BIR: A Method for Selecting the Best Interpretable Multidimensional Scaling Rotation using External Variables

Marion, Rebecca; Bibal, Adrien; Frénay, Benoît

Published in:
Neurocomputing

DOI:
10.1016/j.neucom.2018.11.093

Publication date:
2019

Document Version
Early version, also known as pre-print

BIR: A Method for Selecting the Best Interpretable Multidimensional Scaling Rotation using External Variables

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

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

Abstract

Interpreting nonlinear dimensionality reduction models using external features (or external variables) is crucial in many fields, such as psychology and ecology. Multidimensional scaling (MDS) is one of the most frequently used dimensionality reduction techniques in these fields. However, the rotation invariance of the MDS objective function may make interpretation of the resulting embedding difficult. This paper analyzes how the rotation of MDS embeddings affects sparse regression models used to interpret them and proposes a method, called the Best Interpretable Rotation (BIR) method, which selects the best MDS rotation for interpreting embeddings using external information.

Keywords: Interpretability, Dimensionality Reduction, Multidimensional Scaling, Orthogonal Transformation, Multi-View, Sparsity, Lasso Regularization

1. Introduction

Dimensionality reduction consists of mapping instances from a certain space into a lower-dimensional space. For *nonlinear dimensionality reduction* (NLDR), this mapping is nonlinear, meaning that the new representation of the instances is not a linear transformation of the instances in the original space. NLDR is especially useful when the relationship between features is not linear, for instance in psychology [2] and ecology [3]. However, the nonlinear mapping of instances from high to low dimension makes it difficult to interpret the resulting embedding, whose axes do not have an easily apparent meaning.

In many cases, interpretability is essential to the use of machine learning models [4]. In the context of NLDR, the model of interest is the nonlinear mapping function, which is sometimes interpreted based on an additional set of features. By studying the relationship between the NLDR output and this set of features, the model that generated the output can be interpreted. For example, in implicit measure studies in psychology [5], data describing a given set of instances are collected in two, often independent, experiments. The instances from one experimental dataset are mapped into a reduced space using multidimensional scaling, and then the features from the other dataset are used to interpret the mapping by finding trends with linear functions.

Using a second set of features to interpret an NLDR embedding is also a popular approach in ecology [3]. For instance, a collection of abiotic features – such as soil acidity, temperature and altitude – may be used to interpret similarities and differences between sampling sites in terms of species abundance. A dataset of species abundance for a variety of sampling sites is mapped to a lower-dimensional space using an NLDR method, and then a dataset of abiotic features for these same sites is used to identify a link between abiotic environmental conditions and species abundance.

This approach to the interpretation of NLDR is an example of *multi-view learning*, also known as *data fusion*, or *coupled*, *linked*, *multiset*, *multiblock* or *integrative data analysis* [6], where different feature sets are used to solve a machine learning problem [7]. In this particular case, one view (the m-dimensional NLDR embedding of n instances) is interpreted using another view (d features of the same n instances, i.e. "exter-

*Corresponding authors. Both authors contributed equally. Name order reversed with respect to [1].

Email addresses: rebecca.marion@uclouvain.be (Rebecca Marion), adrien.bibal@unamur.be (Adrien Bibal), benoit.frenay@unamur.be (Benoît Frénay)

nal" variables, which were not used to compute the embedding). Similar two-view problems are encountered in a variety of fields, including, but not restricted to, psychology [2], epidemiology [8], ecology [9], biology [10] and chemometrics [11].

In this work, we are interested in NLDR methods whose objective function is rotation-invariant, particularly *multidimensional scaling* (MDS) [12]. MDS is an NLDR technique that takes an $n \times n$ (dis)similarity or distance matrix $\mathbf{D}$ as input and outputs an embedding (or configuration) $\mathbf{X}$ of these n instances in an m-dimensional space, with $m \ll n$ [12, 13]. More precisely, MDS finds a matrix $\mathbf{X}$ such that (dis)similarities d_{ij} in $\mathbf{D}$ can be mapped to distances between m-dimensional vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ with minimal loss.

In order to find this mapping, the MDS algorithm must minimize a loss function often called the *stress function*. This stress function can take many forms, but one of the most frequently used functions is Kruskal's stress function [12, 13]:

$$\text{stress} = \sqrt{\frac{\sum_{i,j}[d_{ij} - \text{dist}(\mathbf{x}_i, \mathbf{x}_j)]^2}{\sum_{i,j} d_{ij}^2}}. \qquad (1)$$

One particular property of this stress function is that by preserving the relative distances between each pair of instances, the stress function is invariant to a variety of transformations of $\mathbf{X}$. Indeed, the same stress score can be obtained under transformations such as translation, reflection and rotation [13]. The indetermination of the embedding rotation is the motivation for this work.

In practice, MDS is used in psychology and other fields as a means of projecting data into a viewable space, often in two or three dimensions [14]. MDS is also useful for processing data that is stored as (dis)similarity pairs, and it can handle ordinal (dis)similarity values (processed with non-metric MDS) or continuous ones (processed with metric MDS). The widespread use of MDS is supported by its implementation in various social science tools such as SPSS and ANTHROPAC. As the purpose of MDS in practice is to understand data, interpreting the MDS embedding is a crucial step, which is carried out by experts, machine learning techniques or both.

However, the MDS embedding rotation is an issue when the arbitrarily oriented MDS axes must be interpreted. This paper, which is an extended version of [1], analyzes how the rotation of MDS embeddings affects their interpretation and proposes a method for handling this rotational indeterminacy.

This paper is structured as follows. Section 2 reviews how MDS embeddings are interpreted in the literature. Section 3 exposes issues related to embedding orientation when the embedding axes are interpreted using multiple regression models. Section 4 presents several machine learning and statistical methods that can be used to solve such a problem. The *Best Interpretable Rotation* (BIR) selection method that we developed to select the best MDS embedding orientation for interpretation is described in Section 5. Section 6 presents the results of two experiments evaluating the performance of BIR and shows how it compares to the methods listed in Section 4. Discussions about these results are presented in Section 7. Finally, we conclude our paper and provide directions for future work in Section 8.

2. Interpreting an MDS Embedding

Two different and complementary uses of multidimensional scaling (MDS) stand out: *exploratory* and *confirmatory* uses [13, 14]. For the former, the MDS embedding is used as a means of discovering hidden structures in (dis)similarity data [2]. Expert knowledge is therefore needed for analyzing the MDS embedding. For the latter use, the MDS embedding is used to confirm hypotheses the researcher has in mind *a priori* [14]. In this case, *external features* (or *external variables*) are used to discover patterns in the embedding. As the confirmatory process must remain objective, the user lets machine learning techniques find the patterns for him.

For each of these two purposes, there are two main ways to interpret MDS embeddings: *neighborhood interpretation* and *dimensional interpretation* [12]. Clustering (or cluster analysis) is the machine learning problem associated with the first type of interpretation. The goal of clustering is to group instances in a given dataset. The groups found by clustering algorithms are called *clusters*. For instance, Lebel et al. [15] use an agglomerative hierarchical clustering technique for exploring their MDS embedding. They then ask experts to provide an interpretation for each cluster found, as well as each dimension. Therefore, they combine neighborhood interpretation and dimensional interpretation for the purpose of exploration. For hypothesis confirmation, the clustering of instances based on external features is not often used in the literature.

In the context of hypothesis confirmation, the most frequently used technique to link external features with an embedding is linear regression [12, 14]. More precisely, let $\mathbf{X}$ be an $n \times m$ embedding and $\mathbf{F}$ an $n \times d$ matrix of external features. The goal is to estimate the weights (or parameters) $\mathbf{W}$ in

$$\mathbf{F} = \mathbf{X}\mathbf{W} + \mathbf{E}, \qquad (2)$$

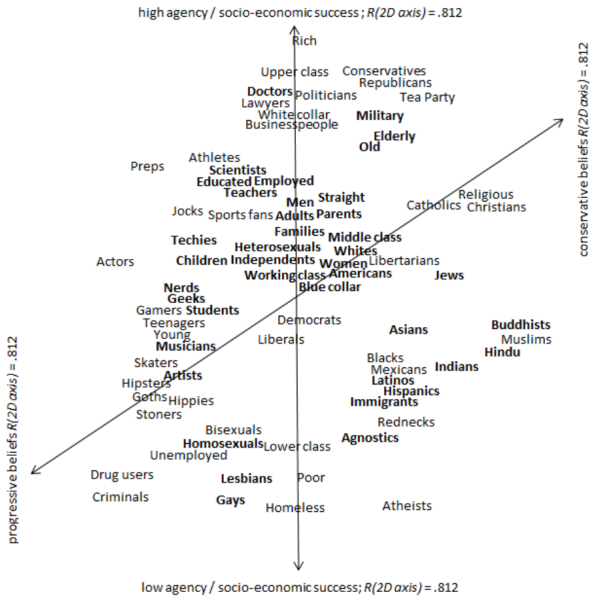


Figure 1: Figure reproduced from Koch et al. [17] presenting two stereotype trends in an MDS embedding of social groups: socio-economic success (vertical line) and beliefs (oblique line).

with $\mathbf{E}$ being an error term [12]. In most cases, $m = 2$ to allow visualization of the embedding $\mathbf{X}$. Indeed, a line representing the trend explained by a given feature $\mathbf{f}_j$ can be drawn in a 2D plot of the embedding $\mathbf{X}$. This line is given by the unit vector $\hat{\mathbf{w}}_j$, whose m elements are normalized versions of w_{jk}, also called direction cosines $\hat{w}_{jk}$, where k is a given dimension of embedding $\mathbf{X}$ [12]:

$$\hat{w}_{jk} = \frac{w_{jk}}{\sqrt{w_{j1}^2 + w_{j2}^2 + ... + w_{jm}^2}}, \tag{3}$$

with m being the total number of dimensions in $\mathbf{X}$.

In the literature, such an approach is often called PROFIT [16]. PROFIT stands for *PROperty FITting*, with the external features understood as properties. Many articles in the literature use this kind of approach for interpreting MDS embeddings (e.g. [17, 18, 19, 20, 21]). Often, the coefficient of determination R^2 is used to select the fitted properties to keep. Figure 1 shows an example of the regression of external features onto an MDS embedding. The two stereotype trends "socio-economic success" and "type of beliefs" are drawn on an MDS embedding containing social groups as instances.

One drawback of such an approach is that each feature is independently regressed onto the MDS embedding, making it impossible to relate the MDS dimensions to combinations of features. This is problematic because each MDS dimension might best be described by a linear combination of features rather than an individual feature. In order to address this issue, *principal component analysis* (PCA) is often run on the external feature matrix $\mathbf{F}$ in order to extract principal components that are then interpreted as meta-features. For instance, Koch et al. [17] in Figure 1 extract their "agency/socio-economic success" stereotype feature from a linear combination of six other stereotypes: powerless-powerful, dominated-dominating, low status-high status, poor-wealthy, unconfident-confident and unassertive-competitive.

As a complement to this, rotation of these components can overcome some limitations of PCA. Indeed, rotation may be useful for either achieving a more understandable distribution of the features in the PCA components (with e.g. a *varimax rotation* [22]) or, if orthogonality of the PCA components is not desired or required, for breaking the orthogonality of the components (with e.g. an *oblimin rotation* [23]).

Nonetheless, the interpretation problem is not fully addressed by these approaches, as the combination of features is not optimized with respect to the information in the MDS embedding. It would be more appropriate to find the best combination of external features for explaining the embedding. The next section presents the problem of reversing the regression direction in order to account for linear combinations of external features, as well as subsequent issues raised by this problem.

3. Problem Statement

In this paper, we are interested in using a *multi-view learning* approach (see Section 1) in order to interpret an MDS mapping model. In particular, a matrix of external features (view 1) is used to interpret the dimensions of an MDS embedding (view 2). In this context, it seems natural to model each MDS dimension as a linear combination of these external features, rather than modeling the features as linear combinations of the MDS dimensions, as was seen in Section 2. The problem of interest is thus to estimate $\mathbf{W}$ in

$$\mathbf{X} = \mathbf{FW} + \mathbf{E}, \tag{4}$$

where $\mathbf{X}$ is an $n \times m$ MDS embedding, $\mathbf{F}$ is an $n \times d$ matrix of external features and $\mathbf{W}$ is a $d \times m$ matrix of regression weights. The following sections focus on a two-dimensional embedding ($m = 2$) in order to simplify the optimization of the proposed method, and we assume that $d > m$.

As seen in Section 1, the MDS solution is only uniquely determined up to some transformations, including orthogonal transformations such as rotation. The orientation of MDS embeddings is thus arbitrary, and as a consequence, the magnitude of the weights in $\mathbf{W}$ is also arbitrary. Let $\mathbf{W}$ be the ordinary least squares (OLS) weights for a model where $\mathbf{X}$ is not rotated, and let $\mathbf{W}^\theta$ be the OLS weights for a model where $\mathbf{X}$ is rotated by any angle $\theta \in [0, 360]$ degrees. We have that

$$\mathbf{W}^\theta = \mathbf{W}\mathbf{R}^\theta, \tag{5}$$

where $\mathbf{R}^\theta$ is an orthogonal rotation matrix defined as

$$\mathbf{R}^\theta = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}. \tag{6}$$

This follows from the fact that the OLS objective function is invariant to rotation. Indeed,

$$\arg\min_{\mathbf{W}} \|\mathbf{X}\mathbf{R}^\theta - \mathbf{F}\mathbf{W}\mathbf{R}^\theta\|_F^2 = \arg\min_{\mathbf{W}} \|\mathbf{X} - \mathbf{F}\mathbf{W}\|_F^2. \tag{7}$$

Let $\mathbf{M} = \mathbf{X} - \mathbf{F}\mathbf{W}$. By expressing the Frobenius norm as a trace, and using the fact that $\mathbf{R}^\theta\mathbf{R}^{\theta^\top} = \mathbf{I}$ and that the trace is invariant under cyclic permutation, we can show that

$$\begin{aligned} &\|\mathbf{X}\mathbf{R}^\theta - \mathbf{F}\mathbf{W}\mathbf{R}^\theta\|_F^2 \\ &= \|\mathbf{M}\mathbf{R}^\theta\|_F^2 \\ &= \text{trace}(\mathbf{R}^{\theta^\top}\mathbf{M}^\top\mathbf{M}\mathbf{R}^\theta) \\ &= \text{trace}(\mathbf{R}^\theta\mathbf{R}^{\theta^\top}\mathbf{M}^\top\mathbf{M}) \quad \textit{cyclic permutation} \\ &= \text{trace}(\mathbf{M}^\top\mathbf{M}) \qquad\qquad \mathbf{R}^\theta\mathbf{R}^{\theta^\top} = \mathbf{I} \\ &= \|\mathbf{M}\|_F^2 \\ &= \|\mathbf{X} - \mathbf{F}\mathbf{W}\|_F^2. \end{aligned} \tag{8}$$

As shown in Figure 2, rotating the matrix $\mathbf{X}$ results in weight magnitudes that are a sinusoidal function of the rotation angle θ. While the model error remains constant for all rotations, some rotations yield models that are easier to interpret than others (i.e. rotation angles yielding more model weights equal to zero). This means that the arbitrary rotation of an embedding generated by MDS may not be the best rotation for interpretation. Thus, modeling the MDS dimensions as a function of the feature matrix introduces a new problem: the determination of a non-arbitrary rotation that facilitates the interpretation of the MDS dimensions.

The analyses in this paper are applied to MDS embeddings, but the rotation problem exists for any $\mathbf{X}$ generated using an NLDR method with a rotation-invariant objective function (e.g. t-SNE [24]).

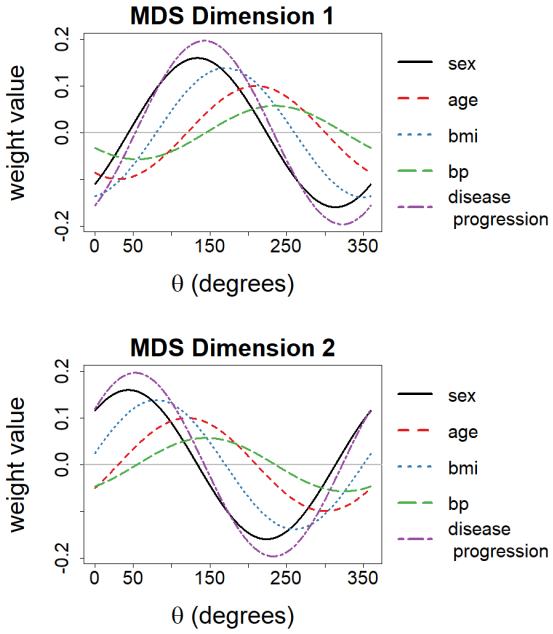


Figure 2: Example of regression weights for OLS models estimated when a 2D MDS embedding is rotated with different angles θ.

4. Existing Methods for View Rotation

The MDS objective function preserves Euclidean distances between all pairs of points, which makes it invariant to orthogonal transformations. As such, any orthogonal transformation of a given MDS embedding is an equally valid solution to the MDS problem. While this paper is primarily concerned with the problem of rotating an MDS embedding, any orthogonal transformation, including rotation and/or reflection, could be applied to an embedding. This section presents several approaches from the statistics and machine learning literature for orthogonally transforming a data "view" – in this case, an embedding generated by MDS. In what follows, $\mathbf{R}$ is an orthogonal transformation matrix of any kind, not exclusively a rotation matrix.

4.1. Principal Component Analysis

The most well known and frequently used single-view rotation method is principal component analysis (PCA). As mentioned in Section 2, PCA can be applied to the matrix of features $\mathbf{F}$ to generate principal components that are then regressed onto an MDS embedding $\mathbf{X}$. However, in this work, we are primarily interested in an orthogonal transformation of $\mathbf{X}$, not $\mathbf{F}$.

In this context, the goal of PCA is to find an orthogonal transformation matrix $\mathbf{R}$ ($m \times m$) that maximizes the

variance in successive columns of $\mathbf{Z} = \mathbf{XR}$. As such, $\mathbf{Z}$ is a rotation of $\mathbf{X}$ such that each successive column of $\mathbf{Z}$ captures a maximum of the variance in $\mathbf{X}$ not already represented in the previous columns.

4.2. Orthogonal Procrustean Transformation

Procrustean transformation [13] is one of the most frequently used MDS embedding transformations. This transformation aims to align an MDS embedding with another matrix. Most of the time, it is used to align two 2D or 3D embeddings in order to visually compare them and remove indeterminacies linked to their orientation or dilation. However, the problem can be generalized to the case where the two matrices do not have the same dimensionality, e.g. by adding columns of zeros [13, 25].

Let $\mathbf{X}$' be the concatenation of $\mathbf{X}$ and a matrix with $d - m$ columns of zeros, such that $\mathbf{X}$' $(n \times d)$ and $\mathbf{F}$ $(n \times d)$ have the same dimensionality. In the orthogonal Procrustes problem, $\mathbf{X}$' is transformed with a matrix $\mathbf{R}$ in order to minimize the squared distance between $\mathbf{X}$'$\mathbf{R}$ and the $n \times d$ target matrix $\mathbf{F}$ [13]:

$$\begin{aligned} \underset{s,\mathbf{R}}{\arg\min}\ & \operatorname{tr}\left[(\mathbf{F} - s\mathbf{X}'\mathbf{R})^\top(\mathbf{F} - s\mathbf{X}'\mathbf{R})\right] \\ & \text{s.t. } \mathbf{R}^\top\mathbf{R} = \mathbf{I}, \end{aligned} \tag{9}$$

where $\mathbf{R}$ is the $d \times d$ Procrustean transformation matrix and s is a scaling factor. The trace calculated is the sum of the squared distances between each point i in $\mathbf{F}$ and the corresponding point i in $\mathbf{X}'\mathbf{R}$, which are found in the diagonal of $(\mathbf{F} - s\mathbf{X}'\mathbf{R})^\top(\mathbf{F} - s\mathbf{X}'\mathbf{R})$ [13].

4.3. Eigenvector Partial Least Squares (PLS)

Eigenvector partial least squares (PLS) [11], also known as Bookstein PLS [26], is a two-view matrix factorization method. The goal of eigenvector PLS is to find orthogonal transformation matrices $\mathbf{P}$ and $\mathbf{R}$ such that the covariance between $\mathbf{T} = \mathbf{FP}$ and $\mathbf{Z} = \mathbf{XR}$ is maximal. Both $\mathbf{T}$ and $\mathbf{Z}$ are of dimension $n \times p$, where $p = \min(m, d)$, d is the number of features in $\mathbf{F}$ and m is the number of columns in $\mathbf{X}$.

4.4. Eigenvector PLS Regression (PLS-R)

Eigenvector PLS Regression (PLS-R) is an extension of eigenvector PLS to regression. Orthogonal transformation matrices $\mathbf{R}$ and $\mathbf{P}$ are first found using eigenvector PLS. Then, the matrix $\mathbf{Z} = \mathbf{XR}$ is regressed onto $\mathbf{T} = \mathbf{FP}$ using ordinary least squares (OLS). The model is defined as

$$\mathbf{Z} = \mathbf{TB} + \mathbf{E}, \tag{10}$$

where $\mathbf{E}$ is an error term and $\mathbf{B}$ is a matrix of regression weights, calculated as follows:

$$\mathbf{B} = (\mathbf{T}^\top\mathbf{T})^{-1}\mathbf{T}^\top\mathbf{Z}. \tag{11}$$

The orthogonally transformed view $\mathbf{XR}$ can thus be interpreted as a linear combination of the features in $\mathbf{F}$:

$$\mathbf{XR} = \mathbf{TB} + \mathbf{E} = \mathbf{FW} + \mathbf{E}, \tag{12}$$

where $\mathbf{W} = \mathbf{PB}$ is a matrix of regression weights describing the linear relationship between each feature in $\mathbf{F}$ and each dimension of $\mathbf{XR}$.

4.5. Sparse Reduced Rank Regression (SRRR)

Unlike eigenvector PLS-R, Sparse Reduced Rank Regression (SRRR) [27] introduces a constraint to encourage $\mathbf{W}$ to be sparse. Both $\mathbf{R}$ $(m \times p)$ and $\mathbf{W}$ $(d \times p)$ are constrained to have rank $p \leq \min(m, d)$, p being a hyperparameter that must be selected. $\mathbf{R}$ and $\mathbf{W}$ are found by optimizing the objective function

$$\begin{aligned} \underset{\mathbf{R},\mathbf{W}}{\arg\min}\ & \|\mathbf{XR} - \mathbf{FW}\|_F^2 + \gamma \sum_{j=1}^{d} \|\mathbf{w}_j\|_2 \\ & \text{s.t. } \mathbf{R}^\top\mathbf{R} = \mathbf{I}, \end{aligned} \tag{13}$$

where $\mathbf{w}_j$ is the j^{th} row of $\mathbf{W}$ and $\gamma > 0$. The second term in Equation (13) is a type of *group regularization*, as groups of weights are penalized together. Note that the L_2 norm $\|\mathbf{w}_j\|_2$ is not squared, and as a result, it forces the elements of $\mathbf{w}_j$ to be either all zero or non-zero (see [28] for more details). As γ increases, more and more rows of $\mathbf{W}$ are set to zero, meaning that the associated features are no longer active in the model. Equation (13) is optimized by alternating between the optimization of $\mathbf{R}$ for fixed $\mathbf{W}$ and $\mathbf{W}$ for fixed $\mathbf{R}$.

4.6. Summary and Shortcomings

The most frequently used orthogonal transformation, PCA, maximizes explained variance by considering only the matrix to which the transformation is applied, making it a single-view method. For our problem setting, the multi-view methods presented above are more appropriate than PCA because the transformation of $\mathbf{X}$ with respect to $\mathbf{F}$ is directly optimized: it is learned using the external feature matrix $\mathbf{F}$ that will later be used to build the model linking the two views.

While orthogonal Procrustean transformation considers both matrices $\mathbf{X}$ and $\mathbf{F}$ for transforming the former, it requires the two matrices to have the same dimensionality. If this is not the case, the number of dimensions

in the smaller matrix must be artificially increased before the transformation is applied. In our case, $m < d$, meaning that both the augmented matrix **X**' and its transformed version **X**'**R** have d columns. Because of this, it is difficult to compare Procrustean transformation to the other methods in this section, which find a m-dimensional orthogonal matrix **R**.

Eigenvector PLS aligns two matrices **X** and **F** using orthogonal transformations such that the dimensionality of **X** ($n \times m$) is preserved. However, for eigenvector PLS-R, the weights linking the two matrices are not sparse, making the interpretation of **XR** difficult in most cases.

SRRR yields a more easily interpretable model than the other multi-view methods because it encourages sparsity in the matrix of regression weights. However, when features are included in the model, they have non-zero-valued weights for each dimension of **XR** due to the group penalty. This is problematic for the interpretation of the MDS axes in **X**, because this group penalization implies that all of the axes are explained by the same features.

Thus, while a few existing multi-view methods are able to find orthogonal transformations adapted for subsequent regression problems (eigenvector PLS-R and SRRR), the sparsity of the models generated is insufficient, in the case of eigenvector PLS-R, and the distribution of non-zero-valued weights is ill adapted to the problem at hand, in the case of SRRR.

5. Proposed Method: BIR Selection

Among all possible MDS embedding rotations, the rotation that interests us is the one making it possible to understand the embedding. In order to do so, some methods, such as SRRR, presented in Section 4, regularize the regression model used to understand the embedding. However, as observed in Section 3, regression weights change depending on the chosen rotation, which implies different possible interpretations of these weights. For a better understanding of how rotation affects regularized regression weights, Section 5.1 analyzes weight changes for Ridge regularization. Note that this type of penalization may not be adapted to our problem because it shrinks all weight values towards zero, yielding small but non-zero weight values. Section 5.2 analyzes weight changes for sparse regression performed using Lasso regularization. After having analyzed various rotation effects, Section 5.3 presents the Best Interpretable Rotation (BIR) selection method, and Section 5.4 presents an extension of BIR, BIR Lasso regression (BIR-LR), which learns a sparse regression model based on the rotation chosen by BIR.

5.1. Effect of Rotation on Ridge Regularization

Ridge regression adds a term to the OLS objective function that penalizes weight values through a squared Euclidean norm (also called the L_2 norm):

$$\arg\min_{\mathbf{W}} \|\mathbf{XR} - \mathbf{FW}\|_F^2 + \lambda \sum_{j=1}^{d} \|\mathbf{w}_j\|_2^2, \tag{14}$$

where the hyperparameter λ controls the balance between error and regularization. The squared L_2 norm shrinks weight values towards zero.

As this work is concerned with the rotation of **X** and its effect on a subsequent regression model, Figure 3a shows how Ridge regression weights depend on rotation angle. As with OLS, the Ridge objective function is rotation invariant, so the regression weights are a sinusoidal function of the rotation angle θ. Note that $\sum_{j=1}^{d} \|\mathbf{w}_j\|_2^2$ can be rewritten using a squared Frobenius norm, $\|\mathbf{W}\|_F^2$, which is rotation invariant. Using the same logic as in Equation (8), we can show that

$$\begin{aligned} &\arg\min_{\mathbf{W}} \|\mathbf{XR}^\theta - \mathbf{FWR}^\theta\|_F^2 + \lambda\|\mathbf{WR}^\theta\|_F^2 \\ &= \arg\min_{\mathbf{W}} \|\mathbf{X} - \mathbf{FW}\|_F^2 + \lambda\|\mathbf{W}\|_F^2. \end{aligned} \tag{15}$$

As with OLS, $\mathbf{W}^\theta$, the weights for a given rotation θ, are equal to $\mathbf{WR}^\theta$, and the error is constant for all rotations.

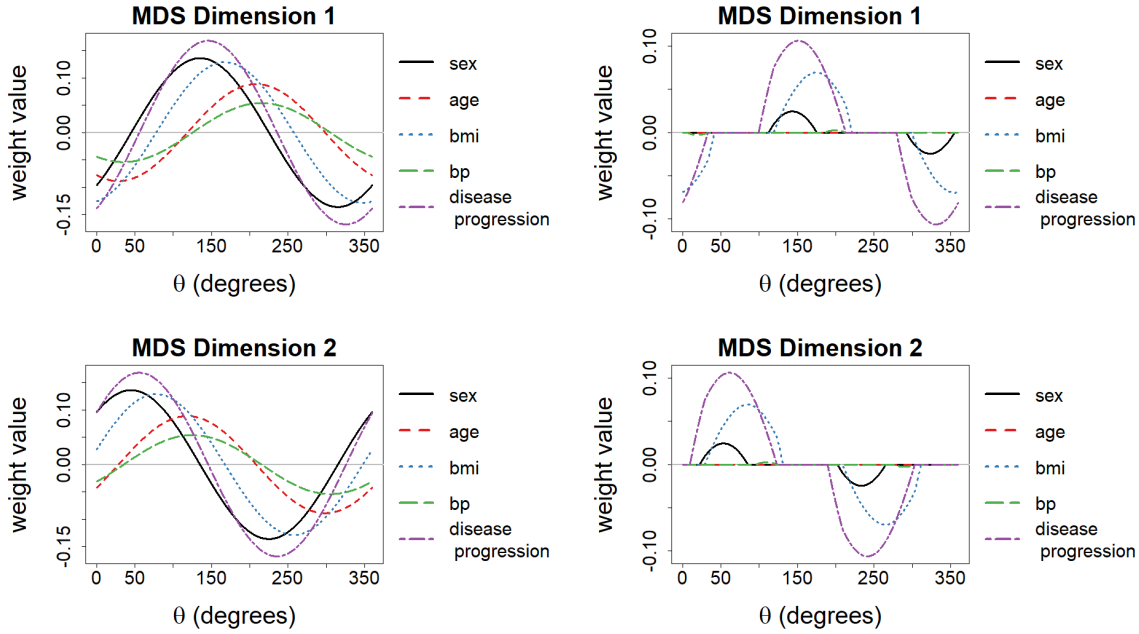
5.2. Effect of Rotation on Lasso Regularization

Another famous penalty is the Lasso penalty. The Lasso penalty regularizes regression weights using an L_1 norm:

$$\arg\min_{\mathbf{W}} \|\mathbf{XR} - \mathbf{FW}\|_F^2 + \lambda \sum_{j=1}^{d} \|\mathbf{w}_j\|_1. \tag{16}$$

The Lasso induces a thresholding effect based on λ, setting all weights under this λ-dependent threshold to zero. This effect can be observed in Figure 3b, where, for certain rotation angles, several features have zero-valued weights. As the Lasso simultaneously sets many weights to zero, the regression model is generally less complex, and thus more interpretable, than OLS and Ridge models.

However, in contrast to OLS and Ridge, the Lasso objective function is not invariant to rotation. Indeed, as shown in Figure 4, the error and the number of weights set to zero change as the MDS embedding is rotated. This means that failing to rotate the MDS embedding before applying the Lasso may yield a model that is suboptimal in terms of model error and sparsity. If one wants to use the Lasso to interpret an MDS embedding, the embedding orientation should be carefully selected.



(a) Example of regression weights for Ridge models estimated when a 2D MDS embedding is rotated with different angles θ ($\lambda = 0.15$).

(b) Example of regression weights for Lasso models estimated when a 2D MDS embedding is rotated with different angles θ ($\lambda = 0.15$).

Figure 3: Effect of rotation on weight values for Ridge and the Lasso.

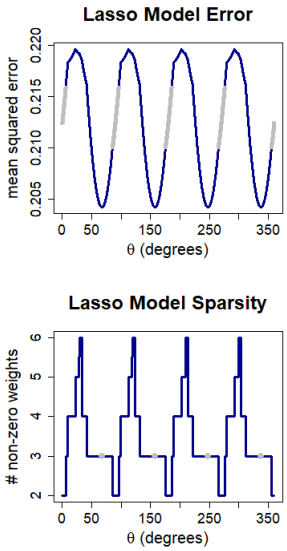


Figure 4: Error and sparsity of Lasso models estimated when a 2D MDS embedding is rotated with different angles θ ($\lambda = 0.15$). Line segments highlighted in gray indicate θ values minimizing sparsity (top plot) or minimizing model error (bottom plot). The different minima for model sparsity and error do not overlap. Model weights are represented in Figure 3b.

.

5.3. *Selecting the Best Rotation for Interpretation*

Among all possible MDS embedding orientations, we are interested in selecting the one yielding a Lasso regression model with the best balance between error and interpretability. In what follows, we measure model error using the mean squared error (MSE), and we quantify interpretability by counting the number of non-zero-valued weights (or active features) in the model (L_0 norm). This leads to the Best Interpretable Rotation (BIR) selection criterion, which selects the best angle θ^* as

$$\begin{aligned}\theta^* &= \arg\min_{\theta} \frac{1}{2n}\|\mathbf{X}\mathbf{R}^{\theta} - \mathbf{F}\mathbf{W}^{\theta}\|_F^2 + \lambda \sum_{k=1}^{2} \|\mathbf{w}_k^{\theta}\|_0 \\ &= \arg\min_{\theta} \sum_{k=1}^{2} \left(\frac{1}{2n}\|\mathbf{X}\mathbf{r}_k^{\theta} - \mathbf{F}\mathbf{w}_k^{\theta}\|_2^2 + \lambda \|\mathbf{w}_k^{\theta}\|_0 \right),\end{aligned} \tag{17}$$

where $\mathbf{R}^{\theta}$ is a rotation matrix dependent on θ, $\mathbf{w}_k^{\theta}$ is the weight vector obtained when the Lasso is applied to the k^{th} column of an embedding rotated by an angle θ and λ strikes a balance between the MSE and the L_0 norm. The solution θ^* is then used to calculate a rotation matrix $\mathbf{R}$ based on Equation (6).

5.4. Lasso Regression based on a BIR-Selected Angle

Similar to some of the methods summarized in Section 4, the BIR selection method finds an orthogonal transformation matrix $\mathbf{R}$ for a view $\mathbf{X}$, given another view $\mathbf{F}$. However, in contrast to methods like eigenvector PLS-R and SRRR, the model used for interpreting $\mathbf{XR}$ based on the external feature view $\mathbf{F}$ is not learned.

The purpose of BIR Lasso regression (BIR-LR) is to learn a sparse linear model linking these two views by applying Lasso regression to a target matrix $\mathbf{X}$ rotated by the angle θ^* found using BIR. Note that the optimization of $\mathbf{R}^\theta$ using BIR involves the L_0 norm in Equation (17), while the Lasso involves the L_1 norm when optimizing $\mathbf{W}$ (see Equation (16)).

BIR-LR is similar to SRRR (see Section 4) in that the optimization of the regularized weight matrix $\mathbf{W}$ depends on the transformation matrix $\mathbf{R}$. For SRRR, $\mathbf{W}$ is regularized using an L_2 penalty,

$$\gamma \sum_{j=1}^{d} \|\mathbf{w}_j\|_2, \tag{18}$$

whereas the $\mathbf{W}$ optimized in BIR-LR is regularized using an L_1 (Lasso) penalty,

$$\lambda \sum_{j=1}^{d} \|\mathbf{w}_j\|_1. \tag{19}$$

The disadvantage of using the L_2 penalty is that each given feature has non-zero-valued weights for all dimensions of the MDS or none of them. Using the L_1 penalty makes it possible to learn models where a given feature has a non-zero-valued weight for one dimension and a zero-valued weight for another, which greatly simplifies interpretation.

6. Evaluation of the BIR Selection Method

This section presents our evaluation procedure and results. The problem at hand involves two tasks: (i) finding an optimal orthogonal transformation matrix $\mathbf{R}$ for interpreting an MDS embedding and (ii) learning an interpretable model $\mathbf{W}$ that accurately relates external features to the orthogonally transformed embedding. Two experiments are run to compare the performance of different methods with respect to these two tasks.

The purpose of the first experiment is to evaluate whether the orthogonal transformation $\mathbf{R}$ found using the BIR selection method yields a better Lasso solution than the orthogonal transformations found using other methods (task 1). PCA, eigenvector PLS and SRRR are used to generate the competitor transformation matrices (see Section 4). The purpose of the second experiment is to compare BIR-LR with two existing methods that combine view transformation with regression: SRRR and eigenvector PLS-R (tasks 1 and 2).

We compare the performance of each method with respect to two baselines: (i) the performance of the *least sparse rotation*, calculated as the average performance of the Lasso for the set of rotation angles yielding the least sparse solution and (ii) the expected performance of a *random rotation*, calculated as the average performance of the Lasso for all rotation angles $\theta \in \Theta = \{0.1, 0.2, ..., 360\}$ degrees.

6.1. Datasets and Pre-Processing

The performance of BIR and the other methods is evaluated using seven popular, publicly available datasets that can easily be split into two meaningful, distinct views: Hepatitis, Dermatology, Heart (Statlog) from [29], Insurance Company Benchmark from [30, 29], Community and Crimes from [31, 32, 33, 29], Pima Indians Diabetes from [34] and Diabetes from [35]. As an example, the features in Diabetes are divided into a view containing blood serum measurements – such as glucose and cholesterol levels – and another view composed of simple patient traits – such as age, sex and disease progression (see Table 1 for all split details). For each dataset, instances with missing values are removed, and non-ordinal categorical features are binarized using one-hot encoding. The total number of instances in each dataset, as well as the number of features in each view, is summarized in Table 2.

For each dataset, a view containing interpretable features is used as the external feature set $\mathbf{F}$. A dissimilarity matrix $\mathbf{D}$ of pairwise Euclidean distances between instances is constructed based on the other view $\mathbf{Q}$, which has been normalized. A 2D metric MDS embedding $\mathbf{X}$ is calculated using $\mathbf{D}$. All MDS embeddings $\mathbf{X}$ are centered and all external feature matrices $\mathbf{F}$ are normalized.

6.2. Evaluation Procedure

For both experiments, 10-fold cross-validation is used to assess the average performance of the different methods for each of the seven datasets. The MDS embeddings $\mathbf{X}$ are trained using all instances in $\mathbf{Q}$, then the instances in $\mathbf{X}$ and $\mathbf{F}$ are split into 10 folds. For each instance in $\mathbf{X}$ assigned to a given fold, the corresponding instance in $\mathbf{F}$ is assigned to the same fold.

Dataset	Features in Q	External Features in F
Hepatitis	**Histopathological features:** bilirubin, alk.phosphate, sgot, Albumin, protime	**Patient clinical information:** hist, age, sex, steroid, antivirals, fatigue, malaise, anorexia, big.liver, firm.liver, spleen.palp, spiders, ascites, varices, class
Dermatology	**Features measured through microscope analysis:** melanin, eosinophils, PNL, fibrosis, exocytosis, acanthosis, hyperkeratosis, parakeratosis, clubbing, elongation, thinning, spongiform, munro.microabcess, hypergranulosis, dis.granular, vacuolisation, spongiosis, saw.tooth, follic.horn.plug, perifolli.parakeratosis, inflam.monoluclear, band.like	**Patient clinical information:** erythema, scaling, def.borders, itching, koebner, polyg.papules, follic.papules, oral.musocal, knee.elbow, scalp, family.hist, age, disease
Heart	**Features measured at a consultation:** rest.BP, cholest, fast.sugar, rest.ECG, max.HR, ex.angina, ST.depress, ST.slope, blood.vessels, thal	**Patient clinical information:** age, sex, pain.type, disease
Diabetes	**Blood serum measurements:** s1, s2, s3, s4, s5, s6 (hdl, ldl, glucose, etc.)	**Patient clinical information:** age, sex, body mass index, blood pressure, disease.prog
Pima	**Features measured at a consultation:** glucose, pressure, triceps, insulin	**Patient clinical information:** pregnant, mass, pedigree, age, diabetes
Crimes	**Criminality features:** e.g. murders, robberies, autoTheft, arsons, etc.	**Socio-demographic features:** e.g. householdsize, racePctWhite, medIncome, RentMedian, etc.
Insurance	**Insurance product usage features:** e.g. PPERSAUT (contribution car policies), ALEVEN (number of life insurances), etc.	**Socio-demographic features:** e.g. MHKOOP (home owners), MRELGE (married), MINKGEM (average income), etc.

Table 1: Division of dataset features into two views: **Q**, which contains the features used for computing the MDS, and **F**, the set of external features used to interpret the MDS. For datasets with more than 50 features (Crimes and Insurance), only a few feature examples are provided.

Dataset	Instances	Features		
		Total	Q	F
Hepatitis	80	20	5	15
Dermatology	358	35	22	13
Heart	270	14	10	4
Diabetes	442	11	6	5
Pima	768	9	4	5
Crimes	302	142	18	124
Insurance	5822	134	43	91

Table 2: Dimensions of evaluation datasets.

6.3. Experiment 1: Orthogonal Transformations

In this experiment, the quality of different orthogonal transformations is studied by evaluating the sparsity and test error of Lasso models where **F** is the feature matrix and transformed embedding **XR** is the target. As the Lasso is used for all evaluated methods, only the quality of the embedding transformation (with respect to the learned Lasso model) is measured.

BIR, PCA, eigenvector PLS and SRRR, as well as the baseline rotations (least sparse and random rotations), are applied to each training fold to produce orthogonal transformation matrices **R**. Then, Lasso models with varying values of λ are trained on the same folds. For SRRR, 25 γ values in the interval $[1, 3000]$, equally spaced in logarithmic scale, were tested. In the evaluated datasets, two distinct trends were observed among the SRRR transformation matrices trained with these γ values. In what follows, the γ values 1 and 208 have been selected because they yield models representative of the two observed trends for all of the datasets. Thirty equally spaced λ values in the interval $[0.01, 0.45]$ are used for the Lasso models. This range was chosen in order to cover a large range of sparsity degrees (as calculated using Equation (20)).

Several evaluation criteria are calculated based on the Lasso model weights **W**: the degree of sparsity

$$s = \sum_{k=1}^{2} \|\mathbf{w}_k\|_0, \tag{20}$$

and the mean squared error (MSE) of prediction on the

test fold, calculated as

$$\text{MSE} = \frac{1}{2n}\sum_{k=1}^{2}\|\mathbf{X}\mathbf{r}_k - \mathbf{F}\mathbf{w}_k\|_2^2, \tag{21}$$

where $\mathbf{X}$ and $\mathbf{F}$ contain only instances in the test fold.

Figure 5 shows the results of the first experiment. For each method, the average MSE over 10 folds is plotted against the average number of non-zero-valued weights (over the same folds). For all datasets except Hepatitis, and for a given number of non-zero-valued weights, BIR angles result in model error that is less than or equal to all other methods. In contrast to BIR, the least sparse rotation always has the worst test error for these datasets, probably because of overfitting during training. Hepatitis is the only exception, where, for non-sparse models, the average MSE of the least sparse case is the smallest and the average MSE of BIR is the largest. However, we argue that the most interesting models for ease of interpretation are the ones with few non-zero-valued weights, in which case BIR yields smaller model error than the other methods. Overall, BIR outperforms all other methods and baseline rotations for sparse and interpretable regression models (left part of the plots).

Transforming the MDS embedding based on information in only one view, the MDS embedding itself, seems less optimal for subsequent regression. For Hepatitis and Dermatology, the MDS orientation selected by PCA is always worse than a random rotation on average, and for the other datasets, the results are sometimes better and sometimes worse than a random rotation (but always worse than BIR).

The same conclusion can be drawn for eigenvector PLS and SRRR. Indeed, despite using both matrices $\mathbf{X}$ and $\mathbf{F}$ to find an orthogonal transformation of $\mathbf{X}$, the performance of the regression of $\mathbf{XR}$ onto $\mathbf{F}$ fluctuates. As with PCA, the results for eigenvector PLS and SRRR are sometimes better than a random rotation and sometimes worse, while always being worse than BIR, except for the non-sparse solutions for Hepatitis. Note that eigenvector PLS is the main competitor of BIR for some datasets (e.g. Hepatitis, Dermatology and Insurance), while SRRR is its principal competitor for other datasets (e.g. Pima and Crimes). This suggests that the transformation quality of eigenvector PLS and SRRR depends heavily on the dataset, while BIR consistently provides good transformations for all datasets tested.

6.4. Experiment 2: Multi-View Regression Models

In this experiment, BIR-LR weights (Lasso weights computed on rotations selected by BIR) are compared to the weights estimated using methods that simultaneously transform the target and estimate weights with a regression method other than the Lasso. These weights are also compared to Lasso weights computed on the least sparse and random baseline rotations. The same set of λ values from the first experiment is used for BIR-LR and the baseline rotations. A sequence of 25 values in the interval [1, 3000], equally spaced in logarithmic scale, is used for the hyperparameter γ in SRRR.

Like in the first experiment, for each method, both the orthogonal transformation matrix $\mathbf{R}$ and the matrix of regression weights $\mathbf{W}$ are estimated using the training folds. However, the weights $\mathbf{W}$ for the methods from the literature are estimated using the specific regression approach of these methods, rather than by applying the Lasso. The degree of sparsity s is calculated for each $\mathbf{W}$, and the MSE of prediction is calculated for instances in the test fold (see Equations (20) and (21)).

Figure 6 shows the results of the second experiment. Note that eigenvector PLS-R has only one data point because it has no extra hyperparameters. Eigenvector PLS-R, which does not explicitly encourage model sparsity, appears to the far right in all plots. Despite its low average MSE, the obtained weights, which are all non-zero-valued, do not meet our need for interpretable solutions. For all datasets except Hepatitis and Insurance, SRRR has a greater average MSE than a random rotation or BIR-LR for all γ values tested. For the Hepatitis dataset, SRRR is only better than a random rotation for complex models with at least 18 active features. For the Insurance dataset, SRRR is comparable to a random rotation. For all datasets, BIR-LR has a lower average MSE than the other methods and baseline rotations for models with fewer than 10 non-zero-valued weights. Note that the weights for BIR-LR, as well as the least sparse and random rotations, were obtained using the Lasso, so they are the same as in the first experiment.

6.5. Analysis of Model Interpretability

In this section, we compare models from experiment 2 in order to assess the interpretability of BIR-LR. To simplify visualization of the models, we focus on the three datasets with the smallest number of external features: Diabetes, Heart and Pima. For BIR-LR, the hyperparameter λ selected for each dataset is the point (average MSE, average degree of sparsity) in the elbow of the corresponding plot in Figure 6 (i.e. the point closest to the origin). For the other methods, we select the hyperparameters yielding an average MSE closest to the chosen BIR-LR average MSE. All models are trained on all instances in $\mathbf{X}$ and $\mathbf{F}$.

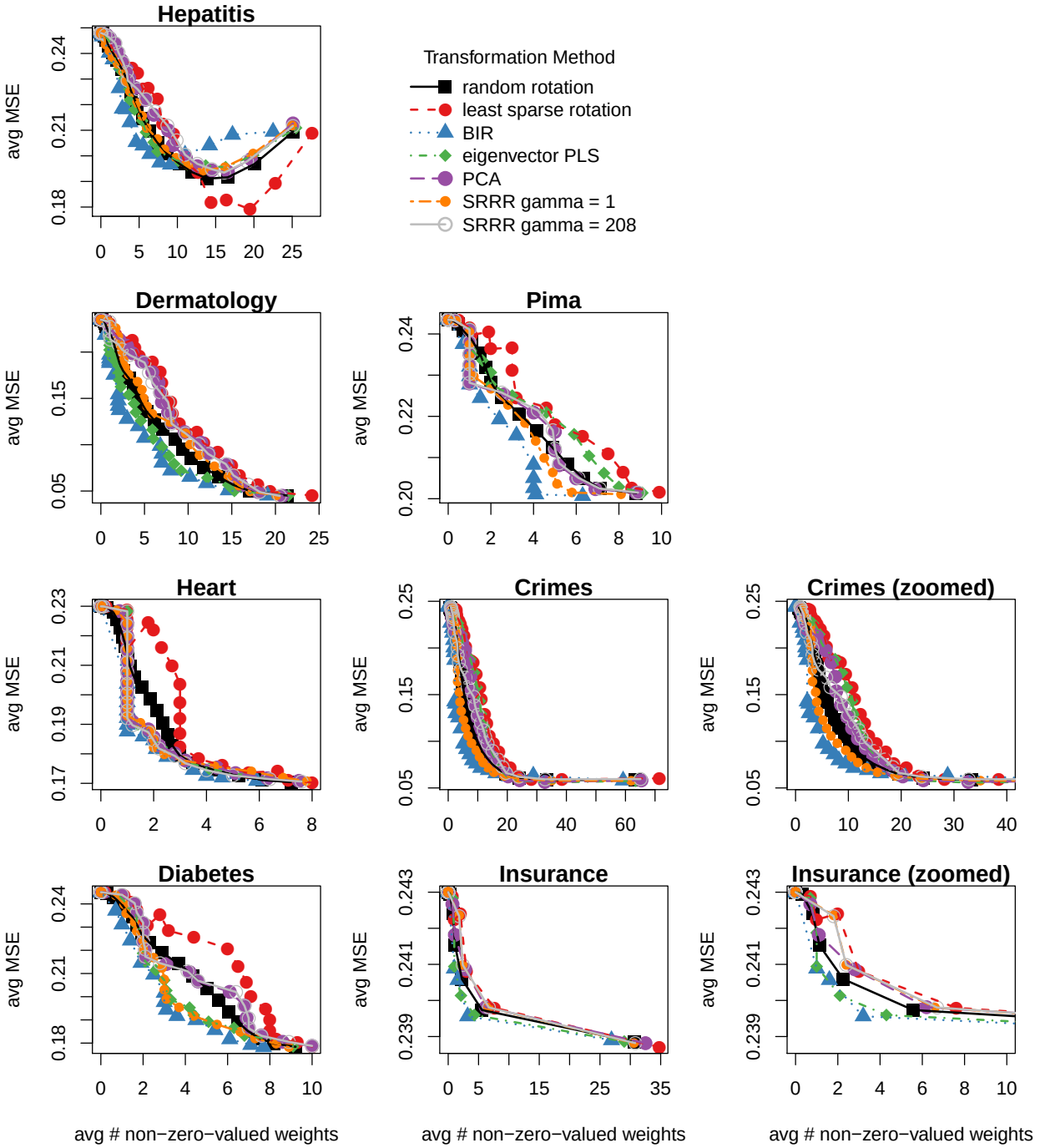


Figure 5: **Experiment 1**. Mean squared error (MSE) and degree of sparsity for Lasso models learned based on different embedding transformations. Each point represents an average value over 10 folds for a given λ, where λ is the hyperparameter used when training the Lasso models. Two SRRR curves are shown here, each representing different γ values. See the text for more details on the selection of γ. Crimes (zoomed) (resp. Insurance (zoomed)) is a zoomed version of the Crimes plot (resp. Insurance plot), showing the average number of non-zero-valued weights in the interval [0, 40] (resp. [0, 10]).

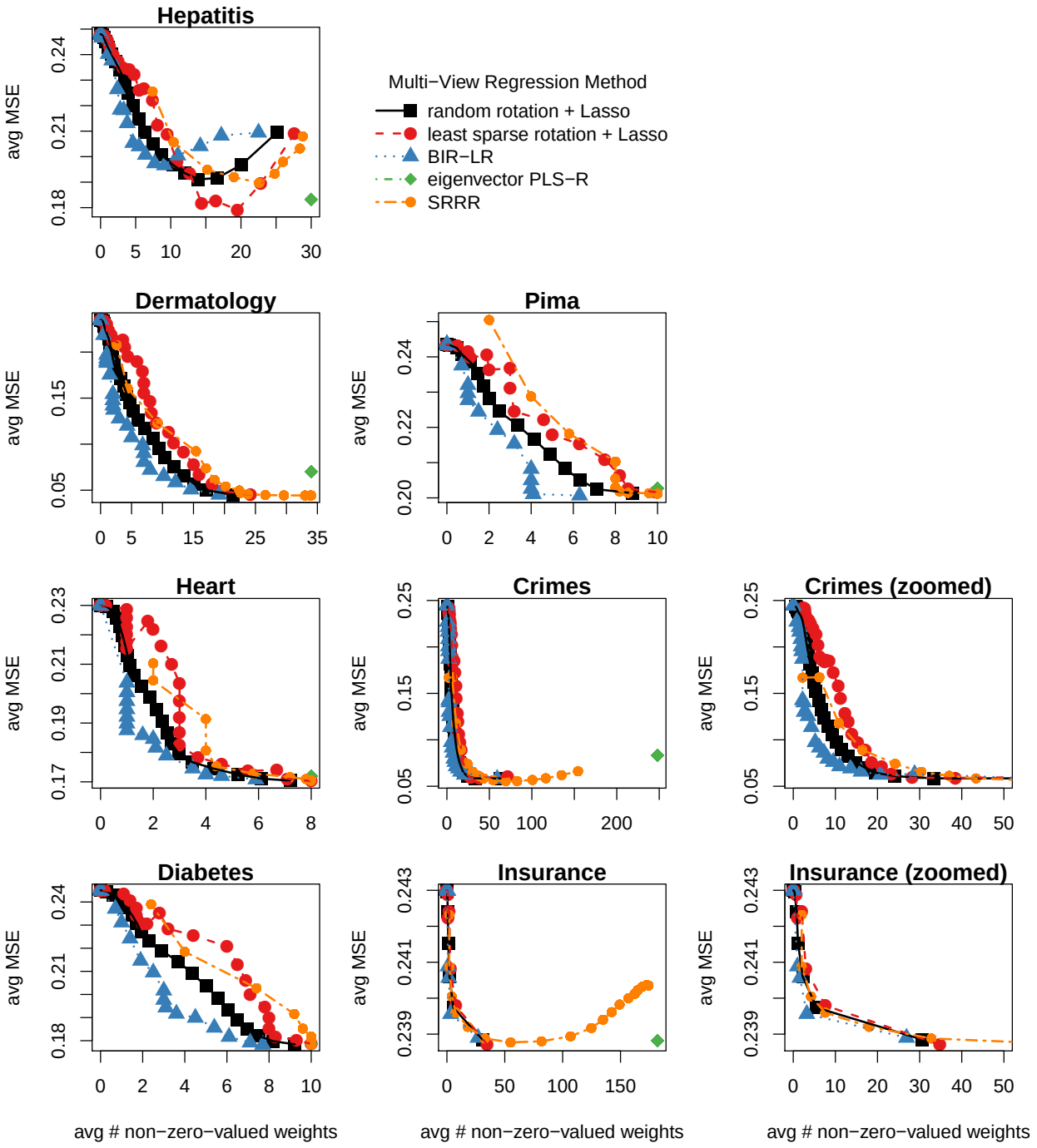


Figure 6: **Experiment 2**. Mean squared error (MSE) for different degrees of sparsity. Each point represents an average value over 10 folds for a particular hyperparameter setting, e.g. a value λ for the Lasso model. Note that eigenvector PLS-R does not have any hyperparameters. Crimes (zoomed) and Insurance (zoomed) are zoomed versions of the Crimes and Insurance plots, showing the average number of non-zero-valued weights in the interval [0, 50].

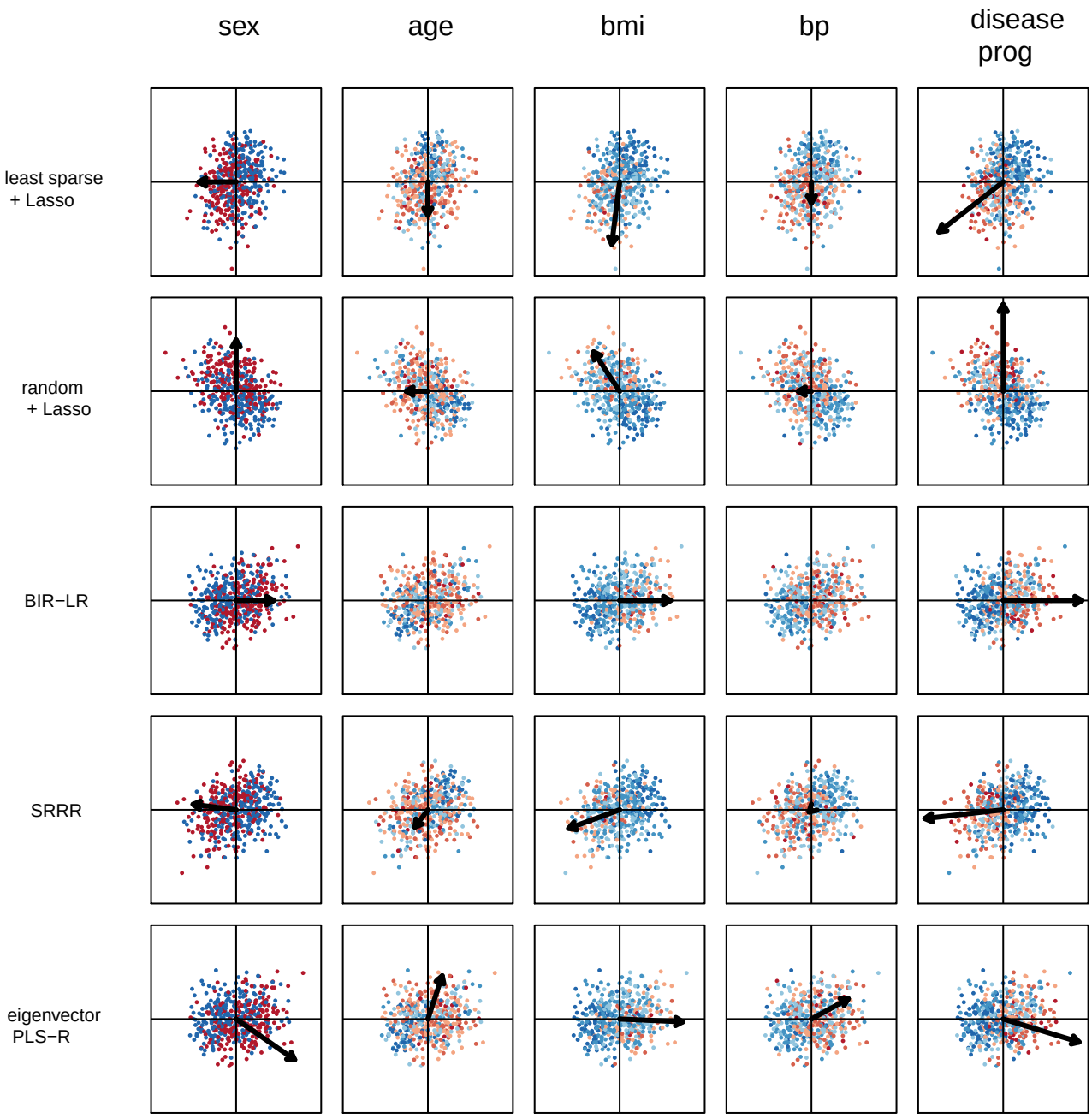


Figure 7: **Experiment 2 (Diabetes)**: Each row represents a specific multi-view regression method and each column corresponds to a feature in the Diabetes dataset. Each scatterplot depicts an MDS embedding transformed by the method in the corresponding row. Each instance in the scatterplots is colored according to its value for the feature in question, using a scale from blue (minimum) to red (maximum). Finally, each arrow direction represents the regression weights $\mathbf{w}_j$ for the corresponding column feature. The arrow length is proportional to the L_2-norm of $\mathbf{w}_j$. Note that the "least sparse" (resp. "random") row presents a single example of a rotation yielding the least (resp. average) model sparsity. Eigenvector PLS-R does not have any hyperparameters, so the error level for this method does not necessarily match the others. It is included here to be consistent with previous figures.

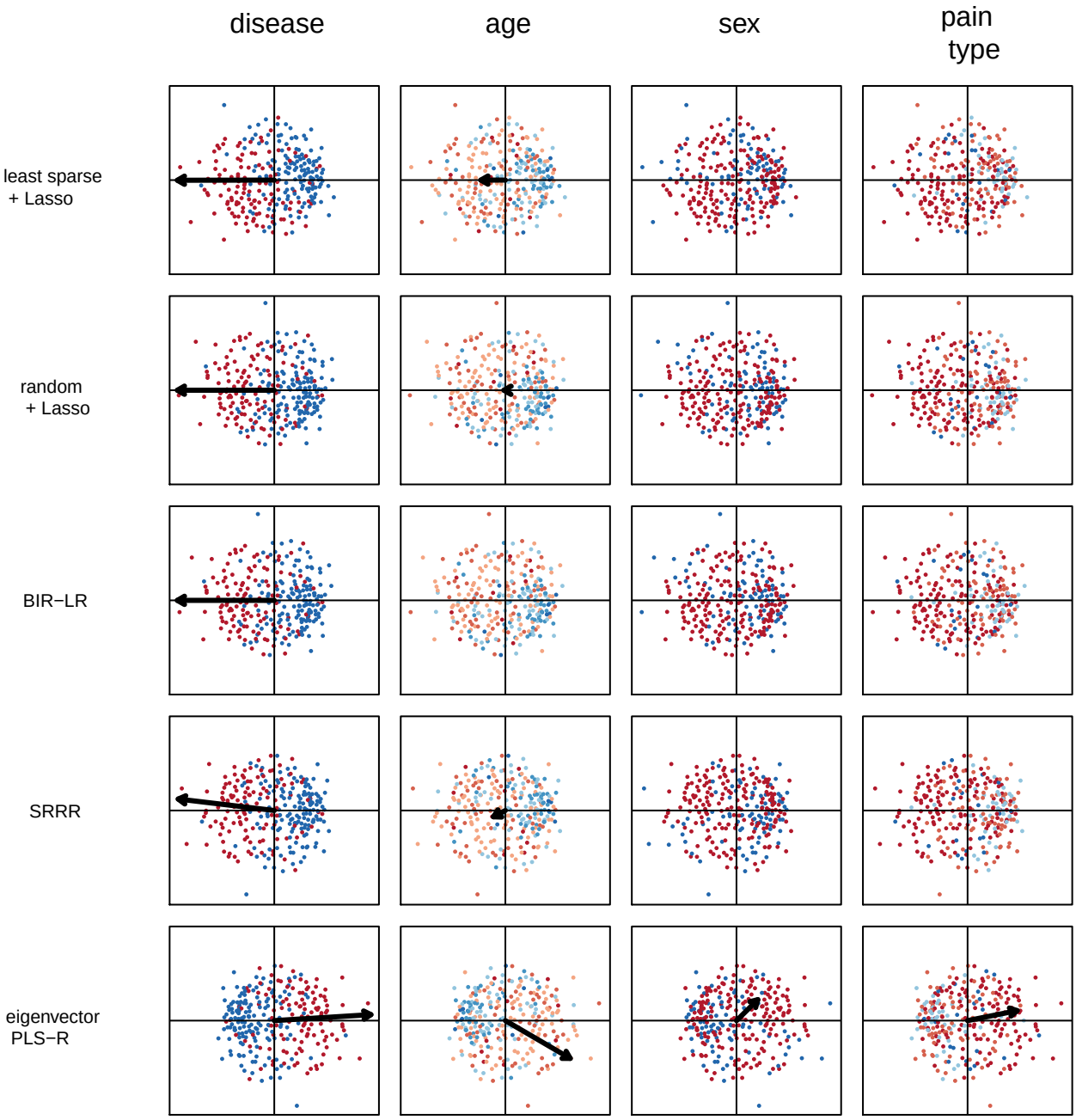


Figure 8: **Experiment 2 (Heart)**: Each row represents a specific multi-view regression method and each column corresponds to a feature in the Heart dataset. Each scatterplot depicts an MDS embedding transformed by the method in the corresponding row. Each instance in the scatterplots is colored according to its value for the feature in question, using a scale from blue (minimum) to red (maximum). Finally, each arrow direction represents the regression weights $\mathbf{w}_j$ for the corresponding column feature. The arrow length is proportional to the L_2-norm of $\mathbf{w}_j$. Note that the "least sparse" (resp. "random") row presents a single example of a rotation yielding the least (resp. average) model sparsity. Eigenvector PLS-R does not have any hyperparameters, so the error level for this method does not necessarily match the others. It is included here to be consistent with previous figures.

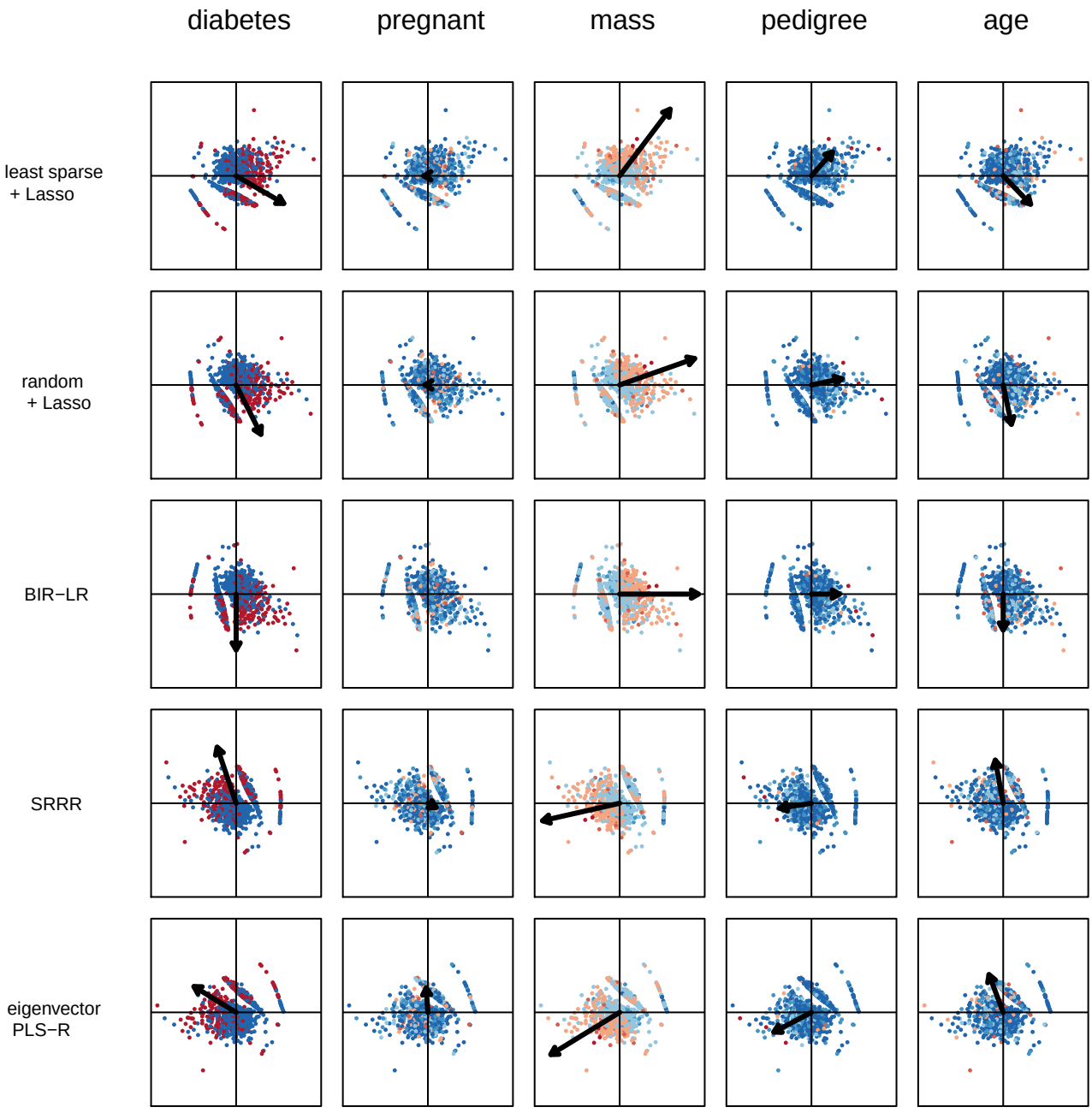


Figure 9: **Experiment 2 (Pima)**: Each row represents a specific multi-view regression method and each column corresponds to a feature in the Pima dataset. Each scatterplot depicts an MDS embedding transformed by the method in the corresponding row. Each instance in the scatterplots is colored according to its value for the feature in question, using a scale from blue (minimum) to red (maximum). Finally, each arrow direction represents the regression weights $\mathbf{w}_j$ for the corresponding column feature. The arrow length is proportional to the L_2-norm of $\mathbf{w}_j$. Note that the "least sparse" (resp. "random") row presents a single example of a rotation yielding the least (resp. average) model sparsity. Eigenvector PLS-R does not have any hyperparameters, so the error level for this method does not necessarily match the others. It is included here to be consistent with previous figures.

Figures 7, 8 and 9 present the transformations and weights learned by different multi-view regression methods applied to Diabetes, Heart and Pima, respectively. Each row contains scatterplots of an MDS embedding transformed by a given method, and each column contains a different coloration of the instances for each feature. Each instance is colored based on the value of the feature for that instance, where dark blue is the minimum value and dark red is the maximum value. For instance, the first scatterplot in the second column of Figure 7 is an MDS embedding rotated by an angle yielding the least sparse solution, and it is colored according to the age of each patient in the scatterplot.

The arrows in the figures represent the weight vectors $\mathbf{w}_j$ for the different features j. Their length is proportional to $\|\mathbf{w}_j\|_2$. Arrows that are vertical or horizontal indicate that the feature in question is used to explain only the vertical or horizontal dimension of the MDS. For example, the first scatterplot in the second column of Figure 7 has a vertical arrow, meaning that age is not used to explain the horizontal dimension. For the third scatterplot in the second column of Figure 7, which corresponds to the age weights in the BIR-LR model, there is no arrow, meaning that the weights for age are equal to zero for both of the rotated MDS dimensions.

These figures allow us to show that, for the same level of error as the other methods, BIR-LR models often have more zero-valued weights (vertical or horizontal arrows, or no arrows at all). This means that the two dimensions can often be interpreted using small, disjoint sets of features. In these figures, we observe that BIR-LR finds rotations resulting in models that only include the features displaying a strong relationship with one of the rotated dimensions. For instance, in Figure 7, clear visual color trends are observed for the features sex, bmi and disease progression, which are the only features selected by BIR-LR. Because the weights for these features take the value zero for the vertical dimension, the resulting model is much sparser than for the other methods. Features like age, for which blue and red instances are mixed in all directions, are not selected by BIR-LR. Thus, for this dataset (Diabetes), the BIR-LR model suggests that a horizontal trend can be captured by three distinct features but that no feature in $\mathbf{F}$ can explain the vertical axis in the MDS embedding. This seems to be confirmed by the observation that no top-down color change is apparent for this orientation.

Similar observations can be made for Figures 8 and 9. Based on these three figures, we demonstrate that BIR-LR can provide models facilitating the interpretation of MDS embeddings, thanks to rotations resulting in sparse models.

7. Discussion on the Performance of BIR

As discussed in Section 5 and demonstrated by the experiments in Section 6, choosing an angle for interpreting an MDS embedding with sparse regression is important. Bibal, Marion and Frénay [1] have shown that choosing an angle at random leads to a worse solution on average than BIR-LR in terms of both model sparsity and error.

In this paper, we have shown that selecting the orientation of an MDS embedding using single-view rotation methods such as PCA, or two-view orthogonal transformation methods in the case of eigenvector PLS and SRRR, does not necessarily lead to the most interpretable regression models. Indeed, for all datasets evaluated, except Hepatitis, BIR-LR proposes solutions that have lower or equal test error for all degrees of sparsity. For the Hepatitis dataset, we observe that these conclusions may only hold for sparser solutions. However, for the purpose of interpretation, these may be precisely the solutions that are most desirable.

The degree of sparsity in BIR-LR weights is controlled by the hyperparameter λ, which must be selected by the user according to his needs. One possible heuristic for choosing this hyperparameter could be the elbow method, but this should only be used in cases where the plot of average MSE with respect to degree of sparsity has an elbow shape. The chosen λ would be the point with the smallest distance from the origin. The selection of an optimal λ is beyond the scope of this paper, but several potential strategies can be found in [36].

8. Conclusion and Future Works

This paper was concerned with the problem of interpreting a nonlinear dimensionality reduction (NLDR) model using a set of external features. In particular, we studied the use of linear regression to model multidimensional scaling (MDS) embedding dimensions as linear combinations of external features. This approach makes it possible to explain how the MDS model mapped instances into to the new, lower-dimensional space as a linear function of external features.

As MDS embeddings are only uniquely determined up to certain transformations, including rotation, we studied how the rotation of an MDS embedding affects subsequent linear regression models. While Lasso regression generally yields a model that is sparser and more interpretable than ordinary least squares (OLS) or Ridge regression, its model error and sparsity are both dependent on the rotation angle of the MDS embedding.

Thus, when using the Lasso to model the linear relationship between an MDS embedding and a set of external features, the rotation of the MDS embedding should not be chosen arbitrarily.

In this paper, we proposed the Best Interpretable Rotation (BIR) selection method for choosing an angle that rotates a 2D embedding such that a subsequent Lasso model strikes a balance between model error and sparsity. BIR Lasso regression (BIR-LR), which consists of Lasso regression where the target is rotated with the angle found using BIR, was also introduced. Using BIR-LR to interpret an MDS embedding model is an example of *post hoc interpretation* [37]. Indeed, sparse linear regression is used after the MDS embedding has been generated in order to interpret the way in which the instances were mapped into the embedding.

We compared BIR and BIR-LR to methods in the machine learning and statistics literature that also search for an orthogonal transformation, either based on information in the two data views available (two-view transformation) or information in only one of the two views (single-view transformation). For sparse models (i.e. models with fewer than 10 non-zero-valued weights in our experiments), BIR-LR had smaller test error than all methods tested, for all datasets.

The proposed BIR-LR method does not depend on visualization for interpretation, meaning that it would be possible to extend it to the case of more than two dimensions. In future work, the restriction to the case of two-dimensional embeddings could be lifted.

Finding an objective function integrating the optimization of θ and the weights $\mathbf{W}$ is also a subject of future work. For the moment, the best rotation is found on the basis of possible Lasso solutions. However, an optimal or near optimal rotation could be found while simultaneously learning a sparse regression model.

Finally, in this work and in the literature, constraints are used to encourage overall sparsity s; however, other definitions of sparsity could be developed that are more directly related to model interpretability. Furthermore, a more nuanced measure of interpretability that goes beyond sparsity is a subject of future research.

Acknowledgment

The authors would like to thank Nathan Nguyen from the Université catholique de Louvain for having pointed to the need for this kind of method in psychology, as well as Prof. Bernadette Govaerts and Prof. Rainer von Sachs from the Université catholique de Louvain for their insights on the subject. The first author gratefully acknowledges financial support from the Belgian Fund for Scientific Research (F.R.S.-FNRS, FRIA grant).

References

[1] 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.

[2] N. Jaworska, A. Chupetlovska-Anastasova, A review of multidimensional scaling (MDS) and its utility in various psychological domains, Tutorials in Quantitative Methods for Psychology 5 (1) (2009) 1–10.

[3] P. Legendre, L. Legendre, Numerical ecology: second English edition, Vol. 20 of Developments in Environmental Modeling, Elsevier Science, 1998.

[4] 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.

[5] B. Gawronski, J. De Houwer, Implicit measures in social and personality psychology, Handbook of Research Methods in Social and Personality Psychology 2 (2014) 283–310.

[6] I. Van Mechelen, A. K. Smilde, A generic linked-mode decomposition model for data fusion, Chemometrics and Intelligent Laboratory Systems 104 (1) (2010) 83–94.

[7] S. Sun, A survey of multi-view machine learning, Neural Computing and Applications 23 (7-8) (2013) 2031–2038.

[8] J.-a. Lin, H. Zhu, R. Knickmeyer, M. Styner, J. Gilmore, J. G. Ibrahim, Projection regression models for multivariate imaging phenotype, Genetic Epidemiology 36 (6) (2012) 631–641.

[9] J. Thioulouse, Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods, The Annals of Applied Statistics (2011) 2300–2325.

[10] W. Lee, D. Lee, Y. Lee, Y. Pawitan, Sparse canonical covariance analysis for high-throughput data, Statistical Applications in Genetics and Molecular Biology 10 (1) (2011) 1–24.

[11] K. Varmuza, P. Filzmoser, Introduction to multivariate statistical analysis in chemometrics, CRC press, 2016.

[12] J. B. Kruskal, M. Wish, Multidimensional scaling, Sage, 1978.

[13] I. Borg, P. J. Groenen, Modern multidimensional scaling: Theory and applications, Springer, 2005.

[14] M. C. Hout, M. H. Papesh, S. D. Goldinger, Multidimensional scaling, Wiley Interdisciplinary Reviews: Cognitive Science 4 (1) (2013) 93–103.

[15] A. Lebel, M. Cantinotti, R. Pampalon, M. Thériault, L. A. Smith, A.-M. Hamelin, Concept mapping of diet and physical activity: uncovering local stakeholders perception in the Quebec City region, Social Science & Medicine 72 (3) (2011) 439–445.

[16] 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).

[17] 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.

[18] S. Pattyn, Y. Rosseel, A. Van Hiel, Finding our way in the social world, Social Psychology 44 (2013) 329–348.

[19] P. I. Armstrong, S. X. Day, J. P. McVay, J. Rounds, Holland's RIASEC model as an integrative framework for individual differences, Journal of Counseling Psychology 55 (1) (2008) 1–18.

[20] G. M. Levine, J. B. Halberstadt, R. L. Goldstone, Reasoning and the weighting of attributes in attitude judgments, Journal of Personality and Social Psychology 70 (2) (1996) 230–240.

[21] D. Farrell, Exit, voice, loyalty, and neglect as responses to job dissatisfaction: A multidimensional scaling study, Academy of Management Journal 26 (4) (1983) 596–607.
[22] H. F. Kaiser, The varimax criterion for analytic rotation in factor analysis, Psychometrika 23 (3) (1958) 187–200.
[23] H. H. Harman, Modern factor analysis, University of Chicago Press, 1976.
[24] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (Nov) (2008) 2579–2605.
[25] E. R. Peay, Multidimensional rotation and scaling of configurations to optimal agreement, Psychometrika 53 (2) (1988) 199–208.
[26] H. Abdi, Partial least squares regression and projection on latent structure regression (PLS regression), Wiley Interdisciplinary Reviews: Computational Statistics 2 (1) (2010) 97–106.
[27] 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.
[28] 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.
[29] M. Lichman, UCI machine learning repository (2013). URL `http://archive.ics.uci.edu/ml`
[30] P. van der Putten, M. van Someren, Coil challenge 2000: The insurance company case, Tech. rep., Leiden Institute of Advanced Computer Science (2000).
[31] D. o. C. Bureau of the Census, Census Of Population And Housing 1990 United States: Summary Tape File 1a & 3a, U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan (1992).
[32] D. o. J. Bureau of Justice Statistics, Law Enforcement Management and Administrative Statistics, U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan (1992).
[33] D. o. J. Federal Bureau of Investigation, Crime in the United States (1995).
[34] J. W. Smith, J. Everhart, W. Dickson, W. Knowler, R. Johannes, Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, in: Proceedings of the Annual Symposium on Computer Application in Medical Care, Washington, D.C., USA, 1988, pp. 261–265.
[35] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, The Annals of Statistics 32 (2) (2004) 407–499.
[36] D. Homrighausen, D. J. McDonald, A study on tuning parameter selection for the high-dimensional lasso, Journal of Statistical Computation and Simulation (2018) 1–28.
[37] Z. C. Lipton, The mythos of model interpretability, in: ICML Workshop on Human Interpretability of Machine Learning, New York, USA, 2016.

Rebecca Marion is a Ph.D. student at the Université catholique de Louvain (UCLouvain) in Belgium, under the supervision of Professors Rainer von Sachs and Bernadette Govaerts. She received an M.S. in Statistics, concentration in Biostatistics, from UCLouvain in 2016. Her Ph.D. thesis is on multi-view learning and feature selection for data with grouped features.

Adrien Bibal is a Ph.D. student at the Université de Namur (Belgium) under the supervision of Professor Benoît Frénay. He received an M.S. degree in Computer Science and an M.A. degree in Philosophy from the Université catholique de Louvain (Belgium) in 2013 and 2015 respectively. His Ph.D. thesis in machine learning is on the interpretability of dimensionality reduction models.

Benoît Frénay is associate professor at the Université de Namur. He received his M.S. and Ph.D. degrees from the Université catholique de Louvain (Belgium) in 2007 and 2013, respectively. His main research interests in machine learning include interpretability, interactive machine learning, dimensionality reduction, label noise, robust inference and feature selection. In 2014, he received the Scientific Prize IBM Belgium for Informatics for his PhD thesis on Uncertainty and Label Noise in Machine Learning.