EXTRASOLAR PLANET TAXONOMY: A NEW STATISTICAL APPROACH

S. Marchi

Dipartimento di Astronomia, Università di Padova, Vicolo dell’Osservatorio 2, 35122 Padova, Italy; simone.marchi@unipd.it

Received 2007 February 1; accepted 2007 May 4

ABSTRACT

In this paper we present the guidelines for an extrasolar planet taxonomy. The discovery of an increasing number of extrasolar planets showing a vast variety of planetary parameters, such as Keplerian orbital elements, and environmental parameters, such as stellar masses, spectral types, and metallicity, prompts the development of a planetary taxonomy. In this work, via principal component analysis followed by hierarchical clustering analysis, we report the definition of five robust groups of planets. We also discuss the physical relevance of such analysis, which may provide a valid basis for disentangling the role of the several physical parameters involved in the processes of planet formation and subsequent evolution. For instance, we were able to divide the hot Jupiters into two main groups on the basis of their stellar masses and metallicities. Moreover, for some groups, we find strong correlations among metallicity, semimajor axis, and eccentricity. The implications of these findings are discussed.

Subject headings: planetary systems — planetary systems: formation — planets and satellites: general

Online material: color figures, machine-readable table

1. INTRODUCTION

With the discovery of the first extrasolar planet (EP) orbiting a solar-like star (Mayor & Queloz 1995), a new field of astronomical research started. So far, more than 200 extrasolar planets have been discovered. Despite severe observational biases—only partially overcome by the development of more sophisticated techniques and facilities—the most striking fact is the vast variety of EPs and the remarkable difference with respect to the planets of our solar system. Traditional theories of planetary formation have been seriously challenged by these discoveries, and there are many aspects that still need to be understood in detail, such as planetary migrations (disk-embedded vs. planet-planet interactions) and the influence of the metallicity on the planetary formation processes, just to mention some. The heterogeneity of EPs has been analyzed in several papers, and some important trends have been uncovered (see, e.g., Zucker & Mazeh 2002; Santos et al. 2003; Udry et al. 2003; Eggenberger et al. 2004; Sozzetti 2004; Fischer & Valenti 2005). On the other hand, planetary formation is a very complex process in which a number of parameters have important effects, as shown by several theoretical works. It is also likely that some parameters act simultaneously in a complex way, thus motivating the need for a multidimensional approach going beyond the search for simple correlations among individual parameters. In other words, the planetary formation process occurs in a multidimensional space. In addition to Keplerian orbital elements and stellar properties, other relevant parameters may be discovered in the future, as the formation theories improve.

We adopt a multivariate approach to EPs in order to uncover underlying trends that may provide important information about the planet formation processes. The first step is the development of a robust taxonomy for EPs. Taxonomy, just as in other fields of research, may be a precious tool in defining clusters of EPs, which in turn may highlight differences in the formation processes and subsequent evolution.

This will be of particular importance with the increasing number of EP discoveries that are expected to occur in the near future, thanks to several space missions (e.g., COROT and Kepler) and ground-based surveys. For the moment, the number of EPs is about 200, which may prevent firm statistical conclusions. Nevertheless, here we propose the guidelines for an EP taxonomy, to be refined as more data become available.

In §2 we describe the parameters used for this study, and we perform a statistical approach of the data for dimensional scaling purpose. Then, in §3, cluster analysis is performed. We caution from the beginning that these analyses—much like all statistical analyses—are not unique, since several different criteria may be used for characterizing the data. There is no a priori best criterion: the choice can indeed vary according to the kind of data under analysis. We shall provide a step-by-step justification for all the choices made. Finally, in §4, the solution is discussed along with its physical interpretation.

2. MULTIVARIATE ANALYSIS AND DIMENSIONAL SCALING OF EPs

The inputs to our model are the elements provided by the interactive EP catalog maintained by J. Schneider. These are planetary projected mass ($M_p$), orbital period ($P$), semimajor axis ($a$), eccentricity ($e$), inclination ($i$), stellar mass ($M_*$), and stellar metallicity ([Fe/H]). Other possibly important parameters, such as stellar spectral type and stellar age, have not been considered at this stage, but will be the subject of further refinements. Only objects having simultaneous estimates for {$M_p$, $P$, $a$, $e$, $M_*$, [Fe/H]} have been used. Note that the period and semimajor axis are obviously correlated; thus, only one of them is used (see the following). We refer to them as the input variables. Therefore, each planet can be regarded as a point in a fivefold space. The following analysis has been carried out in the IDL language, and it is structured so as to be easily updatable as more data become available.

We consider 183 EPs (updated at 2006 November 8). To them, the solar system planet Jupiter has been added. This has been done because Jupiter-like bodies are approaching the observability limit in extrasolar planetary systems, thanks to recent improvement in

\footnote{See http://exoplanet.eu.}

\footnote{We chose to use simply $M_p$ and $a$ instead of performing some transformation of the data prior to the statistical analysis, as done by some authors who made recourse to logarithms, because there is no strong physical reason for that. In any case, we plan to exploit this possibility in further works.}
the surveys, and also to have a direct comparison with our own solar system. We decided not to add the other solar system planets, as they are still well below the detection limits. 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 (PCA). As this is a standard technique in multivariate data analysis, details will not be discussed here (interested readers may consult Everitt & Dunn 2001). The basic idea of PCA is to combine the input variables in such a way as to show those of most importance. This is done by describing the data with a number of new variables, \( \text{pc}_1, \ldots, \text{pc}_l \) for \( l = 5 \), ordered in terms of decreasing variance. The \( \text{pc}_l \) are chosen in such a way as to be uncorrelated with each other. This is done in practice by building an \( n \times l \) input data matrix (where \( n \) is the number of planets). From this matrix, the \( l \times l \) correlation or covariance matrix is computed, where the correlation matrix is more appropriate whenever the input variables are measured with disparate units. The \( \text{pc}_l \) are eigenvectors for those matrices. Considering that input variables have different units and that some span orders of magnitudes (e.g., \( M_p \)) while others span a limited range (e.g., \( M_s \)), we opted to perform variable standardization of the input variables, i.e., to scale input variables in such a way as to obtain a mean and variance of 1. This has the advantage of allowing the use of the covariance matrix, for which the eigenvalues \( \lambda_i \) represent the portion of the variance of the original data that the corresponding eigenvectors \( \{ \text{pc}_l \} \) account for. PCA can thus provide a useful means for finding variables of little significance. On the basis of the variance attained by each \( \text{pc}_l \) we may reject some of them. This procedure has the advantage of using only the variables that are important, allowing a simpler description of the data set with only a minor loss of information.

As a first step, we used PCA to decide whether the period or the semimajor axis is more suitable for the statistical analysis (see also § 3.2 on this point). PCA was then performed on the \( \{ M_p, P, e, M_s, [\text{Fe/H}] \} \) and \( \{ M_p, a, e, M_s, [\text{Fe/H}] \} \) spaces. The choice was made by requiring that the variance of the resulting \( \text{pc}_l \) is more concentrated in the first principal components. The five-fold \( \{ M_p, a, e, M_s, [\text{Fe/H}] \} \) space turned out to be slightly more appropriate. Now, no commonly accepted rules exist for deciding which variables, if any, can be safely thrown away. One criterion suggests keeping variables that account for 70%–90% of the total variance (Everitt & Dunn 2001). Thus, it seems reasonable to keep only the first three principal components, which account for 73% of the total variance. The fact that we can use only three principal components allows a graphical representation and a better control of the results. Moreover, dimensional scaling has the effect of simplifying the clustering procedure. This is why we employed PCA, but we are aware it is not strictly necessary for the development of a taxonomy. In Figure 1 the first three principal components are shown.

To the accepted levels of variance we may write the following decomposition formulae,

\[
\begin{align*}
\text{pc}_1 &= -0.206(M_p - 2.64) - 0.543(a - 1.19) - 3.417(e - 0.25) \\
&\quad - 1.823(M_s - 1.03) + 0.067([\text{Fe/H}] - 0.1), \\
\text{pc}_2 &= -0.083(M_p - 2.64) - 0.030(a - 1.19) - 0.335(e - 0.25) \\
&\quad + 3.080(M_s - 1.03) + 3.921([\text{Fe/H}] - 0.1), \\
\text{pc}_3 &= 0.090(M_p - 2.64) - 0.329(a - 1.19) + 1.795(e - 0.25) \\
&\quad - 1.748(M_s - 1.03) + 2.063([\text{Fe/H}] - 0.1),
\end{align*}
\]

where each \( \text{pc}_i \) is expressed in terms of a linear combination of all the input variables. Note that \([\text{Fe/H}]\) has little influence on \( \text{pc}_1 \), while \( a \) and \( e \) basically do not contribute to \( \text{pc}_2 \) (consider that input variables span different ranges, so for instance, the term containing \( e \) in eq. [2] has a lower importance than the term involving \( M_p \)). These formulae may be used to add new planets to the database, without repeating the PCA (this is useful when we want to add only single objects that do not alter the overall statistics; otherwise, PCA has to be repeated).

Finally note that we also checked for the presence of overall correlations within the database. We found significant correlation (see also below) between \( M_p,a,e,\) and \( M_s,a \). However, overall correlations have been widely analyzed elsewhere and are not discussed here (e.g., Santos et al. 2005 and references therein).

3. CLUSTER ANALYSIS

Cluster analysis is a standard technique used in a variety of research fields, from social sciences to geology to engineering, and so on. The main purpose of such analysis is to find clusters in a given data set such that elements belonging to the same cluster have a certain degree of homogeneity, while elements of different clusters have to be as different as possible. As with PCA, there is no unique way of performing the analysis, and a number of clustering algorithms are available. 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 is not known a priori, as in our case, hierarchical clustering is more suitable; thus, we adopt this technique (Everitt et al. 2001). The major advantage of this technique is that it provides a classification that consists of a series of nested partitions, which go from a single cluster containing all objects to \( n \) clusters each containing a single object. This process is based on a proximity parameter, usually a measure of distance among objects. Moreover, this classification may be represented by a two-dimensional diagram known as a dendrogram, which illustrates the nested nature of the hierarchical partition. 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. In the following we refer to the variables used for the clustering (namely the \( \{ \text{pc}_i \} \)) as clustering variables. The general guidelines for hierarchical
cluster analysis are the following (for more details see Everitt et al. 2001):

1. Standardization of the clustering variables.
2. Computation of the proximity matrix: that is, evaluate the degree of proximity among data set members.—This is usually done by computing the distance in the clustering space (in our case the threefold \{pc1, pc2, pc3\} space). Note that several metrics may be used for this purpose.
3. Definition of intergroup proximity measures.—These specify the method used to quantify the proximity level of two clusters. Several methods are available, such as single linkage and centroid linkage.

4. Computation of the distortion of the dendrogram.—The dendrogram itself is a representation of the original data, in which the separation of two members is specified by the minimum distance between the two clusters that contain the objects. Thus, the quality of a dendrogram can be evaluated by comparing the original distance (i.e., in the clustering space) of members with the distance assigned to them by the dendrogram. This is formally done by the so-called cophenetic coefficient (\(\hat{c}\)). It normally ranges from 0.6 to 0.95, where higher values correspond to a lower distortion.

5. Definition of the best partition.—In standard analysis this means stopping the hierarchy at a given distance, in other words, to cut a dendrogram at a particular distance (or height, \(h\)). Again, no unique rule exists to define the best cut. Here we follow the procedure suggested by Mojena (1977): the so-called upper tail rule. In detail, we first evaluate the distribution of the heights \(N(h)\); then, the best-cut height is determined by

\[
\hat{h}_c = \bar{h} + \gamma \sigma_h,
\]

where \(\gamma\) is a coefficient that ranges from 1.25 to 3.5 (Milligan & Cooper 1985), \(\bar{h}\) is the mean of the height distribution, and \(\sigma_h\) is its standard deviation. A possible way to define the best \(\gamma\) is to find the values of \(\gamma\) for which the maximum separation between two consecutive solutions is achieved.

6. Testing of the quality of the solution against the absence of clusters.

3.1. Finding the Best Solution

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 intergroup proximity measures. We explored different metrics and the effects of different algorithms of intergroup merging. The metrics used are Euclidean, City Block, Chebyshev, and correlative (based on the Pearson correlation coefficient). The intergroup algorithms considered are single linkage, complete linkage, weighted pairwise average, and weighted centroid. The full set of possibilities has been investigated by using the cophenetic coefficient and analyzing the corresponding dendrograms. The value of \(\hat{c}\) ranged from about 0.55 to 0.84. Although useful, the cophenetic coefficient cannot be the only parameter for finding the best solution; very different hierarchies may have the same \(\hat{c}\)-value. Here we also considered other features in order to define the best solution. Among them, we first select the dendrograms having \(N(h)\) distributions with the lowest variance, i.e., those more peaked for low heights. Then, we required the tails of the \(N(h)\) distributions to be very “discrete” for large heights; that is, the solutions with \(k\) and \(k + 1\) clusters (for low values of \(k\) ) had to be well separated. Both conditions help to define a robust solution that is closely nested for small heights and stable against errors (e.g., observational errors) on the position of the EPs in the clustering space.

Taking into account these constraints, the best solution was obtained with the Pearson correlation distance\(^4\) and weighted centroid merging. This solution corresponds to \(\hat{c} = 0.83\). According to the upper tail rules, the best cut produces five robust clusters. Remarkably, the interval of heights for which the solution is stable corresponds to 7% of the maximum height (note that the average separation between two consecutive clusters is of the order of the maximum height divided by \(n\), or \(\sim 0.5\%\)). The corresponding best solution dendrogram is shown in Figure 2. The dendrogram is well structured: for low heights the clusters are closely nested (i.e., a small variation of the height results in a significant change in the number of clusters), while for higher values the clusters are well separated.

It is interesting to note that traditional metrics (such as the Euclidean) and traditional cluster merging (such as single linkage) produce, in general, bad results. It turns out that they are not able to find structures for EPs. This is in agreement with the analysis of Jiang et al. (2006), although it was performed with a very different approach to that adopted here. The main problem of these hierarchies is that either they produce just one or two big clusters and some outliers, or the solution is so nested that the definition of cutoff is too arbitrary and meaningless.

3.2. Testing the Solution

Before analyzing in detail the best solution found, we have to check for its robustness against the absence of clusters. We recall that cluster analysis will always produce a solution. The test for the absence of structure is done by an accurate evaluation of the intrachannel separations. In general, unless we are dealing with data with very well separated clusters, the clusters will tend to have some degree of overlap. The overlap between two clusters can be estimated by comparing the distance of two clusters with respect to their volume in the clustering space. In practice, we estimate the center of each cluster (\(c_i\)) and its radius (\(r_i\)), defined as the sphere that contains a given percentage of the cluster members (both \(c_i\) and \(r_i\) are computed with a Euclidean metric). Thus, the degree of overlap between clusters \(i\) and \(j\) is set to \(\omega_{ij} = (r_i + r_j)/\|c_i - c_j\|\). For \(\omega_{ij} < 1\) the clusters \(i\) and \(j\) are well separated. The overall overlap of the solution is defined as

\[
\omega = \sum_{i<j} \omega_{ij}
\]

if \(\omega_{ij} > 1\). Higher values of \(\omega\) imply larger overlap. Note that \(\omega\) is a cumulative parameter and is sensitive to the number of single intrachannel overlaps, as it should be. To clarify this point, imagine a solution with a given degree of overlap between two clusters, and another one with the same degree of overlap but among three clusters. The average degree of overlap will be the same, but the \(\omega\) of the second solution will be higher than that of the first. Indeed, the first solution is better than the second, since only two clusters are partially merged. For the best solution, we obtained \(\hat{\omega} = 12.3\).

\(^3\) To compute \(\hat{c}\), proceed as follows. First, compute the distance between two planets (indicated by indices \(i\) and \(j\)) assigned by the dendrogram. Let \(d_{ij}\) be this distance, and \(d_i\) the average value of all \(d_{ik}\). Let \(d_{kj}\) be the real distance between planet \(i\) and \(j\), and \(r\) the average value of all \(r_k\). Thus, \(\hat{c} = [\sum \sum (d_{ij} - d_i)(d_{kj} - r)^2]^{1/2} / (\sum d_{ij})\).

\(^4\) The distance between two items, \(i\) and \(j\), is defined as \(d_{ij} = [(1 - \phi)2]^{1/2}\), where \(\phi\) is the Pearson correlation coefficient.
One way of testing the absence of clusters is with Monte Carlo simulations. We then performed a number of simulations (with $10^3$ solutions each), generating $n$ random points in the clustering space. The points are generated with a uniform distribution along each axis. Then we run the clustering analysis using the same procedure described before, varying also the volume of the Monte Carlo–generated points in the clustering space, and the percentage of objects used for defining the cluster’s radii. The probability of obtaining $\omega < \tilde{\omega}$ is less than 10% in all cases. Thus, we may safely reject the possibility of the absence of structures. For the sake of completeness, we also perform additional tests to see whether our initial choice of using $a$ instead of $P$ is indeed a good one in terms of taxonomy. Performing clustering analysis using $P$ instead of $a$, we obtain a solution that has $\omega = 20$ and is compatible with the absence of structures. A possible explanation of this behavior could be due to the fact that $P$, $a$, and $M_z$ are related. As a result, the use of the pair $(a, M_z)$ is better than the pair $(P, M_z)$. This further strengthens the case for the use of the semimajor axis.

Moreover, we also tested the solution with respect to the presence of observational errors, which can be very large in some cases. To do this, we generated a random error in the input variables around the nominal values for all the EPs. We then repeated the PCA and cluster analysis of these fictitious samples (one for each of the input variables). The solution obtained has been compared with the best solution. For errors of a few percent, we find that some planets may move to other clusters, but this is limited to a few objects. Increasing the errors, sometimes a cluster may split into two subclusters. For errors larger than 10% the solution may alter considerably. In general, the best solution does not change much for errors up to $\pm 10\%$ for each of the input variables. This is very important, since it means that the solution is quite stable, in particular considering that we are dealing with projected planetary masses and not real masses. The process developed for EP taxonomy is sketched in Figure 3.

4. ANALYZING THE CLUSTERS

Our best solution is composed of five robust clusters. In this section we present the properties of each cluster. Figure 4 shows the position of clusters in the clustering space. We checked for intercorrelation among the input variables within each cluster. It is 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 that deal with the whole EP data set. On the contrary, if cluster separation has something to do with the formation and evolutionary processes and is not just a mere classification, it becomes important to look for trends within each cluster. In the following, we report only highly significant (with a two-tailed probability less than 5%) intracluster correlations.

4.1. Cluster C1

Containing 11 EPs (see Tables 1 and 2 for a detailed description), C1 is the least populated cluster. We can define a prototype planet, that is, the object closest to the center $c_i$ of the cluster. The prototype is HD 41004Ab. Figure 5 shows the distribution of the input variables for each cluster. The EPs of this cluster are
characterized by a wide range of planetary masses, from about 0.2 to 18.4 $M_J$. The planetary semimajor axis ranges from 0.018 to 1.97 AU. The eccentricities are quite spread, from 0.08 to 0.63. The stellar masses are remarkably subsolar (except for HD 8574b, which has 1.04 $M_{\odot}$). Finally, star metallicity is very spread, from $-0.5$ to 0.28 dex.

C1 contains several peculiar EPs. HD 41004Bb, HD 162020b, and HD 114762b are very massive planets, with 18.4, 13.7, and 11 $M_J$, respectively (see Tables 1 and 2). The first two are also hot Jupiters (note that HD 162020b could be a brown dwarf, according to Udry et al. 2002). This cluster contains six EPs in multiple star systems (MSSs); see Tables 1 and 2 (data on MSSs have been taken from Desidera & Barbieri 2007). Moreover, HD 41004Bb, HD 162020b, HD 114762b, and HD 111232b orbit low-mass, low-metallicity stars. Despite this variety, the objects in this cluster have to be considered “similar” in terms of clustering analysis and “different” from the other EPs. We underline that the cluster analysis has been done in the $p_c$ space; therefore, the input variables may vary considerably within a cluster.

### TABLE 1

| Cluster | Prototype | Members | Corr. | HJ | T | MSS | MPS | SS |
|---------|-----------|---------|-------|----|---|-----|-----|----|
| C1      | HD 41004Ab| 11      | $M_p$,$e$, $M_p$-$M_e$ | 2  | ... | 6   | ... | ...|
| C2      | HD 69830c | 46      | $a$-$e$, $a$-$[Fe/H]$ | 17 | 5  | 4   | 13  | ...|
| C3      | HD 11964b | 48      | $a$-$[Fe/H]$, $e$-$[Fe/H]$, $M_p$-$[Fe/H]$ | 23 | 4  | 11  | 8   | Jupiter |
| C4      | HD 142022Ab | 48 | $M_p$-$e$, $M_p$-$M_e$ | ... | ... | 12  | 12  | ...|
| C5      | HD 117207b | 31 | ... | 1  | ... | 7   | 13  | ...|

**Notes.** —“Corr.” corresponds to all significant intracluster correlations, i.e., those having two-tailed probability less than 5%, with C2, C3, and C4 being the most important ones. We show them in Figs. 7–9. “HJ” corresponds to hot Jupiters. “T” corresponds to transiting planets. “MSS” corresponds to planets in multiple star systems. “MPS” corresponds to multiple planetary systems. “SS” corresponds to solar system planets.
Significant intracluster correlations exist between $M_p$-$e$ and $M_p-M_*$ (see Fig. 6). Note that $M_p$ is anticorrelated with $M_*$ This is somehow unexpected, since for a higher $M_*$ we would expect a higher dust surface density for the protoplanetary disks (Ida & Lin 2005) and hence more massive planets (consider also that here we have subsolar stellar masses; on this point, see also C4). We may argue that this has something to do with the peculiar way these EPs formed, but no firm conclusion can be drawn yet.

4.2. Cluster C2

This cluster contains 46 EPs (see Tables 1 and 2). The prototype is HD 69830e. This cluster is characterized by an $M_p$ distribution with an average of about 1 $M_J$ and a standard deviation of 1.2 $M_J$, clearly peaked at low masses. Most of the planets have masses below 1.5 $M_J$. The semimajor axis distribution is characterized by two distinctive groups, one peaked at very small $a$, and the second, far less numerous, centered at $a \sim 1.6$ AU. The overall distribution has an average of 0.46 AU and a standard deviation of 0.7 AU. Planetary eccentricities are moderate to low, 0.11 ± 0.11, and stellar masses are remarkably subsolar, 0.83 ± 0.2 $M_*$. Star metallicities are around solar, with an average of −0.04 dex and a standard deviation of 0.22 dex (see Fig. 5). Gl 581b, Gliese 876b/c/d, and GJ 436b, which orbit low-mass stars (with, respectively, 0.31, 0.32, and 0.41 $M_*$, they are the lowest $M_*$ in the sample) and have low metallicities (−0.33, −0.12, and −0.32 dex, respectively), belong to this cluster. C2 contains 17 hot Jupiters (that is, 37% of its members), and four EPs belong to MSSs (see Tables 1 and 2). It also contains five transiting EPs (the total number of transiting EPs is 14—as of 2006 December—but only nine are involved in the present analysis; see Burrows et al. 2007). These five transiting EPs do not seem to have any particular properties, except to have the highest stellar ages of the sample (but we warn that those ages may not be well constrained). Finally, C2 contains 13 planets in multiple planet systems (MPSs).

The significant intracluster correlations are $a-e$ and $a-[Fe/H]$ (see Fig. 7). The first one is very interesting because $a$ is anticorrelated with $e$. Thus, planets farther away from the stars have higher eccentricities. In other words, either the excitation of $e$ is more effective farther away from the central star (assuming that planets form in circular orbits) and/or e-dumping is more effective for lower $a$. This result is consistent with tidal circularization of planets with small $a$; however, note that at $a \sim 0.02$ AU the average $e$ is about 0.1, that is, still considerably nonzero (we caution that low $e$-values could be affected by biases due to fitting procedures; Ford 2005). Moreover, with a notably steep trend, $a$ anticorrelates with $[Fe/H]$. Thus, either the planetary migration is more pronounced for high $[Fe/H]$ (which is in agreement with simulations; e.g., see Livio & Pringle 2003) or giant planets of this cluster may form close to stars ($\leq 1$ AU) in high-metallicity environments. We recall that for $[Fe/H] > 0$, the most accredited formation theory is core accretion (Fischer & Valenti 2005; Santos et al. 2005). Note also that the distribution of the MPS planets in the $a-[Fe/H]$ plane is somehow opposite to the overall observed trend: without them the correlation would be even more pronounced. Both trends are very important because, for the first time, they show significant dependence between metallicity and orbital parameters (on the possible existence of a period-metallicity correlation, see Sozzetti 2004). These findings also suggest that lower $[Fe/H]$ planets tend to have larger orbits, making them difficult to detect (see, e.g., Boss 2002). In turn, this may affect the probability of forming planets with respect to the metallicity (Santos et al. 2004; Fischer & Valenti 2005).

4.3. Cluster C3

Containing 48 EPs (see Tables 1 and 2), this cluster, along with C4, is the most populated. The prototype is HD 11964Ab, and it contains Jupiter. This cluster is characterized by an $M_p$ distribution peaked at low masses: basically all EPs are below 2.5 $M_J$. The semimajor axis distribution is very peaked at low values: most of the bodies are below 0.25 AU, with a second, less numerous
peak at about 1 AU. Eccentricities are below about 0.3, having an average of 0.096 and a standard deviation of 0.091. Stellar masses span from 0.98 to 1.24 $M_\odot$. Metallicities tend to be supersolar and vary from 0.02 to 0.35 dex (see Fig. 5).

The significant intracluster correlations are $M_p-M_\ast$, $M_p-[\text{Fe/H}]$, $a-M_\ast$, $a-[\text{Fe/H}]$, $e-[\text{Fe/H}]$, and $M_\ast-[\text{Fe/H}]$. However, the first three are due to outliers—without them the correlations do not hold. Therefore, they are not considered interesting. On the contrary, the latter three—all involving the metallicity—are robust (see Fig. 8). First of all, $a$ is anticorrelated with $[\text{Fe/H}]$. Thus, either the planetary migration is more pronounced for high $[\text{Fe/H}]$ (see also C2) or planets with lower $[\text{Fe/H}]$ form at larger distances. A striking result is that Jupiter fits very well in this cluster: its large $a$ would be the result of its formation in a solar-like metallicity.

Fig. 5.—Comparison of the distributions of the input variables for each cluster. Within each panel the distributions have been vertically shifted for clarity. The vertical distance between two consecutive major ticks corresponds to five units. [See the electronic edition of the Journal for a color version of this figure.]
Moreover, $e$ is anticorrelated with $[\text{Fe/H}]$. In other words, higher metallicities correspond to lower eccentricities. Thus, either the excitation of $e$ is more effective, or dumping is less effective, for lower metallicities. In order to understand the meaning of these correlations, two further points have to be considered. First, the existence of $a-[\text{Fe/H}]$ and $e-[\text{Fe/H}]$ correlations does not imply a correlation between $a$ and $e$. Indeed, they are not correlated to the level of confidence adopted here. Moreover, the members of this cluster have very supersolar metallicities. These correlations, if related to the formation processes, may give important indications about the origin of high eccentricities. Many theories have been proposed (Holman et al. 1997; Murray et al. 2002; Goldreich & Sari 2003; Zakamska & Tremaine 2004; Namouni 2005; Adams & Laughlin 2006; D'Angelo et al. 2006; Fregéau et al. 2006), but none has observational support yet. Here we suggest that the metallicity acts in some way in determining $e$. For instance, since high $[\text{Fe/H}]$ produces a faster migration, the low values of $e$ observed for high $[\text{Fe/H}]$ may be the result of the migration process, e.g., tidal circularization (see also Halbwachs et al. 2005). Alternatively, the condition for the pumping up of $e$ during planet-disk interactions (Sari & Goldreich 2004; Matsumura & Pudritz 2006) is not achieved in high-$[\text{Fe/H}]$ environments: indeed higher $[\text{Fe/H}]$ implies higher disk viscosity and hence a lower probability of $e$-excitation (see eqs. [8] and [9] of Sari & Goldreich 2004). If confirmed, these trends suggest that planetary orbital parameters are mainly controlled by disk properties (e.g., metallicity) rather than affected by external factors such as perturbation due to interactions with a companion star or star encounters.

Finally, $M_*$ is anticorrelated with $[\text{Fe/H}]$. It is not easy to understand this correlation, and in particular if it has something to do with the planetary formation process. We note that EPs in this cluster (which are quite confined in terms of $a$ and $M_*$) orbit stars whose metallicities tend to increase as stellar masses decrease. If we assume that a higher $M_*$ implies higher protoplanetary disk masses from which EPs formed, this implies that to form planets in lower metallicity environments a more massive disk is required, in agreement with the core accretion theory.

Eleven EPs are in MSSs (see Table 2). It also contains 23 hot Jupiters (48% of the members) and 8 MPS planets. They are well spread and seem not to affect the $a-[\text{Fe/H}]$, $e-[\text{Fe/H}]$ correlations. However, all MPS planets and most of the hot Jupiters lie above the linear fit in the $M_*-[\text{Fe/H}]$ plane. Note also that the MPS planets of this cluster have $e < 0.2$ and $[\text{Fe/H}] > 0.2$, and for this reason they differ from those of C2 (which have $e < 0.3$, $[\text{Fe/H}] < 0.2$) and those of C4 (which have $e > 0.3$).
4.4. Cluster C4

This cluster contains 48 EPs (see Tables 1 and 2). The prototype is HD 142022Ab. C4 is characterized by rather flat distributions of the input variables (see Fig. 5). In terms of planetary masses, it contains very massive bodies, basically all having $M_p > 2.0 M_J$. Apart from very few exceptions, it contains all the EPs of the data set having a mass greater than 5 $M_J$. The average and standard deviations are 5.45 and 3.92 $M_J$, respectively. The semimajor axis distribution is also quite flat, ranging from about 0.5 to 5 AU. Mean and standard deviations are 1.98 and 1.27 AU, respectively. Eccentricities are remarkably moderate to high, spanning from 0.3 to 0.8 (mean and standard deviation are 0.49 and 0.18, respectively). Stellar masses are mainly around 1 $M_{\odot}$, with a slight overabundance of supersolar mass objects. Mean and standard deviations are 1.06 and 0.13 $M_{\odot}$, respectively. Metallicities span from −0.25 to 0.3 dex, having an average and standard deviation of 0.14 and 0.17 dex, respectively. Thus, despite its wide distribution of metallicities, EPs are mainly supersolar.

Two significant correlations exist for this cluster: $M_p-e$ and $M_p-M_s$ (see Fig. 9). The first implies that lower mass EPs have higher $e$-values; thus, the mechanisms for the pumping-up of the eccentricity are more active in low-mass planets, at least for the high semimajor axes and moderate positive metallicities of this cluster. Moreover, EP masses are correlated with stellar masses. This may confirm the fact that a higher $M_s$ implies a larger protoplanetary disk surface density and hence a larger $M_p$ (Ida & Lin 2005).

C4 contains 12 EPs in MSSs and 12 in MPSs (see Tables 1 and 2). Note that the MSS planets may be responsible of the $M_p-e$ correlation, as many of them have a low $M_p$ and high $e$.

4.5. Cluster C5

This cluster contains 31 EPs (see Tables 1 and 2), and the prototype is HD 117207b. Planetary masses have intermediate values, with a mean of 2.16 $M_J$ and a standard deviation of 1.24 $M_J$, respectively. The semimajor axis distribution is rather flat, spanning from 0.37 to 3 AU, with a few bodies around 4 AU. Eccentricities are peaked at 0.2–0.3 and range from 0.2 to 0.5. Stellar masses are supersolar, having a mean and standard deviation of 1.22 and 0.21 $M_{\odot}$, respectively. Stellar metallicities are also remarkably supersolar, having a mean and standard deviation of 0.15 and 0.15 dex, respectively (see Fig. 5).

The formally significant correlations are $M_p-M_s$, $M_p-[Fe/H]$, and $a-e$. However, they are all due to outlier planets and thus cannot be considered real (note that if we do not consider the outliers, the correlation $a-e$ is very close to the 5% significance level; thus, it may become significant as more objects are added to this cluster by future discoveries).

C5 contains seven EPs in MSSs and only one hot Jupiter (namely, HD 118203b). It is interesting to understand why this hot Jupiter is in this cluster. The peculiarity of HD 118203b is that it has the highest eccentricity (0.309) among hot Jupiters. Note that HD 185269b (in C3) also has a very high eccentricity.
of 0.3. All the other input variables are the same, except for the planetary mass. HD 118203b, with a mass of 2.13 $M_p$, is one of the most massive hot Jupiters. This explains why HD 118203b has been put in this cluster. This cluster contains 13 EPs in MPSs. Despite the lack of any correlation, few comments may be added to uncover the nature of this cluster. C4 and C5 have some similar traits. They have spread and rather flat distributions of input variables. Moreover, both have large semimajor axes and eccentricities. However, C5 contains objects with higher stellar masses, lower eccentricities, and lower masses than C4.

5. DISCUSSION AND CONCLUSION

In this paper we develop the basis for an extrasolar planet taxonomy. We use as many inputs as possible for this analysis, in particular the planetary mass, semimajor axis, eccentricity, stellar mass, and stellar metallicity. We identify the best procedure to follow: a multivariate statistical analysis (PCA) to find the most important variables and then hierarchical cluster analysis. We analyze the solutions via canonical means (through the cophenetic coefficient) and by analysis of the distribution of the dendrogram’s heights. The best result is achieved with nontraditional 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 test the stability of the solution against observational errors of the input variables. The procedure we followed is able to provide a robust extrasolar planet taxonomy even if the number of planets is still low. 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 ($M_\ast$ and [Fe/H]) to discriminate between otherwise similar planets, and also to merge together different bodies (like EPs in MSSs and orbiting single stars). For instance, we were able to divide the hot Jupiters into—at least—two main groups (see Table 1). This division is mainly due to the stellar mass and metallicity. Those belonging to C3 basically orbit around stars with supersolar masses and high metallicities; those of C2 orbit mostly subsolar mass stars with moderate (both positive and negative) metallicities. This may reflect differences in the formation processes of these EPs.

Jupiter belongs to cluster C3. Much has been speculated about the similarity of our solar system and extrasolar systems (e.g., Beer et al. 2004), in particular concerning the formation histories. With the help of cluster analysis we may identify those EPs that are more similar—in the threefold clustering space—to Jupiter. We suggest that the actual large semimajor axis of Jupiter is the result of its formation in a solar-like metallicity disk.

We also analyzed the intracluster 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 that may be hidden in the “blind” statistical analysis performed on the whole EP database. The most important correlations found are those for C2, C3, and C4 (see Table 1). Remarkably, for C2 and C3 we find important trends between metallicity and orbital parameters. We find that [Fe/H] has very important effects on the semimajor axis (and thus on the migration processes) and the eccentricity. It may also happen that the same variables correlate in an opposite way between two different clusters (see the $M_p$-$M_\ast$ correlations for C1 and C4). 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.

In addition to these five main clusters, we may see the position of the pulsar planets in the clustering space. Obviously, these planets were not included in the previous analysis because we do not have $M_\ast$ and [Fe/H]. However, we may use as test values $M_\ast = 10 M_\odot$ for the progenitors of both PSR 1257+12 and PSR B1620–26. As for the metallicity, we assume 0 and $-1.05$ (the first is an indicative value, the latter is the average for M4 stars), respectively. Using equations (1)–(3), we find that these planets are very far from all the other EPs in the clustering space and hence for each pulsar we have a single cluster. This is consistent with the very likely different origin of the pulsar planets with respect to other EPs.

The author wishes to thank C. Barbieri for a careful reading of the paper and financial support. The paper has been funded on MIUR grant PRIN 2006. Many thanks to S. Ortolani and P. Paolicchi for helpful comments and discussion, which improved the quality of the paper, and to the anonymous referee for useful comments. I wish to also thank N. Schneider (on sabbatical at the University of Padova) for discussions. Thanks to S. Magrin for IDL support and M. Clemens for corrections to the English.
REFERENCES

Adams, F. C., & Laughlin, G. 2006, ApJ, 649, 992
Beer, M. E., King, A. R., Livio, M., & Pringle, J. E. 2004, MNRAS, 354, 763
Boss, A. P. 2002, ApJ, 567, L149
Burrows, A., Hibeny, I., Budaj, J., & Hubbard, W. B. 2007, ApJ, 661, 502
D’Angelo, G., Lubow, S. H., & Bate, M. R. 2006, ApJ, 652, 1698
Desidera, S., & Barbieri, M. 2007, A&A, 462, 345
Eggenberger, A., Udry, S., & Mayor, M. 2004, A&A, 417, 353
Everitt, B. S., & Dunn, G. 2001, Applied Multivariate Data Analysis (2nd ed.; London: Arnold)
Everitt, B. S., Landau, S., & Leese, M. 2001, Cluster Analysis (4th ed.; London: Arnold)
Fischer, D. A., & Valenti, J. 2005, ApJ, 622, 1102
Ford, E. B. 2005, AJ, 129, 1706
Fregeau, J. M., Chatterjee, S., & Rasio, F. A. 2006, ApJ, 640, 1086
Goldreich, P., & Sari, R. 2003, ApJ, 585, 1024
Halbwachs, J. L., Mayor, M., & Udry, S. 2005, A&A, 431, 1129
Holman, M., Touma, J., & Tremaine, S. 1997, Nature, 386, 254
Ida, S., & Lin, D. N. C. 2005, ApJ, 626, 1045
Jiang, I.-G., Yeh, L.-C., Hung, W.-L., & Yang, M.-S. 2006, MNRAS, 370, 1379
Livio, M., & Pringle, J. E. 2003, MNRAS, 346, L42
Matsumura, S., & Pudritz, R. E. 2006, MNRAS, 365, 572
Mayor, M., & Queloz, D. 1995, Nature, 378, 355
Milligan, G. W., & Cooper, M. C. 1985, Psychometrika, 50, 159
Mojena, R. 1977, Computer J, 20, 359
Murray, N., Paskowitz, M., & Holman, M. 2002, ApJ, 565, 608
Namouni, F. 2005, AJ, 130, 280
Santos, N. C., Benz, W., & Mayor, M. 2005, Science, 310, 251
Santos, N. C., Israelian, G., & Mayor, M. 2004, A&A, 415, 1153
Santos, N. C., Israelian, G., Mayor, M., Rebolo, R., & Udry, S. 2003, A&A, 398, 363
Sari, R., & Goldreich, P. 2004, ApJ, 606, L77
Sozzetti, A. 2004, MNRAS, 354, 1194
Udry, S., Mayor, M., Naef, D., Pepe, F., Queloz, D., Santos, N. C., & Burnet, M. 2002, A&A, 390, 267
Udry, S., Mayor, M., & Santos, N. C. 2003, A&A, 407, 369
Zakamska, N. L., & Tremaine, S. 2004, AJ, 128, 869
Zucker, S., & Mazeh, T. 2002, ApJ, 568, L113