ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Finding the Most Interpretable MDS Rotation for Sparse Linear Models based on External Features

Adrien Bibal[1]*, Rebecca Marion[2]* and Benoît Frénay[1]

1- NADI Institute - PReCISE Research Center
University of Namur - Faculty of Computer Science
Rue Grandgagnage 21, 5000 Namur - Belgium

2- ISBA - Université catholique de Louvain
Voie du Roman Pays 20, 1348 Louvain-la-Neuve - Belgium

Abstract. One approach to interpreting multidimensional scaling (MDS) embeddings is to estimate a linear relationship between the MDS dimensions and a set of external features. However, because MDS only preserves distances between instances, the MDS embedding is invariant to rotation. As a result, the weights characterizing this linear relationship are arbitrary and difficult to interpret. This paper proposes a procedure for selecting the most pertinent rotation for interpreting a 2D MDS embedding.

1 Introduction

In many applications, the usability of machine learning techniques depends on their interpretability [1]. This paper deals with the problem of understanding, or interpreting, a multidimensional scaling (MDS) embedding using features that were not used to compute the MDS (i.e. "external" features). This is a kind of *multi-view learning* task based on data from multiple sources [2]. The goal here is to characterize the relationship between two views: one taking the form of (dis-)similarities between instances and the other expressing features of these instances.

For example, in psychology, two independent experiments are sometimes run where one is used to interpret the result of the other. This is the case for implicit measure studies, which aim to understand human decisions encoded in one database by using another database. A first database is composed of similarity ratings for a set of instances, whereas the second database contains characterizations of the same instances with respect to a set of features. The research question is then: how can the feature matrix be used to explain the comparisons in the first database? Another field of application is the medical sciences, where clinical features can be used to interpret patient similarity with respect to gene expression, protein abundance, etc.

This work proposes an approach that strikes a balance between interpretability and performance: it finds an optimal rotation of an MDS embedding that can be used to identify a small subset of features necessary for accurately explaining that embedding. In this work, we focus on 2D MDS embeddings, constraining the rotation to revolve around a single axis.

*Both authors have contributed equally.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

2 State of the Art

The problem of interpreting an MDS representation of a set of instances is frequently encountered in the social sciences (see, e.g., [3]). Let $\mathbf{Y}$ $(n \times K)$ be a matrix resulting from the application of MDS to an $n \times n$ (dis-)similarity matrix. Some authors interpret this embedding by clustering the instances in $\mathbf{Y}$ [4]. For 2D MDS embeddings ($K = 2$), another more popular approach is to regress a set of external features $\mathbf{f}_j$, $j : 1, .., d$, onto the MDS matrix $\mathbf{Y}$ through *property fitting* [5]: $\mathbf{f}_j = \mathbf{Y}\mathbf{w}_j + \boldsymbol{\xi}_j$, where $\mathbf{w}_j$ is a vector of weights and $\boldsymbol{\xi}_j$ is an error vector. A subset of features important for explaining the MDS dimensions are identified based on some measure of model fit, such as the coefficient of determination R^2. If the model for a given feature $\mathbf{f}_j$ has a sufficiently adequate fit with respect to some threshold, its line of fit is plotted in the MDS space. As a result, the MDS can be interpreted based on a subset of external features.

Unfortunately, because each feature is regressed separately onto $\mathbf{Y}$, potential dependence between features is ignored. In order to account for all features at once, some authors apply Principal Component Analysis (PCA) to a feature matrix $\mathbf{F}$ $(n \times d)$, then regress each principal component l onto the MDS matrix: $\text{PCA}(\mathbf{F})_l = \mathbf{Y}\mathbf{w}_l + \boldsymbol{\xi}_l$, for $l : 1, ..., q$, where q is the total number of principle components. Extra processing steps have also been proposed in order to allow the PCA components to be non-orthogonal (see [3] for an applied example).

While the weights for each dimension of $\text{PCA}(\mathbf{F})$ are still estimated independently of each other, this method has the advantage of accounting for dependence between features: each component regressed onto $\mathbf{Y}$ is a linear combination of features. However, the PCA components of $\mathbf{F}$ are estimated independently of the MDS embedding $\mathbf{Y}$. This means that the PCA components are not optimal, in terms of precision, for a regression onto the MDS space. In addition, the solution does not necessarily improve model interpretability, as a single principle component l could depend on all of the features in $\mathbf{F}$.

3 Proposed Approach

As seen in Section 2, there is a need for a method that identifies a small subset of features that best explain two MDS dimensions $\mathbf{y}_1$ and $\mathbf{y}_2$ while accounting for dependence between features $\mathbf{f}_j$. In order to allow features to jointly explain the MDS dimensions, we propose performing a linear regression where the MDS dimensions $\mathbf{y}_1$ and $\mathbf{y}_2$ are response variables, rather than predictors, and thus the predictors are the features in $\mathbf{F}$. This section presents our motivation and goals, as well as our proposed approach, which is then evaluated in Section 4.

3.1 Motivation and Goals

Let the multivariate regression model be defined as $\mathbf{Y} = \mathbf{F}\mathbf{W} + \boldsymbol{\Xi}$, where $\mathbf{W}$ $(d \times 2)$ is a matrix containing the regression weights to be estimated and $\boldsymbol{\Xi}$ $(d \times 2)$ is an error matrix. Variables with non-zero weights for a given dimension of $\mathbf{Y}$ are considered to be explicative of the corresponding axis in the 2D MDS space.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Unfortunately, the orientation of the MDS embedding $\mathbf{Y}$ is arbitrary, meaning that the weights $\mathbf{W}$ are also arbitrary, and thus difficult to interpret. Indeed, $\mathbf{Y}$ is found by minimizing a measure of the degree to which distances between n instances in the $n \times n$ (dis-)similarity space are preserved in the new $n \times 2$ space. A popular measure of this kind is the Kruskal stress [6]. Minimizing this criterion results in MDS solutions with arbitrary orientations because distances between instances in the resulting space remain the same for any rotation.

Rather than simply regressing the arbitrarily rotated MDS solution $\mathbf{Y}$ onto $\mathbf{F}$, it could be more relevant to find a rotation of $\mathbf{Y}$ that optimizes some criterion related to the analysis goals at hand. Let the 2D rotation matrix $\mathbf{R}^\theta$ for a given angle θ be defined as

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

The regression model of interest is thus: $\mathbf{YR}^\theta = \mathbf{FW}^\theta + \mathbf{\Xi}$, where $\mathbf{W}^\theta$ is a weight matrix that depends implicitly on the rotation angle θ.

For our particular case, we are interested in finding the rotation angle θ that optimizes some trade-off between interpretability and model error. We assume that the model is most interpretable when the number of non-zero weights in $\mathbf{W}^\theta$ is minimal, i.e. the model is "sparse."

Without considering sparsity, the ordinary least squares (OLS) solution for $\theta = 0$ is given by $\mathbf{W}^0 = (\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{Y}$. It can be shown that the OLS solution for any θ is $\mathbf{W}^\theta = (\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{YR}^\theta = \mathbf{W}^0\mathbf{R}^\theta$. Thus, the effect of rotating $\mathbf{Y}$ is to rotate $\mathbf{W}^0$ with the same angle, and the mean squared error (MSE) of the model, which is the sum of the MSE for $\mathbf{y}_1$ and $\mathbf{y}_2$, is invariant under rotation.

The OLS solution, however, does not guarantee interpretability as defined above. In order to encourage interpretability, some model constraint must be included so that unimportant variables are excluded from the model. A natural constraint for this purpose is the L_0 norm, which counts the number of non-zero weights in the model. The function to minimize is

$$\frac{1}{2n}\sum_{k=1}^{2}||\mathbf{Yr}_k^\theta - \mathbf{Fw}_k^\theta||_2^2 + \sum_{k=1}^{2}\lambda||\mathbf{w}_k^\theta||_0, \tag{1}$$

where λ is a tuning parameter that controls the trade-off between model error and interpretability. Optimizing Eq. (1) with respect to $\mathbf{W}^\theta$ is an NP-Hard problem [7], so in practice, the L_1 norm is often used as an approximation [8]:

$$\frac{1}{2n}\sum_{k=1}^{2}||\mathbf{Yr}_k^\theta - \mathbf{Fw}_k^\theta||_2^2 + \sum_{k=1}^{2}\lambda||\mathbf{w}_k^\theta||_1. \tag{2}$$

For a given θ, the solution $\mathbf{W}^\theta$ is found using any Lasso implementation. However, in contrast to the OLS solution, for a given λ, the model error and sparsity of the Lasso solution depend on the rotation angle (see Section 4.3). The optimal rotation angle θ^* being unknown, it must be optimized.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

3.2 Finding the Best Rotation with the L_0 Norm

The proposed procedure for finding the best interpretable rotation (BIR) provides an optimal angle θ^* and associated weight matrix $\mathbf{W}^{\theta^*}$ for which the number of non-zero weights and the model error are minimized. In an approach inspired by [9], the procedure finds an angle θ whose corresponding Lasso solution $\mathbf{W}^{\theta}$ minimizes Eq. (1). This procedure is formalized by

$$\theta^* = \arg\min_{\theta} \sum_{k} \left(\frac{1}{2n} ||\mathbf{Y}\mathbf{r}_k^{\theta} - \mathbf{F}\mathbf{w}_k^{\theta}||_2^2 + \lambda ||\mathbf{w}_k^{\theta}||_0 \right), \tag{3}$$

where $\mathbf{W}^{\theta} = \text{Lasso}(\mathbf{F}, \mathbf{YR}^{\theta}, \lambda)$, which is found by minimizing Eq. (2). The univariate function to minimize in Eq. (3) being non-convex, any generic solver for non-convex optimization may be used.

4 Evaluation

This section evaluates the performance of the Lasso solution when the matrix $\mathbf{Y}$ is rotated with the angle found using the BIR selection procedure. This is then compared to (i) the average performance of angles resulting in the least sparse Lasso solutions, as well as (ii) the estimated performance when $\mathbf{Y}$ is rotated with a random angle from the set $\Theta = \{0.1, 0.2, ..., 360\}$ degrees. The first case demonstrates the worst case scenario and the second represents the estimated expected performance obtained for an arbitrary MDS orientation.

4.1 Data and Pre-Processing

We evaluated the performance of the proposed BIR selection procedure on five popular datasets: Hepatitis, Dermatology, Heart (Statlog), and Pima Indians Diabetes from [10] and Diabetes from [11]. These datasets were chosen because their features can be easily split into two different, meaningful data views. For example, Hepatitis can be split into a view with basic clinical features (e.g. age, family history, etc.) and another view with more complex histopathological features (e.g. melanin incontinence, etc.). For each dataset, we removed all instances with missing values. We used the view with the most complex features to compute a dissimilarity matrix based on Euclidean distances, then applied 2D metric MDS. We used the other view (normalized) to interpret the MDS space.

4.2 Evaluation Criteria

We evaluated the BIR procedure using two criteria. The first criterion, referred to as s^{θ}, measures the degree of model sparsity (i.e. interpretability), and is calculated as $\sum_{k=1}^{2} ||\mathbf{w}_k^{\theta}||_0$, the number of non-zero weights in $\mathbf{W}^{\theta}$. $\text{Prob}(s^{\theta}) = \frac{1}{|\Theta|} \left| \left\{ \theta' \in \Theta \mid \sum_{k=1}^{2} ||\mathbf{w}_k^{\theta'}||_0 = s^{\theta} \right\} \right|$ represents the approximate probability that Lasso obtains a degree of sparsity s^{θ} when θ is chosen at random. The second criterion is the overall model error $MSE = \frac{1}{2n} \sum_{k=1}^{2} ||\mathbf{Y}\mathbf{r}_k^{\theta} - \mathbf{F}\mathbf{w}_k^{\theta}||_2^2$.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Dataset	**Angle Selection**	θ (°)	s^θ	**Prob**(s^θ)	$\boldsymbol{MSE}$
Hepatitis	least sparse case		11	8.9%	0.169
$d = 15$	average case		9.1		0.170
30 weights	BIR procedure	59.8	6	2.1%	0.168
Dermatology	least sparse case		12	3.1%	0.098
$d = 17$	average case		9.5		0.092
34 weights	BIR procedure	36.4	7	12.7%	0.086
Heart	least sparse case		3	71.9%	0.180
$d = 4$	average case		2.7		0.180
8 weights	BIR procedure	0.9	2	28.1%	0.180
Diabetes	least sparse case		7	42.5%	0.195
$d = 5$	average case		5.7		0.194
10 weights	BIR procedure	68.2	3	10.1%	0.191
Pima	least sparse case		5	8.6%	0.220
$d = 5$	average case		3.4		0.219
10 weights	BIR procedure	20.3	2	0.7%	0.224

Table 1: Comparison of BIR selection with the least sparse and average cases. The total number of weights is twice the number of external features ($= 2 \times d$).

4.3 Results

For each dataset, the results presented here correspond to (i) the least sparse rotations, which highlights the importance of choosing an appropriate angle, (ii) the expected value estimated by averaging the criterion values for all θ in the set $\Theta = \{0.1, 0.2, ..., 360\}$ degrees, and (iii) the rotation chosen by the BIR procedure. The relative performance of i-iii was similar for a variety of λ values. Experimental results for one of these values, $\lambda = 0.1$, can be found in Table 1.

4.4 Discussion

For all datasets, the BIR procedure yields a solution that is 1.5-2.5 times more sparse than the least sparse solution with a negligible computational cost (a few seconds). Selecting a random angle results, on average, in models that are also less sparse than for the BIR procedure, with a greater or equal error for all but one dataset. These results suggest that using a rotation selection procedure is advantageous for someone requiring interpretability. Furthermore, the probability of randomly choosing a solution with the least sparsity can be high relative to a sparser solution. For Diabetes, there is a 42.5% chance of randomly selecting a rotation yielding 7 non-zero weights, whereas a solution with only 3 non-zero weights can be found using the BIR procedure.

5 Conclusion

This paper demonstrates the importance of choosing a rotation angle for a 2D MDS embedding that makes it easier to interpret. A procedure is provided for

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

selecting a rotation and estimating a sparse linear regression model that finds a compromise between interpretability and model error.

In the current procedure, an optimal rotation angle θ^* is chosen by minimizing a function that depends on Lasso solutions $\mathbf{W}^\theta$. In a future work, it would be interesting to develop a more direct and simultaneous optimization of the angle and weight matrix. Another extension would be to tackle the problem of rotating an MDS space with more than two dimensions, which would require the optimization of a vector $\boldsymbol{\theta}$. Moreover, a more nuanced definition of interpretability could be used to encourage both overall sparsity and an equal distribution of non-zero weights among the MDS dimensions.

Acknowledgment

The authors want to thank Nathan Nguyen from the Université catholique de Louvain for having pointed to the need for this kind of procedure in psychology. The second author gratefully acknowledges financial support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy), the Fonds spécial de recherche (Fédération Wallonie-Bruxelles) and the Belgian Fund for Scientific Research (F.R.S.-FNRS, FRIA grant).

References

[1] A. Bibal and B. Frénay. Interpretability of machine learning models and representations: an introduction. In *Proceedings of ESANN*, pages 77–82, 2016.

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

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

[4] M. L. Davison and S. G. Sireci. Multidimensional scaling. In *Handbook of applied multivariate statistics and mathematical modeling*, pages 323–352. Elsevier, 2000.

[5] J. J. Chang and J. D. Carroll. How to use PROFIT, a computer program for property fitting by optimizing nonlinear or linear correlation. *Unpublished manuscript, Bell Laboratories*, 1968.

[6] J. B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Publications, 1978.

[7] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

[8] C. Ramirez, V. Kreinovich, and M. Argaez. Why L1 is a good approximation to L0: A geometric explanation. *Journal of Uncertain Systems*, 7, 2013.

[9] W. Herlands, M. De-Arteaga, D. Neill, and A. Dubrawski. Lass0: sparse nonconvex regression by local search. *NIPS Workshop on Optimization*, 2015.

[10] M. Lichman. UCI machine learning repository, 2013.

[11] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.