.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 [6], Koller et al. [4], Højsgaardand [3] et al. and Russel et al. [8]. Furthermore we will provide also a rapid overview of the main concepts of information theory, following the Cover [1] and MacKay [5] 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 [7] into the Appendix A

2.1 MARKOV RANDOM FIELDS

Lets start by supposing that we would represent compactly a joint distribution such as [6]:

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

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 [6]:

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

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 [9]. 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 [6]:

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

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 [6]:

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

In this case a first order Markov chain is obtained, where the transtion tensor is of second order [9]. 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 indipendence 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 [6]:

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 [6]:

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

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 [6]:

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

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[6]:

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

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 $\mathbf{y} = |i_1, i_2, ..., i_n\rangle$ [6]:

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

If we also consider an external field [6]:

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

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 energy. Thus according to the Boltzmann or Gibbs distribution we have [6]:

[1] In physics

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

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 $\Sigma^{-1} = \mathbf{W}$, $\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{b}$ and $c = \frac{1}{2}\mu^T\boldsymbol{\Sigma}^{-1}\mu$ we obtain a Gaussian [6]:

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

In general we refer to Gaussian Markov random fields for a joint distribution that can be decomposed in the following way [6]:

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

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

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

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

(this last expression can be reconducted to the multivariate gassian if one consider $\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{-1}$ 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 [6]:

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

thus we can define the log-likehood as [6]:

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

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

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

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

In the first term **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 [6]:

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

Therfore at the optimum we will have [6]:

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

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 [6] and [4]. 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 [3]:

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

Which can be rewritten in the exponential family form [3]:

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

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 [3]:

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

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 [3].

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 [1] and MacKay [5] 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) [1] : and probability mass function $p(x) = \Pr\{X = x\}$ $x \in \chi$ we define the entropy of X as $H(X) = -\sum_{x \in X} p(x) \mathrm{log} p(x)$ 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 [1]:

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

Furthermore, we can define also the conditional entropy as [1]:

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

The joint entropy and the conditional entropy are related by the chain rule [1]:

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

Such rule can be extended to to the following from [1]:

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

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 [1, 5]:

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

As stated by the Gibbs inequality [1, 5]:

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

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 [1]:

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

As for the joint distribution also in this case we have a chain rule [1]:

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

Finally we would report the data process inequality theorem that connects the information theory with the Markov chain: if we have a Markov chains, $X \rightarrow Y \rightarrow Z$ 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 [1, 5]. Otherwise we would have a clear violation of the second principle of thermodynamics (see for instance the Maxwell's demon [2])

3 DATASET DESCRIPTION

4 RESULTS

5 CONCLUSION

6 APPENDIX A: CONCEPTS OF CELESTIAL MECHANICS

BIBLIOGRAPHY

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

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

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

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

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

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

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

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

[9] 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.