three dimensions with the stereotypes wealthy, traditional/conventional and not smart. For the 4D embedding (second column of Table 4), BIOT provides an explanation of the fourth dimension by roughly separating traditional and conventional into two dimensions. Finally, for the 5D embedding (third column of Table 4), BIOT explains the new fifth dimension as a political dimension through the conservative-liberal stereotype.

Table 4: BIOT weights for three embeddings of the Stereotypes dataset (embedding dimensions in rows, BIOT weights for the transformed dimensions in parentheses). The most important features for each dimension are in bold.

$m = 3$	$m = 4$	$m = 5$
wealthy (0.28)	**wealthy** (0.26)	**wealthy** (0.22)
scientific (0.06)		powerful (0.05)
uniformity (0.05)	uniformity (0.06)	uniformity (0.01)
traditional (0.17)	traditional (0.04)	**traditional** (0.08)
religious (0.01)	**religious** (0.15)	**religious** (0.09)
	comfort (0.04)	comfort (0.04)
	prevention (0.02)	prevention (0.04)
conventional (0.15)	**conventional** (0.22)	**conventional** (0.14)
loyalty (0.05)	loyalty (0.07)	loyalty (0.01)
familiarity (0.01)		communal (0.01)
		friendly (0.01)
not smart (0.16)	**not smart** (0.13)	**not smart** (0.16)
egoistic (0.07)	egoistic (0.05)	egoistic (0.01)
masculine (0.06)	masculine (0.09)	masculine (0.04)
competitive (0.06)		competitive (0.05)
typical (0.04)	typical (0.03)	typical (0.03)
	intolerant (0.02)	
	familiarity (0.01)	
		conservative (0.14)
		masculine (0.03)
		preservation (0.03)

The advantage of analyzing more than two dimensions is that higher-dimensional embeddings have less reconstruction error, which is quantified by Kruskal's stress (see Section 3.2). Moreover, with BIOT, it is possible to observe the way in which low dimensional embeddings approximate trends from higher-dimensional embeddings. For example, when changing from 4D to 3D (0.169 to 0.207 in stress), BIOT associates the traditional/religious and conventional stereotypes with a single dimension, rather than two. This combination may explain the increase in stress in the 3D embedding, as two orthogonal trends are approximated by a single trend.

By adding dimensions, it is also possible to identify trends that are not apparent in lower-dimensional embeddings. While the original study did not identify smartness as a relevant stereotype, the 3D analysis with BIOT identifies it as important for explaining a third dimension in the data. Indeed, social groups such as criminals and red necks are identified as egoistic, masculine and not smart by the participants. At the same time, the two other dimensions explained by BIOT correspond to the findings in the original paper,

where MDS embeddings were explained by two quasi-orthogonal trends: the socio-economical status (represented here by wealthy) and the type of beliefs (represented here by the stereotypes conventional and traditional) [8].

This case study shows how insightful it is to use BIOT to analyze MDS embeddings with more than two dimensions. Given BIOT's sparsity and MSE performance, increasing the number of dimensions (and therefore reducing the stress) does not make the new embedding much more difficult to understand. Indeed, each new dimension is explained by small, generally disjoint sets of external features. The results of this case study were presented to the main investigator of the original study [8], who found them coherent with current theory in psychology, while providing interesting insights.

7. Conclusion

In this paper, we proposed a method, called BIOT, that makes it possible to explain MDS embeddings of any number of dimensions. BIOT is based on an iterative optimization of two parameter matrices: a weight matrix $\mathbf{W}$ and an orthogonal transformation matrix $\mathbf{R}$.

BIOT was evaluated on datasets corresponding to real-world problems. We demonstrated that BIOT outperforms competitive methods with respect to the interpretability of solutions. The analysis of MSE-sparsity curves revealed that, for the same level of MSE, BIOT provides models that are more sparse, and thus easier to interpret. In order to demonstrate BIOT's ease of use, a case study based on a dataset from a psychological experiment on stereotypes was presented.

In future work, a grouping-penalty could be added to BIOT to encourage groups, rather than individual features, to be selected for each embedding dimension. By grouping features in a meaningful way, even models with many features could be easily interpreted. This would be advantageous for datasets where mostly non-sparse models have the best test MSE or for applications where feature grouping is desired.

Acknowledgements

The authors would like to thank Alex Koch, assistant professor at the University of Chicago, for his feedback on the application of BIOT to his dataset. We also thank Reviewer 2 for proposing the proof in Eq. 9. The work of R. Marion was supported by the Belgian Fund for Scientific Research (F.R.S.-FNRS, FRIA grant).

References

[1] A. Bibal, B. Frénay, Interpretability of machine learning models and representations: an introduction, in: Proceedings of the European Symposium on Artificial Neural Networks, Bruges, Belgium, 2016, pp. 77–82.

[2] Z. C. Lipton, The mythos of model interpretability, in: ICML Workshop on Human Interpretability of Machine Learning, New York, USA, 2016.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys 51 (5) (2018) 1–42.

[4] R. E. Bellman, Adaptive control processes: a guided tour, Princeton university press, 1961.

[5] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag New York, 2009.

[6] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (Nov) (2008) 2579–2605.

[7] J. B. Kruskal, M. Wish, Multidimensional Scaling, Sage, 1978.

[8] A. Koch, R. Imhoff, R. Dotsch, C. Unkelbach, H. Alves, The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion, Journal of Personality and Social Psychology 110 (5) (2016) 675–709.

[9] T. Næs, P. B. Brockhoff, O. Tomic, Statistics for sensory and consumer science, John Wiley & Sons, 2011.

[10] J. A. Lee, M. Verleysen, Nonlinear Dimensionality Reduction, Springer, 2007.

[11] J. J. Chang, J. D. Carroll, How to use PROFIT, a computer program for property fitting by optimizing nonlinear or linear correlation, Unpublished Manuscript, Bell Laboratories (1968).

[12] R. Marion, A. Bibal, B. Frénay, BIR: A method for selecting the best interpretable multidimensional scaling rotation using external variables, Neurocomputing 342 (2019) 83–96.

[13] A. Bibal, R. Marion, B. Frénay, Finding the most interpretable MDS rotation for sparse linear models based on external features, in: Proceedings of the European Symposium on Artificial Neural Networks, Bruges, Belgium, 2018, pp. 537–542.

[14] T. Hastie, R. Tibshirani, M. Wainwright, Statistical Learning with Sparsity: the Lasso and Generalizations, Chapman and Hall/CRC, 2015.

[15] J. C. Gower, G. B. Dijksterhuis, Procrustes Problems, Oxford University Press, 2004.

[16] G. H. Golub, C. F. Van Loan, Matrix Computations, Johns Hopkins University Press, 2013.

[17] L. Chen, J. Z. Huang, Sparse reduced-rank regression for simultaneous dimension reduction and variable selection, Journal of the American Statistical Association 107 (500) (2012) 1533–1545.

[18] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (1) (2006) 49–67.

[19] G. Obozinski, B. Taskar, M. Jordan, Multi-task feature selection, Statistics Department, UC Berkeley, Tech. Rep 2 (2.2).

[20] S. Lee, J. Zhu, E. P. Xing, Adaptive multi-task lasso: with application to eqtl detection, in: Advances in neural information processing systems, 2010, pp. 1306–1314.

[21] F. Nie, H. Huang, X. Cai, C. H. Ding, Efficient and robust feature selection via joint 2, 1-norms minimization, in: Advances in neural information processing systems, 2010, pp. 1813–1821.

[22] F. Bunea, Y. She, M. H. Wegkamp, et al., Joint variable and rank selection for parsimonious estimation of high-dimensional matrices, The Annals of Statistics 40 (5) (2012) 2359–2388.

[23] K. Varmuza, P. Filzmoser, Introduction to Multivariate Statistical Analysis in Chemometrics, CRC press, 2016.
[24] H. Abdi, Partial least squares regression and projection on latent structure regression (PLS regression), Wiley Interdisciplinary Reviews: Computational Statistics 2 (1) (2010) 97–106.
[25] D. M. Witten, R. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (3) (2009) 515–534.
[26] I. Wilms, C. Croux, Sparse canonical correlation analysis from a predictive point of view, Biometrical Journal 57 (5) (2015) 834–851.
[27] X. Suo, V. Minden, B. Nelson, R. Tibshirani, M. Saunders, Sparse canonical correlation analysis, arXiv preprint arXiv:1705.10865.
[28] Q. Mai, X. Zhang, An iterative penalized least squares approach to sparse canonical correlation analysis, Biometrics.
[29] H. Chun, S. Keleş, Sparse partial least squares regression for simultaneous dimension reduction and variable selection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (1) (2010) 3–25.
[30] J. Verneaux, Cours d'eau de franche-comté (massif du jura). recherches écologiques sur le réseau hydrographique du doubs. essai de biotypologie., Ph.D. thesis, Université de Besançon (1973).
[31] D. Borcard, P. Legendre, P. Drapeau, Partialling out the spatial component of ecological variation, Ecology 73 (3) (1992) 1045–1055.
[32] D. Borcard, P. Legendre, Environmental control and spatial structure in ecological communities: an example using oribatid mites (acari, oribatei), Environmental and Ecological Statistics 1 (1) (1994) 37–61.
[33] N. Smeenk-Enserink, P. Van Der Aart, Correlations between distributions of hunting spiders (lycosidae, ctenidae) and environmental characteristics in a dune area, Netherlands Journal of Zoology 25 (1) (1974) 1–45.