

# Discussion on "Exploiting Non-Linear Structure in Astronomical Data for Improved Statistical Inference" by Ann B. Lee and Peter E. Freeman

Didier Fraix-Burnet

**Abstract** Both dimensionality reduction and classification seek a reduced simpler form of the data. The first one works with the parameter space, while classification works with the object space. Ideally, one wishes to find a parameter space in which the points are naturally gathered into distinct groups and, as a physicist more particularly, data points can fit our model curves. I want to point out that dimensionality reduction methods and classification approaches are highly complementary and should even be carried out together. Astrophysical objects are complex, so that numerical simulations are now a common tools to do physics. Model fitting has thus become a comparison between populations (the observed ones and the synthetic ones) rather than plotting a curve onto data points. This is exactly the role of statistics.

**Key words:** Classification; Dimensionality Reduction; Populations

## 1 Structures in the Data Space and Classification

The paper by Ann Lee and Peter Freeman deals with the difficulty of inferring anything meaningful from astrophysical data that are complex and of high-dimensionality (and non-standard). Dimensionality reduction aims at easing statistical inference and simplifying interpretation through a simpler form of the data. In astrophysics, where technological achievements provide us with a growing number of different kinds of observables, extracting the most influential parameters also serves as a guide for future investigations and even telescope/detector design. A reduced parameter space is essential for modeling especially if analytical calculations are carried out. However, the numerical simulations become most often unavoid-

---

Université Joseph Fourier - Grenoble 1 / CNRS, Institut de Planétologie et d'Astrophysique de Grenoble, BP 53, F-38041 Grenoble cedex 9, France, e-mail: fraix@obs.ujf-grenoble.fr

able because of the complexity of the astrophysical objects. Then, the number of parameters here also must be synthesized to the most important ones.

The general purpose of classification is to ease memory and discover the relationships between classes. It is easier to recall properties for tens of classes rather than a million objects. It is also much easier (and less computer intensive) to fit models on a limited number of representatives of classes than to many not so different objects. But obtaining classes is not sufficient if we are not able to understand why they are composed as they are and why they are different. Finding relationships is thus essential.

Dimensionality reduction reduces  
the number of parameters

	Par1	Par2	Par3	Par4	...	
Classification reduces the number of objects	Object1	.	.	.	.	...
	Object2	.	.	.	.	...
	Object3	.	.	.	.	...
	Object4	.	.	.	.	...
	...	...	...	...	...	...

In summary, both dimensionality reduction and classification share the same goal. In simple words, the common ideal objective is to find a parameter space in which the points are naturally gathered into distinct groups and data points can fit our models. Ann Lee has shown us how dimensionality must care about structures in the data space. I would like to show that classification is also very concerned with these same structures.

Traditional classification in astrophysics makes heavy use of scatter plots and hard limits, most often linear. Parameters are chosen according to the observational means (infrared or radio galaxies, X-ray objects...), their "obviousness" (elliptical, Lyman- $\alpha$  or compact galaxies...) or an a priori understanding of the underlying physics (star-forming or massive galaxies...). Such classifications are thus limited by the use of very few properties and cannot reflect the real complexity of astrophysical objects.

Multivariate classifications are just beginning to be used in astrophysics [1, 2, 3]. Clustering analyses are generally based on distance matrices, principally using euclidian distances, thus assuming a linear multivariate space. More sophisticated methods use a priori knowledge to implement a particular geometry of the data space and use an adapted distance definition. On the contrary parameter-based (or character-based) approaches, using the coordinates of the objects and not their pairwise distances, explore the geometry of the data space. As one can easily understand, distance-based methods are generally much more computationally efficient.

It appears to me that the diffusion maps technique described in the paper by Ann Lee and Peter Freeman, and the spectral connectivity analysis more generally, is of the second kind. These methods explore the geometry of the data space even though they assume an euclidian metrics *locally* (any curved geometry can be locally ap-

proximated by euclidian spaces). This works well because the data space is expected to be sparse due to the physical relations that explain the diversity of objects.

Transformation processes that cause properties to evolve are all continuous in astrophysics. The distribution of data points in a multivariate space is thus mainly continuous. For sparsity to occur, that is for structures to be differentiated with voids in between, the variables must be constrained by some underlying phenomena.

Classifying objects in a continuous data space is not that easy because fuzziness is unavoidable: limits cannot be hard and overlaps are possible. Even if gaps are observed, it is generally impossible to guarantee that they will not be filled by newly discovered objects. So classification in a continuous data space must be understood as an ordered organisation. Distance-based or character-based methods establish relationships between the objects, most easily depicted on a hierarchical representations like trees or split-networks (a generalization of trees). The relationships so revealed allow for a flexible classification, the number of groups depending on the level where the tree is cut.

However, when does a parameter matrix or a distance matrix be represented on a tree-like scheme? It can be shown [4] that this is the case when the objects define a convex structures in the data space This is very similar to the salesman problem, a classical question in algorithmics that seeks to optimize the journey of a salesman through different cities. The solution is easy when the cities are arranged on a single convex hull, then the tree is linear. When several complex hulls are present, the tree becomes more complicated and can take the form of a split-network.

Hence, the geometry of the data space is crucial to organize the objects in an intelligible way. This data space cannot be any, it is defined by the parameters with which the convex hulls appear.

In conclusion, to reduce the number of objects, we need to be in the right data space. We thus need to select the right parameters, To do that objectively and extensively, the methods to reduce the dimensionality are extremely useful since they can identify the most discriminant axes of variability. But they must preserve the main geometrical properties of the data space. This is a quality of the spectral connectivity analysis method used by Ann Lee and Pete Freeman.

## 2 Finding the right data space

There is thus a parallel and complementary search of the right data space both by using dimensionality reduction techniques, to probe the parameters, and by using multivariate classification, to probe the robustness and the interest of the groups that can be defined from these parameters. Starting from the initial parameter space, one constructs a sub-parameter space with the first kind of approach, and then check whether a classification can be obtained. From this second analysis, some information is gathered on the structuring properties of the parameters, then further iterations can lead to a final sub-parameter space from which a final classification is proposed. Then, and only then, the interpretation can begin.

### 3 Model fitting and populations

Would we envisage to put living organisms into equations and follow their evolution? Biologists rather use statistical laws to model the evolutions and relationships of *populations*.

Model fitting in astrophysics still often means plotting a curve onto data points. Unfortunately, the observations and their parameters are too many, so that most scatter plots are merely clouds of points in which many curves can fit equally well. In addition, without a proper classification, the chance is weak that the right population of galaxies has been picked up for the test.

But there is more. Ann Lee presents an application of the spectral connectivity analysis to obtain prototypes of synthetic galaxy spectra. The reason is that it would take too much time to find the best values for the many variables of the models by fitting each of the million observed spectra. It is simpler to only use a limited number of model prototypes selected from the synthetic population of models. We have here a good example where the search of the most influencing parameters (reduction of dimensionality in the model space) leads to a classification (the prototypes).

I however find it amusing to use individual observed objects against prototypes of models, and not using "prototypes" of observed objects. This reflects the radical evolution of contemporary astrophysics. On one side we have a huge amount of observations, with many objects described by many parameters. On the other side, computers allow us to investigate a detailed and complicated physics. Numerical simulations produce huge populations of synthetic objects. The question is how use them to compare with the observed populations?

Model fitting nowadays clearly appears as a comparison between populations, not any more fitting a curve for an individual galaxy. Classification becomes crucial, but not with the old fashioned way of segregating objects according to their most obvious properties. This is real statistics that astronomers must use. Physicists in general are not formed at all to this way of thinking, of doing Science. This is cultural, and certainly explains why astrostatistics is still not widely popular in astrophysics. It will certainly take some time, but change is coming.

### References

- [2] Chattopadhyay, A.K., Chattopadhyay, T., Davoust, E., Mondal, S., Sharina, M.: Study of NGC 5128 globular clusters under multivariate statistical paradigm. *Astrophysical Journal* **705**, 1533-1547 (2009) (arXiv:0909.4161)
- [1] Ellis, S.C., Driver, S.P., Allen, P.D., Liske, J., Bland-Hawthorn, J., De Propris, R.: The millennium galaxy catalogue : on the natural sub-division of galaxies. *Monthly Notices of the Royal Astronomical Society* **363**, 12571271 (2005) (astro-ph/0508365).
- [3] Fraix-Burnet, D., Dugué, M., Chattopadhyay, T., Chattopadhyay, A. K., Davoust, E.: Structures in the fundamental plane of early-type galaxies. *Month. Not. Royal. Astron. Soc.* **407**, 2207-2222 (2010) (arXiv:1005.5645)
- [4] Thuillard M., Fraix-Burnet D.: *Evolutionary Bioinformatics*, **5**, 33-46 (2009) (arXiv:0905.2481)