ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Efficient unsupervised clustering for spatial bird population analysis along the Loire river

Aurore Payen[1], Ludovic Journaux[1,2], Clément Delion[1], Lucile Sautot[1,3], Bruno Faivre[3] *

1- AgroSup Dijon
26 Boulevard Docteur Petitjean, 21000 Dijon

2- Université de Bourgogne - LE2I
Avenue Alain Savary 21000 Dijon - France

3- Université de Bourgogne - Biogéosciences
6 Boulevard Gabriel 21000 Dijon - France

Abstract. This paper deals with application and comparison of Nonlinear Dimensionality Reduction (NLDR) methods on natural high dimensional bird communities dataset along the Loire River (France). In this context, biologists usually use the well-known PCA in order to explain the upstream-downstream gradient. Unfortunately this method was unsuccessful on this kind of nonlinear dataset. This paper aims at comparing recent NLDR methods coupled with different data transformations in order to find out the best approach. Results show that Multiscale Jensen-Shannon Embedding (Ms JSE) outperforms all the other methods in this context.

1 Introduction

Longitudinal distribution pattern of organisms along rivers is a major research topic in ecology, which was initiated by Illies on invertebrates, by Huet on fishes and by Frochot & *al.*[3] on bird population. In this latest work, ornithologists analyze spatiotemporal distribution of bird communities in order to study river zonation by detecting ecological discontinuities due to geomorphology of landscapes (discontinuities resulting from bird species assemblage). Biologists usually use Principal Component Analysis (PCA) in order to explain the longitudinal distribution pattern and to find discontinuities along the upstream-downstream gradient of the river. But on our dataset, PCA shows a strong limitation. To overcome this problem, it is interesting to use Non Linear Dimensionality Reduction (NLDR) methods to transform high-dimensional data into a meaningful low dimension representation. Numerous studies have aimed at comparing NLDR algorithms, usually using synthetic data such as the swissroll [7], but less with natural data. So, this paper explores and compares recent NLDR methods with different data transformations in order to find out the best approach in this ecological context. This paper is organized as follows. Section 2 presents the real-life dataset used in this context, the data transformations

*Data acquisition received financial support from the FEDER Loire, Etablissement Public Loire, DREAL de Bassin Centre (Etude des oiseaux nicheurs de la Loire et de l'Allier sur l'ensemble de leurs cours) to BF, and from the Région Bourgogne (PARI, Projet Agrale 5) to BF. Data analysis received support from the French Ministry of Agriculture (Bourse FCPR to LS).

and an NLDR methods overview with a comparison based on a quality assessment. Section 3 presents and discusses experimental results. Section 4 draws the conclusions.

2 Materials and methods

2.1 Real-life dataset and transformation

Our dataset comes from the census birds STORI[1] program for nesting birds along the Loire River (France) [3]. STORI aims to observe spatiotemporal changes into bird populations along rivers.198 census points were defined along the Loire. At each point birds are identified with the PAI (Punctual Abundance Index) method [3] during four census campaigns. Bird abundances were described by a semi-quantitative abundance index. One of the main objectives is to study global/local factors that explain bird abundances changes. Finally, we consider 140 bird species along the 198 census points. In practice, ornithologists capped PAI to 5 even if there are more than 5 couples of birds.

So the number of couples could be underestimated. This fact leaded us to try different data transformations and watched their impact on a quality criterion. In section 3.1 we use this quality criterion to evaluate the data correction obtained by using the square root and the Anscombe transformation (AT).

2.2 Overview of different methods of dimensional reduction

NLDR methods can be classified according to different characteristics: (i) **Scale analysis (local/global/multi-scale)**: this reflects the kind of properties the transformation does preserve; (ii) **Distance metrics/similarity**: this shows the distance used to estimate if two data points are close. We retained 7 NLDR methods completed by to 2 linear methods: the Classical Multidimensional Scaling (CMDS) and the Non-metric Multidimensional Scaling (NMDS) [7].

Nonlinear Mapping (NLM) (Sammon's mapping) [7] tries to preserve the neighborhood topology of data by preserving distances between points according to the following stress function:

$$J_{NLM} = \frac{1}{\sum_{i,j=1}^{n} d_{i,j}^{m}} \left(\sum_{i,j=1}^{n} \frac{(d_{i,j}^{m} - d_{i,j}^{p})^2}{d_{i,j}^{m}} \right)$$

With d_{ij}^m and d_{ij}^p are the distances between points i^{th} and j^{th}, in $\mathbb{R}^m$and $\mathbb{R}^p$.

Curvilinear Component Analysis (CCA)[7] is an evolution of NLM. Instead of the optimization of a reconstruction error, CCA aims at preserving the distance matrix while projecting data onto $\mathbb{R}^p$ dimension, giving priority to low distances. The use of similarities in NLDR is recent[5].This approache is based on sparse matrices of similarities defined in $\mathbb{R}^m$, such as in ***Stochastic Neighbor Embedding (SNE)*** [1, 4] where distances are converted into probabilities

[1] Temporal Monitoring of Nesting Birds in River Valley

which represent similarities. SNE aims to preserve similarities in $\mathbb{R}^m$ and $\mathbb{R}^p$. In this context ***t-distributed Stochastic Neighbor Embedding (t-SNE)***[2] and ***Neighbor Retrieval Visualizer (NeRV)***[9] are SNE evolutions. t-SNE is based on Student t-distribution to calculate similarities while a Gaussian distribution is used in SNE. Both SNE and t-SNE try to reduce the Kullback–Leibler divergence (KLD) as a cost function. NerV uses two dual KLD instead of a single KLD. From this KLD approach, different refinements have been proposed. First by replacing KLD by Jensen-Shannon divergence in the ***Jensen-Shannon Embedding (JSE)***[6]. Secondly, to overcome one of the major drawbacks fixed size of neigborhood, [5] proposed to take into account different sizes of neigborhood, thanks to a log scale, in ***Multiscale Jensen-Shannon Embedding (Ms JSE)***.

2.3 Objective comparison based on quality assessment

Several quantitative evaluation measures for NLDR have been proposed including techniques which rely on neighborhood ranking. We based our quality criterion on the intrusion/extrusion diagram proposed by Lee & Verleysen. Fore more details on quality assessment see on[8]. This criterion is the Area Under Curve (AUC), a scale-independent criterion got by calculating the area under the curve of R_{NX} function, which gives the percentage of improvement of neighborhood preservation compared to a random projection, depending on the size of the neighborhood.

3 Results and discussions

3.1 The choice of a transformation

The Table 1 shows AUC results for each transformation in order to measure their impacts on data. The Figure 1 shows R_{NX} functions of the square root of the data, on which AUC results of the square root of the data are calculated. Results show for every NLDR methods that the square root gets better results than the data without transformation or AT. Moreover, the best result is obtained with MS JSE with 57.2%. Finally, we select the square root, which corrects best the data.

3.2 The choice of the more efficient NLDR method

Figure 1 presents the quality curves obtained with the different methods. Colored curves are R_{NX} functions for each method. They represent the improvement of neighborhood conservation compared to a random projection. Percentages of neighborhood conservation (Q_{NX} function) are given by dotted lines. For instance, t-SNE almost reaches 70% of neighborhood preservation when K=1, as we see that its army green curve almost crosses the 70% dotted line. At local scale (K¡40), Ms JSE, JSE and t-SNE clearly outperform the other methods. At K¿80, NeRV and SNE get as good as Ms JSE, JSE and t-SNE. At about K=300, methods that were until now clearly outperformed become as effective

Without transformation	Square root	Transformation of Anscombe
· 21.5 CMDS	· 27.3 CMDS	· 26.6 CMDS
× 30.7 NMDS	× 34.3 NMDS	× 33.0 NMDS
• 26.1 NLM	• 27.5 NLM	• 28.0 NLM
○ 29.5 CCA	○ 31.2 CCA	○ 31.8 CCA
+ 33.8 SNE	+ 43.6 SNE	+ 43.0 SNE
◇ 42.8 *t*-SNE	◇ 50.0 *t*-SNE	◇ 46.2 *t*-SNE
▽ 38.0 NeRV	▽ 46.8 NeRV	▽ 45.6 NeRV
◇ 45.9 JSE	◇ 52.8 JSE	◇ 50.9 JSE
□ 49.4 Ms. JSE	□ 57.2 Ms. JSE	□ 55.1 Ms. JSE

Table 1: AUC results for data transformation

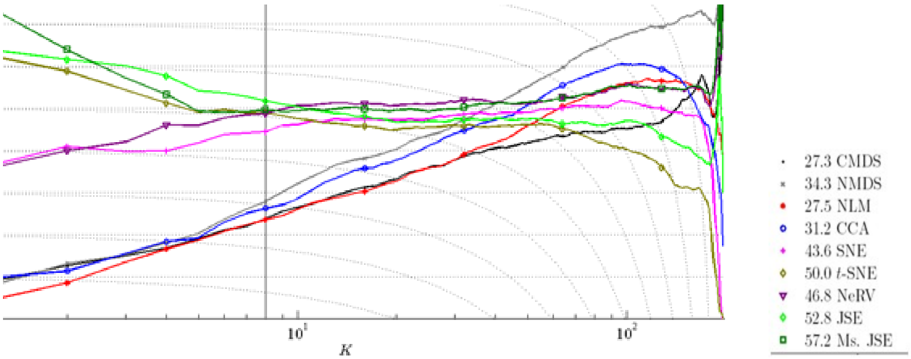

Figure 1: Quality curves of R_{NX} function on the square root of the data

as the others in conserving neighborhood (NMDS, NLM,...). For higher values of neighborhood, NMDS becomes the most efficient method (85% of improvement for 1000 neighbors). The AUC results can be seen in the caption and show that Ms JSE is the best method on all scale. What makes Ms JSE the best method despite all of that, is that:

1. All the methods reach high percentage of neighborhood conservation at global scale: at K=1000, all the methods reach or exceed 70% of neighborhood conservation.
2. At middle scale, the methods have roughly the same efficiency.
3. But at small neighborhood, the methods are under 40% of neighborhood conservation, except Ms JSE, JSE and t-SNE that exceed 60%.
4. Ms JSE has better results than t-SNE and JSE at large neighborhoods.

That's why Ms JSE will be used to analyse our dataset.

3.3 Resulting projection of the data with NLDR methods

Census points are projected with methods (Figure 2). The projection comparison shows that NLDR based on similarity outperform over methods in clustering context. The Figure 3 focuses on Ms JSE. We can see that global/local organization of data is respected. At global scale we observe the upstream-downstream

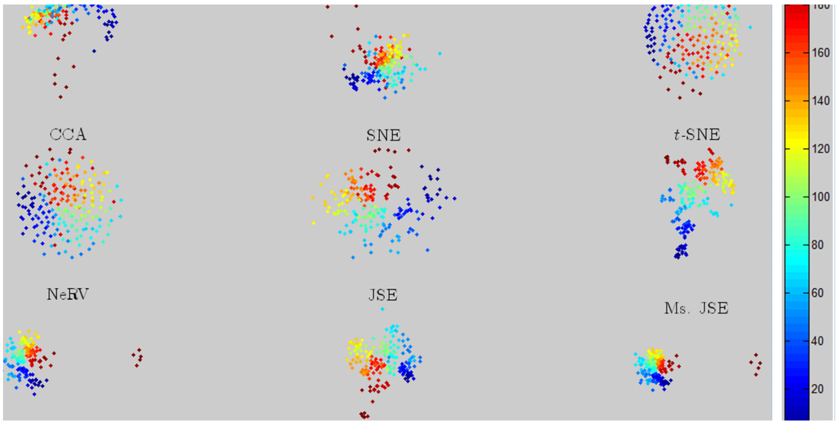

Figure 2: Projection of the square root of the data, with all the methods (blue: upstream of Loire river, red brown: downstream)

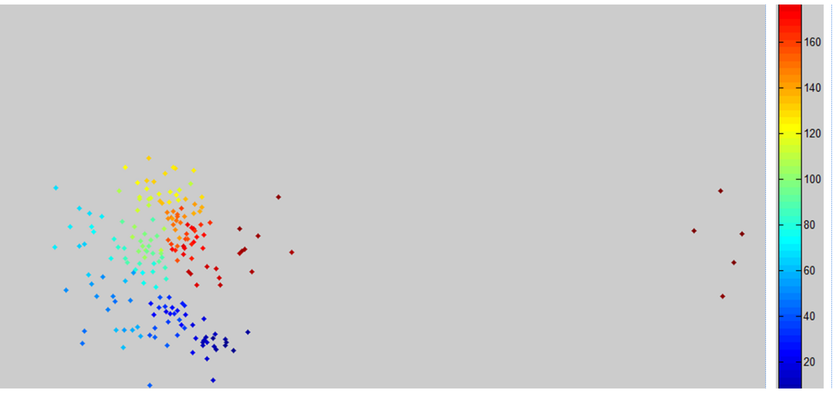

Figure 3: Census points gradient with Ms JSE and square root transformation

gradient relationship between census points. At local scale, Ms JSE is able to give an efficient clustering, grouping the data which have the same characteristics in terms of birds species assemblage. In the context of river zonation, each cluster represents a different bird species assemblage depending on environmental features and each distance between clusters represent an ecological discontinuity which is not detectable with linear approach. Moreover, Ms JSE has the characteristic of preserving outliers. The five last census points are distinct from the others: they are the five last census points, located in the downstream of the Loire River (next to the Atlantic Ocean). These census points are indeed very different from the others considering their bird population. This difference with the other census points isn't clear with local NLDR methods, such as NLM, CCA, SNE and t-SNE.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

4 Conclusion

This paper explores and compares recent NLDR methods with data transformations in order to find the best method on a real-life ornithological application. Results highlight that Ms JSE with square root transformation is the most efficient method. The global organization of census points reveals the upstream-downstream gradient and the local clustering highlights discontinuities. These results outperform traditional PCA in this context. In order to generalize these results, more tests on other nonlinear natural datasets should be made to confirm the ability of Ms JSE to make efficient projections in ecological context.

References

[1] Kerstin Bunte, Sven Haase, and Michael Biehl et al. Stochastic neighbor embedding (sne) for dimension reduction and visualization using arbitrary divergences. *Neurocomputing*, 90:23–45, 2012.

[2] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[3] B. Frochot, M.C. Eybert, L. Journaux, J. Roché, and B. Faivre. Nesting birds assemblages along the river loire: result from a 12 years-study. *Alauda*, 71(2):179–190, 2003.

[4] G. Hinton and S. Roweis. Stochastic neighbor embedding. *Advances in Neural Information Processing Systems*, 15:833–840, 2002.

[5] John A. Lee, Diego H. Peluffo-Ordonez, and Michel Verleysen. Multiscale stochastic neighbor embedding: Towards parameter-free dimensionality reduction. In *Proceedings of 22st European Symposium on Artificial Neural Networks, Computational Intelligence And Machine Learning (ESANN)*, 2014.

[6] John A Lee, Emilie Renard, Guillaume Bernard, Pierre Dupont, and Michel Verleysen. Type 1 and 2 mixtures of kullback–leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 112:92–108, 2013.

[7] John Aldo Lee and Michel Verleysen. *Nonlinear Dimensional Reduction.* Information Science and Statistics. Springer, 2007.

[8] John Aldo Lee and Michel Verleysen. Quality assessment of dimensionality reduction: rank-based criteria. *Neurocomputing*, 72:1431–1443, 2009.

[9] Jarkko Venna, Jaakko Peltonen, and Kristian Nybo et al. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490, 2010.