Diffusion Maps and Local Models for Wind Power Prediction

Ángela Fernández, Carlos M. Alaíz, Ana M. González,
Julia Díaz and José R. Dorronsoro

Departamento de Ingeniería Informática and Instituto de Ingeniería del Conocimiento
Universidad Autónoma de Madrid, 28049, Madrid, Spain
{a.fernandez,carlos.alaiz,ana.marcos,julia.diaz,jose.dorronsoro}@uam.es

Abstract. In this work we will apply Diffusion Maps (DM), a recent technique for dimensionality reduction and clustering, to build local models for wind energy forecasting. We will compare ridge regression models for K–means clusters obtained over DM features, against the models obtained for clusters constructed over the original meteorological data or principal components, and also against a global model. We will see that a combination of the DM model for the low wind power region and the global model elsewhere outperforms other options.

1 Introduction

Local models are a natural and attractive option when trying to approach processes with high variance data or whose underlying phenomena are known to possibly correspond to quite different settings. However, to identify the appropriate local feature areas may be quite difficult, particularly for high dimensional data that do not lend themselves easily to such a task. Unsupervised clustering methods, such as K–means, appear as an attractive option. However, clustering is often more an art than a technology and while many methods have been proposed, simple approaches are usually followed in practice, in particular K–means which is applied assuming an Euclidean distance in the feature space. Besides fixing the number K of clusters, an adequate sampling is also an important issue when working with high dimensional data as samples are then bound to be very sparse. Moreover, the features to be used may not be homogeneous, something probably better to be handled outside the chosen clustering procedure.

In this paper we will address the above issues in the context of wind energy prediction. Wind power clearly presents wide, fast changing fluctuations, certainly at the individual farm level but also when the production of much larger areas is considered. This is the case of Spain, the world's fourth biggest producer of wind power, where wind is currently the third source of electricity. The well known, sigmoid–like structure of wind turbine power curves clearly shows different regimes at low, medium and high wind speeds. Compounded with this are wind speed frequencies, that follow a Weibull distribution, that is, a stretched exponential with low wind having large frequencies. While the above does not directly apply when a wide area is considered, different regimes also appear. Wind energy forecasting for large areas also implies high dimensional features as the predictive variables, that are the outputs of numerical weather prediction (NWP) models such as the ECMWF or GFS ones given for large grids that cover the areas under study. Global models may find it difficult to handle these regimes and local models are natural alternatives [1, 9].

This high dimension suggests to precede clustering by some dimensionality reduction (DR) technique, preferably one that is likely to yield an Euclidean metric for the new features. Diffusion Maps (DM) [5], a novel spectral technique for DR, is particularly suited to these requirements. In fact, there is a natural diffusion metric in the original feature space that corresponds with Euclidean metric in the embedded space. This means that clustering methods that rely on Euclidean metrics, particularly K–means, should work well on the new features. DM also allows to control to some extent the effects of the underlying data distribution and, moreover, it allows to work with heterogeneous variables. In other words, DM can be a powerful tool for finding informative clusters in high dimensional, heterogeneous data.

Of course, DM is not the only option. Straight K–means clustering can certainly be used. Moreover, NWP variables for a large area usually show high correlation among different grid points. This may suggest that variance–based DR methods such as Principal Component Analysis (PCA) may be a useful alternative. We shall consider these three options here in order to, first, identify local clusters and then to construct local models to be compared against a global one. Many paradigms can be considered for model building but here we will concentrate on the simplest alternative, ridge regression, i.e., regularized linear least squares, certainly not the strongest possible method but a good option to measure the usefulness of local methods against a global one.

The paper is organized as follows. We will review in Sect. 2 DM from a general point of view, as well as its use over heterogeneous data. In Sect. 3 we will consider K–means on DM, PCA and the original features, we will compare local ridge regression models on these clusters, we will discuss their effectiveness and we will conclude on how to combine local and global models for better predictors. Section 4 ends this paper with a brief discussion and conclusions.

2 Diffusion Maps Review

The key assumption in Diffusion Maps (DM) is that the data to be studied lie in a low–dimensional manifold whose geometry can be described through a Markov chain diffusion metric. To capture this intrinsic geometry, the first step is to build a connectivity graph using the sample points $\mathcal{S} = \{x_1, \dots, x_n\}$ as graph nodes and defining a symmetric weight matrix $W_{ij} = w(x_i, x_j)$. The most common way to build this matrix is to use the Gaussian Kernel and define $w(x_i, x_j) = \exp\left(-||x_i - x_j||^2/\sigma^2\right)$, where σ determines the radius of the neighborhoods centered at individual sample points. We start with this matrix towards defining a Markov chain over this graph. We first choose a parameter $\alpha \in [0, 1]$

Algorithm 1 Diffusion Maps Algorithm.

```
Input: $\mathcal{S} = \{x_1, \dots, x_n\}$, the original data set.
Output: $\{\Psi^t(x_1), \dots, \Psi^t(x_n)\}$, the embedded data set.
Construct $G = (\mathcal{S}, W)$ where $W$ is a symmetric distance matrix, $W_{ij} = w(x_i, x_j)$.
Define the initial density function as $q(x_i) = \sum_{j=1}^{n} w(x_i, x_j)$.
Normalize the weights by the density, $w^{(\alpha)}(x, y) = \frac{w(x,y)}{q(x)^\alpha q(y)^\alpha}$.
Let $g^{(\alpha)}(x_i) = \sum_{j=1}^{n} w^{(\alpha)}(x_i, x_j)$ be the graph degree. Define the transition probability $P_{ij}^t = p^{(\alpha),t}(x_i, x_j) = \frac{w^{(\alpha)}(x_i,x_j)}{g^{(\alpha)}(x_i)}$.
Obtain the eigenvalues $\{\lambda_r^t\}_{r \geqslant 0}$ and eigenfunctions $\{\psi_r\}_{r \geqslant 0}$ of $P^t$.
Compute the embedding dimension using a threshold $d = \max\{l : |\lambda_l^t| > \delta|\lambda_1^t|\}$.
Formulate Diffusion Map, $\Psi^t(x) = (\lambda_1^t \psi_1^\tau(x), \dots, \lambda_d^t \psi_d^\tau(x))^\tau$.
```

that is used to control the combined effects of manifold geometry and sample distribution and define $w^{(\alpha)}(x_i, x_j) = \frac{w(x_i,x_j)}{q(x_i)^\alpha q(x_j)^\alpha}$, where $q(x_i) = \sum_{j=1}^{n} w(x_i, x_j)$ is the degree at the i–th node of the W matrix. We define now the new α–degree at x_i as $g^{(\alpha)}(x_i) = \sum_{j=1}^{n} w^{(\alpha)}(x_i, x_j)$ and arrive at the transition probability $p^{(\alpha)}(x_i, x_j) = \frac{w^{(\alpha)}(x_i,x_j)}{g^{(\alpha)}(x_i)}$. Notice that when $\alpha = 0$, we are essentially defining the weight matrix typically used in spectral dimensionality reduction [3]. In this case, the infinitesimal generator L_0 of the resulting Markov chain acts on an f as $L_0(f) = \frac{\Delta(fq)}{q} - \frac{\Delta(q)}{q} f$ [5], with Δ the manifold's Laplace–Beltrami operator. However, when $\alpha = 1$ the infinitesimal generator L_1 verifies $L_1(f) = \Delta f$ and it is not influenced by the underlying density q (this will not be the case for $\alpha = 0$ unless q is uniform). We will consider here the case $\alpha = 1$ and write just $p^t(x_i, x_j)$ if a t–step Markov chain is used. We will denote by P^t the matrix of transition probabilities in t steps ($P_{i,j}^t = p^t(x_i, x_j)$).

Let $\lambda_i, \psi_i(x)$, $i = 0, \dots, n-1$, be the eigenvalues and eigenvectors of P, where we assume $1 = \lambda_0 \geqslant \dots \geqslant \lambda_{n-1}$; P^t has then eigenvalues λ_i^t and the same eigenvectors $\psi_i(x)$. To select for a given t the embedding dimension $d = d(t)$ we may fix a precision δ and choose $d = \max\{l : |\lambda_l^t| > \delta|\lambda_1^t|\}$. The embedding projection is then $\Psi^t(x) = (\lambda_1^t \psi_1^\tau(x), \dots, \lambda_d^t \psi_d^\tau(x))^\tau$, with τ the transpose operator. The previous steps are summarized in Algorithm 1.

The Euclidean distance $\|\Psi^t(x) - \Psi^t(z)\|^2 = \sum_j \lambda_j^{2t}(\psi_j(x) - \psi_j(z))^2$ in the embedding coincides with the diffusion distance $D_t^2(x, z) = ||p^t(x, \cdot) - p^t(z, \cdot)||_{L^2(\frac{1}{\phi_0})}^2$, where ϕ_0 is the stationary distribution of the P–Markov process. In other words, if the diffusion distance D_t approximates the manifold metric, we get the original data embedded in a lower dimension space for which Euclidean distance captures the original local geometry, something very convenient if we want to apply K–means. Once we have obtained K clusters $\{C_1, \dots, C_k\}$ over the embedded features, they can be projected back into clusters $\{A_1, \dots, A_k\}$ in the original space $\mathcal{S}$ defined as $A_i = \{x_j | \Psi_t(x_j) \in C_i\}$.

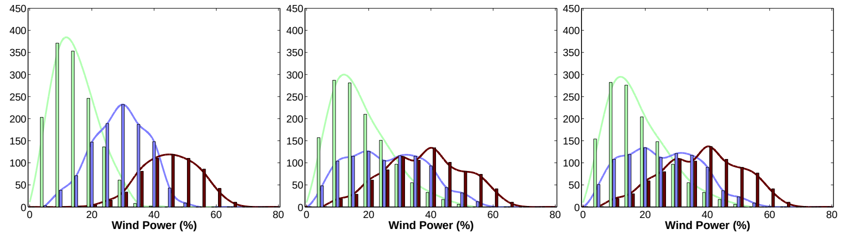
A limitation of the above scheme is that it implicitly assumes the attributes to be homogeneous; however, real–life datasets are frequently heterogeneous, something that often cannot be handled just by normalizing the data. In [10] a method is proposed to adapt DM to work with heterogeneous features just by dealing separately with groups of attributes that are deemed to be homogeneous. More precisely, assume that we have M such groups; we then split each pattern x_i into M new, lower dimensional ones x_i^m and build the corresponding sample sets $\{\mathcal{S}_m\}_{m=1}^M$. We now apply DM as described before to each $\mathcal{S}_m$ obtaining M embeddings $\{\Psi_m\}_{m=1}^M$ that capture the geometry associated to each feature subset. Now, these Ψ_m are given by eigenvalue–eigenvector products, with the eigenvectors being comparable across the embeddings since they have unit norm. We can make eigenvalues also comparable if we re–scale them as $\lambda_{m,i}^{\text{new}} = \frac{\lambda_{m,i}}{\sum_j \lambda_{m,j}}$. Thus the union of the normalized features gives a set of homogeneous features that still represent the intrinsic geometry of our original data and we can simply apply DM again to this new dataset to get the final lower dimensional embedding.

In summary, DM makes possible low dimensional embeddings of heterogeneous data while transforming the original space metric into an Euclidean one. However, they require proper choices for the parameters σ and δ. Moreover, a main drawback (as it also happens in spectral DR) is the difficulty to apply the computed DM projection to new, unseen patterns. There are several proposals for this such as Nyström formulae [4] or Laplacian Pyramids [10], but this is still an area where further work is needed.

3 Experiments

In this section we will apply K–means clustering based on DM to build local models for predicting the wind energy production in Spain and compare it with the results of K–means applied to either the original full dimensional data or to PCA lower dimensional features. Once clusters are defined, we will use Ridge Regression (RR) [7] for model building. Recall that RR adds a ℓ_2 regularization term to an Ordinary Least Square (OLS) regression, so the optimization problem becomes $\min_w \|X\mathbf{w} - \mathbf{y}\|_2^2 + \gamma \|w\|_2^2$. This prevents over–fitting in plain OLS but requires a procedure to compute the penalty term γ. While stronger models could be considered [2, 8], our primary interest here is whether DM–based local models improve on either other local models or global ones. If so, stronger models should also benefit from this, although we will not consider them in this work.

We will use as inputs NWPs from the European Centre for Medium–Range Weather Forecasts (ECMWF) [6] and consider five surface variables: wind speed (V), its horizontal and vertical components (V_x and V_y), pressure (p) and temperature (T), which we normalize component–wise to zero mean and unit variance. These variables are available over a rectangular 522–point, $0.5°$ resolution grid that covers the Iberian peninsula. Pattern dimension is thus $2,610 = 5 \times 522$. Two year data will be considered, the first one for training purposes and the second for testing. Since eight forecasts are given daily, training sample size is thus $2,910 = 365 \times 8$, close to feature dimension and hence making regularization mandatory.



(a) $\mathrm{LM_{DM}}$ clusters. (b) $\mathrm{LM_{Or}}$ clusters. (c) $\mathrm{LM_{PC}}$ clusters.

Fig. 1: Wind power histograms for the clusters obtained using the 3 approaches.

We will use NWPs and wind energy production data y for cluster definition. Wind power is obviously unknown for the test dataset so we will use as a proxy the wind power forecast of a global model. We already mentioned the difficulties associated with the application of DM to test patterns. We will skip on them by building the DM features and clusters, as well as the plain and PCA clusters, over the full two year dataset. This confers some advantage to the local models over the global one, partially compensated by the global model influencing cluster definition. In any case, and as mentioned before, the computation of DM features for new patterns is an area of active research.

We consider wind power production and the NWP variables as heterogeneous and build first DM features separately on the y, V, V_x, V_y, T and p variables. In all of them we define the graph's weight matrix using a Gaussian Kernel with bandwidth σ equal to the dataset diameter. We arrived at this value heuristically after visually analyzing the structure of the resulting embeddings. We also work with $t = 1$, i.e., considering the one–step diffusion distance on the original feature space and final embedding dimension was obtained using a $\delta = 0.1$ precision parameter. Embedding dimensions for the above variables were 1, 6, 3, 5, 1 and 1 respectively and the final dimension for the DM embedding is 5. Therefore, we also considered a $2,610$ to 5 dimension reduction for PCA. Finally, the choice of K is always difficult. We will consider 3 clusters that hopefully capture high, medium and low ranges of wind power. While initial centroids are randomly chosen in K–means, we found that the DM parameters used lead to very stable cluster structures that are essentially independent of centroid initialization. Figure 1 gives the cluster histograms of the local wind power distributions for each approach. As it can be seen, the 3 DM clusters offer a more clear–cut structure while the other two methods seem to differentiate less between wind energy regimes.

Once DM, PCA and original feature clusters are defined, we build a global model and also three local RR models, one per cluster, that we denote as GM, $\mathrm{LM_{DM}}$, $\mathrm{LM_{PC}}$ and $\mathrm{LM_{Or}}$ respectively. Prior to model building we select the optimal regularization parameters for all the RR models by grid search for γ in the interval $[10^{-2}, 10^4]$, with a logarithmic step of 0.1 and using as validation

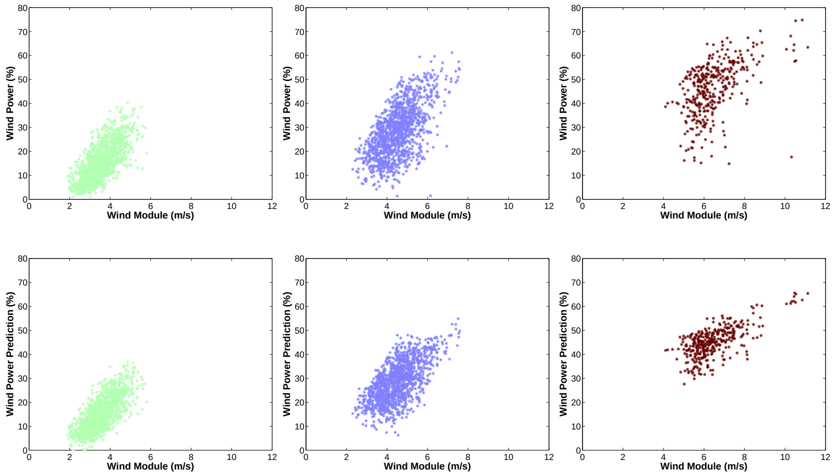


Fig. 2: Average wind versus actual power (top) and average wind versus predicted power (bottom) for the 3 clusters.

set the last 20% patterns of the first year data clusters. As usually done in wind energy, we measure model performance by the mean absolute error (MAE) and the relative mean absolute error (RMAE). The MAE is defined as the mean of the differences between the predictions and the real values. The RMAE computes the mean for the ratio of the absolute errors over actual wind power. Table 1a contains local model errors per cluster as well as the cluster errors of the global model. As we can see, the local models beat the global one in the first, low wind power cluster C_1 but GM beats them in C_2 and particularly in the high wind power cluster C_3. A reason for this can be seem in Fig. 2 that depicts for the 3 LM_{DM} clusters the relationships between average wind and power (top) and between average wind and predicted power (bottom). Cluster C_3 has the fewest number of points but presents several marked outliers; these two facts clearly penalize the local C_3 models. Table 1a also gives values for the standard deviations of MAE and RMAE, although they are rather conservative (assuming independence for these errors would lead to divide the values given by the square root of sample size and, hence, much smaller values).

These facts suggest to build predictors combining a local model on the C_1 cluster and the global one on the other two. Table 1b contains the MAE and RMAE errors of the individual GM, LM_{DM}, LM_{Or} and LM_{PC}, and of the combined models $CM_{DM;G}$, $CM_{Or;G}$ and $CM_{PC;G}$. It shows that there is a clear advantage of the combined models over the global one and that the gain is largest for the $CM_{DM;G}$ model. While modest at first sight (a MAE of 3.37% against 3.48% for GM), such gains may have a large economic impact, as wind energy represents about 16% of Spain's electricity demand.

Table 1: (a) Errors per cluster (top). (b) Global errors (bottom).

	MAE		**RMAE**		**stdAE**		**stdRAE**	
	LM$_{DM}$	**GM**	**LM$_{DM}$**	**GM**	**LM$_{DM}$**	**GM**	**LM$_{DM}$**	**GM**
C$_1$	**2.29**	2.53	**19.54**	22.98	1.88	1.94	31.49	40.66
C$_2$	4.01	**3.83**	19.93	**18.34**	3.20	3.16	72.67	65.63
C$_3$	5.94	**5.73**	15.67	**13.93**	4.76	4.94	23.13	21.20

	MAE		**RMAE**		**stdAE**		**stdRAE**	
	LM$_{Or}$	**GM**	**LM$_{Or}$**	**GM**	**LM$_{Or}$**	**GM**	**LM$_{Or}$**	**GM**
C$_1$	**2.52**	2.69	**18.76**	20.10	1.95	2.14	22.94	23.78
C$_2$	3.72	**3.65**	20.52	**19.37**	3.11	3.15	38.75	40.73
C$_3$	5.40	**4.77**	24.14	**20.36**	4.39	4.26	94.33	93.43

	MAE		**RMAE**		**stdAE**		**stdRAE**	
	LM$_{PC}$	**GM**	**LM$_{PC}$**	**GM**	**LM$_{PC}$**	**GM**	**LM$_{PC}$**	**GM**
C$_1$	**2.53**	2.66	**18.99**	20.01	1.95	2.10	23.24	23.84
C$_2$	3.68	**3.64**	20.39	**19.62**	3.08	3.12	37.37	40.48
C$_3$	5.21	**4.78**	22.91	**20.11**	4.31	4.28	93.32	92.68

	GM	**LM$_{DM}$**	**LM$_{Or}$**	**LM$_{PC}$**	**CM$_{DM;G}$**	**CM$_{Or;G}$**	**CM$_{PC;G}$**
MAE	3.48	3.47	3.56	3.53	**3.37**	3.40	3.42
RMAE	19.89	19.24	20.52	20.34	**18.35**	19.32	19.46
stdAE	3.16	3.19	3.21	3.17	2.80	2.87	2.88
stdRAE	51.60	52.80	51.25	50.87	44.98	44.12	44.37

4 Conclusions

Local models are obviously useful in many applied problems, being wind energy forecasting a clear example. The main obstacle for their construction usually is how to define the local regions on which models will be built. A natural option is K–means clustering that requires to choose an adequate metric, something always difficult and more so when we also have to deal with the high dimensional, heterogeneous features that arise in wide area wind energy forecasting. In this work we have applied to this task Diffusion Maps (DM), a novel dimensionality reduction technique that lends itself naturally to work with heterogeneous data and that has the very important property that Euclidean metric in the projected space is naturally related to a diffusion distance on the original features. This distance is in turn related to a Markov process whose infinitesimal generator is just the Laplace–Beltrami operator of the underlying manifold. We can expect that the Euclidean metric in the reduced features captures the original space metric and, thus, standard K–means on the embedding results in meaningful clusters for the original features.

We have compared this approach with clusters obtained by straight Euclidean K–means on the full features and on PCA features with the same dimension as the DM ones, building local ridge regression models that, in turn, are compared with a global one. The local models beat the global one over a low wind power cluster, with DM a clear winner, but the global model performs better on the other medium and high wind power clusters. This suggests to define a mixed model, using the DM local model for the low wind power cluster and the global one for the other two. This model outperforms the others.

We can conclude that DM dimensionality reduction and clustering is an effective tool for local model building, although further work is needed. In fact, DM features are derived after an spectral analysis of the sample distance matrix. As it is also the case with spectral dimensionality reduction and clustering, this makes costly to assign new, unseen patterns to already defined clusters. Tools to alleviate this are Nyström formulae or Laplacian Pyramids. We are currently doing research on this for wind energy and other applied problems.

Acknowledgement.

With partial support from grant TIN2010-21575-C02-01 of Spain's Ministerio de Economía y Competitividad and the UAM–ADIC Chair for Machine Learning in Modelling and Prediction. The first author is also supported by an FPI-UAM grant and kindly thanks the Applied Mathematics Department of Yale University for receiving her during a visit. The second author is supported by the FPU-MEC grant AP2008-00167. We also thank Red Eléctrica de España, Spain's TSO, for providing historic wind energy data.

References

1. Alaíz, C., Barbero, A., Fernández, A., Dorronsoro, J.: High wind and energy specific models for global production forecast. In: Proceedings of the European Wind Energy Conference and Exhibition (EWEC–09). Marseille, France (March 2009)
2. Barbero, A., López, J., Dorronsoro, J.: Kernel methods for wide area wind power forecasting. In: Proceedings of the European Wind Energy Conference and Exhibition (EWEC–08). Brussels, Belgium (April 2008)
3. Belkin, M., Nyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15(6), 1373–1396 (2003)
4. Bengio, Y., Delalleau, O., Roux, N.L., Paiement, J., Vincent, P., Ouimet, M.: Learning eigenfunctions links spectral embedding and kernel pca. Neural Computation 16(10), 2197–2219 (2004)
5. Coifman, R., Lafon, S.: Diffusion maps. Applied and Computational Harmonic Analysis 21(1), 5–30 (2006)
6. European center for medium–range weather forecasts (2005), http://www.ecmwf.int/
7. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(12), 55–67 (1970)
8. Monteiro, C., Bessa, R., Miranda, V., Botterud, A., Wang, J., Conzelmann, G.: Wind power forecasting: State–of–the–art 2009. Tech. rep., INESC Porto and Argonne National Laboratory (2009)
9. Pinson, P., Nielsen, H., Madsen, H., Nielsen, T.: Local linear regression with adaptive orthogonal fitting for the wind power application. Statistics and Computing 18(1), 59–71 (2009)
10. Rabin, N., Coifman, R.: Heterogeneous datasets representation and learning using diffusion maps and laplacian pyramids. In: Proceedings of the 12th SIAM International Conference on Data Mining (SDM12). Anaheim, California, USA (April 2012)