Unveiling exoplanet families

Simone Marchi and Sergio Ortolani

Dipartimento di Astronomia, University of Padova,
I-35122 Padova, Italy
email: `simone.marchi@unipd.it`, `sergio.ortolani@unipd.it`

Abstract. The discovery of an increasing number of extrasolar planets (EPs) prompts the development of a planetary taxonomy. Such analysis, as in many other fields of research, is useful to identify groups of objects sharing similar traits. When applied to extrasolar planets, the taxonomy may provide a valid support for disentangling the role of the several physical parameters (semimajor axis, metallicity etc.) involved in the planetary formation processes and subsequent evolution. We present the state-of-the-art for exoplanets taxonomy obtained with hierarchical algorithms and the definition of robust clusters of planets (this is an update of the taxonomy published in Marchi 2007). The physical relevance of the exoplanet clusters along with their implications for the formation theories are also discussed. Finally, we comment on the future improvements of such analysis taking into account new algorithms and new input variables.

Keywords. planetary systems, planetary systems: formation

1. Multivariate analysis and dimensional scaling of EPs

The results presented here are an update of the taxonomy presented in Marchi (2007). Before going into the details of the present results, we recall few basic concepts about the method.

The inputs to our model are the following: planetary projected mass (M_p), semimajor axis (a), eccentricity (e), stellar mass (M_s), stellar metallicity ([Fe/H])†. Only objects having simultaneously estimates for $\{M_p, a, e, M_s, [\text{Fe/H}]\}$ have been used.

We consider 209 EPs (updated at 13 June 2007). To them, the Solar System planet Jupiter has been added. The first step is to perform a statistical analysis in order to find out if there are useless -or less significant- input variables. This is done using principal component analysis (Everit & Dunn 2001). The basic idea of principal component analysis (PCA) is to combine the input variables in such a way as to show the most important ones. This is done by describing the data with a number of new variables pc_i, ordered in terms of decreasing variance.

On the basis of the variance attained by each pc_i we may reject some of them. This procedure has the advantage of using only the variables which are important, allowing a simpler description of the data set with only a minor loss of information. According to general criteria, it seems reasonable to keep only the first three principal components which account for 81% of the total variance (notice we use the logarithms of M_p and a since this gave an higher variance of the first principal components).

† They are obtained from the interactive extrasolar planets catalog mantained by J. Schneider, see http://exoplanet.eu/
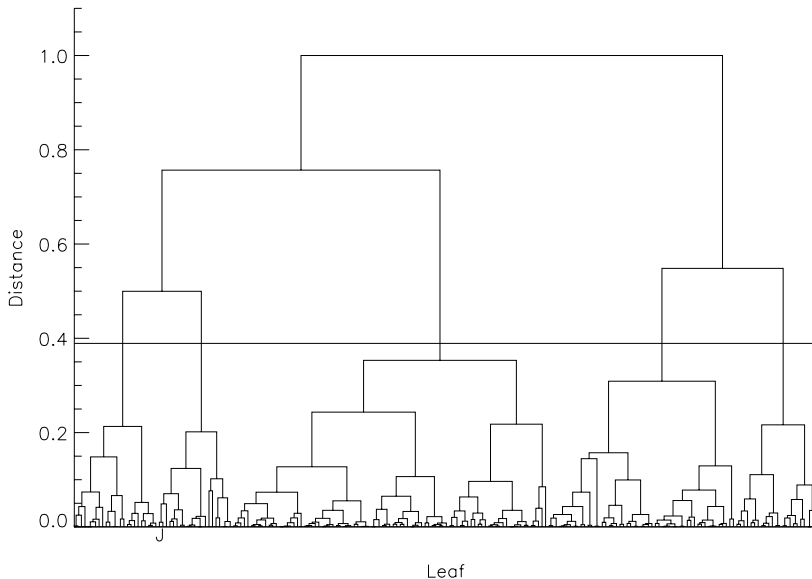
Figure 1. Dendrogram of the best solution. The vertical line correspond to the best cutoff which identify the five cluster (see text for futher details).

2. Cluster analysis

The choice of the clustering technique is quite arbitrary and it relies mostly on the kind of description of the data we are interested in. When the number of clusters are not known a priory, like in our case, hierarchical clustering is more suitable (Everit et al. 2001). One of the advantages of this technique is that it provides a classification which consists of a series of nested partitions, which is well illustrated by a two-dimensional diagram known as *dendrogram*. However, there are a number of possible ways to perform the analysis, and an accurate step-by-step evaluation of the process has to be performed.

First we decided not to standardize the clustering variables as this may reduce the difference among members, making the identification of clusters more difficult. Moreover, as a general rule, the same metrics should be used for the proximity matrix and the inter-group proximity measures. We explored different metrics and the effects of different algorithms of inter-group merging. The full set of possibilities has been investigated by using the traditional means (e.g. cophenetic coefficient) and analyzing the corresponding dendrograms.

We therefore identify a robust solution which is closely nested for small heights and stable against errors (e.g. observational errors) on the position of the EPs in the clustering space. The best solution was obtained with the Pearson correlation distance† and weighted centroid merging. It has five robust clusters.

It is interesting to note that traditional metrics (like the Euclidean) and traditional cluster merging (like single linkage) produce in general bad results. It turns out that they are not able to find structures for EPs.

---

† The distance between two items i, j is defined as $d_{ij} = [(1-\phi)/2]^{1/2}$, where ϕ is the Pearson correlation coefficient.
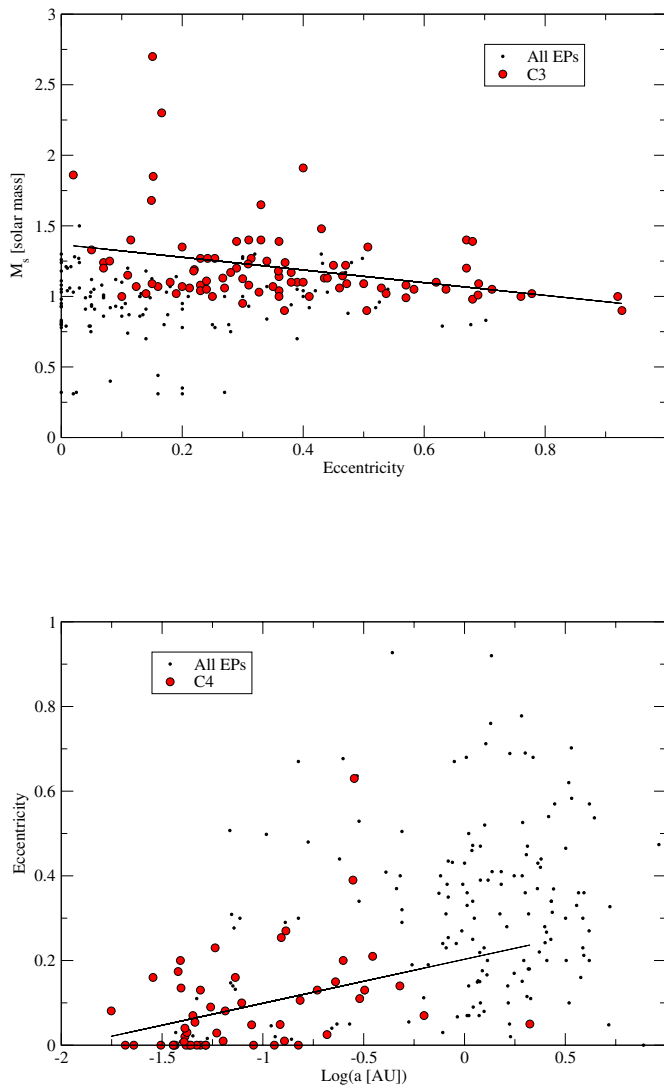
Figure 2. Examples of correlation found within clusters. Here we show the $e - M_s$ correlation for $\mathcal{C}3$ and $a - e$ correlation for $\mathcal{C}4$. We also overplot all the EPs.

Moreover, we tested the solution against the absence of clusters with Monte Carlo simulations. We also tested the solution with respect to the presence of observational errors. We find that the solution is stable with respect to both tests.

3. Analysing the clusters

In this section we present a briefly overview of the properties of the five clusters identified.

A first important point is that, although the clusters have been identified using pc_i, we may nevertheless find differences among the input variables. We may summarize these differences in this way: $\mathcal{C}1$ has sub-solar [Fe/H] and super-jovian M_p (notice that Jupiter belong to this cluster); $\mathcal{C}2$ has sub-solar [Fe/H] and sub-solar M_s; $\mathcal{C}3$ has super-solar [Fe/H] and M_s, $\mathcal{C}4$ has sub-jovian M_p and low e; $\mathcal{C}5$ has super-solar [Fe/H] and M_s. Moreover, the hot jupiters (HJs; i.e. planets with period less than 10 days) belongs to $\mathcal{C}4$ and $\mathcal{C}5$: we thus identify two main types of HJs. Most notably all the so-called hot Neptunes belong to cluster $\mathcal{C}4$. In other words, our cluster analysis is able to identify two types of hot "planets", possibly indicating different origin and evolution. We shall deal with this aspect in a forthcoming paper. As for the multiplicity, we find that all clusters have a similar number of multiple planet systems (MPS) and multiple stellar systems (MSS) except for $\mathcal{C}1$ which has no MSS and only a few MPS.

Finally, we also checked for inter-correlation among the input variables within each cluster. The reason for that, is that commonly accepted that planets may form in different ways (core accretion vs disk instability), and that their evolution is affected by several parameters (disk density, stellar types, opacity etc). The EP database may reflect such complexity, however the signature of these processes may be blurred in statistical analyses which deal with the whole EP dataset. On the contrary, if cluster separation has something to do with the formation and evolutionary processes and not being just a mere classification, it becomes important to look for trends within each cluster. In the following we report only highly significant (i.e. having a 2-tailed probability less then 5%) intra-cluster correlations. These correlations are: $a - M_s$, M_p-[Fe/H] for $\mathcal{C}1$; $M_p - M_s$ for $\mathcal{C}2$; $e - M_s$ for $\mathcal{C}3$ (see fig. 2); $a - e$ for $\mathcal{C}4$ (see fig. 2); $a - M_s$ for $\mathcal{C}5$. Therefore the stellar mass seems to play an important role in several clusters, while the [Fe/H] seems to be by far less important. This is somehow in contadiction with Marchi (2007), where both variable where found to be important.

It is important to realize that such correlations rise considering the present population of discovered exoplanets. In this respect two important points have to be considered. First, some of the correlations may be affected by observational biases. On the other hand, some of them are probably affected by planetary evolution. This is clearly the case for the cluster $\mathcal{C}4$, where the $a - e$ correlation is likely the results of tidal circularization. For the purpose of the present proceeding we do not tackle with the physical interpretation of the clusters and the correlations, which will be the topic of a detailed forthcoming paper.

Another relevant point, is that we find a significant correlation between [Fe/H] and M_s for several clusters and also for the whole EPs sample. This clearly implies that the two parameters are linked in the field stars selected for planetary search purpose. Indeed for a given age this is expected from the theoretical models of stellar evolutions (Girardi et al. 2002): we estimated an increase of 0.5 $M_\odot$/dex at 5 Gyr for a star of about 1 $M_\odot$ (see fig. 3).

4. Discussion and conclusion

In this paper we develop the basis for an extrasolar planet taxonomy. We identify the best procedure to follow: a multivariate statistical analysis (PCA) to find the most important variables, and then hierarchical clustering analysis. The best result is achieved with non traditional metric and merging algorithms, namely the Pearson correlation metric and weighted centroid cluster merging. We reject the absence of clustering structure with Monte Carlo simulations, and also tested the stability of the solution against
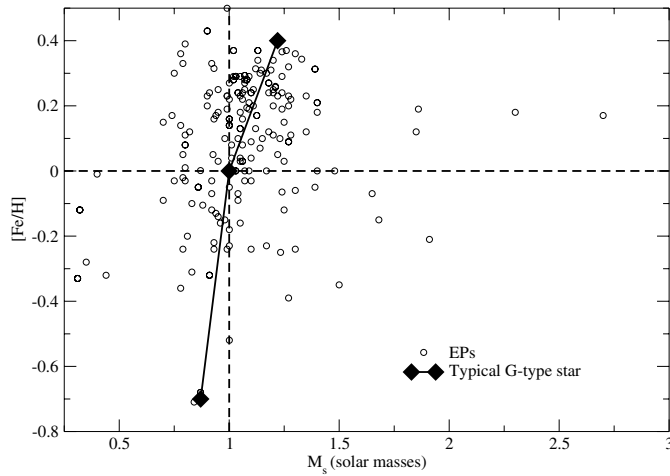
Figure 3. Metallicity vs stellas mass scatter plot for all EPs. Apart for few outliers, the overal trend is a correlation between [Fe/H] and M_s. The theoretical trend for a typical, 5 Gyr old, G-type star is also overplotted.

observational errors of the input variables. The procedure we followed is able to provide -even if the number of planets is still low- a robust extrasolar planet taxonomy. The general traits of the taxonomy developed here will be updated as more planets become available.

Our best solution consists of five clusters. We discuss their properties with respect to the physically relevant input variables. We show the importance of including the environmental variables (in particular of M_s) to discriminate between otherwise similar planets; and also to merge together different bodies (like EP in MSSs and orbiting single stars). For instance, we were able to divide the hot Jupiters into -at least- 2 main groups. This division is mainly due to the metallicity and -to a lesser extent- to stellar mass. There other interesting features about the HJs, but we will focus on them in a forthcoming paper.

We also analyzed the intra-cluster correlations, since this may provide important information about the formation and evolution of bodies within a cluster. This is crucial in order to uncover information which may be hidden in the "blind" statistical analysis performed on the whole EP database. The most important correlations found are those for $\mathcal{C}1$, $\mathcal{C}3$, $\mathcal{C}4$ and $\mathcal{C}5$. Remarkably, for $\mathcal{C}3$ and $\mathcal{C}5$ we find important trends between stellar mass and orbital parameters. It results that M_s has very important effects on the semi-major axis (and thus on the migration processes) and the eccentricity. Moreover, we also studied the distribution of planets in multiple star systems in each cluster. They do not seem to play a particular role in the corresponding cluster correlations. Similar considerations apply also for multiple planet systems.

Future improvements will include more parameters in the taxonomical analysis like the stellar age, chemical peculiarities and stellar/planetary multiplicity.

References

Everitt, B. S. & Dunn, G. 2001, Applied multivariate data analysis, second edition, Arnold Ed., London

Everitt, B. S., Landau, S., & Leese, Morven. 2001, Cluster analysis, fourth edition, Arnold Ed., London

Girardi, L., Bertelli, G., Bressan, A., Chiosi, C., Groenewegen, M. A. T., Marigo, P., Salasnich, B., & Weiss, A. 2002, A&A, 391, 195

Marchi, S. 2007, ApJ, 666, 475