.5ex 2em

HAZARDOUS ASTEROIDS FORECAST VIA MARKOV RANDOM FIELDS

Project for the course Probabilistic modelling (DSE)

Marzio De Corato

This day may possibly be my last: but the laws of probability, so true in general, so fallacious in particular, still allow about fifteen years.
Edward Gibbon (1737-1794)

ABSTRACT

CONTENTS

1 INTRODUCTION

2 THEORETICAL FRAMEWORK

In this section we are going to review the theoretical concepts that underlies to the probabilistic methods here used: we will expose them following the approaches of Murphy [21], Koller et al. [15], Højsgaardand [14] et al. and Russel et al. [25]. Furthermore we will provide also a rapid overview of the main concepts of information theory, following the Cover [6] and MacKay [20] approaches , since we used some of its concepts in the preliminary analysis of the dataset. On the other side the concepts related to the celestial mechanics here used will be described, following the Murray approach [22] into the Appendix A

2.1 MARKOV RANDOM FIELDS

Lets start by supposing that we would represent compactly a joint distribution such as [21]:

$$p(x_1, x_2, ..., x_n) \tag{2.1}$$

that can represent for instance words in a documents or pixels of an image. Firstly we know that using the chain rule, we can decompose it, into the following form [21]:

$$p(x_{1:V}) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1)...p(x_V|x_{1:V-1}) \tag{2.2}$$

where V is the number of variables and 1:V stands for $1, 2, ..., V$. This decomposition makes explicit the conditional probability tables, or in other terms the transition probability tensors [29]. As one can point out the number of parameter is cumbersome as the number of variables grows: indeed the number of parameter required scales as $\mathcal{O}(K^V)$. Such formidable problem can be attacked by considering the concept of conditional independence. This is defined as [21]:

$$X \perp Y|Z \iff p(X, Y|Z) = p(X|Z)p(Y|Z) \tag{2.3}$$

A particular case of this definition is the Markov assumption, by which *the future is independent from the past given the present* or in symbols [21]:

$$p(\mathbf{x}_{1:V}) = p(x_1) \prod_{t=1}^{V} p(x_t|x_{t-1}) \tag{2.4}$$

In this case a first order Markov chain is obtained, where the transition tensor is of second order [29]. Given this formalism we are interested in finding a smart way to plot such joint distribution into an intuitive way: the graph theory provide the answer to this quest. In particular the random variables can be represented by nodes and presence of conditional independence for two random variables by the lack of an edge that interconnects them. Bayesian networks consider directed edges, while Markov random fields (MRF) only undirected. As consequence, while the the concept of topological ordering, by which the parents n nodes are labelled with a lower with respect to their children, is well defined for Bayesian network, for MRF is not. In order to solve this issue it is useful to consider the Hammersley-Clifford theorem as stated in [21]:

Theorem 1 (Hammersley-Clifford)**.** *A positive distribution p(**y**)>0 satisfies the CI properties of an indirect graph G iif p can be represented as a product of factor, one per maximal clique, i.e.*

$$p(\boldsymbol{y}|\theta) = \frac{1}{Z(\theta)} \prod_{c \in C} \psi_c(\boldsymbol{y}_c|\theta_c) \tag{2.5}$$

where C is the set of all the (maximal) cliques of G, and Z(θ) is the partition function given by

$$Z(\theta) := \sum_{y} \prod_{c \in C} \psi_c(\boldsymbol{y}_c|\theta_c) \tag{2.6}$$

Note that this partition function is what ensures the overall distribution sums to 1

Such theorem allows to represent a probability distribution with potential functions for each maximal clique in the graph. A particular case of these is the Gibbs distribution [21]:

$$p(y|\theta) = \frac{1}{Z(\theta)} \exp\left(-\sum_{c} E(y_c|\theta_c)\right) \tag{2.7}$$

where $E(y_c) > 0$ represent the energy associated with the variables in the clique c. This form can be adapted to a UGM with the following expression [21]:

$$\psi_c(y_c|\theta_c) = exp\left(-E(y_c|\theta_c)\right) \tag{2.8}$$

Finally in order to reduce the computational cost, one can consider only the pairwise interaction instead of the maximum clique. This is the analogue of what is usually performed in solid state physics (but surely not always) when only the interaction between the first neighbour is considered. Another example is the Ising model: here we have a lattice of spins that can be or in $|+\rangle$ or in $|-\rangle$ and their interaction is modelled by[21]:

$$\psi_{st}(y_s, y_t) = \begin{pmatrix} e^{w_{st}} & e^{-w_{st}} \\ e^{-w_{st}} & e^{w_{st}} \end{pmatrix} \tag{2.9}$$

where $w_{st} = J$ represent the coupling strength between two neighbour site. The collective state is described by

$$|i_1, i_2, ..., i_n\rangle = |i_1\rangle \otimes |i_2\rangle \otimes ... \otimes |i_n\rangle \tag{2.10}$$

where $\otimes$ is the tensor product. If this parameter is associated with a positive finite value we have an associative Markov network: basically collective states in which all sites have the same configuration is favoured. Thus we will have two collective states: one for which we have all $|+\rangle$ and another in which we have all $|-\rangle$ Such situation would model, in principle, the ferromagnet materials where the external magnetic field induce into the material a magnetic filed with the same direction. On the other side if the magnetization of the material is opposite with respect to the external field, and thus $J < 0$, we have an anti-ferromagnetic system in which frustrated states are present. Furthermore lets consider the unnormalized log probability of a collective state $\mathbf{y} = |i_1, i_2, ..., i_n\rangle$ [21]:

$$\log \tilde{\mathbf{p}}(y) = -\sum_{s\sim t} y_s w_{st} y_t \tag{2.11}$$

If we also consider an external field [21]:

$$\log \tilde{\mathbf{p}}(y) = -\sum_{s\sim t} y_s w_{st} y_t + \sum_s b_s y_s \tag{2.12}$$

But this is nothing more that the well know [1] Hamiltonian of an Ising system. This is not a simple coincidence: indeed the Hamiltonian of a system represent, rudely speaking, its total

1 In physics

energy. Thus according to the Boltzmann or Gibbs distribution we have [21]:

$$P_\beta(\mathbf{y}) = \frac{e^{-\beta H(\mathbf{y})}}{Z_\beta} \tag{2.13}$$

where β is proportional to the inverse of the system temperature. Coming back the unnormalized probability of a collective state $\mathbf{y}$, if we set $\Sigma^{-1} = \mathbf{W}$, $\mu = \mathbf{\Sigma b}$ and $c = \frac{1}{2}\mu^T\mathbf{\Sigma}^{-1}\mu$ we obtain a Gaussian [21]:

$$\tilde{\mathbf{p}}(y) \sim exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) + c\right) \tag{2.14}$$

In general we refer to Gaussian Markov random fields for a joint distribution that can be decomposed in the following way [21]:

$$p(\mathbf{y}|\theta) \propto \prod_{s\sim t}\psi_{st}(y_s, y_t)\prod_t \psi_t(y_t) \tag{2.15}$$

$$\psi_{st}(y_s, y_t) = exp\left(-\frac{1}{2}y_s\Delta_{st}y_t\right) \tag{2.16}$$

$$\psi_t(y_t) = exp\left(-\frac{1}{2}\Delta_{tt}y_t^2 + \eta_t y_t\right) \tag{2.17}$$

$$p(\mathbf{y}|\theta) \propto exp\left(\boldsymbol{\eta}^T\mathbf{y} - \frac{1}{2}y^T\Delta\mathbf{y}\right) \tag{2.18}$$

(this last expression can be reconducted to the multivariate gassian if one consider $\boldsymbol{\Delta} = \mathbf{\Sigma}^{-1}$ and $\boldsymbol{\eta} = \boldsymbol{\Delta\mu}$. Given the network, we would now move on how the parameters can be achieved. Lets start from a Markov random field in log-linear form [21]:

$$p(\mathbf{y}|\theta) = \frac{1}{Z(\theta)}exp\left(\sum_c \theta_c^T\phi_c(\mathbf{y})\right) \tag{2.19}$$

thus we can define the log-likehood as [21]:

$$\mathcal{L}(\theta) := \frac{1}{N}\sum_i \log p\left(\mathbf{y}_i|\theta\right) = \frac{1}{N}\sum_i\left[\sum_c \theta_c^T\phi_c(y_i) - \log Z(\theta)\right] \tag{2.20}$$

$$\frac{\partial\mathcal{L}}{\partial\theta_c} = \frac{1}{N}\sum_i\left[\phi_c(y_i) - \frac{\partial}{\partial\theta_c}\log Z(\theta)\right] \tag{2.21}$$

$$\frac{\partial \log Z(\theta)}{\partial \theta} = \mathbb{E}\left[\phi_c(\mathbf{y})|\,\theta\right] = \sum_{\mathbf{y}} \phi_c(\mathbf{y}) p(\mathbf{y}|\theta) \tag{2.22}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_c} = \left[\frac{1}{N}\sum_i \phi_c(y_i)\right] - \mathbb{E}\left[\phi_c(\mathbf{y})\right] \tag{2.23}$$

In the first term $\mathbf{y}$ is fixed to its observed values while in the second it is free. Such expression can be recasted in to a more explicative form [21]:

$$\frac{\partial \mathcal{L}}{\partial \theta_c} = \mathbb{E}_{p_{emp}}\left[\phi_c(\mathbf{y})\right] - \mathbb{E}_{p_{(\cdot|\theta)}}\left[\phi_c(\mathbf{y})\right] \tag{2.24}$$

Therfore at the optimum we will have [21]:

$$\mathbb{E}_{p_{emp}}\left[\phi_c(\mathbf{y})\right] = \mathbb{E}_{p_{(\cdot|\theta)}}\left[\phi_c(\mathbf{y})\right] \tag{2.25}$$

From this expression it is clear why this method is called moment matching; it is worth nothing tha such computation is largely expensive from a computational point of view: thus scholar usually consider other techinques or at least stochastic gradient descent method. A full review can be found in [21] and [15]. Finally we consider, as for the dataset in analysed in this work, the case where we have both discrete and continuous variables i.e. $x = (i_1, ..., i_d, y_1, ..., y_q)$ with d discrete variable and q continuos variables. The are called in the literature Mixed Interaction Models. In this case the following density has to be considered [14]:

$$\begin{aligned} f(i, y) =& p(i)(2\pi)^{-q/2}\det(\Sigma)^{-1/2} \\ & \exp\left[-\frac{1}{2}(y - \mu(i))^T \Sigma^{-1}(y - \mu(i))\right] \end{aligned} \tag{2.26}$$

Which can be rewritten in the exponential family form [14]:

$$\begin{aligned} f(i, y) &= \exp\left\{g(i) + \sum_u h^u(i) y_u - \frac{1}{2}\sum_{uv} y_u y_v k_{uv}\right\} \\ &= \exp\left\{g(i) + h(i)^T y - \frac{1}{2} y^T K y\right\} \end{aligned} \tag{2.27}$$

where $g(i)$, $h(i)$ and K are the canonical parameters. These are connected with the parameters of expression 2.26 by the following identities [14]:

$$\begin{aligned} K &= \Sigma^{-1} \\ h(i) &= \Sigma^{-1}\mu(i) \\ g(i) &= \log p(i) - \frac{1}{2}\log\det(\Sigma) \\ &\quad - \frac{1}{2}\mu(i)^{\top}\Sigma^{-1}\mu(i) - \frac{q}{2}\log 2\pi \end{aligned} \tag{2.28}$$

Moreover one can further modify the previous form in order to obtain a particular factorial expansion: such models are referred as homogeneous mixed interaction models [14].

2.2 INFORMATION THEORY

Given an ensemble of random variable, we can find the amount of information that one variable contains of another one: such quantity is called mutual information and it is a key concept within the information theory. This approach, that was implemented by Claude Shannon decades before the probabilistic modelling, represent a complementary way by which one can attack the problem of conditional dependence between random variables. Here we would provide some basic concepts of this theory, following the Cover [6] and MacKay [20] approaches, that allows to properly define the concept of mutual information. The starting concept of information theory is the entropy which express the uncertainty of a random variable. Given a random variable X with alphabet (the accessible states) [6] : and probability mass function $p(x) = Pr\{X = x\}\ x \in \chi$ we define the entropy of X as $H(X) = -\sum_{x\in\chi} p(x)\log p(x)$ where the logarithm has to be considered with basis 2. In analogous way the joint entropy of two random variables (X, Y) with a joint distribution p(x,y) is defined as [6]:

$$H(X,Y) = -\sum_{x\in\chi}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \tag{2.29}$$

Furthermore, we can define also the conditional entropy as [6]:

$$\begin{aligned} H(X|Y) &= \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) \\ &= -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\ &= -\,E\log p(Y|X) \end{aligned} \tag{2.30}$$

The joint entropy and the conditional entropy are related by the chain rule [6]:

$$H(X,Y) = H(X) + H(Y|X) \tag{2.31}$$

Such rule can be extended to to the following from [6]:

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z) \tag{2.32}$$

Given a distribution q and another distribution p, one can quantify how inefficently the second one describe the first one using the concept of relative entropy or Kullback-Leibler distance [6, 20]:

$$D(p||q) = \sum p(x) \log \frac{p(x)}{q(x)} \tag{2.33}$$

As stated by the Gibbs inequality [6, 20]:

$$D(p||q) \geqslant 0 \tag{2.34}$$

this quantity can not be negative: the entropy of a random variable associated to another cannot have a degree of uncertainty lower with respect to the quantity that its aimed to describe. On these basis we are now ready to introduce the concept of mutual information. This is defined as [6]:

$$\begin{aligned} I(X;Y) &= \sum (x,y) \log \frac{p(x,y)}{p(x)p(y)} = \\ &= D(p(x,y)||p(x)p(y)) \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X) \end{aligned} \tag{2.35}$$

As for the joint distribution also in this case we have a chain rule [6]:

$$I(X_1, X_2, ..., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, ..., X_1) \tag{2.36}$$

Finally we would report the data process inequality theorem that connects the information theory with the Markov chain: if we have a Markov chains, $X \to Y \to Z$ then $I(X;Y) \geqslant I(X;Z)$. As for the Gibbs inequality, the underlying idea is that no clever manipulation of the data can improve the inference that can be made on them [6, 20]. Otherwise we would have a clear violation of the second principle of thermodynamics (see for instance the Maxwell's demon [11])

3 DATASET DESCRIPTION

The asteroid dataset was retrieved from Kaggle [1], which reports into an more user-friendly mode the dataset of The Center for Near-Earth Object Studies (CNEOS) [2], a NASA research center. Among the 40 features there were present in the dataset we excluded the ones that were clearly redundant (such as the distance quantity evaluated in miles instead of kilometres), the ones that were connected only to the other name of the asteroid, the one connected to the name of the orbit and the one connected with the orbiting planet, since for all it was the Earth. Thus the features obtained were reduced to 22. Their enumeration and their description will be postponed to Appendix A, since a description of them cannot be disjoint with celestial mechanics concepts. Since only the 16 % of the asteroids were hazardous, we considered to reduce the number of the non-hazardous in order to improve the portion of the hazardous ones. For this purpose we constructed a database in which the proportion hazardous/not hazardous was 1:5: thus all the hazardous one were included, while the not-hazardous were randomly selected [1]. Furthermore, concerning the continuous measures, the dataset was standardised and demeaned.

1 The author is aware that in principle the proportion should be much less unbalanced, however the tool that has to be paid for this operation is an overall reduction of the cases in the dataset that lowers the performances of the algorithms here used. The proportion here used was a trade-off for these two contrasting facts

4 RESULTS

This chapter is organized in the following way: first we are going to report the results concerning the preliminary analysis performed on the dataset. This include the factor analysis of mixed data and the mutual information analysis of the continuous variables of asteroids vs their hazard. Then it will follow the analysis of the dataset performed with the probabilistic methods: after a preliminary analysis on the continuous variables, the mixed interaction model as well the minforest model obtained for the whole dataset will be presented and discussed. Finally the probabilistic models previously obtained will be compared with the outputs and the performances of four machine learning algorithm (Random Forest, Support vector machines Quadratic Discriminant Analysis and Logistic regression).

4.1 PRELIMINARY ANALYSIS

The first inspection that was performed on the dataset was related the density distributions of a selection of continuous features that are known, from the celestial mechanics, to be important for the prediction of the asteroids dangerousness. These are reported in Fig. 4.1. Next we moved on a more systematic analysis with FADM and mutual information. In Fig. 4.2,4.3 and 4.4 are reported the main achivements of FADM (performed with FactoMineR package [18]): in particular from the correlation circle in Fig. 4.3 the different correlations that are given by the celestial mechanics laws can be recognized: for instance, see that there is strong correlation of the mean motion with the semi-major axis. This is due to the Kepler law discussed in Appendix A. Furthermore we see that that the mean motion is correctly almost independent with respect to the the diameter (max or min) of the asteroid. As explained in Appendix A this is an another result of Newton gravitation theory for a two body interaction: the mean motion of an asteroid, as long as the the mass of it is much lower with respect to the sun,

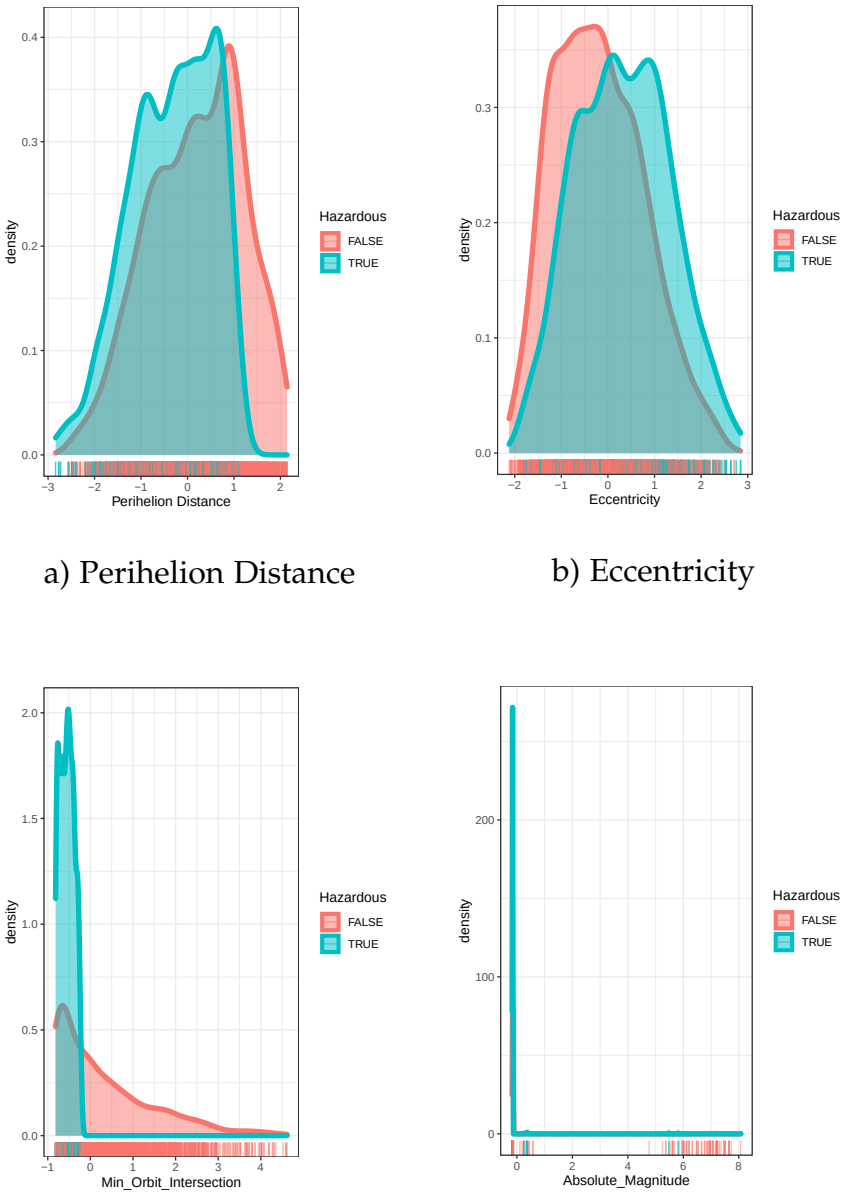
is independent from its mass. On the other side the fact that relative velocity is correlated with the diameter is a spurios correlation. At this point one can ask why the mean motion and the relative velocity are orthogonal: this point will be clarified from a theoretical point of view in appendix A and also with graphical models. Briefly the motion of an object on an ellipse as seen from a focus is not uniform: this is faster as the two bodies approaches. Beside this inspection a further analysis based on the concept of mutual information (summarized in the previous chapter) was performed: its result is reported in Fig. 4.5. This figure summarize the ranking of the features considered in the dataset for the dangerousness classification of the asteroids according to the following expression [16]

$$g(\alpha, \mathbf{C}, \mathbf{S}, f_i) = MI(f_i; \mathbf{C}) - \sum_{f_s \in S} \alpha(f_i, f_s, \mathbf{C}, \mathbf{S}) MI(f_i; f_s) \qquad (4.1)$$

where the first term $MI(f_i; \mathbf{C})$ is called relevance and measures the Mutual Information between the interesting feature set $\mathbf{C}$ (only Hazardous in our case) and the analysed one f_i; the third term $MI(f_i; f_s)$ is called redundancy and measures the MI between the analysed feature and a chosen set of them. Finally the $\alpha(f_i, f_s, \mathbf{C}, \mathbf{S})$ is a normalization function and in our case was set to [16]

$$\alpha(f_i, f_s, \mathbf{C}, \mathbf{S}) = \frac{1}{|\mathbf{S}|} \qquad (4.2)$$

following the Peng. et al approach [23]. We see that the first place in the mutual information ranking, looking to the diagonal element, is taken by minimum orbit intersection: this is correct since this is one parameter used by NASA for deciding if an asteroid is hazardous or not (see Appendix A). The second place is occupied by Epoch date close approach. The third place is the eccentricity value: this parameter is entangled with the minimum orbit intersection and thus it is reasonable that it is important. Then we have two parameter that are related to the dimension of the asteroid: this is meaningful since if the asteroid has a too reduced volume it will be destroyed by the Earth atmosphere. On the other side the other parameters seems to have a too low MI for being interesting in this preliminary analysis. The results obtained for FAMD and MI will provide a useful path-guide, together with the celestial mechanics theory, for the interpretation of the results obtained in the following section



a) Perihelion Distance

b) Eccentricity

c) Min orbit intersection

d) Semi Major Axis

Figure 4.1: Comparison between the density distributions of hazardous (red) and non-hazardous (light blue) asteroids for a selected set of features that, according to the theory, are interesting.

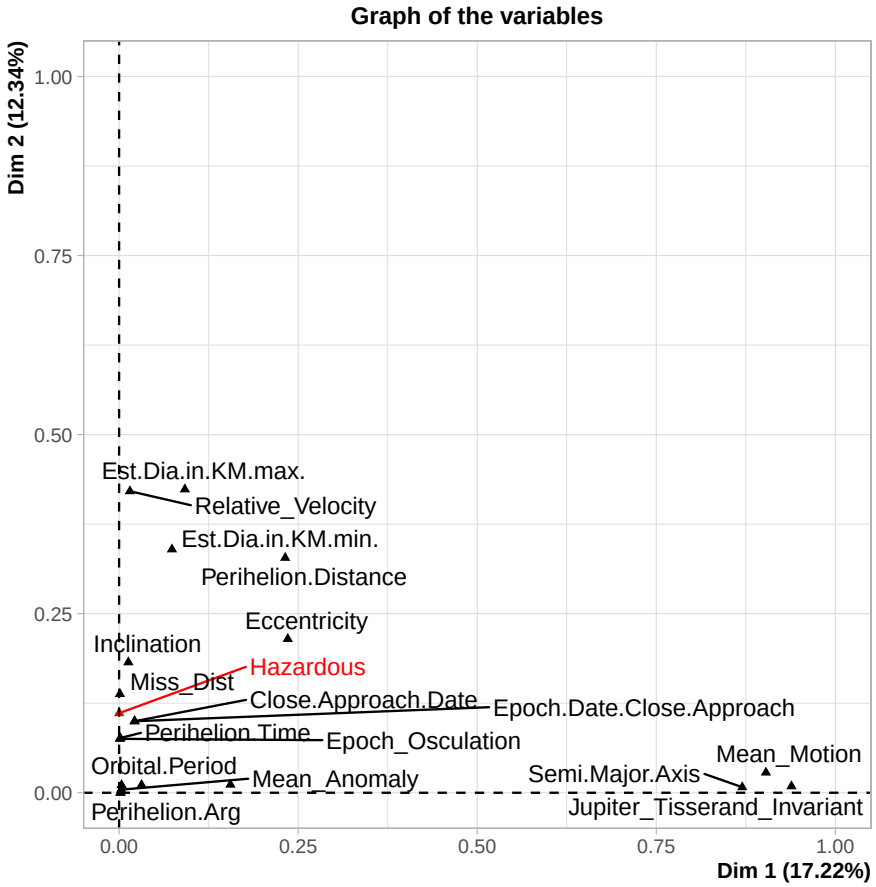


Figure 4.2: The FAMD main plot in which the correlation between the continuous and discrete variables is reported. Plot obtained from FactoMineR package [18]

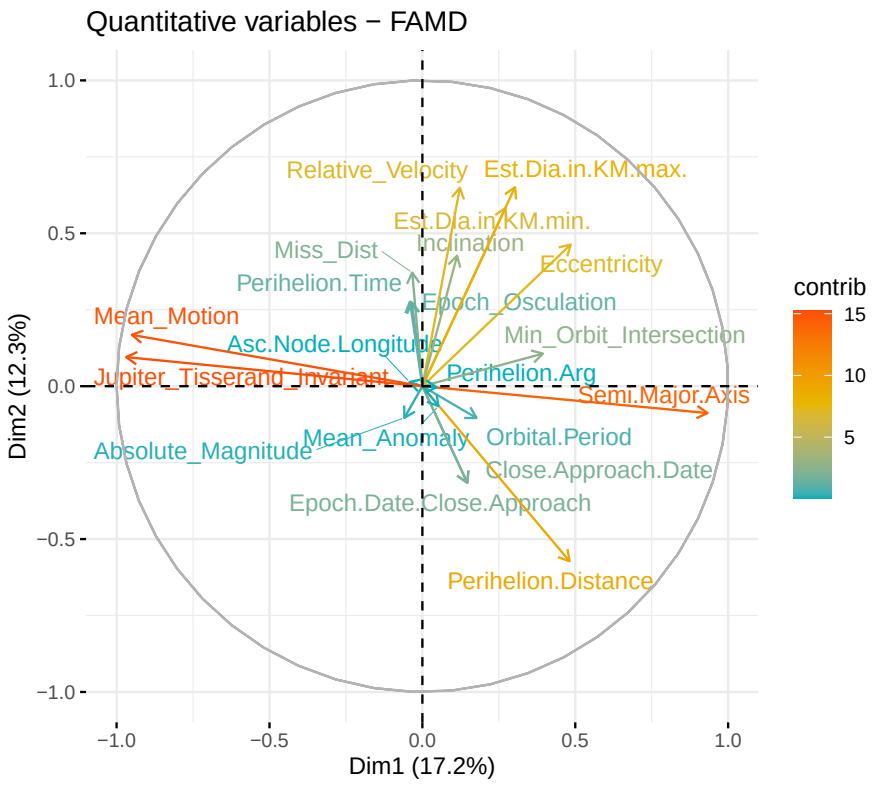


Figure 4.3: The FAMD correlation circle for continuos variables as obtained from FactoMineR package [18]

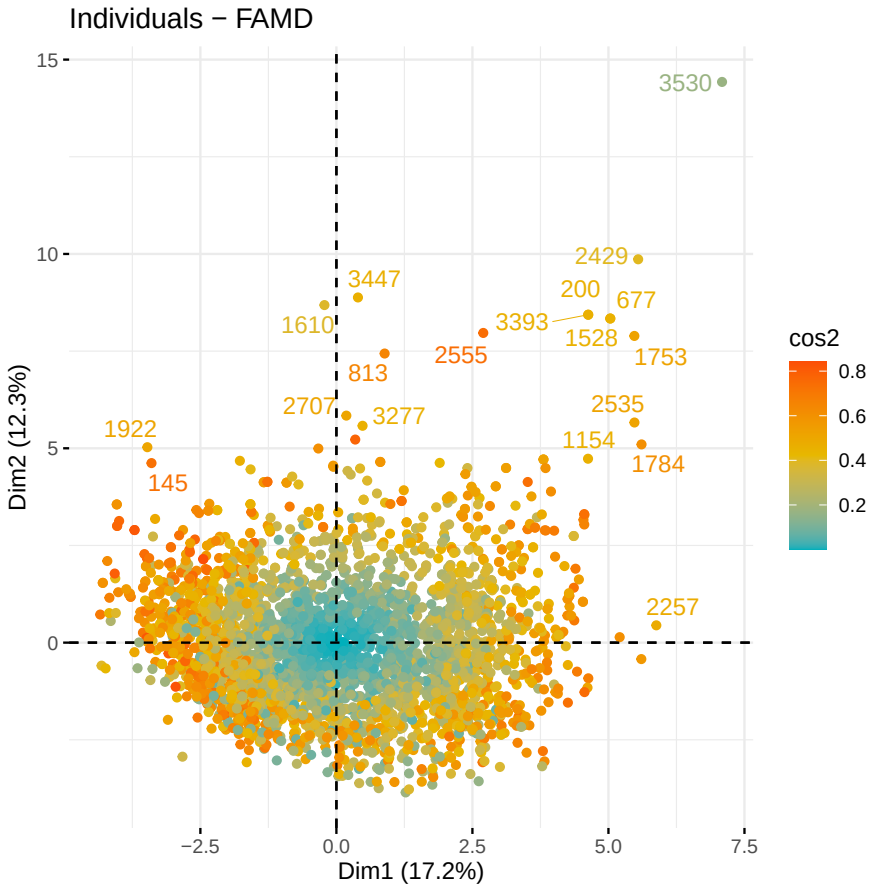


Figure 4.4: Graph of individuals, for the qualititative variables, as obtained from FactoMineR package [18]

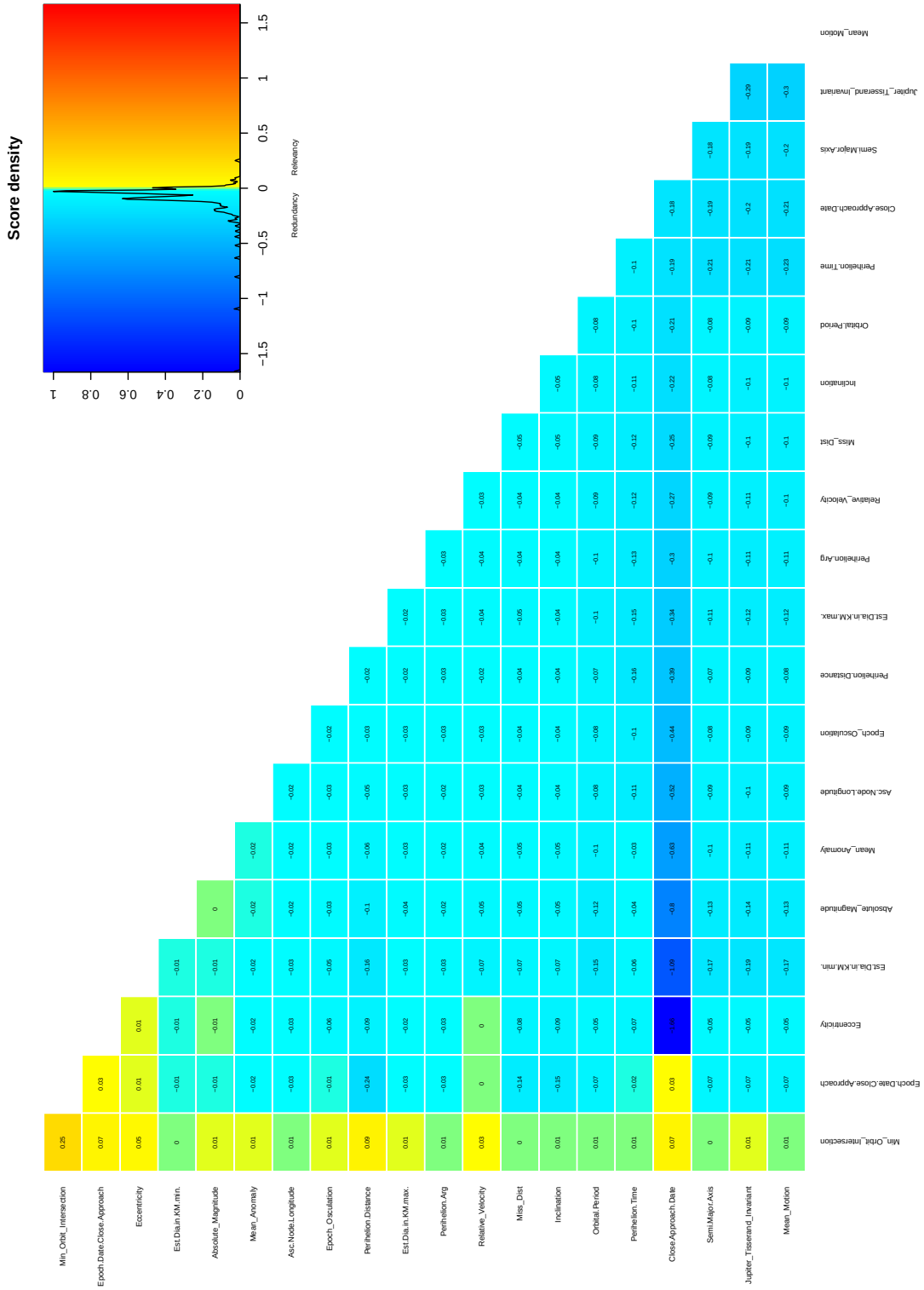
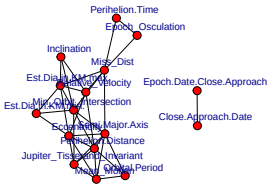


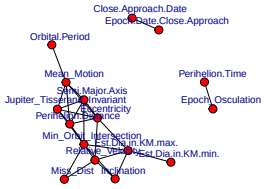
Figure 4.5: Mutual information as obtained with the varrank package [16]

4.2 PROBABILISTIC MODELS

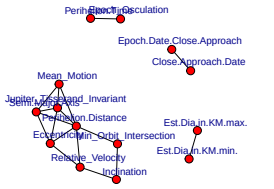
We started the analysis with the graphical models by inspecting the relations between continuous variables in the dataset: thus in this first step the binary variable *Hazardous* was excluded. For this purpose, among the different methods available, we considered the *Graphical least absolute shrinkage and selection operator* GLASSO as implemented in the *glasso* R package [12, 3]. After different tests reported in Fig. 4.6, we considered as final result the graph obtained with the value of ρ (the one that penalize further connections) equal to 0.3. This is because, in this plot, we see that the conditional dependences/independences stated by the Celestial Mechanics are correctly reproduced: for instance we see that the diameter of the asteroids is independent from all features related to its motion. This is definitely meaningful since the mass of the asteroids is largely lower with respect to the mass of earth: thus there is no way by which the asteroid orbit can be modified by its mass as explained in Appendix A. Furthermore we see that the Close approach Date and the Epoch date are dependant each other but in no way from the other features: this is right since the date and epoch are set with an arbitrary scale. Also the Perihelion Epoch and Osculation Time are dependant, as expected, but there is no dependence with the orbit parameters. We see that the mean motion is conditionally independent with respect to relative velocity, but correctly this is dependent from perihelion distance. Lets now move on the mixed interaction model: first we considered its implementation in the mgm package [4, 13]. We choose as k parameter 2 and a cross validation (CV) with ten folds. The result is reported in Fig. 4.7. From this figure we see that the features that are conditionally dependant to the Hazardous features are: the absolute magnitude, the min orbit intersection, and the eccentricity. All these connections are supported by the theory: the min orbit intersection is the main parameter for the hazardous value, the eccentricity on the other side can be tough as a parameter that describes how near the celestial body moves at the perihelion (indeed this parameter is correctly linked with the perihelion distance) and the Absolute Magnitude is also a meaningful parameter for the hazard evaluation since, if the asteroid is too small it will be destroyed by the earth atmosphere. It is worth mentioning that this parameter is correctly not connected with all quantities related to the orbital parameter except for the mean anomaly. Furthermore



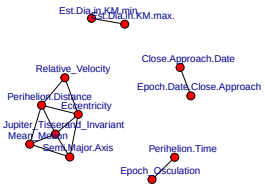
ρ=0.1



ρ=0.2



ρ=0.3



ρ=0.4

Figure 4.6: GLASSO analysis, performed with the glasso package [12, 3], with different ρ parameter (the one that penalize further connections) for the Asteroid dataset without the discrete variable Hazardous. The plots were obtained with the *igraph* package for R [7]

the parameters related to the diameters are also not connected with the orbital parameters. Furthermore we see also that the mean velocity is not directly related to the mean motion but there are the orbital parameters within them as expected. Thus it can be said that the model produced is almost consistent with the astronomic laws. On this basis, we can evaluate the performances of the model in term of confusion matrix, ROC curve and ϕ parameters (we used for this purpose the Caret R package [17, 5]). The confusion matrix obtained is reported in Fig. 4.8: the ROC curve is given in Fig. 4.9 and it corresponds to a $\phi = 0.6$. We will comment these performances in a few when we will present the performances of other algorithms for the same dataset. Before this we would show the graphical models obtained with different implementation of mixed interaction model: we considered the function *mmod()* of gRim package [14] and the *minforest()* of gRapHD (which uses the minForest

method) [8]. Both were obtained with a stepwise algorithm. Their result are reported in Fig. 4.10 and 4.11. We see that in these models the minimum diameter is linked with the hazordous feature, and this is correct, but what is without a physical meaning is that this quantity is also linked to the features connected with the orbital parameters. Indeed, while for the magnitude (as for mgm model) a weak relationship can be admitted since this depends also on the distance between the asteroids and the observer and from the orbit, a volume and thus a mass dependence is not acceptable. In principle one can set these link as forbidden in the algorithm (blacklist), but in the author view it is preferable a model were the result is obtained without boundaries, instead of a one obtained with a large number of boundaries. Thus since a correct model was obtained without the imposition a blacklist, the mgm, we simply reject the mmod and minforest ones in favour of the first one. As consequence this is the reason why we report the performances of mgm model only.

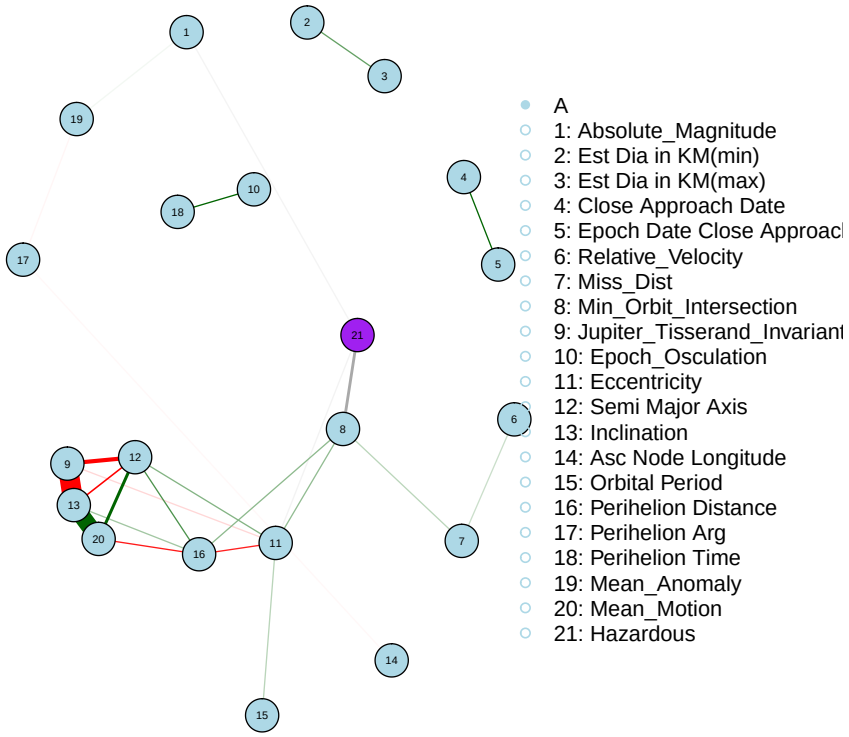


Figure 4.7: The graphical model obtained with the mixed interaction model as implemented in the mgm package [4, 13]. The plot was obtained with the [10] package

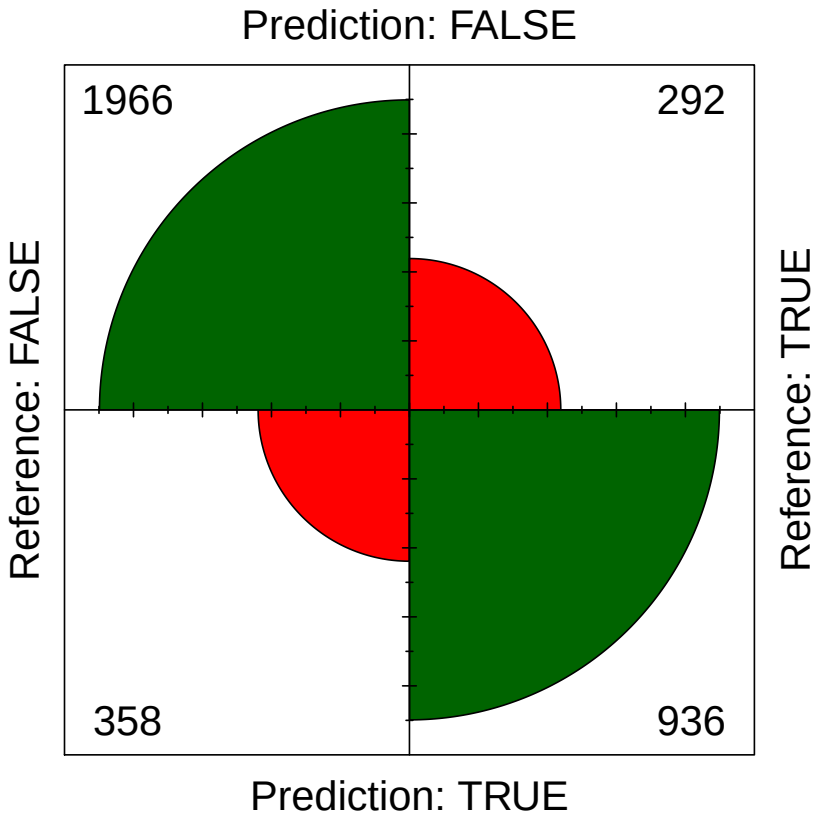


Figure 4.8: The confusion matrix of the graphical model reported in Fig. 4.7 as obtained from the Caret package [17, 5]

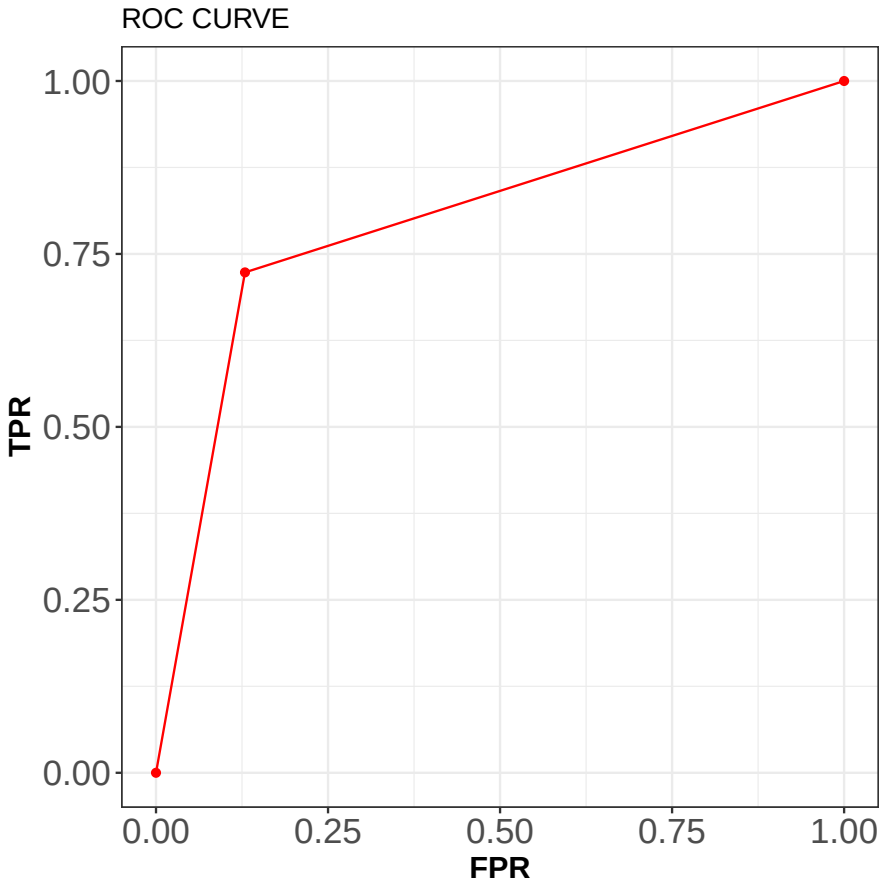


Figure 4.9: The ROC curve as obtained from the ROCR package [26] and ggplot2 [28]. The corresponding ϕ value is 0.6

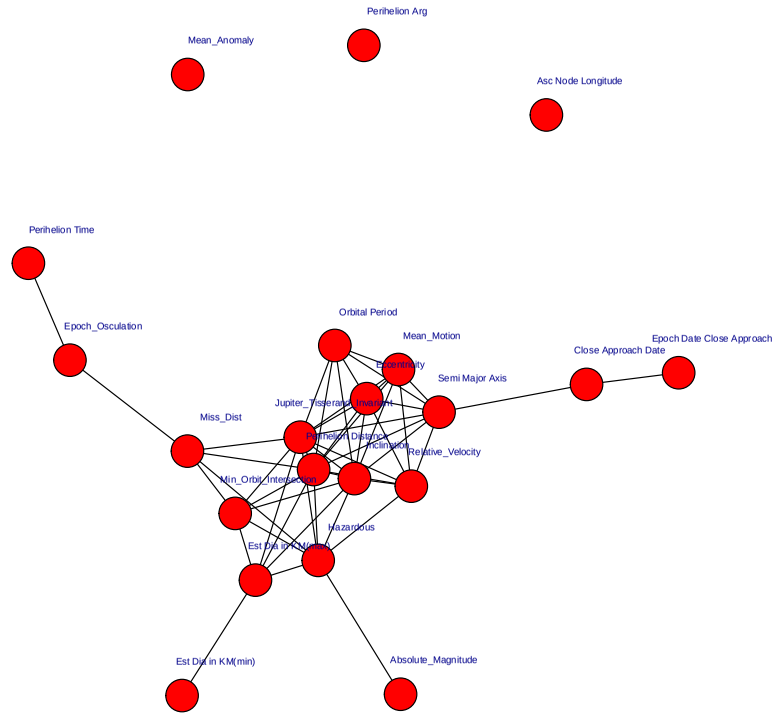


Figure 4.10: The graphical mixed interaction model obtained with the grim package [14]. The plot was obtained with the igraph package for R [7]

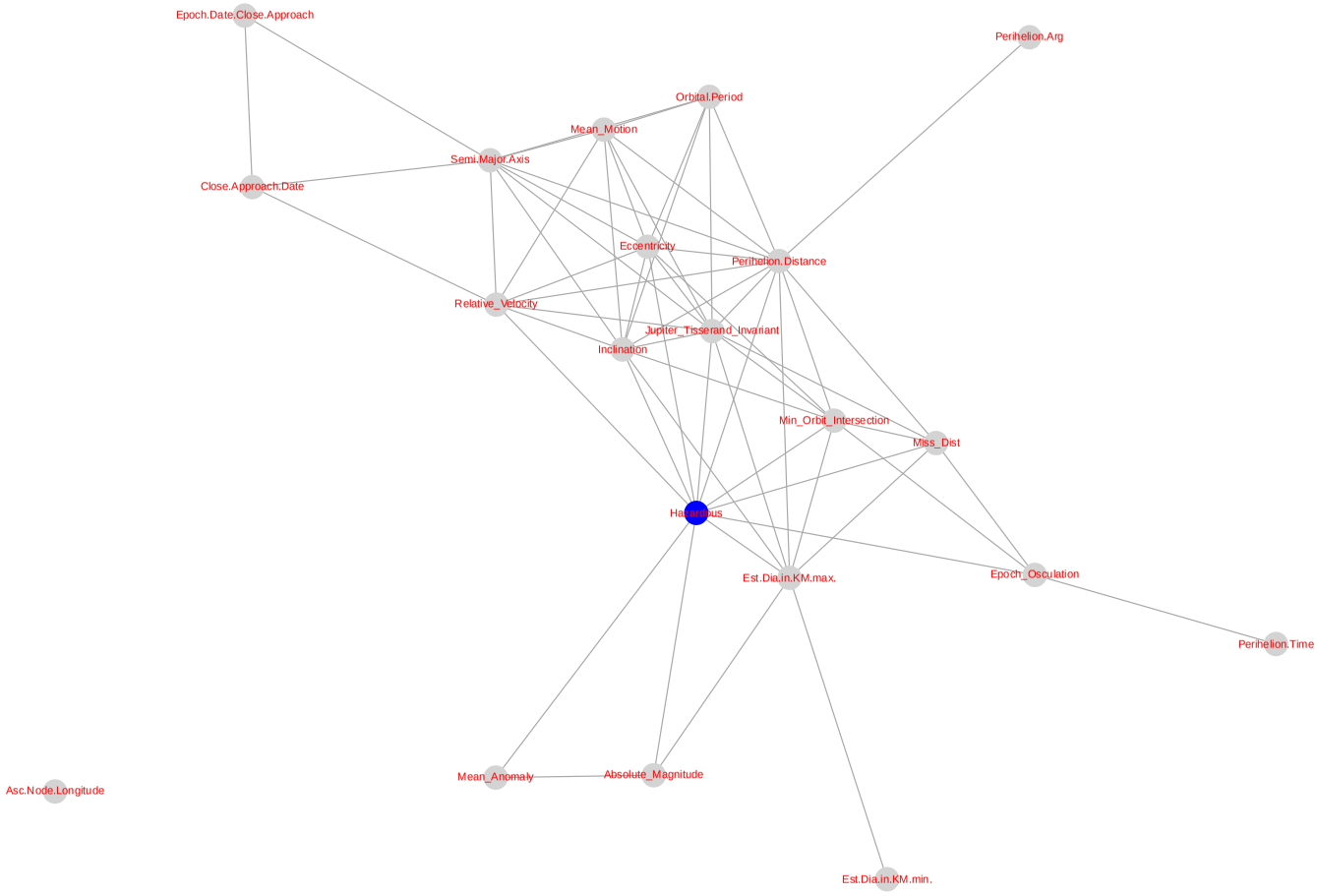


Figure 4.11: The graphical mixed interaction model obtained with the gRapHD package [8] with a stepwise algorithm. The plot was obtained with the [10] package

Table 4.1: ϕ factor for a selected set of ML algorithms as compared with the mgm

Algorithm	ϕ
RF	0.9876
SVM	0.7111
logistic	0.6173
mgm	0.5997
QDA	0.5562

4.3 MACHINE LEARNING ALGORITHMS

Given the mgm model as the one obtained with the probabilistic graphical methods, we would compare its performances with the ones obtained from the following machine learning algorithms: the random-forest (as implemented in the RandomForest package [19]), the support vector machines (as implemented in the e1071 package [9]) , the quadratic discriminant analysis (as implemented in the MASS package [27] and the logistic regression (as implented in stats package [24]). Their performances are reported in Fig. 4.12 and 4.9 as well in the Tab. 4.1. From these comparison we can see that most of these algorithms, expect for the QDA outperform the graphical method mgm or have similar performances as for the logistic algorithm. Thus at this point one can ask what is the advantage of using a graphical method instead of a random forest or a SVM since its performances seems lower. The answer is the interpretability of the model provided: the random forest, at lest, can provide a variable rank importance as reported in Fig. 4.14 (note that apart from the Min Orbit intersection, there is a slight reshuffling in the feature importance with respect to the one established by the MI in fig 4.5; however apart from this is a black-box as the SVM, the logistic regression and the QDA. Differently the probabilistic graph, providing the conditional independences, give to the user a interpretable model whose properties can be also compared, discussed and validated with the theory. Thus the model developed in this way allows a more scientific evaluation with respect to the black-box one. In the author view this characteristic definitely compensates the lack of predictive power with respect to the Random Forest or the SVM methods.

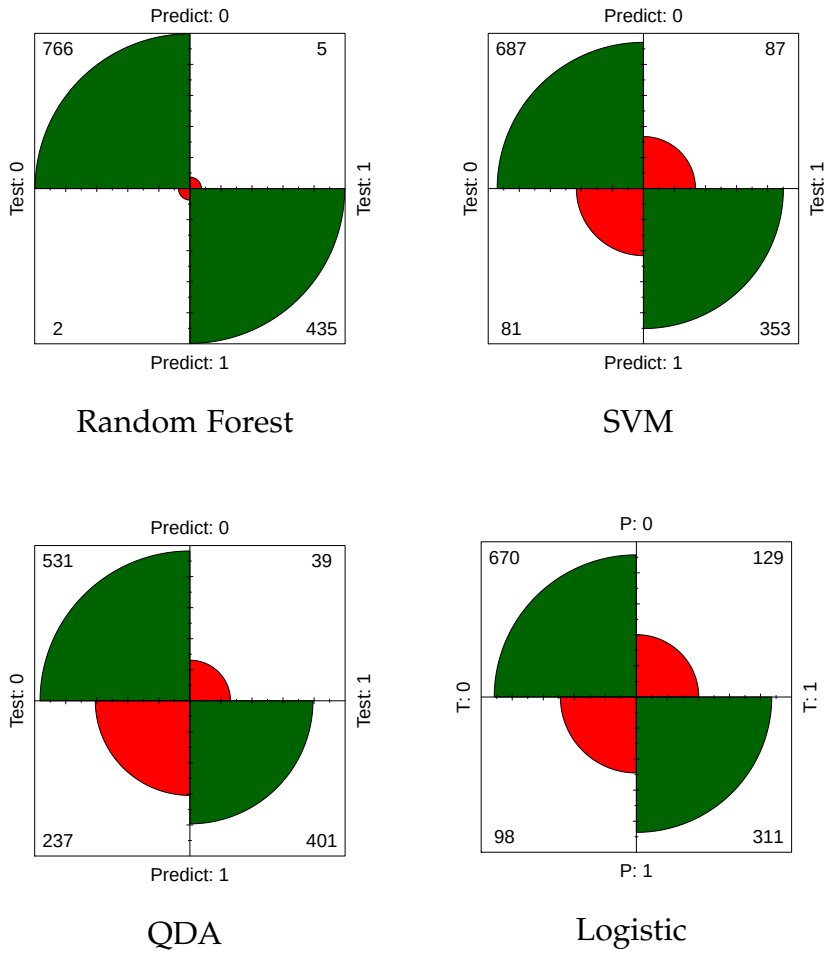


Figure 4.12: Confusion matrices for a selected set of ML algorithms as obtained from the Caret package [17, 5]

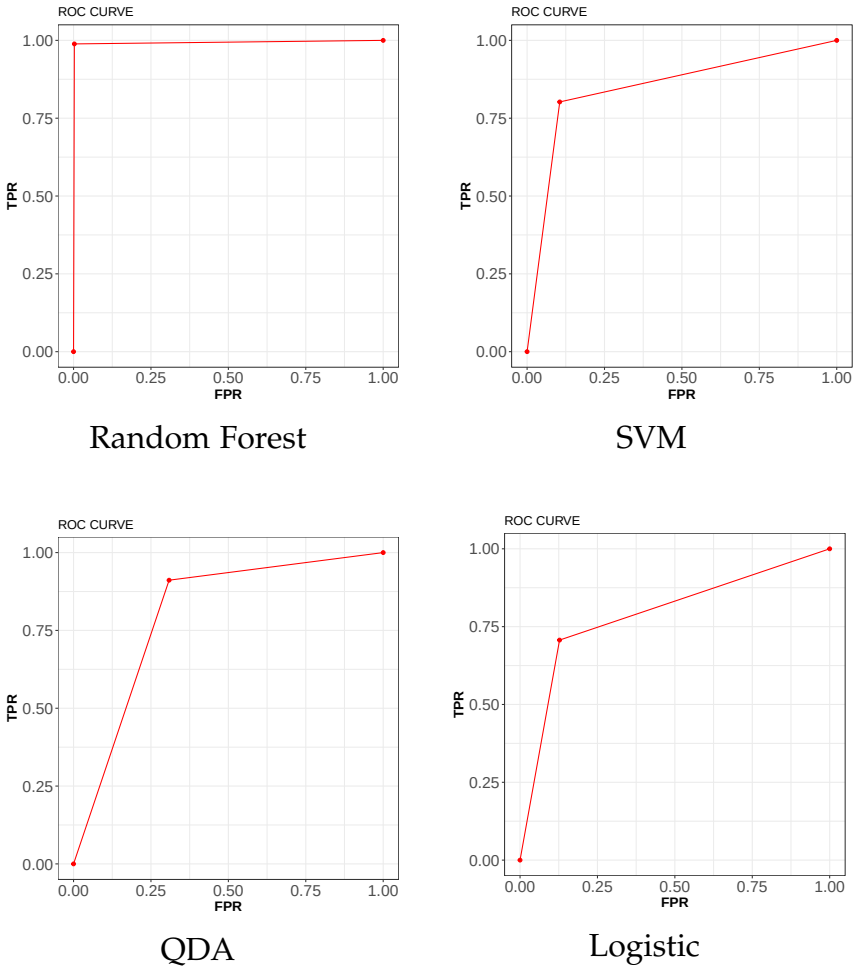


Figure 4.13: ROC curves for a selected set of ML algorithms. Plot obtained with caret and ggplot2 package [17],ggplot2

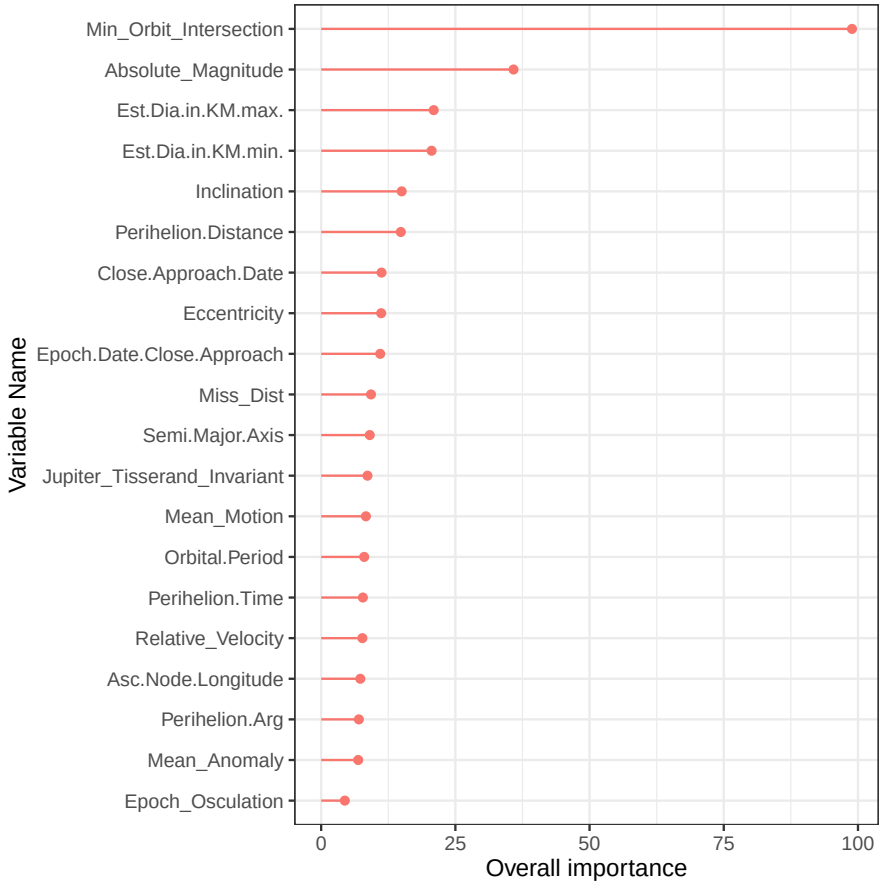


Figure 4.14: Variable importance according to the random forest algorithm as implemented in [19] package. Plot obtained with ggplot2 package [28]

5 CONCLUSIONS

6 APPENDIX A: CONCEPTS OF CELESTIAL MECHANICS

This Appendix is dedicated to review the basics concepts of celestial dynamics, following the Murray approach [22], that are needed to understand the features of the dataset and their connections. Lets start by considering two masses m_1 and m_2, which in the present case will be respectively the planet Earth and the asteroid. Their position will be given, respectively, by two vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ considering the origin O bounded in to an inertial space. Furthermore we can define also the relative position with the vector $\mathbf{r} = \mathbf{r}_2 - \mathbf{r}_1$. Since we suppose that the masses are not interacting with the electromagnetic force, these will be bounded only by the gravitational interaction which is given by the Newton law [22]:

$$\mathbf{F}_1 = \mathcal{G} \cdot \frac{m_1 m_2}{r^3}\mathbf{r} = m_1 \ddot{\mathbf{r}}_1 \tag{6.1}$$

$$\mathbf{F}_2 = -\mathcal{G} \cdot \frac{m_1 m_2}{r^3}\mathbf{r} = m_1 \ddot{\mathbf{r}}_2 \tag{6.2}$$

where $\mathcal{G}$ is universal gravitational constant, and the second equivalence is given by the second Newton law $\mathbf{F} = m \cdot \mathbf{a}$ in which $\mathbf{a}$ is the acceleration calculated as the second derivative of the position vector $\mathbf{r}$. Setting $\ddot{\mathbf{r}} = \ddot{\mathbf{r}}_2 - \ddot{\mathbf{r}}_1$ (thus we consider the motion of the second item with respect to the first one) and $\mu = \mathcal{G}(m_1 + m_2)$ the following differential equation will be obtained from the previous two ones [22]:

$$\frac{d^2\mathbf{r}}{dt^2} + \mu\frac{\mathbf{r}}{r^3} = 0 \tag{6.3}$$

It can be seen that the $\mathbf{r}$ and $\cdot\mathbf{r}$ lies always in the same plane: this is because the product vector $\mathbf{r} \times \ddot{\mathbf{r}} = 0$, thus if one integrates she will get that the product vector $\mathbf{r} \times \mathbf{r} = \mathbf{h}$ where $\mathbf{h}$ is a constant vector. Furthermore the problem can be simplified by using polar coordinates $\hat{\mathbf{r}}$ and $\hat{\theta}$. Indeed since the position, speed and acceleration in polar coordinates have the following form [22]:

$$\mathbf{r} = r\hat{\mathbf{r}} \tag{6.4}$$

$$\dot{\mathbf{r}} = \dot{r}\hat{\mathbf{r}} + r\dot{\theta}\hat{\theta} \tag{6.5}$$

$$\ddot{\mathbf{r}} = \left(\ddot{r} - r\dot{\theta}^2\right)\hat{\mathbf{r}} + \left[\frac{1}{r}\frac{d}{dt}\left(r^2\dot{\theta}\right)\right]\hat{\theta} \tag{6.6}$$

Using these coordinates in the product vector between the speed and the position the following equation will be obtained [22]:

$$\mathbf{h} = r^2\dot{\theta}\hat{\mathbf{z}} \tag{6.7}$$

where $\mathbf{z}$ is a vector perpendicural to the plane, whose module is equal to

$$h = r^2\dot{\theta} \tag{6.8}$$

If we consider the motion of the body m_2 in the time interval δt we have that the area δA illustrated in Fig 6.1 will be [22]:

$$\delta A \approx \frac{1}{2}r(r + dr)\sin(\delta\theta) \approx \frac{1}{2}r^2\delta\theta \tag{6.9}$$

where the Taylor expansion at first order was used. Therfore [22]:

$$\frac{dA}{dt} = \frac{1}{2}r^2\frac{d\theta}{dt} = \frac{1}{2}h \tag{6.10}$$

but we know that h is constant, therefore the first derivative, is constant. This is the 2^{th} Kepler law. The Eq. 6.5 in polar coordinates [22]:

$$\ddot{r} - r\dot{\theta}^2 = -\frac{\mu}{r^2} \tag{6.11}$$

This differential equation can be rewritten as an armonic oscillator with the substitutions $u = \frac{1}{r}$ $h = r^2\dot{\theta}$ [22]:

$$\dot{r} = -\frac{1}{u}\frac{du}{d\theta}\dot{\theta} = -h\frac{du}{d\theta} \tag{6.12}$$

$$\ddot{r} = -h\frac{d^2u}{d\theta^2}\dot{\theta} = -h^2u^2\frac{d^2u}{d\theta^2} \tag{6.13}$$

$$\frac{d^2u}{d\theta^2} + u = \frac{\mu}{h^2} \tag{6.14}$$

$$u = \frac{\mu}{h^2}\left[1 + e\cos(\theta - \phi\right] \tag{6.15}$$

where the integration constants e and ϕ are respectively the amplitude and the phase. Therefore we have [22]:

$$r = \frac{p}{1 + e\cos(\theta - \phi)} \tag{6.16}$$

In this form we can recognize in e the eccentricity and p is the semilatus rectum [22]:

$$p = \frac{h^2}{\mu} \tag{6.17}$$

Depening on the eccentricity we have four possible conics [22]:

- circle: $e = 0 \quad p = a$
- ellipse: $0 < e < 1 \quad p = a$
- parabola: $e = 1 \quad p = 2q$
- hyperbola: $e > 1 \quad p = a(e^2 - 1)$

in which a is the semi-major axis of the conic. The shape of these orbits is reported in Fig. 6.3. All the asteroids considered here an eccentricity $0 < e < 1$, therefore they have an elliptical orbits in which the earth lies in to one of the two focal point. It is worth nothing that this is the first Kepler law. Looking to Fig. we can define the point of minimum distance between m_1 and the orbiting body as the pericentre or perihelion, and the maximum distance as the apocentre or the aphelion. The semi-major axis, here denoted as b on the other side is defined as the distance between the pericentre and the apocentre. Using the following identity [22]:

$$b^2 = a^2(1 - e^2) \tag{6.18}$$

we get [22]:

$$r = \frac{a(1 - e^2}{1 + e \cdot \cos(\theta - \phi} \tag{6.19}$$

Furthermore the third Kepler law, can be quickly obtained considering the area swept in one orbital period T (the time needed to complete a full round of the orbit) $A = \pi ab$. Since we know that this area is equal to $hT/2$ and $h^2 = \mu a(1 - e^2)$ [22]:

$$T^2 = \frac{4\pi^2}{\mu} a^3 \tag{6.20}$$

If we have two bodies, of masses m and m', that orbit around the Earth m_c, we can use the previous equation to obtain [22]:

$$\frac{m_c + m}{m_c + m'} = \left(\frac{a}{a'}\right)\left(\frac{T'}{T}\right)^2 \tag{6.21}$$

But since m,$m' << m_c$ [22]:

$$(a/a')^3 \approx (T/T')^2 \tag{6.22}$$

Therefore [22]:
[22]:
We see that the two orbital parameters T' and a' are independent with respect to orbiting mass. This statement can be extended to the other orbital parameters by considering the previous approximation m,$m' << m_c$. This is why we expect that the mass/volume of the asteoroid can not be conditionally dependant to the orbital parameters. It is useful to define also the mean motion (feature that is also present in the asteroids dataset) as [22]:

$$n = \frac{2\pi}{T} \tag{6.23}$$

Therefore [22]:

$$\mu = n^2 a^3 \tag{6.24}$$

$$h = na^2\sqrt{1 - e^2} = \sqrt{\mu a(1 - e^2)} \tag{6.25}$$

From which we can see that the angular velocity $\ddot{f}$ is function of the longitude. We are now going more in deep with this statement. Lets come back to the Eq. 6.19, this can be rewritten as [22]:

$$\dot{\mathbf{r}} \cdot \ddot{\mathbf{r}} + \mu \frac{\dot{r}}{r^2} = 0 \tag{6.26}$$

whose integration gives [22]:

$$\frac{1}{2}v^2 - \frac{\mu}{r} = C \tag{6.27}$$

In which $v^2 = \dot{\mathbf{r}} \cdot \dot{\mathbf{r}}$, and C the integration costant. This expression express the energy conservation: on the left side we

have the vis-via term (basically the kinetic energy without the mass), on the right side the potential energy (rescaled with the reduced mass). Lets come back to Eq. 6.19, and make the following substitution $f = \theta - \phi$, which is called the true anomaly. If we differentiate it we will obtain [22]:

$$\dot{r} = \frac{r\dot{f}e\sin f}{1 + e\cos f} \tag{6.28}$$

Remembering the the definition of $h = r^2\dot{f}$, from Eq. 6.25 we have [22]:

$$\dot{r} = \frac{na}{\sqrt{1-e^2}} e\sin f \tag{6.29}$$

$$r\dot{f} = \frac{na}{\sqrt{1-e^2}} (1 + e\cos f) \tag{6.30}$$

Therefore [22]:

$$\begin{aligned} v^2 &= \frac{n^2a^2}{1-e^2}\left(1 + 2e\cos f + e^2\right) = \\ &= \frac{n^2a^2}{1-e^2}\left(\frac{2a(1-e^2)}{r} - (1-e^2)\right) \end{aligned} \tag{6.31}$$

$$v^2 = \mu\left(\frac{2}{r} - \frac{1}{a}\right) \tag{6.32}$$

From which we get that the velocity of the asteroid is maximum at the perihelion, and minimum at the aphelion. Their values that are equal respectively to [22]:

$$v_{\text{perihelion}} = na\sqrt{\frac{1+e}{1-e}} \tag{6.33}$$

$$v_{\text{aphelion}} = na\sqrt{\frac{1-e}{1+e}} \tag{6.34}$$

Another quantity that is contained in the asteroids dataset, and is useful to describe their orbits is the mean anomaly. This is defined as [22]:

$$M = n(t - \tau) \tag{6.35}$$

where τ, the time of pericentre passage, increases linearly with time at a costant rate equal to the mean motion. Furthermore is bounded by the following relations for the perihelion and aphelion [22]:

- $M = f = 0 \quad t = \tau$ Perihelion
- $M = f = \pi \quad t = \tau + T/2$ Aphelion

Such boundaries should be intended as periodic for multiple of the orbital period T. The geometrical interpretation of the angle associated to the mean anomaly is given in Fig. 6.4. It can be proven (see [22]) that the value of this angle, which describe the position of the orbiting item, is given by the following expression, know as Kepler equation [22]:

$$M = E - e \sin E \tag{6.36}$$

Finally as move on to space, other two angles are necessary for the description of an orbit: these are shown in Fig. 6.5 and are the inclination of the orbit (I) and the longitude of the ascending node Ω. Given this quantities the Tisserard invariant can be calculated as [22]:

$$T_P = \frac{a_p}{a} + 2\cos I \sqrt{\frac{a}{a_p}(1 - e^2)} \tag{6.37}$$

If Jupiter is considered as perturbing body, we have the Juptier Tisserard Invariant. The underlying reason for this choice is to distinguish the Jupiter family comets ($2 < T_j < 3$) and the asteroids $T_j < 2$.

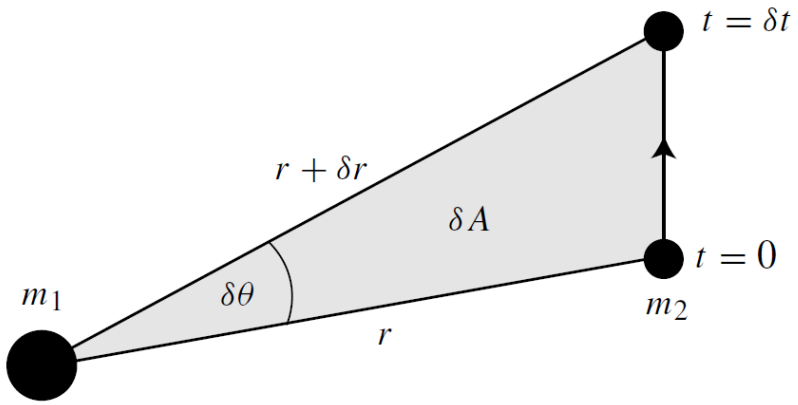


Figure 6.1: The portion of area δA obtained when the position vector moves with an angle $\delta\theta$. Image taken from [22]

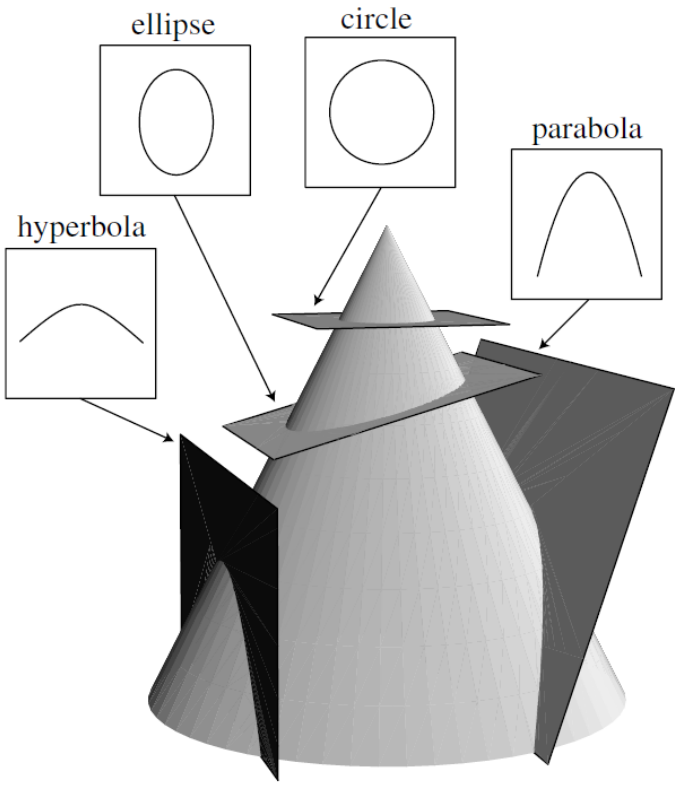


Figure 6.2: The four possible orbits as obtained from a sectionof a cone. Image taken from [22]

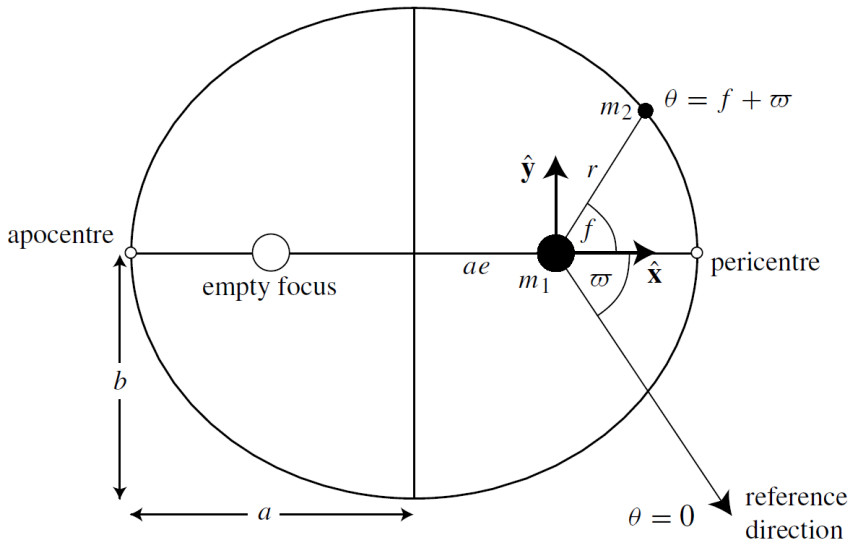


Figure 6.3: Main features of an elliptical orbit. Image taken from [22]

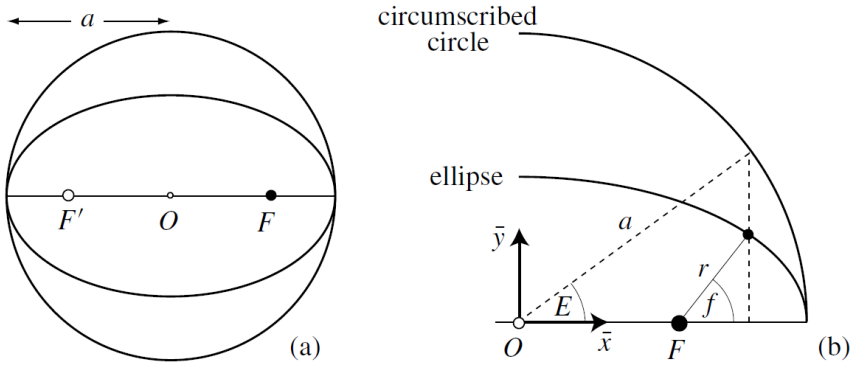


Figure 6.4: The geometrical interpretation of mean anomaly: on the left panel a) is reported how the circumscribed circle should be draw, while on the right panel b) it is shown how the angle associated with the mean anomaly should be interpreted and its relation with the true anomaly angle f. Image taken from [22]

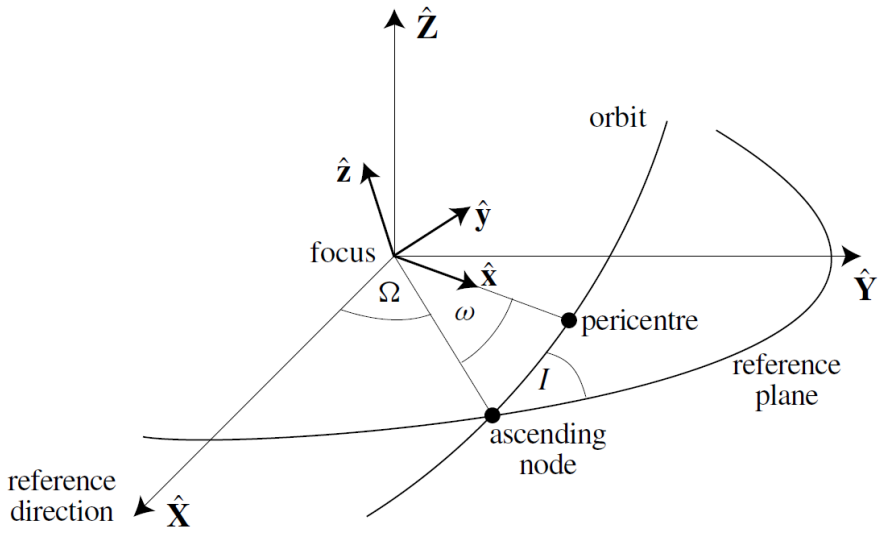


Figure 6.5: The parameters that are necessary for the description of an orbit in three dimension. Image taken from [22]

BIBLIOGRAPHY

[1] https://www.kaggle.com/shrutimehta/nasa-asteroids-classification.

[2] https://cneos.jpl.nasa.gov/.

[3] https://cran.r-project.org/web/packages/glasso/glasso.pdf.

[4] https://cran.r-project.org/web/packages/mgm/mgm.pdf.

[5] https://cran.r-project.org/web/packages/caret/caret.pdf.

[6] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. A Wiley-Interscience publication. Wiley, 2006.

[7] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[8] Gabriel CG de Abreu, Rodrigo Labouriau, and David Edwards. High-dimensional graphical model search with graphd r package. *arXiv preprint arXiv:0909.1234*, 2009.

[9] Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer, and Andreas Weingessel. Misc functions of the department of statistics (e1071), tu wien. *R package*, 1:5–24, 2008.

[10] Sacha Epskamp, Angélique O. J. Cramer, Lourens J. Waldorp, Verena D. Schmittmann, and Denny Borsboom. qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, 48(4):1–18, 2012.

[11] Richard P Feynman, Tony Hey, and Robin W Allen. *Feynman lectures on computation*. CRC Press, 2018.

[12] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[13] Jonas Haslbeck and Lourens J Waldorp. mgm: Estimating time-varying mixed graphical models in high-dimensional data. *arXiv preprint arXiv:1510.06871*, 2015.

[14] Søren Højsgaard, David Edwards, and Steffen Lauritzen. *Graphical models with R*. Springer Science & Business Media, 2012.

[15] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[16] Gilles Kratzer and Reinhard Furrer. varrank: an r package for variable ranking based on mutual information with applications to observed systemic datasets. *arXiv preprint arXiv:1804.07134*, 2018.

[17] Max Kuhn. Building predictive models in r using the caret package. *Journal of statistical software*, 28(1):1–26, 2008.

[18] Sébastien Lê, Julie Josse, and François Husson. Factominer: an r package for multivariate analysis. *Journal of statistical software*, 25(1):1–18, 2008.

[19] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[20] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[21] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[22] Carl D Murray and Stanley F Dermott. *Solar system dynamics*. Cambridge university press, 1999.

[23] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[24] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[25] S. Russell, S.J. Russell, P. Norvig, and E. Davis. *Artificial Intelligence: A Modern Approach*. Prentice Hall series in artificial intelligence. Prentice Hall, 2010.

[26] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940–3941, 2005.

[27] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[28] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[29] Sheng-Jhih Wu and Moody T Chu. Markov chains with memory, tensor formulation, and the dynamics of power iteration. *Applied Mathematics and Computation*, 303:226–239, 2017.