.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 [11], Koller et al. [7], Højsgaardand [6] et al. and Russel et al. [14]. Furthermore we will provide also a rapid overview of the main concepts of information theory, following the Cover [2] and MacKay [10] 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 [12] into the Appendix A

2.1 MARKOV RANDOM FIELDS

Lets start by supposing that we would represent compactly a joint distribution such as [11]:

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

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 [11]:

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

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 [15]. 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 [11]:

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

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 [11]:

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

In this case a first order Markov chain is obtained, where the transition tensor is of second order [15]. 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 [11]:

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 [11]:

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

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 [11]:

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

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[11]:

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

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 $\mathbf{y} = |i_1, i_2, ..., i_n\rangle$ [11]:

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

If we also consider an external field [11]:

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

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 [11]:

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

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 [11]:

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

In general we refer to Gaussian Markov random fields for a joint distribution that can be decomposed in the following way [11]:

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

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

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

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

(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 [11]:

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

thus we can define the log-likehood as [11]:

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

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

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

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

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 [11]:

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

Therfore at the optimum we will have [11]:

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

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 [11] and [7]. 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 [6]:

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

Which can be rewritten in the exponential family form [6]:

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

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 [6]:

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

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 [6].

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 [2] and MacKay [10] 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) [2] : 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 [2]:

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

Furthermore, we can define also the conditional entropy as [2]:

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

The joint entropy and the conditional entropy are related by the chain rule [2]:

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

Such rule can be extended to to the following from [2]:

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

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 [2, 10]:

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

As stated by the Gibbs inequality [2, 10]:

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

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 [2]:

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

As for the joint distribution also in this case we have a chain rule [2]:

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

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 [2, 10]. Otherwise we would have a clear violation of the second principle of thermodynamics (see for instance the Maxwell's demon [4])

3 DATASET DESCRIPTION

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

In Fig. 4.1,4.2 and 4.3 are reported the main achivements of FADM (performed with FactoMineR package [9]): in particular from the correlation circle in Fig. 4.2 we can recognize the different correlations that are given by the celestial mechanics laws: 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 uniforms: 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. 4.4. This figure summarize the ranking of the features considered in the dataset for the dangerousness classification of the asteroids according to the following expression [8]

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

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 [8]

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

following the Peng. et al approach [13]. We see that the first placein 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

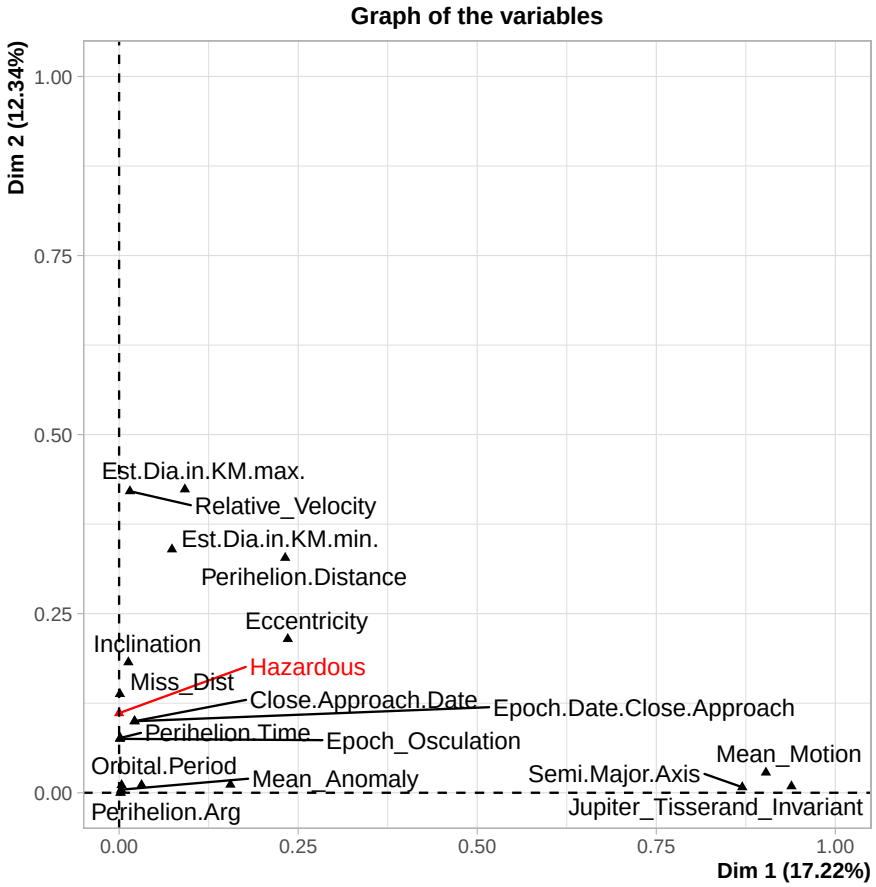


Figure 4.1: The FAMD main plot in which the correlation between the continuous and discrete variables is reported. Plot obtained from FactoMineR package [9]

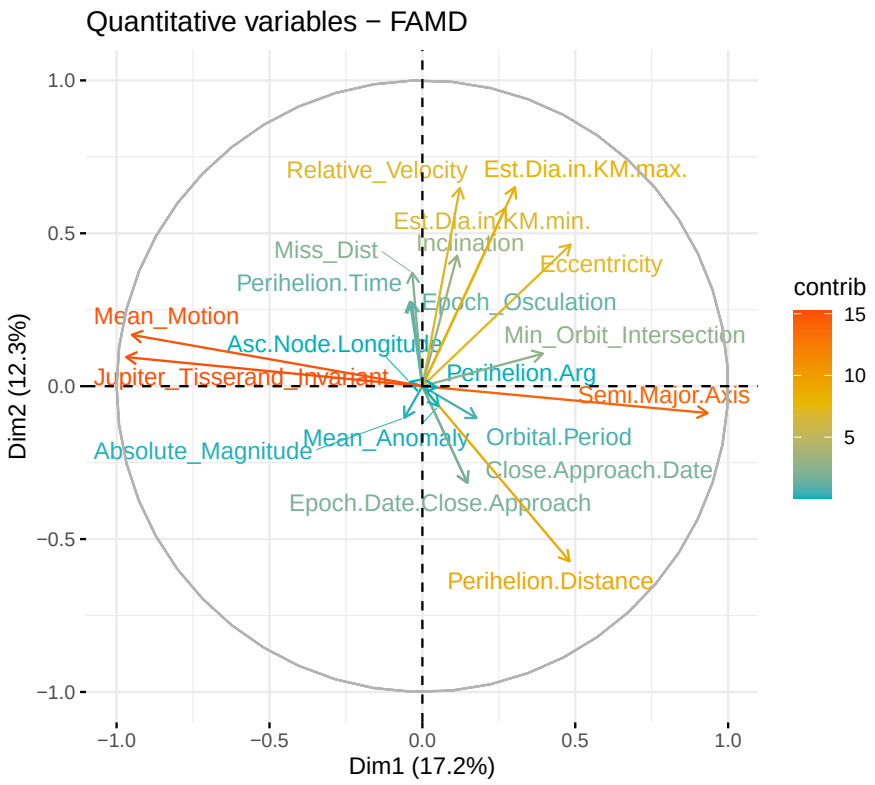


Figure 4.2: The FAMD correlation circle for continuos variables as obtained from FactoMineR package [9]

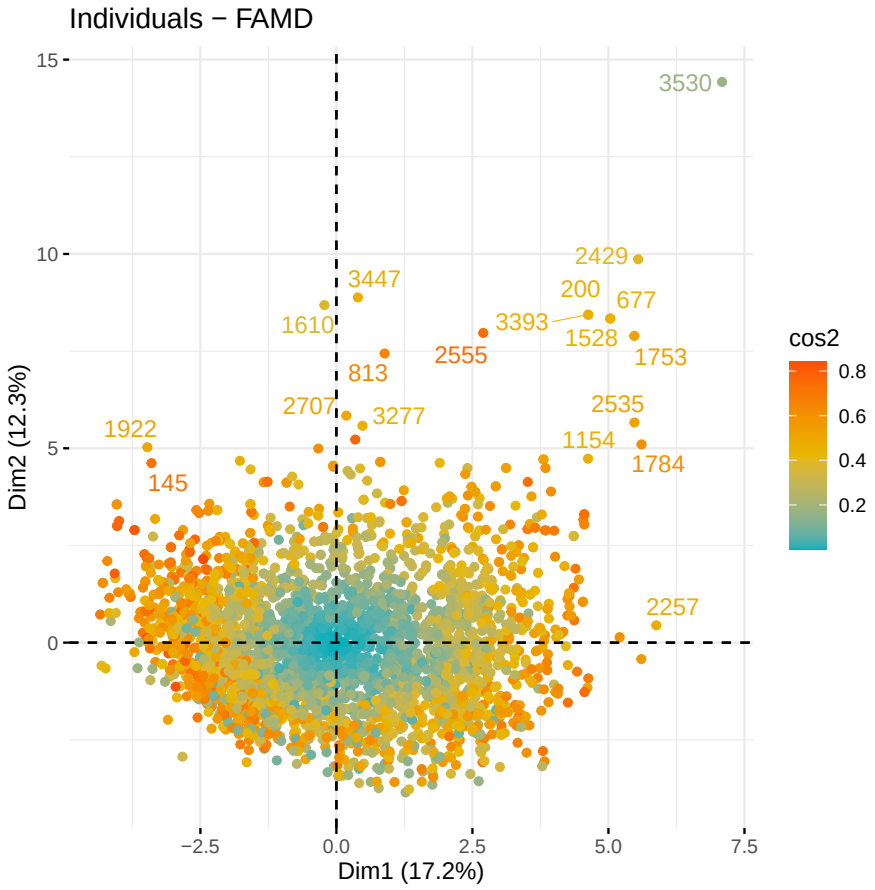


Figure 4.3: Graph of individuals, for the qualititative variables, as obtained from FactoMineR package [9]

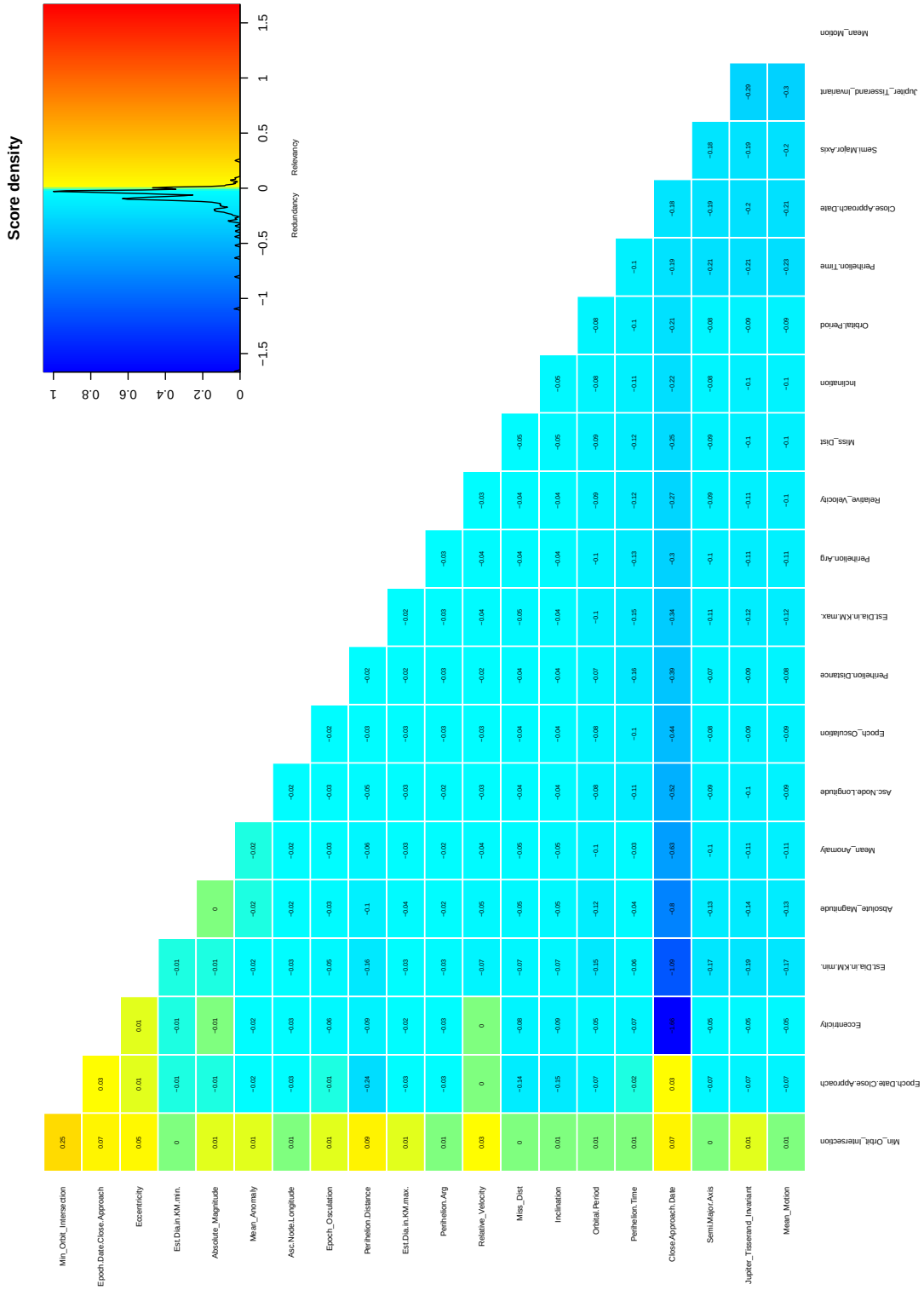
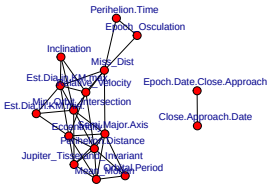


Figure 4.4: Mutual information as obtained with the varrank package [8]

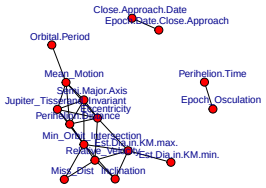
4.2 PROBABILISTIC MODELS

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 [5, 1]. After different tests reported in Fig. 4.5, 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 beacuse in this plot we see that the conditional dependences/independences state by the Celestial Mechanics are correctly reproduced: for instance we see that the diameter of the asteorids is indipendent 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 arbitary scale. Also the Perihelion Epoch and Osculation Time are dependant, as expected, but there is no depencdence with the orbit parameters. We see that the mean motion is conditionally indipendent with respect to relative velocity, but correctly this is dependent from perihelion distance.

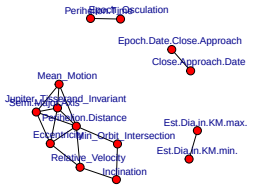
4.3 MACHINE LEARNING ALGORITHMS



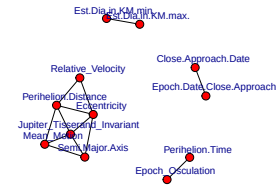
ρ=0.1



ρ=0.2



ρ=0.3



ρ=0.4

Figure 4.5: GLASSO analysis, performed with the glasso package [5, 1], 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 [3]

5 CONCLUSIONS

6 APPENDIX A: CONCEPTS OF CELESTIAL MECHANICS

BIBLIOGRAPHY

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

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

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

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

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

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

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

[8] 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.

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

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

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

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

[13] 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.

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

[15] 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.