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/.

# Diffusion Maps Parameters Selection Based on Neighbourhood Preservation

Carlos M. Alaíz, Ángela Fernández and José R. Dorronsoro *

Universidad Autónoma de Madrid & Instituto de Ingeniería del Conocimiento
Tomás y Valiente 11, 28049 Madrid - Spain
{carlos.alaiz,a.fernandez,jose.dorronsoro}@uam.es

**Abstract**. Diffusion Maps is one of the leading methods for dimensionality reduction, although it requires to fix a certain number of parameters that can be crucial for its performance. This parameter selection is usually based on the expertise of the user, as there are no unified criterion for evaluating the quality of the embedding. We propose to use a neighbourhood preservation measure as the criterion for fixing these parameters. As we shall see, this approach provides good embedding parameters without needing problem specific knowledge.

## 1 Introduction

A common assumption in many problems is that although original data appear to have a very large dimension, they actually lie in a low-dimensional manifold $\mathcal{M}$ of which a suitable representation has to be given. This is the scenario in manifold learning, where the key problem is to identify $\mathcal{M}$ and to derive useful embeddings. The preceding assumption has given rise to a number of methods, among which one of the main representatives is Spectral Clustering or, more generally, Diffusion Maps (DM) [1], a dimensionality reduction technique based on the assumption that the metric of the low-dimensional Riemannian manifold where data lie can be approximated by a certain diffusion metric. Nevertheless, these methods often requires the proper tuning of a number of parameters but, being an unsupervised problem, the lack of an adequate measure of embedding quality makes parameter selection quite difficult. This is indeed the case in DM as there are several parameters, such as the number of Markov steps or the width of the kernel, that have to be selected manually or using some heuristics. As an alternative, we propose in this paper to use the neighbourhood preservation measure of [2] as a criterion for selecting the best parameters. As shown in the experiments, when applied to labelled data, this measure is highly correlated with the classification accuracy, and can thus be used for automated unsupervised parameter selection so as to provide good embeddings.

The paper is structured as follows. Section 2 introduces briefly DM. We review the quality measure and propose how to use it for parameter selection in Sect. 3, and in Sect. 4 we present some experiments to verify its usefulness. Finally, Sect. 5 presents some conclusions and pointers to further work.

*With partial support from Spain's grants TIN2013-42351-P and S2013/ICE-2845 CASI-CAM-CM, and the Cátedra UAM–ADIC in Data Science and Machine Learning. The authors also acknowledge the use of the facilities of Centro de Computación Científica (CCC) at UAM.

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/.

## 2 Diffusion Maps

The first step in DM is to define a weighted graph from the sample using a similarity matrix $w_{ij} = \exp\left(-\|x_i - x_j\|^2 / \left(2\sigma^2\right)\right)$. To take into account the sampling distribution, a parameter $\alpha \in [0,1]$ is introduced to define ${w^{(\alpha)}}_{ij} = w_{ij} / \left({g_i}^{\alpha} {g_j}^{\alpha}\right)$, where $g_i = \sum_j w_{ij}$ is the degree of a vertex $i$ [1]. The degrees corresponding to these new weights are ${g^{(\alpha)}}_i = \sum_j {w^{(\alpha)}}_{ij}$, and we can define the matrix $\tilde{W}^{(\alpha)} = \{{\tilde{w}^{(\alpha)}}_{ij} = {w^{(\alpha)}}_{ij} / {g^{(\alpha)}}_i\}$, which is a Markov transition matrix over the graph. After fixing a number $t$ of Markov process steps, the $t$-step diffusion distance is given by $D^t_{ij} = \|{\tilde{w}^{(\alpha;t)}}_{i,\cdot} - {\tilde{w}^{(\alpha;t)}}_{j,\cdot}\|_{L^2(1/\phi_0)}$, where $\phi_0$ is the stationary distribution of the Markov process and $\tilde{w}^{(\alpha;t)}$ the transition probability in $t$ steps (the $t$-th power of $\tilde{w}^{(\alpha)}$). The eigenanalysis of the Markov matrix $\tilde{W}^{(\alpha)}$ gives [1] an alternative representation of the diffusion distance as $D^t_{ij} = \sum_k \lambda_k^{2t} ({\psi^{(k)}}_i - {\psi^{(k)}}_j)^2$, with $\lambda_k$ the eigenvalues and $\psi^{(k)}$ the left eigenvectors of the Markov matrix $\tilde{W}^{(\alpha)}$, and ${\psi^{(k)}}_i = \psi_k(x_i)$ the eigenvectors' components. If the eigenvalues $\lambda_k$ decay rather fast, we can perform dimensionality reduction by retaining a small number $d$ of the largest eigenvalues and their corresponding eigenvectors. Once fixed, we would thus arrive to the diffusion coordinates $\Psi = (\lambda_1^t \psi_1(x), \ldots, \lambda_d^t \psi_d(x))^\top$, and we can approximate the diffusion distance as $D^t_{ij} \approx \sum_{k=1}^{d} \lambda_k^{2t} ({\psi^{(k)}}_i - {\psi^{(k)}}_j)^2 = \|\Psi(x_i) - \Psi(x_j)\|^2$. In other words, the diffusion distance in the manifold can be approximated by the Euclidean distance in the DM projected space. Moreover, it is worth mentioning that the well-known Spectral Clustering [3] method is a particular case of DM for $\alpha = 0$ and $t = 0$. A detailed discussion of DM can be found in [4].

One of the problems when using DM in a real context is the dependence of the embedding quality on its parameters, namely: 1) $d$, the embedding dimension, 2) $\sigma$, the width of the Gaussian kernel, usually selected as the percentile $p$ of all pattern distances in the original space, 3) $\alpha$, the parameter to control the influence of the sampling density, and 4) $t$, the number of steps. In what follows we will define a framework for the automated selection of these parameters.

## 3 Neighbourhood Preservation Measure for DM

Lee and Verleysen introduced in [2] a quality measure for dimensionality reductions, based on neighbourhood preservation between the original and the reduced space. Specifically, assume a $N$ pattern sample, and let $\nu_i^k$ be the set of $k$ nearest neighbours ($k$NN) of the $i$-th pattern in the original space metric and $n_i^k$ the set of $k$NN using the metric of the embedding. We define the following measures:

$$Q_{\mathrm{NX}}(k) = \frac{1}{kN} \sum_{i=1}^{N} \left|\nu_i^k \cap n_i^k\right| \quad ; \quad R_{\mathrm{NX}}(k) = \frac{(N-1)\, Q_{\mathrm{NX}}(k) - k}{N - 1 - k} \,.$$

Notice that $Q_{\mathrm{NX}}(k)$ measures the average $k$-ary neighbourhood preservation, and $R_{\mathrm{NX}}(k)$ is just a rescaled version that specifies the improvement over a purely random dimensionality reduction [5]. This measure can be integrated to

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/.

get a scalar score [6], namely the area under the curve in logarithmic scale:

$$\mathrm{AUC}^{\mathrm{ln}} = \left(\sum\nolimits_{k=1}^{N-2} \frac{R_{\mathrm{NX}}(k)}{k}\right)\left(\sum\nolimits_{k=1}^{N-2} \frac{1}{k}\right)^{-1} .$$

In this expression all the different neighbourhood sizes are taken into account, although differently weighted according to the scale. Since the final goal of DM is to provide an embedding where the local distances are preserved, considering large neighbourhoods seems pointless. At the same time, when the size is too small the artefacts and instability associated to 1-NN may appear. This is why we propose to modify the previous measure to take into account only those neighbourhoods with sizes between the 5% and 10% of the number of patterns:

$$\mathrm{AUC}^{\mathrm{ln}}_{5:10} = \left(\sum\nolimits_{k=[0.05N]}^{[0.10N]} \frac{R_{\mathrm{NX}}(k)}{k}\right)\left(\sum\nolimits_{k=[0.05N]}^{[0.10N]} \frac{1}{k}\right)^{-1} .$$

We can use $\mathrm{AUC}^{\mathrm{ln}}_{5:10}$ as the fitness function to select an optimum set $\mathcal{P}_{\mathrm{AUC}}$ of DM parameters as one that maximizes this measure amongst a discrete grid of parameter sets, $\mathcal{P}_{\mathrm{AUC}} = \arg\max_{\mathcal{P}} \mathrm{AUC}^{\mathrm{ln}}_{5:10}(\mathrm{DM}_{\mathcal{P}})$, where $\mathrm{DM}_{\mathcal{P}}$ is the DM embedding with parameters $\mathcal{P}$. In other words, we can build several DM embeddings changing the different parameters that configure them, and compute for each of these embeddings the value of $\mathrm{AUC}^{\mathrm{ln}}_{5:10}$. As we will show in Sect. 4, selecting the configuration with maximum $\mathrm{AUC}^{\mathrm{ln}}_{5:10}$ ensures a good embedding. Moreover, the main advantage of this novel approach is that we do not need any kind of additional information about the problem being studied, as the measure is based solely on the characteristics of the resulting embedding and its relation with the original dataset.

## 4 Numerical Experiments

In this section we will confirm that $\mathrm{AUC}^{\mathrm{ln}}_{5:10}$ is a good measure of the quality of an embedding that allows to select the hyper-parameters of DM. We will work first in a supervised setting using the following datasets of three supervised classification problems included in the NLDR Contest of ESANN 2015 [6]: 1) *Coil20*, a dataset with pictures of 20 objects (20 classes), with a total number of 1 440 patterns of dimension 16 384; 2) *MNIST*, a subset of the MNIST image bank of handwritten digits, with 3 000 patterns, 576 dimensions and 10 classes; and 3) *Wine*, a subset of the UCI Wine quality data set with 4 325 patterns of dimension 11 and two clusters (red and white), although we shall consider the 12 classes corresponding to the 6 quality levels of each cluster.

In this way we can validate $\mathrm{AUC}^{\mathrm{ln}}_{5:10}$ as a proper fitness criterion using as a ground truth measure of the embedding quality the accuracy of a $k$NN model. More precisely, all the datasets correspond here to classification problems in which the real class of each pattern is known, providing independent information to measure the suitability of an embedding in terms of the accuracy achievable on a concrete DM representation. Specifically, we will build a $k$-Nearest Neighbours

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/.

($k$NN) model over each DM, and use as measure the accuracy of this model in a straightforward leave-one-out configuration, i.e., for each pattern in the dataset we will predict its label using the nearest $k$ patterns (removing the target pattern itself). The $k$ parameter is selected as the maximizer of this accuracy over the original non-embedded data, and it turns to be $k = 2$ for *Coil20*, $k = 1$ for *MNIST* and $k = 25$ for *Wine*. With this alternative criterion, denoted by $\mathrm{Acc}_{k\mathrm{NN}}$, we can select an "optimal" set of parameters as the maximizer of this functional, i.e., $\mathcal{P}_{\mathrm{Acc}} = \arg\max_{\mathcal{P}} \mathrm{Acc}_{k\mathrm{NN}}(\mathrm{DM}_{\mathcal{P}})$. These "supervised" optimal parameters will be compared with the unsupervised ones $\mathcal{P}_{\mathrm{AUC}}$ that our proposal for neighbourhood-based selection yields.

Both criteria will be used to evaluate a series of DM models obtained under the following different parameter selections: 1) the dimension $d$ will be varied in the set $\{1, 2, 3, 4, 5\}$; 2) the percentile $p$ in $\{0.5, 1, 10, 50, 75, 99, 100, 150, 200\}$ (where 0.5 and 1 represent local models, 99 and 100 global models, and by 150 and 200 we just mean extremely large values of 1.5 and 2 times the maximum pattern distance); 3) the parameter $\alpha$ in $\{0, 0.5, 1\}$; and 4) the parameter $t$ in $\{0, 1, 2\}$. These parameters provide a total of 405 different models.

Figure 1 shows the dependence between $\mathrm{Acc}_{k\mathrm{NN}}$ and $\mathrm{AUC}^{\mathrm{ln}}_{5:10}$ for the embeddings obtained with the distinct parameters. The linear relation between both fitness criteria is clear and the models with a largest $\mathcal{P}_{\mathrm{Acc}}$ (computed using class knowledge) are extremely close to the ones with unsupervised, largest $\mathcal{P}_{\mathrm{AUC}}$.

This analysis is further supported by Table 1, which shows the parameters obtained using the different criteria and the correlation between the $\mathrm{Acc}_{k\mathrm{NN}}$ and $\mathrm{AUC}^{\mathrm{ln}}_{5:10}$ values. In fact, the parameters $\mathcal{P}_{\mathrm{Acc}}$ and $\mathcal{P}_{\mathrm{AUC}}$ only differ slightly and the correlation is above 95% for *MNIST* and *Coil20*. The correlation is a little smaller in *Wine* and the parameters $\mathcal{P}_{\mathrm{AUC}}$ and $\mathcal{P}_{\mathrm{Acc}}$ differ in both $\alpha$ and $p$. Nevertheless, the $\mathrm{DM}_{\mathcal{P}_{\mathrm{Acc}}}$ and $\mathrm{DM}_{\mathcal{P}_{\mathrm{AUC}}}$ embeddings are still almost identical, as shown in Fig. 2. Notice also that all the embeddings take $d = 5$ (the maximum allowed); indeed both $\mathrm{Acc}_{k\mathrm{NN}}$ and $\mathrm{AUC}^{\mathrm{ln}}_{5:10}$ grow with $d$, as a higher $d$ means retaining more information about the original data.

Therefore, $\mathrm{AUC}^{\mathrm{ln}}_{5:10}$ seems to be a good unsupervised measure to evaluate the quality of an embedding and, thus, to select optimal DM hyper-parameters. We have checked this using the Swiss Roll (*SwiRol*) dataset of [6], which contains a rolled 2-dimensional manifold embedded into a 3-dimensional space, with 1 500 patterns (see Fig. 3)[1]. In particular, we are interested in whether the optimal $\mathcal{P}_{\mathrm{AUC}}$ parameters provide a good embedding when $\mathrm{Acc}_{k\mathrm{NN}}$ is not available. Figure 3 includes the original *SwiRol* dataset and the 1- and 2-dimensional embeddings obtained using $\mathcal{P}_{\mathrm{AUC}}$ (parameters are in Table 1) where we fix $d$ to 1 in the 1-dimensional case and allowed to vary in $\{1, 2\}$ in the other (larger values of $d$ do not make sense, as the original data is 3-dimensional). The colour orderings show that both embeddings find the underlying manifold structure and confirm $\mathrm{AUC}^{\mathrm{ln}}_{5:10}$ as an effective unsupervised quality measure.

[1] Although the NLDR Contest proposes also another unsupervised dataset, namely a 3-dimensional sphere, it is not clear which should be the 2-dimensional representation of the manifold, as there exists no homomorphism to unroll the sphere; this is why it is not included.

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/.

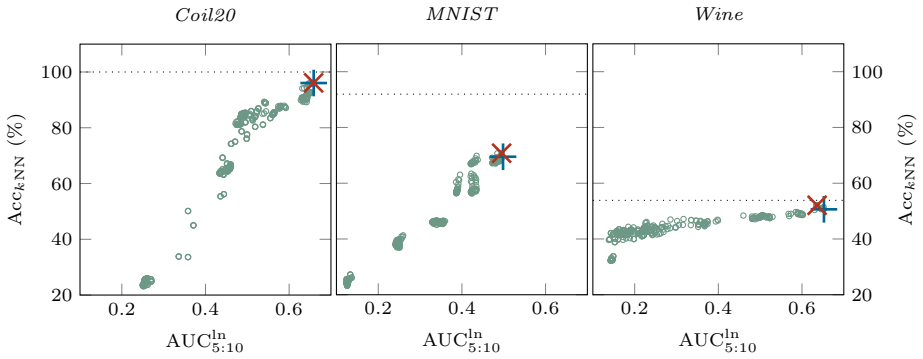


Fig. 1: Relation between $\text{Acc}_{k\text{NN}}$ and $\text{AUC}^{\text{ln}}_{5:10}$: the different models are depicted with ○; $\text{DM}_{\mathcal{P}_{\text{Acc}}}$ and $\text{DM}_{\mathcal{P}_{\text{AUC}}}$ are marked by × and +, respectively; and the dotted black line corresponds to $\text{Acc}_{k\text{NN}}$ over the original data.

| Dataset | Criterion | $d$ | $p$(%) | $\alpha$ | $t$ | $\text{Acc}_{k\text{NN}}$ | $\text{AUC}^{\text{ln}}_{5:10}$ | Corr. (%) |
|---|---|---|---|---|---|---|---|---|
| *Coil20* | $\text{AUC}^{\text{ln}}_{5:10}$ | 5 | 200.0 | 1.0 | 0 | 96.04% | 0.659 | 95.48% |
| | $\text{Acc}_{k\text{NN}}$ | 5 | 200.0 | 0.0 | 0 | 96.11% | 0.659 | |
| *MNIST* | $\text{AUC}^{\text{ln}}_{5:10}$ | 5 | 1.0 | 0.0 | 0 | 69.50% | 0.498 | 97.76% |
| | $\text{Acc}_{k\text{NN}}$ | 5 | 0.5 | 0.0 | 0 | 70.77% | 0.495 | |
| *Wine* | $\text{AUC}^{\text{ln}}_{5:10}$ | 5 | 99.0 | 0.5 | 0 | 50.64% | 0.653 | 85.73% |
| | $\text{Acc}_{k\text{NN}}$ | 5 | 100.0 | 0.0 | 0 | 52.09% | 0.636 | |
| *SwiRol* | $\text{AUC}^{\text{ln}}_{5:10}$ | 1 | 0.5 | 0.0 | 0 | × | 0.374 | × |
| | $\text{AUC}^{\text{ln}}_{5:10}$ | 2 | 0.5 | 0.5 | 2 | × | 0.602 | |

Table 1: Parameters obtained (with corresponding $\text{Acc}_{k\text{NN}}$ and $\text{AUC}^{\text{ln}}_{5:10}$) and correlation between both fitness criteria.

## 5 Conclusions

We have introduced a methodology to select the parameters of Diffusion Maps (DM) that provide a good embedding of the data. It is based on maximizing the measure proposed in [2], i.e., the area under a curve which represents the preservation of the neighbourhood between the original data and the embedding. While a full $x$-range is considered in [2], we integrate this area only for neighbourhoods of sizes between 5% and 10% of sample size, resulting in the criterion $\text{AUC}^{\text{ln}}_{5:10}$ that aims at a good balance between the sharpness and the locality of the embeddings. We have shown that in this unsupervised and parameter-free approach, $\text{AUC}^{\text{ln}}_{5:10}$ is highly correlated with the accuracy of a $k$NN classification model, and that the parameters that maximize $\text{AUC}^{\text{ln}}_{5:10}$ provide embeddings able to find the underlying structure of the data, for example correctly unrolling the classical Swiss Roll.

As further work, we intend to exploit $\text{AUC}^{\text{ln}}_{5:10}$ in concrete DM applications and also to study the relation between $\text{AUC}^{\text{ln}}_{5:10}$ and the embedding dimension

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/.

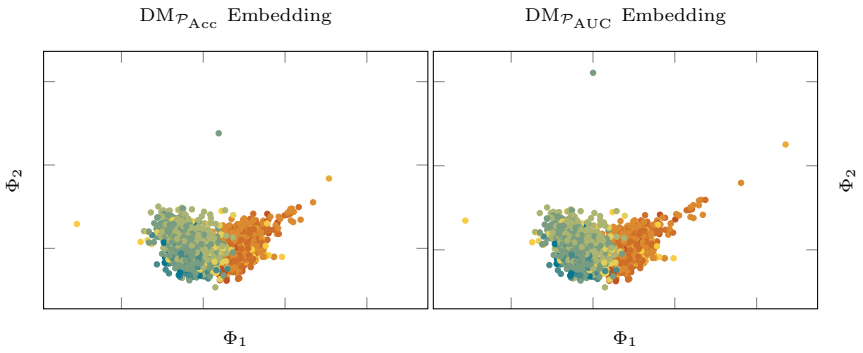


Fig. 2: Embeddings for the *Wine* dataset using $\mathcal{P}_{\mathrm{Acc}}$ and $\mathcal{P}_{\mathrm{AUC}}$.

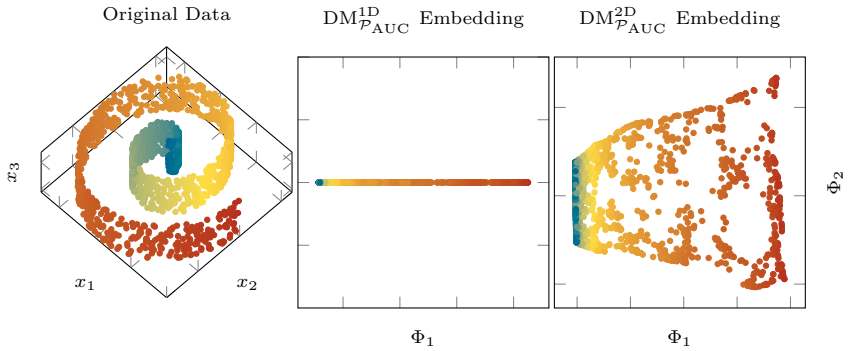


Fig. 3: Original data and embeddings for the *SwiRol* dataset.

$d$. It is to be expected that $\mathrm{AUC}^{\mathrm{ln}}_{5:10}$ increases with $d$, but it is interesting to see if it saturates once the dimension of the underlying manifold is achieved, so that any additional increment in $d$ adds almost no information (in terms of $\mathrm{AUC}^{\mathrm{ln}}_{5:10}$).

## References

[1] R.R. Coifman and S. Lafon. Diffusion Maps. *ACHA Journal*, 21(1):5–30, 2006.

[2] J. A Lee and M. Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7):1431–1443, 2009.

[3] U. Luxburg. A tutorial on spectral clustering. *Statistics and Comp.*, 17(4):395–416, 2007.

[4] Á. Fernández. *Diffusion Methods and Applications*. PhD thesis, Escuela Politénica Superior, Universidad Autónoma de Madrid, Madrid, Spain, July 2014. Available at `http://arantxa.ii.uam.es/~gaa/theses.html`.

[5] J.A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen. Type 1 and 2 mixtures of kullback-leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 112:92–108, July 2013.

[6] J.A. Lee and K. Bunte. ESANN NLDRcontest. `https://sites.google.com/site/nldrcontest/home`, 2015.