Clusterin' Clusters

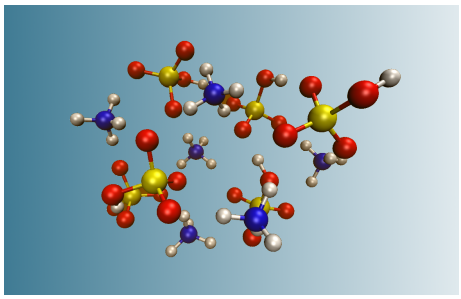
Clusterin'clusters

Vitus Besel, Matias Jääskeläinen, Ilaria Pia

October 22, 2019

1 Introduction

Atmospheric science has become increasingly popular especially in the face of climate change and a growing environmental awareness in society. In order to understand the highly complex processes happening in the atmosphere it is not only necessary to conduct fieldwork and measurements, it is also important to deliver the theoretical framework in order to perform simulations complementing the fieldwork or making large scale predictions.
One subfield of atmospheric sciences deals with New Particle Formation, which is the formation of particles from single gaseous molecules in the atmosphere, which then can grow further into cloud condensation nuclei. As aerosol- and aerosol-cloud interactions still contribute to the highest uncertainties within current climate models (`https://www.ipcc.ch/report/ar5/wg1/anthropogenic-and-natural-radiative-forcing/`) this subfield of New Particle Formation is a especially interesting research object.



1.1 Configurational Sampling

In order to understand how single gaseous molecules form first molecular cluster and then bigger particles, it is necessary to know the structure of these molecular clusters. However, with an increasing number of molecules inside of an atmospheric cluster, the number of possible conformers rises rapdily. More simply put there are e.g. for a molecular cluster made of six sulfuric acid and six ammonia molecules tens of thousands possible ways to form, differing in distances, angles or protonation states of single molecules. Configurational Sampling is a method to obtain the relevant local minima structures - *i.e.* the energetically most favorable ones - by a combination of computational chemistry methods and data analysis. This is needed because mainly these minima structures are present in the atmosphere and therefore relevant for atmospheric simulations.

1.2 ABCluster and GNF-xTB

Our code comes into action after following programs have provided the set-up for the Configurational Sampling: ABCluster utilizes the artificial bee colony algorithm (2) a generic algorithm which takes the structures of single gaseous molecules given by the user and combines them to molecular clusters and optimizes these structures on a Molecular Mechanics level of theory. It does this in a way which samples the whole Configurational Space and produces an amount of local minima defined by the user (typically 2000 - 100 000). Further the semi-empirical method GNF-xTB is used to once again optimize these structures on a better level of theory. We assume that these steps are conducted by the user within the Jammy Key for Configurational Sampling (JKCS))

2 Data cleaning

After ABCluster and GNF-xTB within JKCS the user is left with thousands of output (.log) files and structure (.xyz) files in a certain directory structure. Clusterin' Clusters contains two bash-script, which deal with this data.
CollectAndSortFiles.sh creates a directory with a user specified name and resolves the directory structure from XTB by copying the .log files and .xzy files enumerated into respective directories.
DataExtractor.sh produces the Data_Collection.csv file, which contains the file number, energies and dipoles from the .log files and the paths to the .log and .xzy files.
Both of these script are set-up to be called from the Scripts directory and will work in the specified directory located on level higher than Scripts.

3 Statistical and machine-learning methods

3.1 Data exploration

As a basic data exploration we return a plot of the distributions of the variables energy and dipole and a plot of the correlation among all variables of the dataset as a measure of the linear dependency between them.

3.2 Principal Components Analysis

The Principal components analysis (PCA) is a tool used to reduce the dimensions of the data that tries to preserve as much information as possible. These two goals in PCA are pursued by means of a transformation of the original variables into new variables, called Principal Components (PCs) that consists of a linear combination of the original variables. PCs are uncorrelated and arranged in order of decreasing variance, so that the first PCs account for most of the variation in the sample. Assuming we want to reduce the number of our original p variables: $X_1, \ldots, X_p$ to $k < p$ variables $Z_1, \ldots, Z_k$, denoted Σ the covariance matrix of $X_1, \ldots, X_p$, the PCA method can be formalized as follows:

- 1st PC: determine the coefficients of the linear combination

$$Z_{1j} = a_1^T X = \sum_{i=0}^{p} a_{1i} X_i$$

that maximize $Var(Z_1) = a_1^T \Sigma a_1$ under the constraint $a_1^T a_1 = 1$

- 2nd PC: determine coefficients of the linear combination

$$Z_{2j} = a_2^T X = \sum_{i=0}^{p} a_{2i} X_i$$

that maximize $Var(Z_2) = a_2^T \Sigma a_2$ under the constraint $a_2^T a_2 = 1$ and $Cov(Z_1, Z_2) = 0$

- proceed in a similar fashion for all other components...

The final output is a set of p uncorrelated variables with decreasing variance: $Z_1, \ldots, Z_p$ such that $Var(Z_1) > Var(Z_2) > \cdots > Var(Z_p)$ and $Cov(Z_j, Z_k) = 0$ for $j \neq k$.

We apply PCA to our coordinates variables, and select new variables, that explain at least 80% of the variability of our data. The variables Dipole and Energy are kept unchanged so that we won't lose their intrinsic meaning.

3.3 Clustering with k-Means

In order to reduce the variety of the observations we cluster the data, by using a k-means algorithm, that is a non-hierarchical method of clustering, *i.e.* the number k of groups is assumed to be fixed. The algorithm, introduced by MacQueen in 1967, consists of assigning each datapoint to the cluster whose centroid (*i.e.* vector of means) is the closest one. The metric used to measure the distance among groups is typically the Euclidean one.
It can be represented as follow:

1. Initial partition into K clusters (possibly randomly generated)
2. For each of the K clusters, compute the cluster centroid
3. Assign each observation to the cluster whose centroid is closest
4. Recompute centroids for all clusters
5. Repeat 3. - 4. until reaching a maximum number of iterations or when it is not possible to redistribute observations

As stated above a crucial parameter is the number of groups k. If such parameter is not clearly recoverable from the data itself, a good criteria to chose the optimal one, is to find a balance between low within-cluster variation and number of groups. A common method is to select the number of clusters k that maximize the Calinksi-Harabasz (CH) index:

$$CH(k) = \frac{B(k)}{W(k)} \frac{n-k}{k-1}$$

where n is the number data points, k the number of clusters, $W(k)$ the within cluster variation and $B(k)$ the between cluster variation.
As the resulting clusters depend strongly on the choice of the starting centroids a common practice is to repeat the algorithm several times with different starting centroids, randomly generated.

We apply the k-mean algorithm to our data, selecting a number of clusters $k = 23$ as this is the number of possible permutations of our chemical cluster: there are 23 different ways of adding 5 water molecules, 1 ammonia and 1 sulfuric acid in a way considering all possible proton transfers which result in a neutral cluster. The possibility to select k according to the CH index is also given.

3.4 Visualization with t-SNE

To visualize the multidimensional dataset we use the t-Distributed Stochastic Neighbor Embedding (t-SNE), an unsupervised, non-linear technique of dimensionality reduction, introduced by Laurens van der Maatens and Geoffrey Hinton in 2008. The t-SNE aims to preserve the similarity between the original d-dimensional points and the 2-dimensional points returned as an output. As a measure of similarity we take conditional probability under specific kernels.
The algorithm can be divided in three steps.
First we convert the high-dimensional Euclidean distances between d-dimensional datapoints $\{x_i\}_{i=1,\dots,n}$ into conditional probabilities $p_{i|j}$ that represent the probability that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a properly scaled Gaussian centered at x_i.
In the same way, we convert the Euclidean distances between 2-dimensional points $\{y_i\}_{i=1,\dots,n}$ into conditional probabilities $q_{i|j}$ that give the probability that y_i would pick y_j as its neighbor if neighbors were picked in proportion to their probability density under a Chaucy distribution centered at y_i.
Finally, to measure the difference between the probability distributions of the d-dimensional and the 2-dimensional points we use the Kullback-Liebler divergence (KL) :

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$

T-SNE minimizes the sum of Kullback-Leibler divergences over all datapoints using a gradient descent method.

A plot of the clustered datapoints in the 2-dimensional t-SNE space is given as a final output together with the clustered dataset.

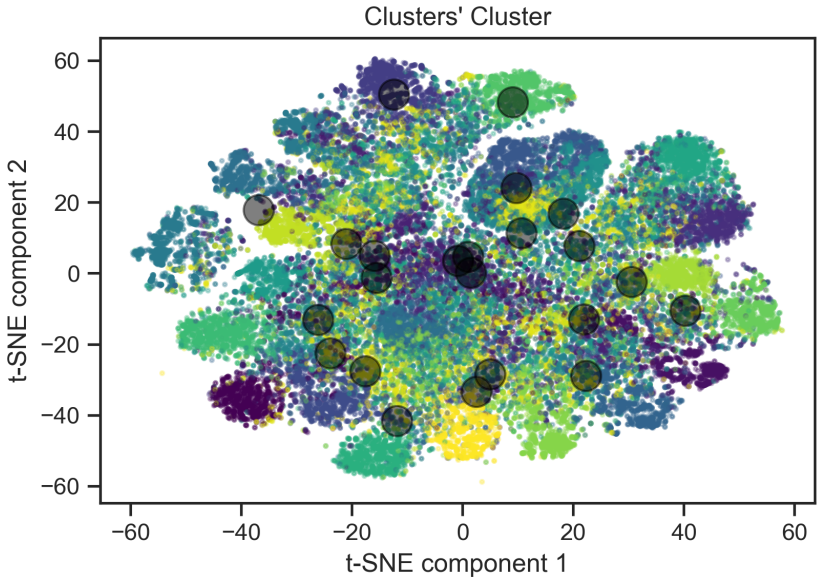


Figure 1: t-SNE representation of the data. The colours correspond to different statistical clusters. The darkened circles are the centroids.

4 Conclusion

Our software can cluster the clusters taken from the raw data. It offers a tool for refining the raw data to something that the data analysis and machine learning can be applied to. We used Clusterin'Clusters on an example dataset consisting of 46 000 molecular clusters made of 1 sulfuric acid, 1 ammonia and 5 water molecules. We then applied k-means clustering algorithm to the PCs dataset. The resulting clusters are not really well separated, because cartesian coordinates are not a good representation, since they are *e.g.* not invariant to translation or rotation. Additionally, the poor separation may just correspond to the physical reality. However, we are able to produce a meaningful representation of our clusters by using t-SNE. Figure 1 shows that there is statistical clustering happening based on geometry, *i.e.* it is possible to identify groups of distinct molecular structures.
We applied Linear Regression on our data in an attempt to predict electronic energies of the molecular clusters with respect to the coordinate variables. However, due to the already mentioned poor representation in cartesian coordinates the outcome did not yield any significant results. Therefore, we did not include Linear Regression in our final product.

As the .xyz-files are most popular way of representing molecules, the next step for this program is to alter the representation of the molecules from cartesian coordinates. Then the machine learning algorithms will be able to learn features from the molecules without being "fooled" by *e.g.* the rotation of the molecule. We are able construct a "hypothetical" molecule corresponding to the centroids of the statistical clustering, which will, once the representation is fixed show the user directly which distinct structures exist within the data.