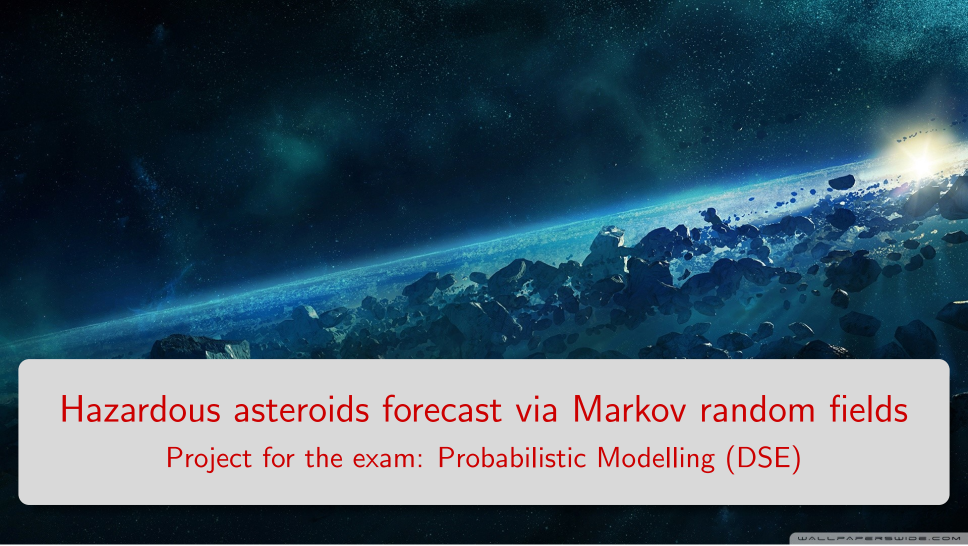


Hazardous asteroids forecast via Markov random fields

Project for the exam: Probabilistic Modelling (DSE)

Marzio De Corato

December 13, 2021

Introduction

- **Final Goal** Assessment of forecasts and interpretability for different machine learning algorithms, including the probabilistic models
- **Method** Use a dataset for which the laws that interconnect the different features are known from general principles
- **Dataset** CNEOS asteroids dataset for more than 3500 asteroids
- **Theoretical laws** Celestial mechanics
- **Algorithms involved - probabilistic models** GLASSO, mgm, minforest, mmod
- **Algorithms involved - others** Random forest, Support Vector Machines, Quadratic Discriminant Analysis, Logistic Regression

Celestial mechanics

Celestial mechanics [14]: equations of motion

The interaction between a planet of mass m_1 at the position r_1 (inertial frame) and an asteroid of mass m_2 at the position r_2 is given by:

$$\mathbf{F}_1 = \mathcal{G} \cdot \frac{m_1 m_2}{r^3}\mathbf{r} = m_1\ddot{\mathbf{r}}_1 \quad \mathbf{F}_2 = -\mathcal{G} \cdot \frac{m_1 m_2}{r^3}\mathbf{r} = m_2\ddot{\mathbf{r}}_2 \tag{1}$$

Where $\mathcal{G}$ is the universal gravitational constant. If we consider the motion of the second item with respect to the first one

$$\ddot{\mathbf{r}} = \ddot{\mathbf{r}}_2 - \ddot{\mathbf{r}}_1 \quad \mu = \mathcal{G}(m_1 + m_2) \tag{2}$$

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

$\mathbf{r} \times \ddot{\mathbf{r}} = 0 \implies \mathbf{r}$ and $\dot{\mathbf{r}}$ lies in the same plane

Celestial mechanics [14]: equations of motion

Integrating $\mathbf{r} \times \ddot{\mathbf{r}} = 0$

$$\mathbf{r} \times \dot{\mathbf{r}} = \mathbf{h} \tag{4}$$

Where $\mathbf{h}$ is a constant of Integration. Using the polar coordinates $\hat{\mathbf{r}}$ and $\hat{\boldsymbol{\theta}}$

$$\mathbf{r} = r\hat{\mathbf{r}} \tag{5}$$

$$\dot{\mathbf{r}} = \dot{r}\hat{\mathbf{r}} + r\dot{\theta}\hat{\boldsymbol{\theta}} \tag{6}$$

$$\ddot{\mathbf{r}} = \left(\ddot{r} - r\dot{\theta}^2\right)\hat{\mathbf{r}} + \left[\frac{1}{r}\frac{d}{dt}\left(r^2\dot{\theta}\right)\right]\hat{\boldsymbol{\theta}} \tag{7}$$

$$\mathbf{h} = r^2\dot{\theta}\hat{\mathbf{z}} \tag{8}$$

$$h = r^2\dot{\theta} \tag{9}$$

Celestial mechanics [14]: 2^{th} Kepler law

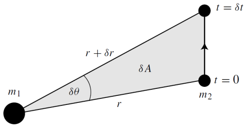


Figure 1: [14]

$$\delta A \approx \frac{1}{2} r(r + dr) \sin(\delta\theta) \approx \frac{1}{2} r^2 \delta\theta \tag{10}$$

$$\frac{dA}{dt} = \frac{1}{2} r^2 \frac{d\theta}{dt} = \frac{1}{2} h \tag{11}$$

h is constant $\implies$ 2^{th} Kepler law

Celestial mechanics [14]: 1^{th} Kepler law

Using the substitution $u = \frac{1}{r}$ $h = r^2\dot{\theta}$

$$\dot{r} = -\frac{1}{u}\frac{du}{d\theta}\dot{\theta} = -h\frac{du}{d\theta} \tag{12}$$

$$\ddot{r} = -h\frac{d^2u}{d\theta^2}\dot{\theta} = -h^2u^2\frac{d^2u}{d\theta^2} \tag{13}$$

$$\frac{d^2u}{d\theta^2} + u = \frac{\mu}{h^2} \tag{14}$$

$$u = \frac{\mu}{h^2}\left[1 + e\cos(\theta - \phi)\right] \tag{15}$$

Celestial mechanics [14]: 1^{th} Kepler law

$$r = \frac{p}{1 + e\cos(\theta - \phi)} \qquad (16)$$

e is eccentricity

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

a is the semi-major axis of the conic

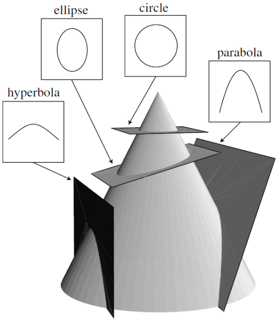


Figure 2: [14]

Celestial mechanics [14]: 3^{th} Kepler law

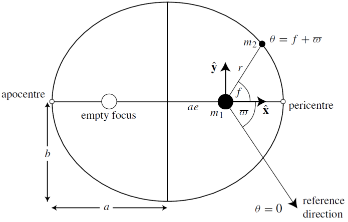


Figure 3: [14]

$$b^2 = a^2(1 - e^2) \tag{17}$$

$$r = \frac{a(1 - e^2)}{1 + e \cdot cos(\theta - \phi)} \tag{18}$$

Area swept in one orbital period T

$$A = \pi ab$$

We know that: $hT/2 \quad h^2 = \mu a(1 - e^2)$

Therefore

$$T^2 = \frac{4\pi^2}{\mu} a^3 \tag{19}$$

Celestial mechanics [14]: 3^{th} Kepler law

Consider two asteroids of mass m and m' orbiting the Earth m_c , with semi-major axes a and a' and orbital periods T and T'

$$\frac{m_c + m}{m_c + m'} = \left(\frac{a}{a'}\right)^3 \left(\frac{T'}{T}\right)^2 \tag{20}$$

But since $m, m' << m_c$

$$\left(\frac{a}{a'}\right)^3 \approx \left(\frac{T}{T'}\right)^2 \tag{21}$$

Remark: The mass of the asteroids is **not** involved

Celestial mechanics [14]: Orbital parameters

Mean motion $n = \frac{2\pi}{T}$

$$v_{perihelion} = na\sqrt{\frac{1+e}{1-e}} \tag{22}$$

$$v_{aphelion} = na\sqrt{\frac{1-e}{1+e}} \tag{23}$$

Remark: The mean motion of an asteroid is different with respect to the the asteroid relative velocity (measured from Earth), since the latter is different at the perihelion an at the aphelion

Celestial mechanics [14]: Orbital parameters

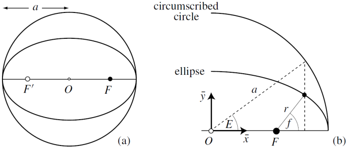


Figure 4: [14]

Mean anomaly

$$M = n(t - \tau) \tag{24}$$

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

$$M = E - e \sin E \tag{25}$$

Jupiter Tisserard invariant

$$T_P = \frac{a_p}{a} + 2 \cos I \sqrt{\frac{a}{a_p}(1 - e^2)} \tag{26}$$

Celestial mechanics [14]: Orbital parameters

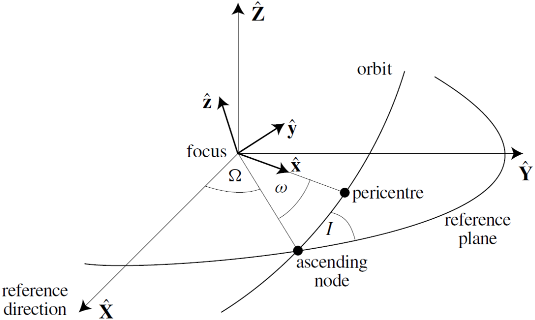


Figure 5: [14]

I: inclination of the orbit

Ω: longitude of the ascending node

Celestial mechanics [14]: Magnitude

$$\Phi = \frac{L}{4\pi d^2} \tag{27}$$

$$m = -2.5 \log_{10} \Phi + C \tag{28}$$

$$m_1 - m_2 = -2.5 \log_{10} \frac{\Phi_1}{\Phi_2} \tag{29}$$

$$M - m = -2.5 \log_{10} \frac{\Phi \cdot d^2}{\Phi \cdot 10^2} \tag{30}$$

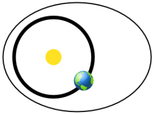
$$M = m + 5 - 5 log_{10} d \tag{31}$$

Where Φ is the flux for a sphere of radius r, m the relative magnitude and M the Absolute magnitude

Celestial mechanics [1]: Classification

Amors

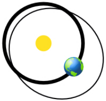
Earth-approaching NEAs with orbits exterior to Earth's but interior to Mars' (named after asteroid (1221) Amor)



$$a > 1.0 \text{ AU}$$
$$1.017 \text{ AU} < q < 1.3 \text{ AU}$$

Apollos

Earth-crossing NEAs with semi-major axes larger than Earth's (named after asteroid (1862) Apollo)



$$a > 1.0 \text{ AU}$$
$$q < 1.017 \text{ AU}$$

Atens

Earth-crossing NEAs with semi-major axes smaller than Earth's (named after asteroid (2062) Aten)

$$a < 1.0 \text{ AU}$$
$$Q > 0.983 \text{ AU}$$

Atiras

NEAs whose orbits are contained entirely within the orbit of the Earth (named after asteroid (163693) Atira)

$$a < 1.0 \text{ AU}$$
$$Q < 0.983 \text{ AU}$$

(q = perihelion distance, Q = aphelion distance, a = semi-major axis)

Celestial mechanics [1]: Classification

- **Potentially Hazardous Asteroids**: MOID ≤ 0.05 au $M \leq 22.0$
  NEAs whose Minimum Orbit Intersection Distance (MOID) with the Earth is 0.05 au or less and whose absolute magnitude (M) is 22.0 or brighter

Dataset

Dataset

- The asteroid dataset was retrieved from Kaggle [2], which reports into a more machine readable form the dataset of The Center for Near-Earth Object Studies (CNEOS) [3], a NASA research centre.
- 3552 Asteroids
- Among the 40 the features, the ones connected only to the other name of the asteroid, or connected only to the name of the orbit and the one connected with the orbiting planet (since for all it was the Earth) were discarted
- The proportion hazardous/not hazardous was set 1:5
- The continuous measures were standardised and demeaned

Features

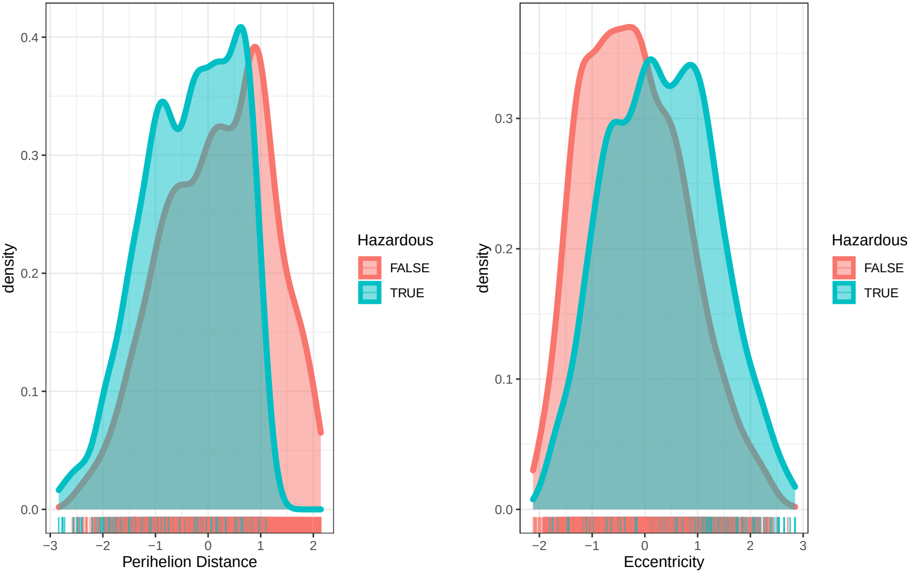
Features	Type
Neo Reference ID	not used
Absolute Magnitude	Continuous
Est Dia in KM (min)	Continuous
Est Dia in KM (max)	Continuous
Close Approach Date	Continuous
Epoch Date Close Approach	Continuous
Relative_Velocity	Continuous
Miss_Dist	Continuous
Min_Orbit_Intersection	Continuous
Jupiter_Tisserand_Invariant	Continuous
Epoch_Osculation	Continuous
Eccentricity	Continuous

Features

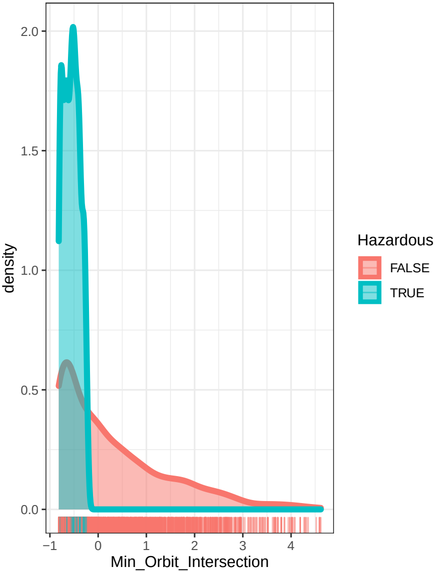
Features	Type
Semi Major Axis	Continuous
Inclination	Continuous
Asc Node Longitude	Continuous
Orbital Period	Continuous
Perihelion Distance	Continuous
Perihelion Arg	Continuous
Perihelion Time	Continuous
Mean_Anomaly	Continuous
Mean_Motion	Continuous
Hazardous	Categorical (Binary)

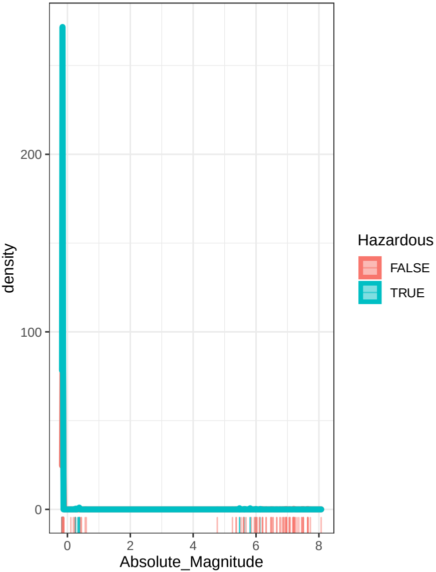
Prelim. analysis

Density Plot

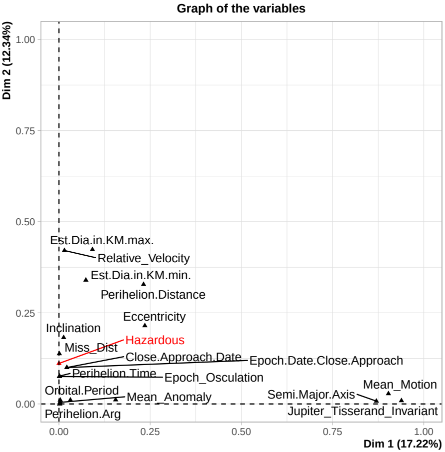


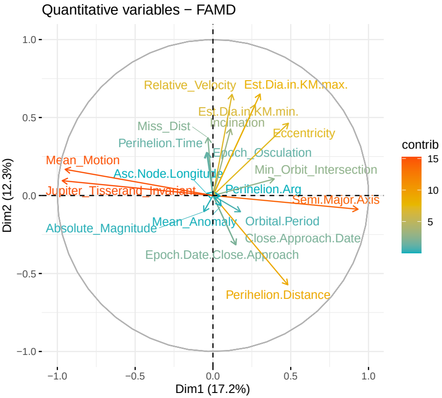
Density Plot





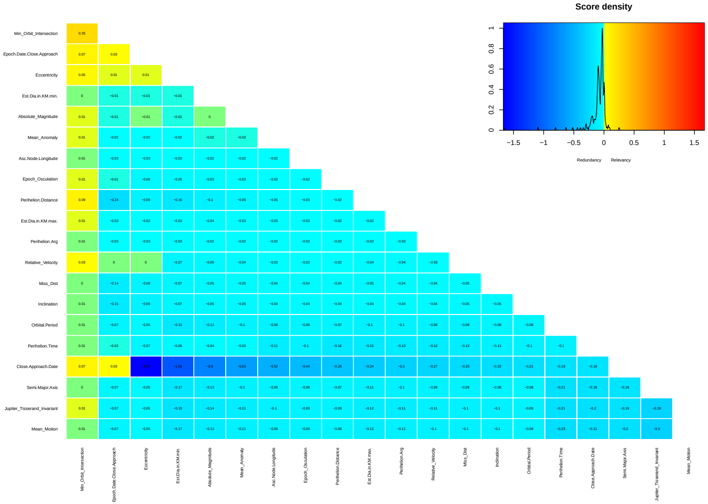
FAMD





Performed with the FactoMineR package [12]

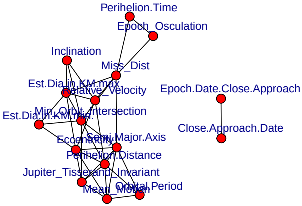
Mutual information analysis



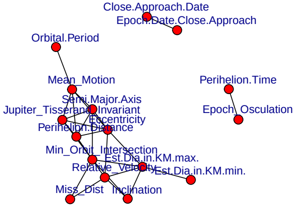
Performed with the varrank package [11]

Probabilistic modelling

GLASSO



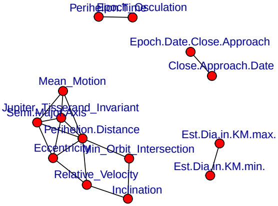
ρ=0.1



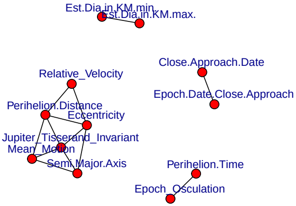
ρ=0.2

Performed with the GLASSO package [4]

GLASSO



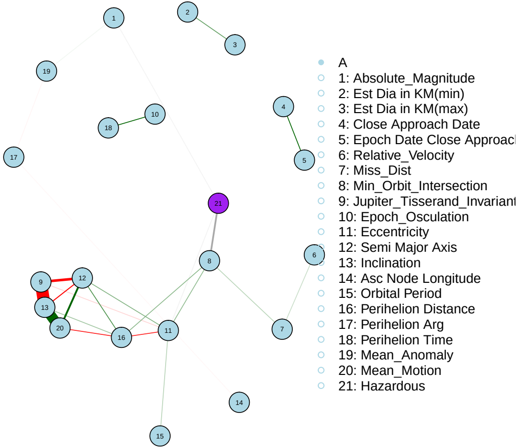
ρ=0.3



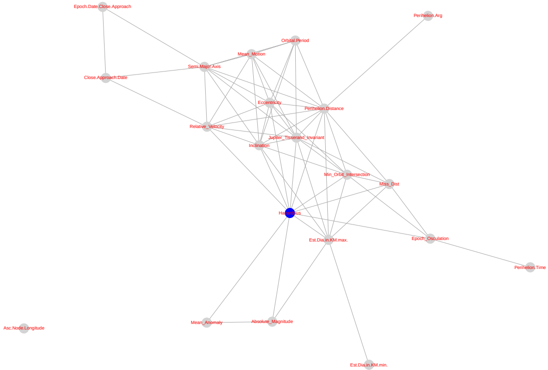
ρ=0.4

Performed with the GLASSO package [4]

Mixed interactions: mgm

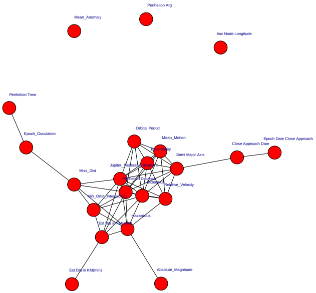


Performed with the mgm package [9]



Performed with the gRapHD package [7]

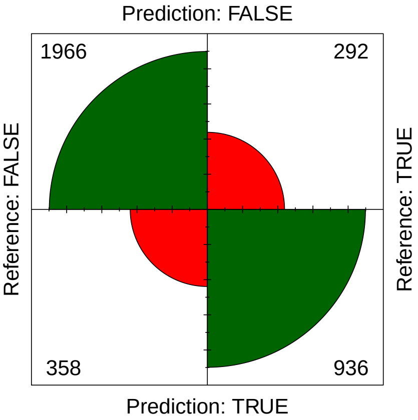
Mixed interactions: mmod

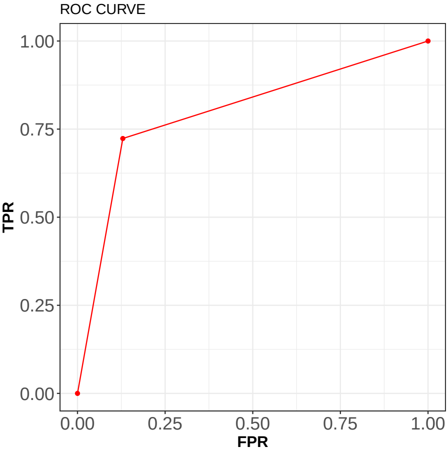


Performed with the gRim package [10]

Mixed interactions

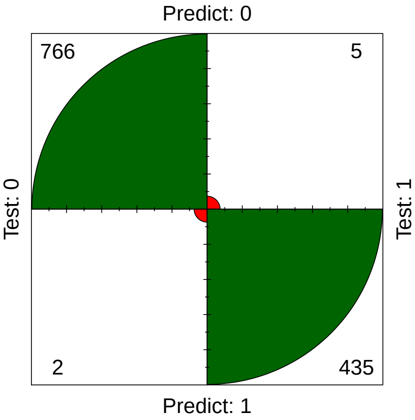
The mgm model is the one that has the list of connection more coherent with the celestial mechanics laws.

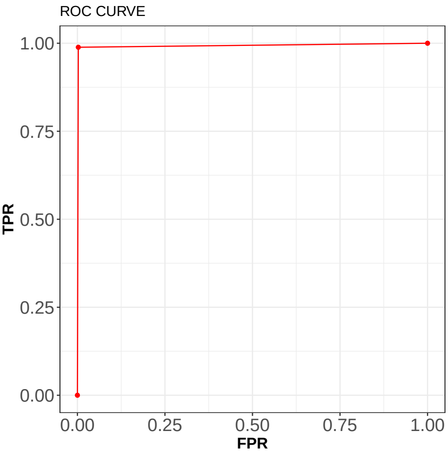




Other ML algorithms

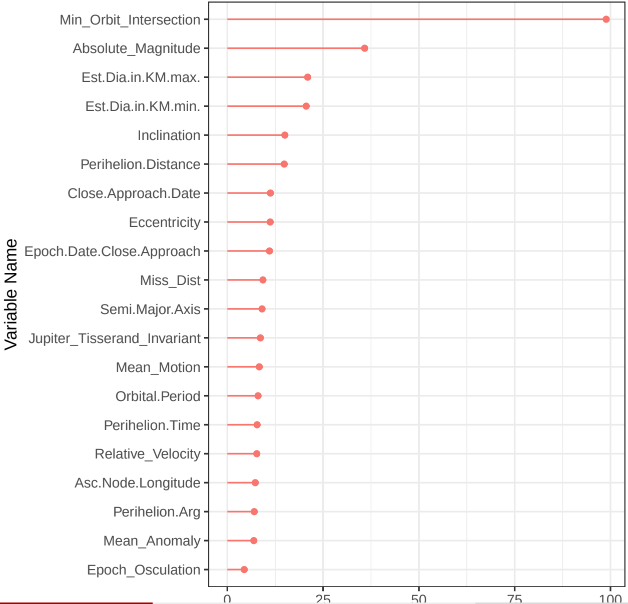
Random Forest



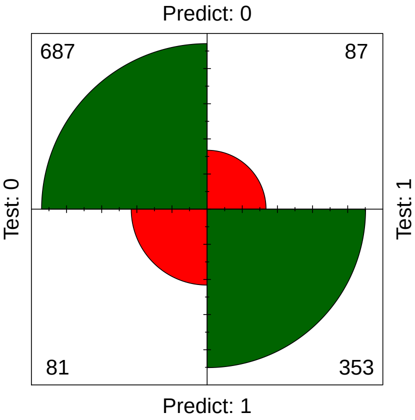


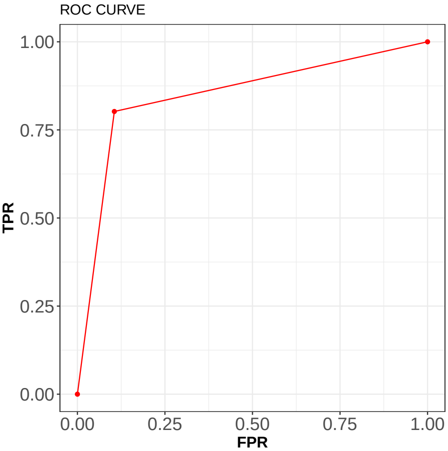
Performed with the rfor package [13]

Random Forest



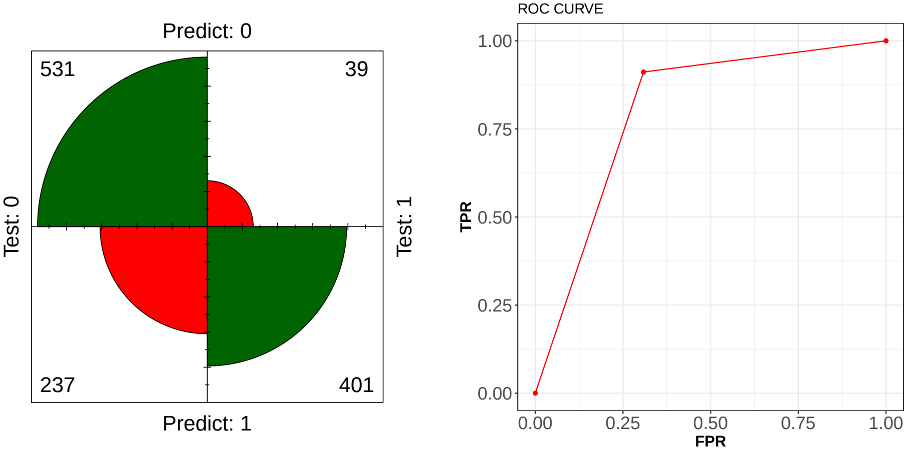
Support Vector Machines





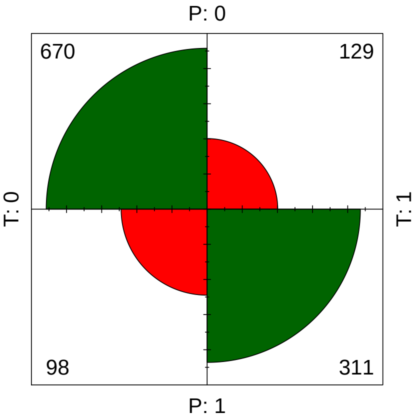
Performed with the e1071 package [8]

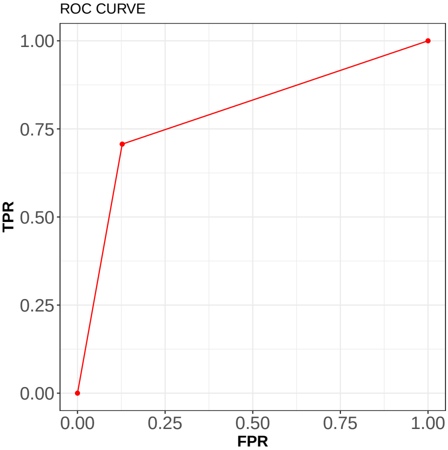
Quadratic Discriminant Analysis (QDA)



Performed with the MASS package [17]

Logistic regression





Performed with the stats package [15]

ϕ coefficient

Table 1: ϕ coefficient (also known as Matthews correlation coefficient)

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

Conclusions

Interpretability and scientific validation

Remark (Interpretability - Tarski definition)

The formal theory T can be translated into S if and only if S can prove the theorem of T in its language [16]

Interpretability and scientific validation

Remark (Scientific method - Einstein definition)

Science uses the totality of the primary concepts, i.e., concepts directly connected with sense experiences, and propositions connecting them. Such a state of affairs cannot, however, satisfy a spirit which is really scientifically minded; **because the totality of concepts and relations obtained in this manner is utterly lacking in logical unity**. *In order to supplement this deficiency, one invents a system poorer in concepts and relations, a system retaining the primary concepts and relations of the first layer as logically derived concepts and relations.* **This new secondary system pays for its higher logical unity by having elementary concepts (concepts of the second layer), which are no longer directly connected with complexes of sense experiences** *[5]*

Conclusions: forecast performances vs intepretability

- The mgm algorithm is not the best one in term of performances, but it provides the connections between the features. On the other side, except for the variable importance in RF, the other are black box one
- The mgm model, as the other graphical model is open to a true scientific validation, the other not.
- The probabilistic models lack in the forecast is definitely compensated by their interpretability
- This is meaningful since this two features are in conflict
- The probabilistic models provide a good trade-off between intepretability and forecast performances, as long as one is interest to produce a really scientific result (e.g if the only aim is the forecast the RF is definitely better. However how long one can trust to the RF result ?)

In its efforts to learn as much as possible about nature, modern physics has found that certain things can never be "known" with certainty. Much of our knowledge must always remain uncertain. The most we can know is in terms of probabilities. Richard P. Feynman (1918-1988)

Bibliography I

[1] https://cneos.jpl.nasa.gov/about/neo_groups.html.

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

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

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

[5] https://www.amacad.org/publication/physics-reality.

[6] https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

[7] Gabriel CG de Abreu, Rodrigo Labouriau, and David Edwards. "High-dimensional graphical model search with graphd R package". In: *arXiv preprint arXiv:0909.1234* (2009).

Bibliography II

[8] Evgenia Dimitriadou et al. "Misc functions of the Department of Statistics (e1071), TU Wien". In: *R package* 1 (2008), pp. 5–24.

[9] Jonas Haslbeck and Lourens J Waldorp. "mgm: Estimating time-varying mixed graphical models in high-dimensional data". In: *arXiv preprint arXiv:1510.06871* (2015).

[10] Søren Højsgaard, David Edwards, and Steffen Lauritzen. *Graphical Models with R*. ISBN 978-1-4614-2298-3. New York: Springer, 2012. DOI: 10.1007/978-1-4614-2299-0.

[11] Gilles Kratzer and Reinhard Furrer. "varrank: an R package for variable ranking based on mutual information with applications to observed systemic datasets". In: *arXiv preprint arXiv:1804.07134* (2018).

Bibliography III

[12] Sébastien Lê, Julie Josse, and François Husson. "FactoMineR: an R package for multivariate analysis". In: *Journal of statistical software* 25.1 (2008), pp. 1–18.

[13] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In: *R News* 2.3 (2002), pp. 18–22. URL: https://CRAN.R-project.org/doc/Rnews/.

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

[15] R Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: http://www.R-project.org/.

[16] Alfred Tarski, Andrzej Mostowski, and Raphael Mitchel Robinson. *Undecidable theories*. Vol. 13. Elsevier, 1953.

Bibliography IV

[17] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: https://www.stats.ox.ac.uk/pub/MASS4/.

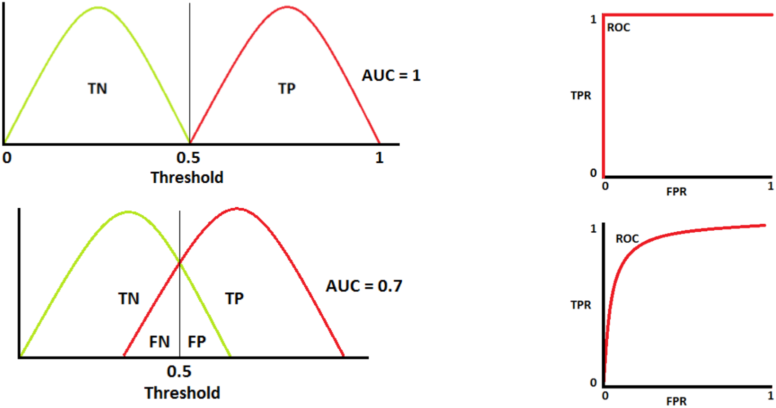
ROC and ϕ factor

ϕ coefficent

	Actual - N	**Actual - P**
Predicted - N	#TP	#FN
Predicted - P	#FP	#TP

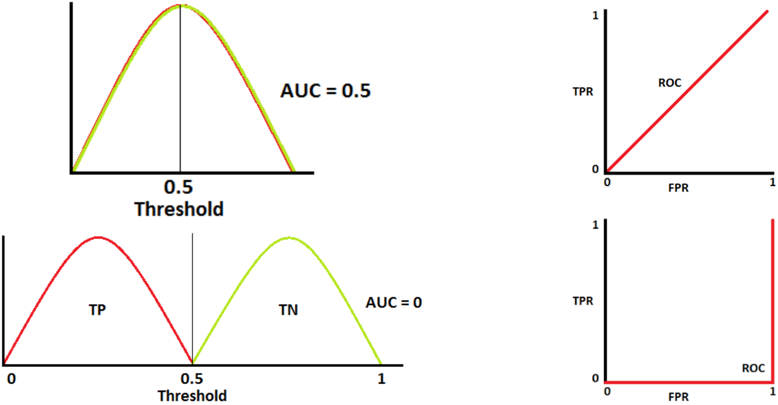
$$\phi = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Receiver operating characteristic



Images taken from [6]

Receiver operating characteristic



Images taken from [6]

FAMD

Factor analysis of mixed data - FAMD

r(z,k) correlation coefficient (z and k quantitative)

η^2(z,q) correlation ratio (z quantitative and q qualitative)

$$\text{PCA} \rightarrow max \sum_k r^2(z,k)$$

$$\text{MCA} \rightarrow max \sum_q \eta^2(z,q)$$

$$\text{FAMD} \rightarrow max \sum_k r^2(z,k) + max \sum_q \eta^2(z,q)$$

Graphical models

Problem: How one can represent in a compact and elegant way the joint distribution $p(\mathbf{x}|\boldsymbol{\theta})$ (where θ are the parameters)

First one can consider to use the chain rule

$$p(x_{1:V}) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1)...p(x_V|x_{1:V-1}) \tag{32}$$

One can represent $p(x_2, x_1)$ as a $\mathcal{O}(K^2)$ where the elements are $p(x_2 = j|x_1 = i) = T_{ij}$ and so on with high order tensors. T is called stochastic matrix

In order to reduce the computational cost $\mathcal{O}(K^2V^2)$ one can consider to take in to account the concept of conditional independence

$$X \perp Y|Z \iff p(X, Y|Z) = p(X|Z)p(Y|Z) \tag{33}$$

Therefore the distribution can be factorized as:

$$p(\mathbf{x}_{1:V}) = p(x_1) \prod_{t=1}^{V} p(x_t|x_{t-1}) \tag{34}$$

Which is a first order Markov chain

Markov random field

Markov random field

Problem: differently with respect to Bayesian Network for an undirected graph there is no topological ordering associated to it. Therefore the chain rule cannot be used

Solution: One can use the potential function associated with maximal clique $\psi_c(\mathbf{y}_c|\boldsymbol{\theta}_c)$

Markov random field

Theorem (Hammersley-Clifford)

A positive distribution $p(\mathbf{y}) > 0$ *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.*

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{c \in C} \psi_c(\mathbf{y}_c|\boldsymbol{\theta}_c) \tag{35}$$

where C is the set of all the (maximal) cliques of G, and $Z(\boldsymbol{\theta})$ *is the partition function given by*

$$Z(\boldsymbol{\theta}) := \sum_{y} \prod_{c \in C} \psi_c(\mathbf{y}_c|\boldsymbol{\theta}_c) \tag{36}$$

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

Markov random field

Gibbs distribution

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(-\sum_{c} E(\mathbf{y}|\boldsymbol{\theta})\right) \tag{37}$$

Where $E(\mathbf{y})$ is the energy associated with the variables in clique c. Therefore

$$\psi(\mathbf{y}_c|\boldsymbol{\theta}_c) = \exp\left(-E(\mathbf{y}_c|\boldsymbol{\theta}_c)\right) \tag{38}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_c} = \frac{1}{N}\sum_i \left[\phi_c(y_i) - \frac{\partial}{\partial \boldsymbol{\theta}_c}\log Z(\boldsymbol{\theta})\right] \tag{39}$$

$$\frac{\partial \log Z(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbb{E}\left[\phi_c(\mathbf{y}) \middle|\, \theta\right] = \sum_{\mathbf{y}} \phi_c(\mathbf{y}) p(\mathbf{y}|\boldsymbol{\theta}) \tag{40}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_c} = \left[\frac{1}{N}\sum_i \phi_c(y_i)\right] - \mathbb{E}\left[\phi_c(\mathbf{y})\right] \tag{41}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_c} = \mathbb{E}_{p_{emp}}\left[\phi_c(\mathbf{y})\right] - \mathbb{E}_{p(\cdot|\boldsymbol{\theta})}\left[\phi_c(\mathbf{y})\right] \tag{42}$$

$$\mathbb{E}_{p_{emp}}\left[\phi_c(\mathbf{y})\right] = \mathbb{E}_{p(\cdot|\boldsymbol{\theta})}\left[\phi_c(\mathbf{y})\right] \tag{43}$$

Markov random field

$$p\left(\mathbf{y}|\boldsymbol{\theta}\right) = \frac{1}{Z(\theta)} \exp\left(\sum_c \boldsymbol{\theta}_c^T \phi_c\left(\mathbf{y}\right)\right) \tag{44}$$

$$\mathcal{L}\left(\boldsymbol{\theta}\right) := \frac{1}{N} \sum_i \log p\left(\mathbf{y}_i|\boldsymbol{\theta}\right) = \frac{1}{N} \sum_i \left[\sum_c \boldsymbol{\theta}_c^T \phi_c(y_i) - \log Z\left(\boldsymbol{\theta}\right)\right] \tag{45}$$

Graphical model algorithms

Gaussian graphical model

$$f_Y(y) = \frac{det(K)^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\mathbf{y}^T K \mathbf{y}\right) \tag{46}$$

$$K = \Sigma^{-1} \tag{47}$$

$$\log L(K, \boldsymbol{\mu}) = \frac{n}{2} \log \det(K) - \frac{n}{2} tr(KS) - \frac{n}{2} \left(\bar{\mathbf{y}} - \boldsymbol{\mu}\right)^T \left(\bar{\mathbf{y}} - \boldsymbol{\mu}\right) \tag{48}$$

GLASSO

$$L_{pen}(K, \hat{\mu}) = log\,\det(K) - tr(K\ S) - \rho||K|| \tag{49}$$

$$K = \Sigma^{-1}$$

S: empirical covariance matrix

mgm algorithm

$$P(X_s|X_{\backslash s}) = \exp\left\{E_s(X_{\backslash s})\phi_s\left(X_s\right) + B_s(X_s) - \Phi\left(X_{\backslash s}\right)\right\} \tag{50}$$

ϕ_s function of sufficient statistics B_s base measure

$$\begin{aligned} P(X) = \exp\Bigg(&\sum_{s\in V}\theta_s\phi_s(X_s) + \sum_{s\in V}\sum_{r\in N(s)}\theta_{s,r}\phi_s(X_s)\phi_r(X_r) \\ &+ \ldots + \sum_{r_1,\ldots,r_k\in C}\theta_{r1,\ldots,rk}\prod_{j=1}^{k}\phi_{rj}\left(X_{rj}\right) + \sum_{s\in V}B_s(X_S) - \Phi(\theta) \end{aligned} \tag{51}$$

$$\hat{\theta} = arg\min_{theta}\{-\mathcal{L}(\theta, X) + \lambda||\theta||_1\} \quad ||\theta||_1 = \sum_{j=1}^{J}|\theta_j| \tag{52}$$

mmod - Homogeneus Mixed Interaction (HMI) models

N observation, d discrete variables and q continuous variable. The observation has the form $x = (i, y) = (i_1, ..., i_d, y_1...y_q)$. The probability of discrete variables falling in the cell i is denoted as $p(i)$. The conditional distribution of continous variables to fall in the cell i is given by the multivariate gaussian $\mathcal{N}(\mu(i), \Sigma)$

$$f(\mathbf{i}, \mathbf{y}) = p(\mathbf{i})(2\pi)^{-q/2} det(\Sigma)^{-1/2} \exp\left[-\frac{1}{2}\left(\mathbf{y} - \mu(\mathbf{i})\right)^T \Sigma^{-1} \left(\mathbf{y} - \mu(\mathbf{i})\right)\right] \tag{53}$$

$$f(\mathbf{i}, \mathbf{y}) = \exp\left[g(\mathbf{i}) + h(\mathbf{i})^T \mathbf{y} - \frac{1}{2} y^T K \mathbf{y}\right] \tag{54}$$

where K is the concentration matrix, $g(i)$ and $h(i)$ are the log-linear expansion of the probability $p(\mathbf{i})$ (canonical parameters)

Decomposable model

$$\hat{f}(x) = \prod_{j=1}^{k} \frac{\hat{f}_{C_j}(x_{Cj})}{\hat{f}_{S_j}(x_{Sj})} \tag{55}$$

Maximized likehood

$$\hat{\mathcal{L}}_s = \sum_i n(i) log\left(\frac{n(i)}{N}\right) - Nq\frac{log(2\pi)}{2} - N\frac{log(det(S))}{2} - N\frac{q}{2} \tag{56}$$

$$\hat{p}(\mathbf{i}) = n(\mathbf{i})/N \quad \hat{\mu}(\mathbf{i}) = \bar{y}(\mathbf{i}) \quad \hat{\Sigma} = S = \sum_i n(i)S_i/N \tag{57}$$

Information theory

Information theory

$$H(X) = -\sum_{x \in X} p(x) logp(x) \tag{58}$$

$$H(X, Y) = -\sum_{x \in \chi} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \tag{59}$$

$$\begin{aligned} H(X|Y) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= -E \log p(Y|X) \end{aligned} \tag{60}$$

$$H(X, Y) = H(X) + H(Y|X) \tag{61}$$

$$D(p||q) = \sum p(x) \log \frac{p(x)}{q(x)} \quad D(p||q) \geq 0 \tag{62}$$

Information theory

$$I(X;Y) = \sum (x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y))$$
$$= H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{63}$$

$$I(X_1, X_2, ..., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, ..., X_1) \tag{64}$$

$$g(\alpha, \mathbf{C}, \mathbf{S}, f_i) = MI(f_i; \mathbf{C}) - \sum_{f_s \in S} \alpha(f_i, f_s, \mathbf{C}, \mathbf{S}) MI(f_i; f_s) \tag{65}$$