Inference, entropies and maximal entropy networks

R. Vilela Mendes
CMAF, University of Lisbon
http://label2.ist.utl.pt/vilela/

April 2016

Contents

- The maximum entropy method (MEM): A tool for inference under incomplete information
- The Khinchin axioms
- Uncertainty functions (entropies)
- Uncertainty measures for complex systems: An ongoing discussion
- MEM, methods and network reconstruction

The maximum entropy method (MEM)

- The MEM states that the probability distribution which best represents the current state of knowledge is the one that incorporates all the known information and has the *maximum uncertainty* about what is not known. That is, the one that is the least biased about unknown factors (Jaynes).
- The first consequence is that application of the MEM needs a reliable measure of uncertainty.
- In current practice *uncertainty* is identified with (Shannon) entropy

$$S = -\sum_i p_i \log p_i$$

Is entropy a consistent measure of uncertainty?

Quantifying uncertainty: The Khinchin axioms

Let X be a discrete random variable, taking k distinct values $(1, 2, ...k)$

(K1). *$S[X]$ depends only on the probability distribution of X.* (That is, we can change the labels of the events as much as we like without changing S.)

(K2). *$S[X]$ is maximal, for a given k, when $p_i = 1/k$ for all i.* (That is, the uniform distribution has maximal S.)

(K3). *If Y is random variable on $1, 2, ...m$, where $m > k$, but $Pr(Y = i) = p_i$ if $i \leq k$, and $Pr(Y = i) = 0$ if $k < i \leq m$, then $S[Y] = S[X]$.* (That is, adding possibilities of probability zero does not change S.)

(K4). *For any random variables X and Y ,*
$S[X, Y] = S[X] + \sum_x \Pr(X = x) S[Y|X = x]$
(That is, our joint S is the sum of the S for one variable, plus the average value of the S of the other variable given the first.)

(K4'). *$S[X, Y] = S[X] + S[Y]$ if X and Y are independent.* Weaker than (4).

Measures of uncertainty

- (K1)+(K2)+(K3)+(K4) $\Longrightarrow S = -\sum_i p_i \log p_i$ (Shannon entropy)
- (K1)+(K2)+(K3)+(K4') $\Longrightarrow S_\alpha = \frac{1}{1-\alpha} \log \sum_i p_i^\alpha$ (Rényi entropies)
  $\alpha \geq 0$
  $\lim_{\alpha \to 1} S_\alpha = S$
- (K4') is quite intuitive. However (K4) is not so reasonably looking. It says that uncertainty is additive in a particular strict way, that the means of conditional uncertainties add up to total uncertainty. In other words *given a composite system the uncertainty is the same no matter how we choose to decompose it and compute the uncertainties.*

  Equivalently: *information (or lack thereof) is independent of the way we choose to collect it*, that is, directly for the compound system or sucessively for the subsystems. Therefore for strongly correlated events (correlated subsystems) S might not be the most appropriate uncertainty measure.

Shannon entropy

- Consider a set of N events all with the same probability and partition the set into R compound events containing r elementary events each. $N = Rr$
  The probabilities are $p_i = \frac{1}{N}$, $P_k = \frac{1}{R}$ and $p\left(j|k\right) = \frac{1}{r}$. Then (K1) and (K4) $\Longrightarrow$

$$S\left(\{p_i\}\right) = S\left(Rr\right) = S\left(R\right) + \sum_{k=1}^{R} \frac{1}{R} S\left(r\right) = S\left(R\right) + S\left(r\right)$$

$$\Longrightarrow S\left(R\right) = c \log R = -c \log \frac{1}{R}$$

- (K3)+(K2) $\Longrightarrow S\left(\left\{\frac{1}{R}, \frac{1}{R}, \cdots, \frac{1}{R}\right\}\right) = S\left(\left\{\frac{1}{R}, \frac{1}{R}, \cdots, \frac{1}{R}, 0\right\}\right) \leq S\left(\left\{\frac{1}{R+1}, \frac{1}{R+1}, \cdots, \frac{1}{R+1}\right\}\right)$

$$\Longrightarrow c > 0$$

$$\text{(K4)} \Longrightarrow S\left(\{p_i\}\right) = -\sum_i p_i \log p_i$$

Rényi entropy

- The Shannon entropy is obtained by a linear average of the uncertainty of the elementary event $-\log p_i$
  A more general way to average a set of quantities $\{x_i\}$ is
  $$g^{-1}\left(\sum_i p_i g\left(x_i\right)\right)$$
  g is called the Kolgomorov-Nagumo function.
- There are two choices that preserve additivity for independent events (K4'),
  $$g\left(x\right)=cx \text{ and } g\left(x\right)=c2^{(1-\alpha)x}$$
  The second choice leads to the Rényi entropies, with limit

$$\lim_{\alpha\to 1} S_\alpha=\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\log\left(\sum_i p_i^{1-\varepsilon}\right)=\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\log\left(\sum_i p_i\left(1-\varepsilon\log p_i\right)\right)=S$$

Other uncertainty measures

- In general
  $$S(p) = \sum_i \phi(p(x_i))$$
  with ϕ a concave function. Is again a maximum when all the probabilities are equal (satisfies the *principle of insufficient reason* - Laplace)
- Examples

$$\begin{aligned} S &= \sum_i \log(p_i) \\ S &= -\sum_i \frac{1}{\log(p_i)} \\ S &= -\sum_i \frac{1}{(\log(p_i))^2} \\ S &= \sum_i \sqrt{\log(p_i)} \end{aligned}$$

A fictitious example: The prisoner in the land of the dice throwers

- A mathematician is in prison in the land of the dice throwers where the jailers spend their time throwing dice, betting and recording the results. To the mathematician, who is in the death row, is proposed the following (sadistic) problem:
- *For a particular dice, after* 100000 *throws, the average was* $3,588886$ *and the number* 2 *appeared* 16700 *times. Problem: Order the dice outcomes in decreasing order according to the number of their appearances*. If the mathematician supplies the right answer he will be free, if not the sentence will be executed immediately.
- Having nothing to loose because he was on the death row anyway, the mathematician decided to accept the challenge. After some thought and a few calculations in the toilette paper available at the cell, the mathematician supplied the answer.

$$6, 5, 4, 2, 3, 1$$

The prisoner in the land of the dice throwers

- To the jailers surprise it was correct and the mathematician was released. How did he manage?
  Answer: By taking into account all the available information and maximizing the uncertainty about what was not known
- Had the dice been a perfect fair dice the number 2 would have appeared 16666 times, not 16700 and the average would have been 3, 5. An estimate of the p_2 probability is
  (a)$p_2 = (16700)/(100000) = 0,167$; and the other constraints are
  (b) $p_1 + p_3 + p_4 + p_5 + p_6 = 1 - 0,167 = 0,833$
  (c) $p_1 + 3p_3 + 4p_4 + 5p_5 + 6p_6 = 3,588886 - 0,334 = 3,254886$
- Maximizing $S = -\sum_k p_k \log p_k + \lambda_1 \left(\sum_k p_k - 1\right) + \lambda_2 \left(\sum_k kp_k - 3.588886\right) + \lambda_3 \left(p_2 - 0.167\right); \left(\frac{\partial S}{\partial p_k} = 0\right) \Rightarrow$
  $p_k = \exp\{\lambda_1 - 1 + \lambda_2 k + \lambda_3\}$ with λ_1, λ_2 and λ_3 obtained from (a), (b) and (c), $x = \exp\{\lambda_2\} = 1,03746159$; $p_1 = 0.833/\left(1 + x^2 + x^3 + x^4 + x^5\right) = 0,15$; $p_k = p_1 \times x^{k-1}$ for $k \neq 2$
  $(p_3 = 0,161449; p_4 = 0,167497; p_5 = 0,1737719; p_6 = 0,18028)$

The continuous case

- When the set of possible states $\{x \in X\}$ forms a continuum, say $\mathbb{R}$, the entropy expression has no natural extension to this case. The expression one would naively write down

  $$-\int p(x) \log p(x)\, dx$$

  for a probability density $p(x)$ has properties which are rather different from those of its discrete counterpart. In particular, probability densities carry a physical dimension, say *probability per unit length*, which gives S the dimension of $\log cm$ which seems somewhat odd. Also this expression is *not invariant under a reparametrization of* X, for example by a change of units. In addition, S may now become negative and is not bounded from above nor below.
- A fruitful way of dealing with the continuum is by replacing the entropy expression by the so called *relative entropy.*

The continuous case

- For the discrete case the relative entropy is defined as

$$S\left(p,\mu\right)=-\sum_{i}p\left(x_i\right)\log\frac{p\left(x_i\right)}{\mu\left(x_i\right)}$$

where the $\mu\left(x_i\right)$'s are positive weights determined by some background measure. In the special case where μ is the counting measure i.e. if $\mu\left(x_i\right)=1\ \forall i$, the relative entropy becomes equal to the absolute entropy. This relative entropy has a natural extension to the continuous case,

$$S\left(P,M\right)=-\int\frac{\partial P}{\partial M}\log\frac{\partial P}{\partial M}dM$$

or, with densities $dP=p\left(x\right)dx; dM=\mu\left(x\right)dx$

$$S\left(p,\mu\right)=-\int p\left(x\right)\log\frac{p\left(x\right)}{\mu\left(x\right)}dx$$

where $\mu\left(x\right)$ is the density of the reference measure.

The continuous case

- The important difference is that if one now partitions the real line in increasingly finer subsets, the probabilities $p(x)$ and the background weights $\mu(x)$ are both split simultaneously and the logarithm of their ratio will generally not diverge. *The relative entropy is non increasing under refinement*.
- One replaces the concept of absolute entropy by that of relative entropy. The MEM is now generalized to the *maximum relative entropy method* (MREM).
- The new rule however is different from the earlier one because it is relative to a choice of the background measure. Different choices of μ will lead to different probability assignments.
  An interpretation of the background measure μ is that it represents a prior distribution that corresponds to our knowledge of the system before the information encapsulated in the constraints. *In this interpretation the MREM becomes a rule for changing or updating a previous probability distribution.*

Pairwise MEM

For L random variables $\overrightarrow{x} = (x_1, x_2, \cdots, x_L)$ for which we know M samples, estimate the probability $P(\overrightarrow{x})$. Maximize

$$\begin{aligned}\mathcal{L} &= -\int P(\overrightarrow{x}) \ln P(\overrightarrow{x})\, d\overrightarrow{x} + \alpha \left(\int P(\overrightarrow{x})\, d\overrightarrow{x} - 1 \right) \\ &\quad + \sum_{i=1}^{L} \beta_i \left(\frac{1}{M} \sum x_i - \overline{x}_i \right) + \sum_{i,j=1}^{L} \gamma_{ij} \left(\frac{1}{M} \sum_{m=1}^{M} x_i x_j - \overline{x_i x_j} \right)\end{aligned}$$

$$\frac{\delta \mathcal{L}}{\delta P(\overrightarrow{x})} = 0 \Longrightarrow -\ln P(\overrightarrow{x}) - 1 + \alpha + \sum_{i=1}^{L} \beta_i x_i + \sum_{i,j=1}^{L} \gamma_{ij} x_i x_j = 0$$

$$P(\overrightarrow{x}) = \exp\left(-1 + \alpha + \sum_{i=1}^{L} \beta_i x_i + \sum_{i,j=1}^{L} \gamma_{ij} x_i x_j \right)$$

The moment problem and reconstruction of processes

- In the classical moment problem a positive density $P(x)$ is sought from knowledge of its power moments

$$\int_a^b x^n P(x) dx = \mu_n$$

- Underdetermined inverse. MEM provides a solution; $\frac{\partial S}{\partial P(x)} = 0$

$$S = -\int P(x) lnP(x) dx + \sum_{n=0}^{N} \lambda_n \left(\int x^n P(x) dx - \mu_n \right)$$

- Reconstruction of processes: Burgh's theorem
  The maximum entropy stochastic process $\{X_i\}$ satisfying the constraints $\mathbb{E}\left[X_i X_{i+k}\right] = \alpha_k, k = 0, 1, ..., p$ for all i is the p-order Gauss–Markov process of the form

$$X_i = -\sum_{k=1}^{p} c_k X_{i-k} + Z_i$$

where the Z_i are i.i.d. $\backsim N(0, \sigma^2)$

Applications of MEM

- There are many applications of MEM, whenever inference under imcomplete information is needed: Image restoration, anomaly detection, sociology, biology,etc.

  MEM is used to compensate for the ignorance about the details. The hope is to eventually obtain robust universal behaviors.

- Before MEM, *pairwise interactions from data* are obtained by correlation or partial correlation

  Pearson correlation

$$C_{ij} = \frac{1}{M} \sum_{m=1}^{M} \left(x_i^{(m)} - \overline{x}_i \right) \left(x_j^{(m)} - \overline{x}_j \right)$$

$$r_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} C_{jj}}}$$

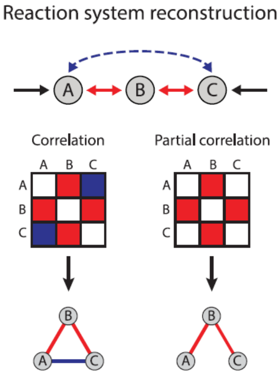
does not distinguish between direct dependence and indirect association

Correlation versus partial correlation

Partial correlation would be a better measure: Example

$$x_i \rightarrow y_i = (x_i - \overline{x}_i) \, / \sqrt{C_{ii}} \Longrightarrow r_{ij} = \overline{y_i y_j}$$

$$r_{AB|C} = \frac{r_{AB} - r_{BC} r_{AC}}{\sqrt{1 - r_{AC}^2}\sqrt{1 - r_{BC}^2}}$$



Reaction system reconstruction

Correlation

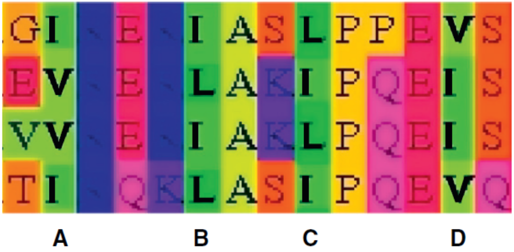
	A	B	C
A	⬜	🟥	🟦
B	🟥	⬜	🟥
C	🟦	🟥	⬜

Partial correlation

	A	B	C
A	⬜	🟥	⬜
B	🟥	⬜	🟥
C	⬜	🟥	⬜

MEM and biological networks

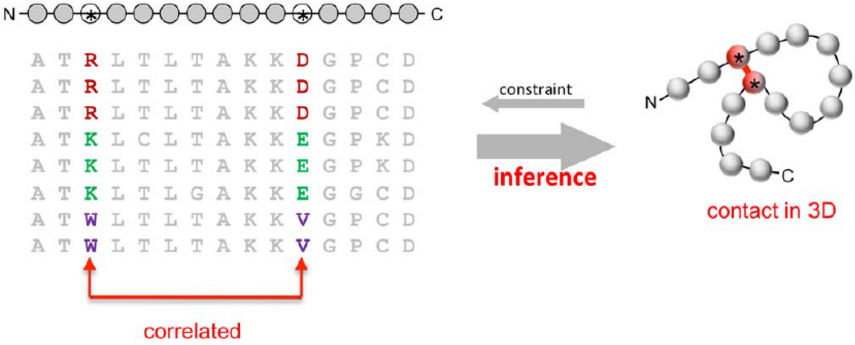
- **Protein 3D structure computed from evolutionary sequence variation**
  The idea behind the search for correlated mutations is that residues in contact would impose constraints on each other, which would lead to a correlation between the substitution patterns in a multiple sequence alignment. Correlated mutations of residues in contact to maintain the function and, by implication, the shape of the protein.



```
GI-E-IASLPPEVS
EV-E-LAKIPQEIS
VV-E-IAKLPQEIS
TI-QKLASIPQEVQ
```

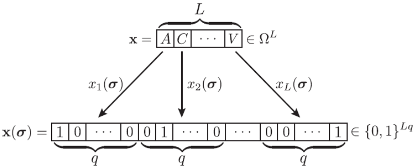
(A↔D) Valine↔Isoleucine (B↔C) Leucine↔Isoleucine

MEM and biological networks



```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

MEM and biological networks

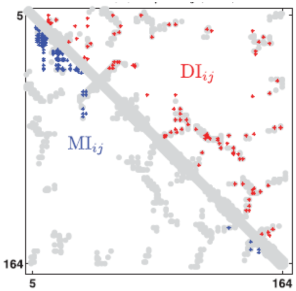


L
$\mathbf{x} = A\ C\ \cdots\ V \in \Omega^L$
$x_1(\boldsymbol{\sigma})$
$x_2(\boldsymbol{\sigma})$
$x_L(\boldsymbol{\sigma})$
$\mathbf{x}(\boldsymbol{\sigma}) = 1\ 0\ \cdots\ 0\ 0\ 1\ \cdots\ 0\ \cdots\ 0\ 0\ \cdots\ 1 \in \{0,1\}^{Lq}$
q
q
q

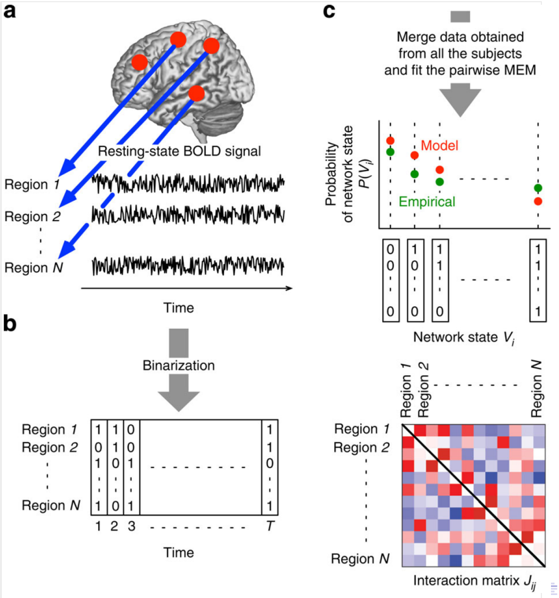
$$MI_{ij} = \sum_{\sigma\omega} \rho_{ij}(\sigma\omega) \ln \frac{\rho_{ij}(\sigma\omega)}{\rho_i(\sigma)\,\rho_j(\omega)}$$

$$DI_{ij} = \sum_{\sigma\omega} P_{ij}(\sigma\omega) \ln \frac{P_{ij}(\sigma\omega)}{\rho_i(\sigma)\,\rho_j(\omega)}$$

$$\begin{aligned} & P(x(\sigma)) \\ = \; & \frac{1}{Z} \exp\left(\sum_{i=1}^{L} \sum_{\sigma\in\Omega} \beta_i(\sigma)\, x_i(\sigma) \right. \\ & \left. + \sum_{i,j=1}^{L} \sum_{\sigma,\omega} \gamma_{ij}(\sigma,\omega)\, x_i(\sigma) x_j(\omega) \right) \end{aligned}$$



MEM and the reconstruction of neuron networks



MEM and traffic anomalies

- Divide packets into multidimensional packet classes according to the packets' protocol information and destination port numbers. For example TCP and UDP packects then subdivided according to the destination ports. *These packet classes serve as the domain of the probability space*.
- The *baseline distribution of the packet classes* is determined by learning a density model from the training data using Maximum Entropy estimation. The training data is a pre-labeled data set with the anomalies labeled by a human and in which packets labeled as anomalous are removed.
- In the detecting phase, an observed network traffic trace is given as the input. *The relative entropy of the packet classes in the observed traffic with respect to the baseline distribution is computed*. The packet classes that contribute significantly to the relative entropy are then recorded. If certain packet classes continue to contribute significantly to the relative entropy, anomaly warnings are generated and the corresponding packet classes are reported.

References

- C. Beck and F. Schlögl; *Thermodynamics of chaotic systems: an introduction*, Cambridge University Press, 1993.
- A. Rényi; *On measures of information and entropy*, Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability 1960.
- T. M. Cover and J. A. Thomas; *Elements of information theory* (2nd edition), Wiley-Interscience 2006.
- N. Wu; *The Maximum Entropy Method*, Springer 1997.
- L. R. Mead and N. Papanicolaou; *Maximum entropy in the problem of moments,* J. Math. Phys. 25 (1984) 2404-2417.
- J. Uffink; *Can the maximum entropy principle be explained as a consistency requirement*, Studies in History and Philosophy of Science Part B 26 (1995) 223-261.

References

- S. Thurner and R. Hanel; *Is There a World Behind Shannon? Entropies for Complex Systems* in A. Sanayei et al. (eds.), ISCS 2013: Interdisciplinary Symposium on Complex Systems, Springer 2014.

- R. Hanel, S. Thurner and M. Gell-Mann; *How multiplicity determines entropy: derivation of the maximum entropy principle for complex systems*, arXiv:1404.5650

- R. R. Stein et al.; *Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models*, PLOS Computational Biology (2015) DOI:10.1371/journal.pcbi.1004182.

- T. Watanabe et al.; *A pairwise maximum entropy model accurately describes resting-state human brain networks*, Nature Communications, 2013, DOI: 10.1038/ncomms2388.

- Y. Gu et al.; *Detecting anomalies in network traffic using maximum entropy estimation, 2005,* Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement, p.32.

Some problems on maximal entropy

- 1) For the problem of the prisoner in the land of the dice throwers, construct an example where the MEM estimate leads to the wrong result
- 2) Consider two independent integer-valued random variables, X and Y . Variable X takes on only the values of the eight integers $\{1, 2, ..., 8\}$ with uniform probability. Variable Y may take the value of any positive integer k, with probabilities $P\{Y = k\} = 2^{-k}$, $k = 1, 2, 3, ...$.

  Which random variable has greater uncertainty?
- 3) What is the maximal entropy distribution of the joint variables x,y with the following marginals. Hint: Use $S\left(x, y\right) \leq S\left(x\right) + S\left(y\right)$

$$\begin{array}{ccccc} & & y & & \\ & p_{11} & p_{12} & p_{13} & \frac{1}{2} \\ & p_{21} & p_{22} & p_{23} & \frac{1}{4} \\ x & p_{31} & p_{32} & p_{33} & \frac{1}{4} \\ & \frac{2}{3} & \frac{1}{6} & \frac{1}{6} & \end{array}$$

Some problems on maximal entropy

- 4) Find the maximal entropy process $\{X_i\}_{-\infty}^{+\infty}$ subject to
  (a) $\mathbb{E}\left[X_i^2\right] = 1$
  (b) $\mathbb{E}\left[X_i^2\right] = 1$ and $\mathbb{E}\left[X_i X_{i+1}\right] = \frac{1}{2}$
  Hint: Use or prove Burgh's theorem