Metodi di massima entropia

Michele Castellana *Laboratoire Physico–Chimie Curie, Institut Curie, PSL Research University; Sorbonne Universités, UPMC Univ. Paris 06*

I metodi di massima entropia (MME) costituiscono uno strumento teorico sistematico per costruire modelli per sistemi fisici: questi modelli devono essere consistenti con un insieme di misure, ma allo stesso tempo avere quanta meno struttura possibile. Questo metodo costituisce, in linea di principio, una strategia priva di *bias* per costruire un modello del fenomeno in considerazione, persino in presenza di una quantità limitata di dati. In quanto segue presenteremo una breve introduzione ai principali aspetti matematici e concettuali dei MME. La seconda parte di questo articolo sarà focalizzata sui possibili punti deboli, sia concettuali che tecnici, dei MME, in modo da fornire al lettore un'analisi critica di questi metodi.

Maximum-entropy methods (MEMs) provide a systematic theoretical framework for building models of physical systems which are consistent with some set of measurements, but otherwise have as little structure as possible. This constitutes, in principle, a bias-free, sensible strategy which allows one to model the phenomenon under consideration, even in the presence of a limited amount of data. In what follows, we will present a short introduction to the main mathematical and conceptual aspects of MEMs. The second part of this review will focus on the potential issues, both conceptual and technical, to which MEMs are subject, so as to provide a critical assessment of the method for the general reader.

Introduzione

Il lavoro pionierisitco di Shannon [1] ha dimostrato che l'entropia di una distribuzione di probabilità può indicare la quantità di `struttura` contenuta nella distribuzione. Shannon considerò un insieme di eventi $i = 1,...,N$ che avvengono con probabilità p_i ed un insieme di criteri intuitivi e minimali per una funzione $S[\boldsymbol{p}]$ che possa rappresentare la quantità di struttura contenuta in $\boldsymbol{p}$, in modo tale che più grande è S, minore è la quantità di struttura. Shannon dimostrò che l'unica quantità che soddisfa questi criteri è

Introduction

The pioneering work by Shannon [1] showed that the entropy of a probability distribution can be regarded as an indicator of the amount of 'structure' contained in such distribution. Shannon considered a set of events $i = 1,...,N$ which happen with probability p_i, and a set of intuitive, minimal criteria for a function $S[\boldsymbol{p}]$ which is required to represent the amount of structure contained in $\boldsymbol{p}$, i.e., the larger S, the smaller the structure. Shannon demonstrated that the only

l'entropia

$$S[\boldsymbol{p}] \equiv -\sum_i p_i \log p_i. \tag{1}$$

In particolare, i criteri per S sono i seguenti:

1. S è una funzione continua di $\boldsymbol{p}$.
2. Se tutti i p_i sono uguali, S è una funzione monotona crescente del numero di eventi N
3. Se un evento j è stato suddiviso in più eventi secondari, l'entropia $S[\boldsymbol{p}]$ originale è uguale alla somma di $S[\boldsymbol{p}']$, dove $\boldsymbol{p}'$ corrisponde alla probabilità degli eventi suddivisi, e dell'entropia degli eventi secondari, dove quest'ultima è pesata con la probabilità p'_j—vedi Fig. 1 per un esempio illustrativo.

I tre criteri di cui sopra possono essere interpretati nel modo seguente. Oltre alla condizione di continuità 1, il criterio 2 riflette il fatto che, se tutti gli eventi si verificano con la stessa probabilità, maggiore è il numero di eventi, N, minore è la quantità di struttura codificata in $\boldsymbol{p}$. Ad esempio, per $N = 1$, $\boldsymbol{p}$ sarebbe una distribuzione altamente strutturata, poiché il processo casuale in esame produrrebbe un singolo evento in modo deterministico. D'altra parte, per N grande, $\boldsymbol{p}$ assegnerebbe la stessa probabilità ad un gran numero di eventi, contenendo quindi una minore quantità di struttura ed essendo quindi meno informativa sull'esito del processo casuale.
Infine, la condizione 3 riflette l'idea che, se uno degli eventi è suddiviso in un insieme di eventi secondari, la quantità di struttura, o entropia, di $\boldsymbol{p}$ deve essere data dalla somma dell'entropia relativo agli eventi, più l'entropia degli eventi secondari.

Il metodo

La definizione quantitativa di cui sopra della struttura contenuta in una distribuzione di probabilità ha consentito diversi importanti progressi nel campo dell'inferenza statistica. Infatti, la relazione (1) ha consentito un'implementazione matematica diretta del principio del rasoio di Ockam [2]. Conosciuto anche come legge della parsimonia, il rasoio di Ockham è un principio che, in generale, afferma che la soluzione più semplice ad un problema è probabilmente quella corretta. A questo proposito, il termine `rasoio` si

quantity which satisfies these criteria is the entropy

$$S[\boldsymbol{p}] \equiv -\sum_i p_i \log p_i. \tag{1}$$

In particular, the criteria for S are the following:

1. S is a continuous function of $\boldsymbol{p}$.
2. If all p_is are equal, then S is a monotonically increasing function of the number of events. N
3. If an event j were broken into multiple subevents, then the original entropy $S[\boldsymbol{p}]$ equals the sum of $S[\boldsymbol{p}]$, where $\boldsymbol{p}'$ corresponds to the probability of the broken events, and the entropy of the subevents, where the latter is weighted with the probability p'_j—see Fig. 1 for an illustrative example.

The three criteria above allow for the following interpretation. In addition to the continuity condition 1, criterion 2 reflects the expectation that, if all events occur with the same probability, the larger the number of events, N, the smaller the amount of structure encoded in $\boldsymbol{p}$. For example, for $N = 1$, $\boldsymbol{p}$ would be a highly structured distribution, implying that the random process under consideration yields a single event in a deterministic way. On the other hand, for larger N, $\boldsymbol{p}$ would assign the same probability to a large number of events, thus being less informative, i.e., bearing a smaller amount of structure, on the outcome of the random process.
Finally, condition 3 reflects the idea that, if one of the events is broken into a set of subevents, then the amount of structure, or entropy, of $\boldsymbol{p}$ must be given by the sum of the entropy relative to the events, plus the entropy of the subevents.

The method

A quantitative definition for the amount of structure encoded in a probability distribution allowed for several important advances in the field of statistical inference. In fact, the relation (1) allowed for a direct, mathematical implementation of the long-standing principle of Ockam's razor [2]. Also known as the law of parsimony, Ockham's razor is a principle which, generally speaking, states that the simplest possible solution to a problem is most likely the correct one. In this regard, the term 'razor' refers to the act of

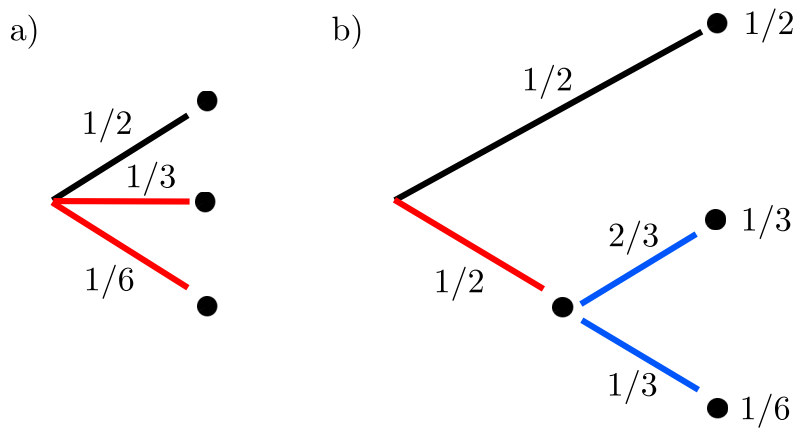

Figura 1: *Uno dei requisiti di Shannon per la funzione entropia. a) Tre eventi (punti) con probabilità* $\boldsymbol{p} = (1/2, 1/3, 1/6)$ *ed entropia* $S[\boldsymbol{p}]$*. b) Gli eventi in a) evidenziati in rosso sono sostituiti da un evento con probabilità* $1/2$*, suddiviso in due eventi secondari con probabilità* $2/3$ *e* $1/3$*. Complessivamente, in b) ci sono tre eventi possibili, che si verificano con probabilità* $(1/2, 1/3, 1/6)$*, come in a). Ponendo* $\boldsymbol{p}' = (1/2, 1/2)$ *l'entropia di b) è scritta come una combinazione lineare dell'entropia di eventi e sottoeventi come* $S[\boldsymbol{p}'] + \frac{1}{2} S[(2/3, 1/3)]$ *e si impone che questa entropia sia uguale a quella di a), ovvero* $S[\boldsymbol{p}]$*.*
One of Shannon's requirement for the entropy function. a) Three events (dots) with probabilities $\boldsymbol{p} = (1/2, 1/3, 1/6)$*, and entropy* $S[\boldsymbol{p}]$*. b) The events in a) highlighted in red are replaced by one event with probability* $1/2$*, which is broken into two subevents with probabilities* $2/3$ *and* $1/3$*. Overall, in b) there are three possible events, which occur with probabilities* $(1/2, 1/3, 1/6)$*, as in a). Setting* $\boldsymbol{p}' = (1/2, 1/2)$ *he entropy of b) is written as a linear combination of the entropy of events and subevents as* $S[\boldsymbol{p}'] + \frac{1}{2} S[(2/3, 1/3)]$*, and it is required to equal that of a), i.e.,* $S[\boldsymbol{p}]$*.*

riferisce all'atto di `radere`, e quindi eliminare, caratteristiche superflue e non necessarie in una soluzione. L'idea del rasoio di Ockham può essere facilmente associata all'entropia. Consideriamo a questo proposito un fenomeno dato da N possibili eventi, che si verificano con una probabilità $\boldsymbol{p}$, che vogliamo determinare. Poiché l'entropia $S[\boldsymbol{p}]$ rappresenta la struttura, o complessità, di $\boldsymbol{p}$, il rasoio di Ockham implica che la distribuzione che è più probabile che sia corretta è quella con l'entropia maggiore

$$\begin{cases} \max_{\boldsymbol{p}} S[\boldsymbol{p}], \\ \text{soggetto a} \\ \sum_i p_i = 1. \end{cases} \tag{2}$$

Nella terza riga di Eq. (2) abbiamo incluso la condizione di normalizzazione per $\boldsymbol{p}$. Ciò costituisce un esempio illustrativo di come il principio del rasoio di Ockham sia implementato in pratica: cerchiamo la soluzione più semplice rimuovendo tutti i presupposti superflui e mantenendo solo quelli strettamente necessari. In questo caso, la condizione di normalizzazione della probabilità è l'unico assunto fondamentale che viene preso in considerazione.
L'equazione (2) è la più semplice formulazione illustrativa di un metodo di inferenza statistica, noto

'shaving off' superfluous and unnecessary features in a solution. The idea of Ockham's razor may be naturally related to the entropy. In this regard, let us consider a phenomenon given by N possible events, which are assumed to occur with a probability $\boldsymbol{p}$, which we want to determine. Because the entropy $S[\boldsymbol{p}]$ represents the amount of structure, or complexity, of $\boldsymbol{p}$, Ockham's razor implies that the distribution which is most likely to be correct is the one with the largest entropy

$$\begin{cases} \max_{\boldsymbol{p}} S[\boldsymbol{p}], \\ \text{subject to} \\ \sum_i p_i = 1. \end{cases} \tag{2}$$

In the third line of Eq. (2), we included the normalization condition for $\boldsymbol{p}$. This constitutes a useful, illustrative example of how the Ockham's razor principle is practically implemented: we seek for the simplest solution by removing all unnecessary assumptions, and keeping only those that are strictly necessary. Here, the normalization condition for the probability is the only fundamental assumption which is taken into account.
Equation (2) is the simplest, illustrative mathematical formulation of a statistical-inference method, known

come metodo di massima entropia (MME). Originariamente introdotto da E. T. Jaynes [3], il MME consiste nel cercare la distribuzione di probabilità meno strutturata—o con la massima entropia—che sia coerente con un insieme di condizioni. In questo esempio, la condizione più semplice appare nella terza riga di Eq. (2) ed è data dalla normalizzazione della distribuzione di probabilità.

In generale, nei MME si possono imporre altre condizioni, a seconda del fenomeno specifico in esame. A questo proposito, si consideri un'osservabile fisica, $\mathscr{O}_i$, che dipende dall'evento i. Ad esempio, se il fenomeno in esame è un dado che viene lanciato ed i etichetta il lato del dado che si trova sulla sua superficie superiore, allora $\mathscr{O}_i$ può indicare il numero scritto sul lato i. Se vengono effettuati T lanci del dado, la media empirica di $\mathscr{O}$ è data da

$$\langle \mathscr{O} \rangle_{\text{exp}} = \frac{1}{T} \sum_{t=1}^{T} \mathscr{O}_{i(t)}, \tag{3}$$

dove $i(t)$ denota il lato del dado ottenuto nel t-esimo lancio. Il MME permette di ricostruire $\boldsymbol{p}$ secondo il seguente principio:

$$\begin{cases} \max_{\boldsymbol{p}} S[\boldsymbol{p}], \\ \text{soggetto a} \\ \langle \mathscr{O} \rangle_{\boldsymbol{p}} = \langle \mathscr{O} \rangle_{\text{exp}}, \\ \sum_i p_i = 1, \end{cases} \tag{4}$$

dove

$$\langle \mathscr{O} \rangle_{\boldsymbol{p}} = \sum_{i=1}^{N} p_i \mathscr{O}_i \tag{5}$$

è la media di $\mathscr{O}$ ottenuta con la distribuzione $\boldsymbol{p}$.

Nell'Eq. (4) abbiamo supposto che, assieme alla condizione di normalizzazione, la media di $\mathscr{O}$ costituisca l'osservabile minimale da includere nel modello. Per imporlo, nella terza riga di Eq. (4) richiediamo che la media di $\mathscr{O}$ ottenuta dal modello, $\langle \mathscr{O} \rangle_{\boldsymbol{p}}$, corrisponda a quella sperimentale, $\langle \mathscr{O} \rangle_{\text{exp}}$.
Procedendo come sopra, la formulazione ME (4) può essere facilmente generalizzata a più osservabili, $\mathscr{O}, \mathscr{Q}, \cdots$, imponendo che le medie del modello corrispondano alle rispettive stime sperimentali.

È importante sottolineare che, se il numero T di campioni sperimentali fosse abbastanza grande, si potrebbe campionare direttamente p_i dai dati empirici, senza ricorrere al MME. Tuttavia, in una varietà di istanze sperimentali, il numero di campioni è suf-

as the maximum-entropy method (MEM). Originally introduced by E. T. Jaynes [3], the MEM consists in seeking the least structured—or maximum-entropy—probability distribution which is consistent with a set of conditions. In this example, the simplest condition appears in the third line of Eq. (2), and is given by the normalization of the probability distribution.

More generally, other conditions may be imposed according to the specific phenomenon under consideration. In this regard, consider a physical observable, or 'feature', $\mathscr{O}_i$, which depends on the event i. For instance, if the phenomenon under consideration is a die which is rolled and i labels the side of the die which lies on its upper surface, then $\mathscr{O}_i$ may denote the number marked on side i. If T experimental trials are made, then the experimental average of $\mathscr{O}$ is given by

$$\langle \mathscr{O} \rangle_{\text{exp}} = \frac{1}{T} \sum_{t=1}^{T} \mathscr{O}_{i(t)}, \tag{3}$$

where $i(t)$ denotes the side of the die obtained in the t-th trial. The MEM allows one to reconstruct $\boldsymbol{p}$ according to the following principle:

$$\begin{cases} \max_{\boldsymbol{p}} S[\boldsymbol{p}], \\ \text{subject to} \\ \langle \mathscr{O} \rangle_{\boldsymbol{p}} = \langle \mathscr{O} \rangle_{\text{exp}}, \\ \sum_i p_i = 1, \end{cases} \tag{4}$$

where

$$\langle \mathscr{O} \rangle_{\boldsymbol{p}} = \sum_{i=1}^{N} p_i \mathscr{O}_i \tag{5}$$

is the average of $\mathscr{O}$ obtained with the distribution $\boldsymbol{p}$. In Eq. (4), we assumed that, together with the normalization condition, the average of $\mathscr{O}$ constitutes the minimal feature that needs to be included in the model. To enforce this, in the third line of Eq. (4) we impose that the average of $\mathscr{O}$ obtained from the model, $\langle \mathscr{O} \rangle_{\boldsymbol{p}}$, matches the experimental one, $\langle \mathscr{O} \rangle_{\text{exp}}$. Proceeding along the same lines as above, the ME formulation (4) can be easily generalized to multiple observables, $\mathscr{O}, \mathscr{Q}, \cdots$, whose model averages are required to match their respective experimental estimates.

It is important to point out that, if the number T of experimental trials were large enough, one would be able to directly sample p_i from the empirical data, with no need to resort to the MEM. However, in a variety of experimental instances, the number of sam-

ficientemente grande per stimare solo le medie di un numero finito di osservabili [4, 5], ad esempio, $\langle \mathscr{O} \rangle_{\rm exp}$, non l'intera distribuzione $\boldsymbol{p}$. Di conseguenza, il MME costituisce un metodo pratico per ricostruire, in modo approssimato, la distribuzione di probabilità in presenza di una quantità limitata di dati, facendo leva sull'ipotesi che le informazioni rilevanti siano contenute in una quantità fondamentale, come ad esempio la media di un'osservabile $\mathscr{O}$.
Dal punto di vista matematico, il MME (4) è un problema di ottimizzazione vincolata rispetto alle variabili $p_1, \cdots, p_N$, risolvibile con il metodo dei moltiplicatori di Lagrange. La funzione Lagrange è

$$\mathscr{L}[\boldsymbol{p}] = S[\boldsymbol{p}] - \lambda \left(\langle \mathscr{O} \rangle_{\boldsymbol{p}} - \langle \mathscr{O} \rangle_{\rm exp} \right) - \mu \left(\sum_i p_i - 1 \right), \tag{6}$$

e le condizioni di stazionarietà di $\mathscr{L}$ rispetto a p_i, λ e μ danno

$$\begin{aligned} p_i &= e^{-\lambda \mathscr{O}_i - \mu - 1}, & (7) \\ \langle \mathscr{O} \rangle_{\boldsymbol{p}} &= \langle \mathscr{O} \rangle_{\rm exp}, & (8) \\ \sum_i p_i &= 1, & (9) \end{aligned}$$

rispettivamente, dove nella prima riga abbiamo usato l'Eq. (5). Risolvendo le Eq.ni (7) - (9) rispetto a $\boldsymbol{p}$, λ e μ e sostituendo la soluzione in Eq. (7), si ottiene la distribuzione di probabilità ME $\boldsymbol{p}$.

L'equazione (7) mostra che la distribuzione di probabilità ME dipende esplicitamente dalla scelta della caratteristica $\mathscr{O}$. Inoltre, osserviamo che la forma esponenziale di $\boldsymbol{p}$ ricorda la distribuzione di probabilità di Boltzmann in meccanica statistica [6]. Infatti, uno degli esempi tipici di utilizzo del MME consiste nella derivazione della distribuzione di Boltzmann stessa. Infatti, è stato dimostrato [3] che, se l'indice i etichetta uno stato di un sistema fisico ed $\mathscr{E}_i$ è la sua energia interna, allora la distribuzione ME coerente con il vincolo che l'energia interna media è uguale ad E è la distribuzione di Boltzmann:

$$p_i = \frac{1}{Z} e^{-\mathscr{E}_i/(k_{\rm B}T)}. \tag{10}$$

Nella relazione qui sopra, $Z = \sum_i e^{-\mathscr{E}_i/(k_{\rm B}T)}$ è la funzione di partizione, $k_{\rm B}$ la costante di Boltzmann, e la temperatura inversa $1/(k_{\rm B}T)$ coincide con il moltiplicatore di Lagrange per il vincolo $\langle \mathscr{E} \rangle_{\boldsymbol{p}} = E$, che mette implicitamente in relazione energia interna e temperatura.

ples is sufficient to estimate only averages of a finite number of features [4, 5], e.g., $\langle \mathscr{O} \rangle_{\rm exp}$, not the entire distribution $\boldsymbol{p}$. As a result, the MEM may be used as a practical tool to approximately reconstruct the probability distribution in the presence of a limited amount of data, by leveraging the hypothesis that the relevant information is encoded into a fundamental, minimal quantity such as the average of a feature $\mathscr{O}$.

From the mathematical standpoint, the MEM (4) is a constrained optimization problem with respect to the variables $p_1, \cdots, p_N$, which can be solved with the method of Lagrange multipliers. The Lagrange function reads

$$\mathscr{L}[\boldsymbol{p}] = S[\boldsymbol{p}] - \lambda \left(\langle \mathscr{O} \rangle_{\boldsymbol{p}} - \langle \mathscr{O} \rangle_{\rm exp} \right) - \mu \left(\sum_i p_i - 1 \right), \tag{6}$$

and the stationarity conditions of $\mathscr{L}$ with respect to p_i, λ and μ yield

$$\begin{aligned} p_i &= e^{-\lambda \mathscr{O}_i - \mu - 1}, & (7) \\ \langle \mathscr{O} \rangle_{\boldsymbol{p}} &= \langle \mathscr{O} \rangle_{\rm exp}, & (8) \\ \sum_i p_i &= 1, & (9) \end{aligned}$$

respectively, where in the first line we used Eq. (5). By solving Eqs. (7)-(9) for $\boldsymbol{p}$, λ and μ, and substituting the solution in Eq. (7), one obtains the ME probability distribution p_i.

Equation (7) shows that the ME probability distribution explicitly depends on the choice of the feature $\mathscr{O}$. In addition, we observe that the exponential shape of $\boldsymbol{p}$ is reminiscent of the Boltzmann's probability distribution in statistical mechanics [6]. In fact, one of the prototypical examples of the use of the MEM consists in the derivation of the Boltzmann's distribution itself. Indeed, it has been shown [3] that, if index i labels a state of a physical systems and $\mathscr{E}_i$ is its internal energy, then the ME distribution consistent with the constraint that the average internal energy is equal to E is the Boltzmann distribution:

$$p_i = \frac{1}{Z} e^{-\mathscr{E}_i/(k_{\rm B}T)}. \tag{10}$$

In the relation above, $Z = \sum_i e^{-\mathscr{E}_i/(k_{\rm B}T)}$ is the partition function, $k_{\rm B}$ is Botlzmann's constant, and the inverse temperature $1/(k_{\rm B}T)$ coincides with the Lagrange multiplier for the constraint $\langle \mathscr{E} \rangle_{\boldsymbol{p}} = E$, which implicitly relates internal energy and temperature.

Avendo applicazioni in molteplici campi, i MME si prestano ad essere utilizzati in particolare nel campo dell'intelligenza artificiale. Infatti, la definizione di un 'agente' intelligente come un'entità che percepisce il suo ambiente ed intraprende azioni in modo da massimizzare la probabilità di raggiungere un obiettivo [7], suggerisce un'analogia diretta con l'inferenza statistica ed i MME. In altre parole, l'agente costruisce un modello della realtà basato su un insieme di input esterni, così come i MME costruiscono il modello minimale (7) basato sui dati sperimentali $\langle \mathscr{O} \rangle_{\text{exp}}$ che vengono inseriti in esso. Successivamente, questo modello può essere utilizzato dall'agente per prevedere il comportamento futuro del sistema in esame e, sfruttando queste previsioni, per prendere una decisione al fine di raggiungere un obiettivo specifico.

Among their applications in multiple fields, a notable use of MEMs concerns the domain of artificial intelligence. In fact, the definition of an intelligent 'agent' as an entity which perceives its environment and takes actions so as to maximize the probability of achieving a goal [7], suggests a direct analogy with statistical inference and MEMs. Namely, the agent builds a model of the reality based on a set of external inputs, in the same way in which MEMs build the minimal model (7) based on the experimental data $\langle \mathscr{O} \rangle_{\text{exp}}$ which is fed into it. Later on, this model may be used by the agent to predict the future behavior of the system under consideration and, by leveraging these predictions, to successfully tailor a decision in order to achieve a specific goal.

Punti deboli e critiche

Basi concettuali

Una critica fondamentale che può essere rivolta ai MME riguarda la loro logica concettuale. In effetti, l'assenza di ipotesi superflue nel principio del rasoio di Ockham viene spesso presentata come un punto di forza del metodo, secondo l'idea che nessun *bias* soggettivo, né alcun ingrediente superfluo, vengano introdotti nel modello, ad eccezione dei i dati sperimentali stessi.
Tuttavia, l'assenza di ipotesi può essere considerata essa stessa un'ipotesi non banale. Questo punto può essere illustrato con il MME per uno stormo di uccelli [8]. In breve, la forza di gravità cui sono soggetti gli uccelli indica che non tutte le direzioni spaziali sono equivalenti per lo stormo: di conseguenza, se non si include esplicitamente il ruolo speciale svolto dalla direzione verticale nelle osservabili del MME, si fa, di fatto, un'ipotesi non banale, che potrebbe influenzare i risultati dell'inferenza.

Scelta delle osservabili

Come abbiamo discusso qui sopra, i MME si basano sull'ipotesi che le informazioni fenomenologiche fondamentali siano incluse nella media di un'osservabile $\mathscr{O}$. Tuttavia, la scelta di questa osservabile è dettata dall'intuizione fisica di colui che studia il fenomeno. Ad esempio, per uno stormo di uccelli che volano coerentemente con direzioni di moto quasi parallele, una possibile scelta è data dalla correlazione e

Issues and criticisms

Conceptual basis

A fundamental criticism which may be adressed to MEMs concerns its conceptual rationale. In fact, the absence of superfluous assumptions in Ockham's razor principle is often presented as a selling point of the method, along with the idea that no subjective bias, nor superfluous ingredients, are introduced in the model except for the data itself.
However, the absence of assumptions may be regarded as a nontrivial assumption itself. This point may be illustrated with the MEM for a flock of birds [8]. In short, the force of gravity to which birds are subject indicates that not all spatial directions are equivalent for the flock: as a result, if one does not explicitly include the special role played by the vertical direction in the MEM features, one may be ultimately making a nontrivial assumption, which could bias the results of the ME analysis.

Choice of features

As we discussed above, the MEM is based on the hypothesis that the key phenomenological information is included in the average of a feature $\mathscr{O}$. However, the choice of this feature is dictated by one's physical intuition on the phenomenon under consideration. For instance, for a flock of birds which fly coherently with nearly parallel directions of motion, a possible choice is given by the velocities' correlation and polarization [8]. For the study of collective behavior in

polarizzazione delle velocità [8]. Per lo studio del comportamento collettivo in reti di neuroni, si possono considerare come osservabili i *firing rates* e la funzione di correlazione per i gli *spikes* [9]. Tuttavia, è importante sottolineare che queste ipotesi non sono uniche e sono dettate non solo dalla prospettiva soggettiva di chi osserva il fenomeno, ma anche dalle caratteristiche fisiche che si vogliono studiare nell'esperimento. Come mostrato in Eq. (7), la distribuzione di probabilità di ME dipende dalla scelta di queste osservabili: ne segue che il risultato del MME deve essere sempre considerato con spirito critico e sottoposto a verifica per valutarne l'affidabilità.
Inoltre, un dato insieme di osservabili minimali per il MME potrebbe non essere sufficiente per descrivere correttamente la fenomenologia. Un semplice esempio illustrativo di questa situazione è costituito dal MME per una popolazione di neuroni. Se solo il *firing rate* di ogni neurone venisse incluso nel metodo come osservabile, la distribuzione ME risultante sarebbe data dal prodotto di distribuzioni di *spikes* indipendenti dei neuroni [9]. Ne segue che tale distribuzione ME non potrebbe descrivere alcun comportamento collettivo della rete. Per descrivere questo comportamento collettivo è necessario includere osservabili aggiuntive, come la correlazione a coppie tra gli *spikes* [9]. Tuttavia va ricordato che, anche in presenza di questa osservabile, i risultati del MME potrebbero essere ulteriormente alterati —sia quantitativamente che qualitativamente—se vi si includessero altre osservabili. In altre parole, i risultati ME andrebbero considerati esatti solo se si considerasse un numero infinito di osservabili indipendenti.

Interpretazione

Osservando l'Eq. (7), si può essere tentati di sfruttare l'equivalenza con una distribuzione di Boltzmann ed interpretare $\mathscr{O}_i$ come l'energia del sistema associata allo stato i. Ad esempio, se gli stati del sistema formassero un insieme continuo e fossero etichettati da una variabile reale x, allora si sarebbe tentati di interpretare $\mathscr{O}(x)$ come l'Hamiltoniana del sistema in esame e di mettere in relazione $d\mathscr{O}/dx$ con una forza. Procedendo lungo questa linea, potremmo essere indotti ad affermare che la dinamica temporale del sistema sia data da un moto Browniano nel potenziale $\mathscr{O}(x)$.
Tuttavia, le interpretazioni qui sopra non sarebbero necessariamente corrette [10]. Ricordiamo infatti che

networks of neurons, one may consider as a feature the correlation function for the neural spikes and the firing rates [9]. However, it is important to point out that these hypotheses are not unique, and they are dictated not only by one's subjective perspective on the phenomenon, but also by the specific physical feature that one aims to characterize in the experiment. As shown in Eq. (7), the ME probability distribution depends on the choice of such features: As a resut, the outcome of the MEM must always be regarded with a critical spirit, and subjected to tests in order to assess its reliability.
In addition, a given set of chosen minimal features included in the MEM may not suffice to correctly describe the phenomenology. A simple, illustrative example of this situation consists in the MEM for a population of spiking neurons. If only the spiking frequencies of each neuron are included as features, then the resulting ME distribution is given by the product of independent spiking distributions of the neurons in the network, and it would fail to describe any collective behavior of the neural network as a whole [9]. In order to describe this collective behavior, additional features need to be included, such as the pairwise correlation between spikes [9]. However, it should be reminded that, even in the presence of this feature, the ME results may be further altered—both quantitatively and qualitatively—if additional features were included, and such results should be considered to be exact only if an infinitely large number of independent features were taken into account.

Interpretation

When we look at Eq. (7), it is tempting to leverage the equivalence with a Boltzmann distribution, and interpret $\mathscr{O}_i$ as the energy of the system associated with state i. For instance, if the states of the system formed a continuum set and were labeled by a real variable x, then one would be tempted to interpret $\mathscr{O}(x)$ as the Hamiltonian of the system under consideration, and to relate $d\mathscr{O}/dx$ to a force. Proceeding along the same lines, one would be induced to state that the system dynamics is given by a Browninan motion in the potential $\mathscr{O}(x)$.
However, the interpretations above need not be correct [10]. In fact, we recall that the function $\mathscr{O}(x)$ which appears in the exponential of the ME distribu-

la funzione $\mathscr{O}(x)$ che appare nell'esponenziale della distribuzione ME (7) è semplicemente il risultato di una costruzione matematica, ovvero l'ottimizzazione vincolata (4) e che essa dipende dalla scelta arbitraria delle osservabili del MME.
Di conseguenza, non è garantito che $\mathscr{O}(x)$ abbia alcun significato fisico, né che essa sia connessa ad un'Hamiltoniana o ad una forza. Inoltre, poiché ci sono infiniti processi dinamici che danno origine alla stessa distribuzione stazionaria [11], come ad esempio (7), la dinamica Browniana nel potenziale $\mathscr{O}(x)$ non rappresenta necessariamente la dinamica fisica del sistema.
Un esempio illustrativo del problema qui sopra è dato dai modelli ME per le reti neurali [9]. L'attività della cellula i è rappresentata da una variabile binaria $\sigma_i = \pm 1$, dove '$+1$' significa che la cellula emette uno *spike* e '-1' che resta silenziosa. Un evento è caratterizzato dalla configurazione $\sigma = (\sigma_1, \sigma_2, \cdots)$ della rete e le osservabili per il MME sono i 'firing rates' delle cellule $\sigma_1, \sigma_2, \cdots$ ed i prodotti $\sigma_1\sigma_2, \sigma_1\sigma_3, \cdots$ su tutte le coppie di cellule. Procedendo sulla falsariga dell'Eq. (4), la distribuzione ME è data da

$$p_\sigma \propto \exp\left(-\sum_i h_i\sigma_i - \sum_{i<j} J_{ij}\sigma_i\sigma_j\right), \qquad (11)$$

dove h_i e J_{ij} sono i moltiplicatori di Lagrange per i vincoli.
L'equazione (11) presenta una forte analogia con il modello di Ising in meccanica statistica [12], dove J_{ij} rappresenta il legame fisico, o interazione, tra gli spin i e j. A causa di questa somiglianza, J_{ij} può essere erroneamente interpretato come un'interazione fisica effettiva tra i neuroni i e j. Tuttavia, questa quantità è semplicemente un moltiplicatore di Lagrange nell'ottimizzazione vincolata e non è garantito che rappresenti—quantitativamente né qualitativamente—un'interazione o una connessione fisica tra cellule, come ad esempio una sinapsi.

Errore sperimentale

Un ulteriore punto delicato dei MME riguarda la presenza di incertezze nei dati sperimentali che vengono inseriti nel modello. Dato che le medie sperimentali $\langle\rangle_{\text{exp}}$ risultano da misure, esse possono essere affette da diversi errori, ad esempio strumentali, procedurali, ambientali, umani ed altri. Ad esempio, se il numero di campioni empirici T è abbastanza piccolo, la media $\langle\mathscr{O}\rangle_{\text{exp}}$ in Eq. (3) sarà affetta da un'incertezza

tion (7) is merely the result of a mathematical construction, i.e., the constrained optimization (4), and that it depends on the arbitrary choice of the ME features. As a result, $\mathscr{O}(x)$ is not guaranteed to bear any physical meaning, nor to be related to a Hamiltonian nor a physical force. On top of this, because there are infinitely many dynamical processes that give rise to the same stationary distribution [11] such as (7), the Brownian dynamics in the potential $\mathscr{O}(x)$ need not to represent the actual physical dynamics of the system.
An illustrative example of the issue above is given by ME models for neural networks [9]. The activity of cell i is represented by a binary variable $\sigma_i = \pm 1$, where $+1$ stands for spiking and -1 for being silent. An event is characterized by the configuration $\sigma = (\sigma_1, \sigma_2, \cdots)$ of the network, and the features for the ME construction are the 'spiking rate' of each cell $\sigma_1, \sigma_2, \cdots$, and the products $\sigma_1\sigma_2, \sigma_1\sigma_3, \cdots$ across all cell pairs. Proceeding along the lines of Eq. (4), the ME distribution reads

$$p_\sigma \propto \exp\left(-\sum_i h_i\sigma_i - \sum_{i<j} J_{ij}\sigma_i\sigma_j\right), \qquad (11)$$

where h_i and J_{ij} are the Lagrange multipliers for the constraints. Equation (11) bears a strong similarity to the Ising model in statistical mechanics [12], where J_{ij} represents the physical bond, or interaction, between spins i and j. Because of this similarity, J_{ij} may be misinterpreted as an actual, physical interaction between neurons i and j. Given that this quantity is merely a Lagrange multiplier in the constrained optimization, it is not guaranteed to represent—quantitatively nor qualitatively—physical interactions or connections between neural cells, such as a synapses.

Experimental uncertainties

A further issue with the MEM concerns the presence of uncertainties in the experimental data which is fed into the model. Given that the experimental averages $\langle\rangle_{\text{exp}}$ result from measurements, they may be affected by different sources of errors, e.g., instrumental, procedural, environmental, human, and others. For instance, if the number of empirical samples T is small enough, then the average $\langle\mathscr{O}\rangle_{\text{exp}}$ in Eq. (3) will be

significativa, data dall'errore standard della media. Di conseguenza, imporre che il vincolo di uguaglianza $\langle \mathscr{O} \rangle_{\boldsymbol{p}} = \langle \mathscr{O} \rangle_{\text{exp}}$ in Eq. (4) sia soddisfatto esattamente costituirebbe un criterio troppo rigido e potrebbe produrre risultati errati [13, 14].
Un esempio illustrativo di questo problema viene dai modelli di ME per la modellizzazione del linguaggio [14]. In questo caso, gli eventi sono dati dall'osservazione di coppie (w, w') di parole consecutive, w e w', in un testo. Date due parole, ad esempio, 'saint' e' George', consideriamo due osservabili $\mathscr{O}$ e $\mathscr{Q}$: La frequenza con cui si verifica 'George' nel testo

$$\mathscr{O}_{w,w'} = \mathbb{I}(w' = \text{George}), \tag{12}$$

e la frequenza della coppia 'saint George':

$$\mathscr{Q}_{w,w'} = \mathbb{I}(w = \text{saint}, w' = \text{George}), \tag{13}$$

dove la funzione indicatrice $\mathbb{I}$ è uguale ad uno se tutte le condizioni nel suo argomento sono soddisfatte mentre vale zero in caso contrario. Se si applica l'analisi ME ad un breve testo in cui la parola 'George' ricorre solo dopo 'saint' e si impongono questi vincoli nella loro forma di uguaglianza sulla falsariga di Eq. (4), è semplice dimostrare che la distribuzione ME soddisfa

$$p_{w,\,\text{George}} = 0 \text{ if } w \neq \text{saint}. \tag{14}$$

Questi eventi a frequenza nulla possono causare instabilità numeriche nella stima ME [13]. È ancora più importante ricordare che tali eventi possono provocare scarse prestazioni del modello ME: per esempio, se la distribuzione di ME (14) fosse usata per un riconoscimento di testo, allora qualsiasi evento (w, George) in cui w è diverso da 'saint' non sarebbe riconosciuto come una coppia di parole.
Una soluzione per superare il problema qui sopra consiste nell'allentare i vincoli di uguaglianza nel MME [14]. Ad esempio, supponiamo che i dati non siano sufficientemente accurati da fornire un valore per la media sperimentale, ma che essi possano stimare solo un intervallo di confidenza dato da un limite superiore e inferiore $\mathscr{O}_+$ e $\mathscr{O}_-$, rispettivamente:

$$\mathscr{O}_- \leq \langle \mathscr{O} \rangle_{\text{exp}} \leq \mathscr{O}_+. \tag{15}$$

A questo punto si può cercare la distribuzione meno strutturata $\boldsymbol{p}$, che è coerente con questa informazione sperimentale, riformulando il metodo ME (4) come

affected by a significant uncertainty, related to its standard error of the mean. As a result, a full satisfaction of the equality constraint $\langle \mathscr{O} \rangle_{\boldsymbol{p}} = \langle \mathscr{O} \rangle_{\text{exp}}$ in Eq. (4) would be too strict of a criterion, and it may produce misleading results [13, 14].
An illustrative example of this issue comes from ME models for language modeling [14]. In this case, the events are given by the observation of pairs (w, w') of consecutive words, w and w', in a corpus of text. Given two words, e.g., 'saint' and 'George', we consider two features $\mathscr{O}$ and $\mathscr{Q}$: The frequency with which 'George' occurs in the text

$$\mathscr{O}_{w,w'} = \mathbb{I}(w' = \text{George}), \tag{12}$$

and the frequency of the bigram 'saint George', i.e.,

$$\mathscr{Q}_{w,w'} = \mathbb{I}(w = \text{saint}, w' = \text{George}), \tag{13}$$

where the indicator function $\mathbb{I}$ is one if all conditions in its argument are satisfied, and zero otherwise. If one applies the ME analysis to a short corpus of text where the word 'George' occurs only after 'saint', and imposes these constraints in their equality form along the lines of Eq. (4), it is straightforward to show that the ME distribution satisfies

$$p_{w,\,\text{George}} = 0 \text{ if } w \neq \text{saint}. \tag{14}$$

These zero-frequency events in the ME model may cause numerical instabilities in ME estimation [13]. More importantly, such events may result in poor performance of the ME model: For instance, if the ME distribution (14) were used for text recognition, then any word pair (w, George) in which w is different from 'saint' would not be recognized as a bigram.
A solution to overcome the issue above consists in relaxing the equality constraints in the MEM [14]. For instance, let us suppose that the data is not accurate enough to provide a value for the experimental average, but can only give a confidence interval given by an upper and lower bound $\mathscr{O}_+$ and $\mathscr{O}_-$, respectively:

$$\mathscr{O}_- \leq \langle \mathscr{O} \rangle_{\text{exp}} \leq \mathscr{O}_+. \tag{15}$$

Then one may seek for the least-structured distribution $\boldsymbol{p}$, which is consistent with this experimental information, by reformulating the ME method (4) as

segue:

$$\begin{cases} \max_{\boldsymbol{p}} S[\boldsymbol{p}], \\ \text{soggetto a} \\ \mathscr{O}_- \leq \langle \mathscr{O} \rangle_{\text{exp}} \leq \mathscr{O}_+, \\ \sum_i p_i = 1. \end{cases} \tag{16}$$

Dal punto di vista matematico, Eq. (16) è un problema di massimizzazione con vincoli sia di uguaglianza che di disuguaglianza, che può essere risolto con metodi matematici noti, introdotti da W. Karush, H. W. Kuhn ed A. W. Tucker (KKT) a metà del ventesimo secolo [15, 16]. L'approccio KKT ha una somiglianza con il metodo Lagrange per i vincoli di uguaglianza, Eq. (6): ad ogni vincolo di disuguaglianza è associato un moltiplicatore KKT non negativo. In breve, se il moltiplicatore svanisce, allora il massimo si trova all'interno di una regione nello spazio delle probabilità $\boldsymbol{p}$, dove i vincoli di disuguaglianza sono soddisfatti: di conseguenza, il vincolo di disuguaglianza è irrilevante per la massimizzazione. D'altra parte, se il moltiplicatore è positivo, allora il massimo si trova sul bordo della regione in cui il vincolo è soddisfatto e la presenza del vincolo di disuguaglianza influenza il valore massimo dell'entropia nell'ottimizzazione.

Discussione

Data la loro generalità, versatilità computazionale e semplicità, negli ultimi anni i metodi di massima entropia (MME) sono stati applicati a una grande varietà di fenomeni. Spaziando dal comportamento animale collettivo, alle sequenze in famiglie di proteine, alle parole in un dizionario o testo, la sorprendente facilità con cui i MME possono essere applicati si è tradotta in un gran numero di modelli, in base ai quali si è cercato di capire i meccanismi che sono alla base dei fenomeni in esame. Dopo aver presentato una breve introduzione ai MME, ne abbiamo fornito un'analisi critica, in modo da guidare il lettore attraverso i potenziali svantaggi e punti deboli di questi metodi di inferenza. Sebbene questo elenco di potenziali problemi non sia esaustivo, il nostro obbiettivo è stato di fornire al lettore un'idea generale di quale tipo di critiche e domande dovrebbero e potrebbero essere indirizzate alle analisi ME e come interpretare correttamente le loro conclusioni sul fenomeno in esame.

follows:

$$\begin{cases} \max_{\boldsymbol{p}} S[\boldsymbol{p}], \\ \text{subject to} \\ \mathscr{O}_- \leq \langle \mathscr{O} \rangle_{\text{exp}} \leq \mathscr{O}_+, \\ \sum_i p_i = 1. \end{cases} \tag{16}$$

From the mathematical standpoint, Eq. (16) is a maximization problem with both equality and inequality constraints, which can be solved with known mathematical methods introduced by W. Karush, H. W. Kuhn and A. W. Tucker (KKT) in the mid-twentieth century [15, 16]. The method bears a similarity to the Lagrange method for equality constraints, Eq. (6): to each inequality constraint is associated a non-negative KKT multiplier. In short, if the multiplier vanishes, then the optimum lies within a region in the space of probabilities $\boldsymbol{p}$, where the inequality constraints is satisfied: as a result, the inequality constraint is irrelevant for the optimum. On the other hand, if the multiplier is positive, then the optimum is located at the boundary of the region where the constraint is satisfied, and the presence of the inequality constraint lowers the optimal value of the objective function.

Discussion

Given their generality, computational versatility and straightforwardness, in recent years maximum-entropy methods (MEMs) have been widely applied to a large variety of phenomena. Spanning from collective animal behavior, to sequences in families of proteins, to words in a dictionary or corpus of text and others, the striking easiness which MEMs can be applied resulted in a plethora of proposed models and interpretations on the underlying mechanisms of the phenomena under consideration. After presenting a short introduction to MEMs, we provided a critical assessment of MEMs, so as to guide the general reader through the potential drawbacks and weak spots of this inference method. While this list of potential issues is not meant to be exhaustive, we aimed at providing the reader with a general flavor of what kind of criticisms and questions should and could be adressed to ME analyses, and how to correctly interpret their results and claims on the phenomenon under consideration.

[1] C.E. Shannon: *A mathematical theory of communication*, Bell System Tech. J. 27 (1987) 379.

[2] E. Sober: *Ockham's Razors: A User's Manual*, Cambridge Univ. Press, Cambridge (UK) (2015).

[3] E.T. Jaynes: *Information theory and statistical mechanics*, Physical Review 106 (1957) 620.

[4] S. Cocco, R. Monasson, M. Weigt: *From Principal Component to Direct Coupling Analysis of Coevolution in Proteins: Low-Eigenvalue Modes are Needed for Structure Prediction*, PLoS Comput. Biol. 8 (2013) e1003176.

[5] M. Weigt, et al., *Identification of direct residue contacts in protein–protein interaction by message passing*, P. Natl. Acad. Sci. USA 106 (2009) 67.

[6] L.D. Landau, E.M. Lifshitz: *Statistical Physics: Course of Theoretical Physics*, Pergamon Press, London (1980).

[7] D. Poole, A. Mackworth, R. Goebel: *Computational Intelligence: a Logical Approach*, Oxford University Press, Oxford (1998).

[8] W. Bialek, et al.: *Flocking is a typical example of emergent collective behavior, where interactions between individuals produce collective patterns on the large scale.*, P. Natl. Acad. Sci. USA 109 (2012) 4786.

[9] E. Schneidman, et al.: *Weak pairwise correlations imply strongly correlated network states in a neural population*, Nature 440 (2006) 1007.

[10] M. Castellana, W. Bialek, A. Cavagna, I. Giardina: *Entropic effects in a nonequilibrium system: Flocks of birds*, Phys. Rev. E 93 (2016), 052416.

[11] P.C. Hohenberg, B.I. Halperin: *Theory of dynamic critical phenomena*, Rev. Mod. Phys. 49 (1977) 435.

[12] K. Huang: *Statistical Mechanics*, Wiley Press, New York (1987).

[13] J. Kazama, J. Tsujii: *Maximum entropy models with inequality constraints: A case study on text categorization*, Mach. Learn. 60 (2005) 159.

[14] S.F. Chen, R. Rosenfeld: *A survey of smoothing techniques for ME models*, IEEE T. Speech Audi. P. 8 (2000) 37.

[15] W. Karush: *Minima of functions of several variables with inequalities as side constraints*, Chicago Univ. (Math. Dept.) (1939).

[16] H.W. Kuhn, A.W. Tucker: *Second Berkeley Symposium on Mathematical Statistics and Probability*, Non-Linear Program. University of California Press, Berkeley (1951).

Michele Castellana: è ricercatore presso il Laboratoire Physico-Chimie Curie di Parigi, un'unità di ricerca del centro nazionale francese di ricerca scientifica (CNRS) che fa parte dell'Institut Curie. Si è laureato in fisica teorica presso l'Università La Sapienza di Roma, con una tesi su approcci di teoria di campo alla gravità quantistica. Per il suo dottorato di ricerca, svoltosi tra l'Università La Sapienza e l'Università Paris Sud, è passato alla fisica statistica e si è concentrato sui metodi di gruppo di rinormalizzazione per sistemi disordinati. Si è poi trasferito all'università di Princeton per il post-dottorato, dove ha lavorato su argomenti tra fisica statistica e biologia.

Michele Castellana: is an associate scientist at Laboratoire Physico-Chimie Curie in Paris, a research unit of the French National Centre for Scientific Research (CNRS) which is part of Institut Curie. He graduated in theoretical physics from Sapienza University of Rome, with a thesis on field-theoretical aspects of quantum gravity. For his Ph.D., joint between Sapienza University and University Paris Sud, he switched to statistical physics, and focused on renormalization-group methods for disordered systems. He then moved to Princeton University as a postdoctoral associate, where he worked on topics at the boundary between statistical and biological physics.