Learning under constraints: inference, selection and validation with maximum entropy models

Giuffrida, F.

CHAPTER 1

Introduction

Picture a lecture room full of students waiting for a class to begin. Some of them wear glasses. The room is divided into two equally sized sectors: the front and the back, each of which is fully occupied. However, even though all students arrived roughly at the same time, there was no tension in choosing where to sit, and everyone appeared satisfied with their position.

Suppose we start wondering whether students wearing glasses tend to prefer the front sector of the room. Intuitively, this seems reasonable: people with impaired vision might wish to sit closer to the blackboard. However, intuition alone is not sufficient. We therefore turn to the data to see whether the arrangement offers any concrete indication of such a preference.

We begin by counting. Among the students who wear glasses, we find that more than half of them are seated in the front sector. At first glance, this appears to support our intuition. Yet this number, by itself, is not very informative. Without a clear reference describing what we should expect in the absence of any seating preference, it is impossible to tell whether such an uneven distribution reflects a genuine tendency or is merely compatible with chance.

To proceed, we must first clarify what we mean by "chance" in this context. Doing so requires us to be explicit about which information we consider relevant to the problem. In this example, assuming that the room will always be fully occupied, we decide that our interpretation of "chance" should still take into account one element: the number of students wearing glasses. All other factors – such as the order of arrival, social interactions among students, or individual preferences unrelated to vision – are deliberately ignored, either because they are unobserved or because we deem them irrelevant to the question at hand. It is worth noting that we do not fix in advance how many students with glasses end up in the front or in the back. We intentionally leave this aspect unconstrained. In this way, the observed arrangement can be compared with what would normally arise under the assumptions we have chosen to encode.

Once this choice has been made, we are left with many possible seating arrangements that are all consistent with the information we have fixed. A fundamental question arises: among all these admissible configurations, how should we assign probabilities? In the absence of any additional information, it would be unjustified to privilege one configuration over another beyond what is imposed by the constraints themselves. The statistical description we seek should therefore reflect the information encoded in the constraints, and otherwise remain maximally random.

These requirements lead us to imagine a large collection of hypothetical copies of the same room, all equally likely to be observed. In each copy, the number of students wearing glasses is the same. What differs from one copy to another is the specific assignment of students to seats, which is performed without introducing any systematic dependence on whether a student wears glasses or not. These hypothetical copies represent the ensemble of configurations generated by a model that incorporates the chosen constraints and no further assumptions.

The role of this ensemble is not merely illustrative. It provides a principled statistical representation of our state of knowledge about the system, given the information we have decided to encode. Any observed seating arrangement can be interpreted only in relation to this reference model. In the classroom example, this means that the observed concentration of students wearing glasses in the front sector acquires meaning only when compared to what is typical within the ensemble of admissible seating configurations.

Once such a model has been constructed, it becomes possible to formulate precise questions about the data. One may ask, for instance, whether the observed arrangement is a typical outcome under the assumption that seating choices are unrelated to vision, or whether it deviates sufficiently from the ensemble's expectations to suggest the presence of an additional organizing mechanism, such as a genuine preference for front seats. In this sense, the problem of students choosing where to sit can naturally be framed as a hypothesis testing problem, whose answer depends explicitly on the statistical model adopted as a reference.

Although presented in the familiar setting of a classroom, this line of reasoning is entirely general. The same questions arise whenever we observe a structured pattern and wish to understand whether it can be explained solely on the basis of partial information, or whether further mechanisms must be invoked. Throughout this thesis, we will follow this strategy systematically: starting from a set of empirically motivated constraints, we construct statistical models that are otherwise maximally random. These models, known as maximum entropy models (MEMs), provide a coherent framework for representing partial knowledge about complex systems and form the basis upon which subsequent analyses, including hypothesis testing, are built.

The logic underlying this construction has a long and well-established history in statistical physics. In his seminal work [1], Edwin T. Jaynes proposed the principle of maximum entropy as a general method for assigning probability distributions from incomplete information. The starting point is the interpretation of entropy as a quantitative measure of uncertainty or randomness in a probability distribution. Among all distributions compatible with a given set of constraints, the distribution that maximizes entropy is therefore the one that remains as random as possible, introducing no additional structure beyond what is strictly implied by the constraints themselves. In this sense, entropy maximization provides a principled way to select, from the many distributions consistent with our partial knowledge, the least biased one.

In equilibrium statistical mechanics, this principle provides a unifying foundation for the familiar ensembles used to describe physical systems. For example, when the total energy of a system of particles is fixed, all microscopic configurations compatible with that energy are considered equally likely. This setting defines the *microcanonical*

ensemble. If instead the temperature is fixed and the energy is allowed to fluctuate around a fixed average value, one obtains the *canonical ensemble*, characterized by the Boltzmann distribution. In both cases, macroscopic constraints encode partial information about the system, while microscopic details remain maximally unconstrained. The corresponding probability distributions are precisely those that maximize entropy subject to the imposed conditions.

Jaynes' key insight was that this reasoning is not specific to physical systems, but applies broadly whenever one seeks a principled statistical description based on limited information. Maximum entropy models can be understood precisely in this spirit: they generalize the ensembles of statistical physics to systems – such as networks and other complex structures – where the constraints take forms very different from energy, yet play an analogous conceptual role in defining what is known and what is left to chance. The link with statistical physics remains evident in both reasoning and mathematical formulation, as well as in notation and terminology: we still speak of *microcanonical* MEMs when constraints are imposed exactly, and of *canonical* MEMs when constraints are enforced on average, even when the system under study has nothing to do with a gas of particles.

Maximum entropy models form the guiding thread of this thesis. This broad class of models has found applications across a wide range of fields, including biology, finance, and the social sciences, and has proven especially fruitful in network science, where they are widely used for reconstruction tasks and for the statistical validation of structural patterns, in both their canonical and microcanonical formulations. In the works composing this thesis, we explore different aspects of inference with MEMs. *Model selection* addresses how to choose, among competing models, the one that offers the "best"[1] description of the data; hypothesis testing evaluates whether the observed data are consistent with the assumptions encoded in a given model; and statistical validation applies hypothesis testing to identify patterns or interactions that cannot be explained by the model and therefore signal the presence of additional structure.

Throughout this work, a recurring theme will be that of the differences and interplays between the microcanonical and canonical formulations of MEMs sharing the same set of constraints. While this difference is very well-known in statistical physics, where it is linked to different experimental settings, it is far less known in other domains, where, nonetheless, MEMs are implemented in both formulations. A further well-known fact in statistical physics is that the two formulations are, in typical settings, asymptotically equivalent, i.e., if the volume of the system of interest is big enough, one will draw the same conclusions by implementing constraints exactly or on average. This phenomenon is called *ensemble equivalence*. When this is not the case, we talk about *ensemble non-equivalence*. Up until a few years ago, non-equivalence of canonical and microcanonical ensembles had been observed in physics only in the presence of long-range interactions or other forms of non-additivity. Recently, non-equivalence has been shown to appear when an extensive number of constraints is enforced, i.e., a number of constraints that scales with the system size. This is of-

[1]Note that the notion of a "best" model is not unique: different practical criteria can lead to different model choices when applied to the same data. This is exemplified by the AIC and BIC criteria discussed in Section 1.2

ten the case for complex systems, such as networks or time series. Returning to the classroom example, notice that the model was specified by a fixed number of constraints, namely one: the total number of students wearing glasses. The number of constraints does not increase if we imagine a larger classroom. In such a setting, a canonical formulation of the model – in which the number of glasses-wearers is allowed to fluctuate slightly across configurations – becomes equivalent to the microcanonical formulation, where this number is fixed exactly, once we let the classroom (still assumed to be completely full) and the number of seats grow. This equivalence would disappear, however, if we changed the set of constraints. For instance, if we decided to fix not only the total number of students wearing glasses but also the number of such students in each row, then the number of constraints would increase as the classroom becomes larger. With a number of constraints that scales with system size, the canonical and microcanonical formulations would no longer coincide in the large-system limit. This is often the case when *local* constraints are enforced. Unlike *global* constraints, which refer to aggregate quantities and remain fixed in number as the system grows, local constraints apply separately to many individual components – for example, to each row in the classroom or to each node in a network. As the system becomes larger, the number of such constraints grows accordingly. When this happens, the canonical and microcanonical formulations no longer need to agree in the large-system limit, and ensemble equivalence may break down. The theme of ensemble non-equivalence plays a central role in the first chapter of this thesis, where it is examined in detail through the lens of model selection. However, the distinction between canonical and microcanonical models and their interplay remains present throughout the entire work: in the design of hypothesis tests, in the comparison of canonical and microcanonical evidence, and in the construction of null models for validating structural patterns.

As our simple classroom example already suggests, maximum entropy models are not only useful as modeling devices but also play a central role in statistical inference. Because they provide principled reference distributions derived directly from the information we choose to encode, they naturally enter problems of statistical inference. These inferential perspectives motivate the three directions explored in this thesis: how to select among competing MEMs using information-theoretic criteria such as the Minimum Description Length principle; how to formulate and assess hypotheses through modern evidence measures such as e-values; and how to extend the framework to non-linear models capable of validating higher-order interactions. Although the tools differ across chapters, they are united by a common theme: understanding how the assumptions encoded in MEMs shape the conclusions we draw from data.

The next section provides a summary of the common background necessary to understand the models, methods, and results developed in the remainder of the thesis, along with an outline of each chapter. We start by providing a formal definition of maximum entropy models.

1.1 Maximum entropy models

We focus on discrete random variables X taking values in a countable set $\mathcal{X}$. Let $\mathbf{c}(X)$ denote a vector of observables defined on $\mathcal{X}$. Given a probability distribution P over $\mathcal{X}$, the *Shannon entropy* is defined as

$$S[P] = -\sum_{X \in \mathcal{X}} P(X) \log P(X). \tag{1.1}$$

The maximum entropy distribution associated with a given vector of observables $\mathbf{c}(X) = (\mathrm{c}_1(X), \ldots, \mathrm{c}_K(X))$ – referred to as the *sufficient statistics* of the model – is the probability distribution P over $\mathcal{X}$ that maximizes the Shannon entropy subject to constraints on the realization of the sufficient statistics. These constraints may be imposed either exactly (microcanonical formulation) or in expectation with respect to the model (canonical formulation). Originally introduced by Gibbs [2] and later reformulated by Jaynes [1], this approach yields probability distributions that are, intuitively speaking, maximally *random*[2] subject to the prescribed constraints, whose values are controlled by the model parameters.

More specifically, maximizing the entropy under the hard constraint $\mathbf{c}(X) = \mathbf{c}$, requiring all realizations to satisfy a fixed value $\mathbf{c}$ of the sufficient statistics, yields the *microcanonical distribution*

$$P_{\mathrm{mic}}(X; \mathbf{c}) = \begin{cases} \dfrac{1}{\Omega(\mathbf{c})} & \text{if } \mathbf{c}(X) = \mathbf{c}, \\ 0 & \text{otherwise}, \end{cases} \tag{1.2}$$

which is uniform over all configurations X satisfying the hard constraint $\mathbf{c}(X) = \mathbf{c}$. The normalization constant $\Omega(\mathbf{c})$ denotes the number of such configurations, while the values $\mathbf{c}$ themselves play the role of model parameters. The corresponding *microcanonical model* is defined as the family of distributions sharing the functional form of Eq. (1.2) and parameters ranging over the set $\mathcal{C}$ of admissible values of the sufficient statistics[3]:

$$\mathcal{M}_{\mathrm{mic}} = \{P_{\mathrm{mic}}(X; \mathbf{c}) \,:\, \mathbf{c} \in \mathcal{C}\}. \tag{1.3}$$

Back to the classroom example of the previous section, the observed classroom corresponds to a specific realization of X, while $\mathcal{X}$ represents the set of all possible fully occupied classrooms of the same size. Choosing the number of students wearing glasses as the sufficient statistic determines the associated microcanonical model. A specific probability distribution is then selected by fixing the parameter c so as to match the observed number of students wearing glasses. From a statistical inference perspective, this choice admits a direct interpretation: selecting the parameter that reproduces the observed value of the sufficient statistic is equivalent to maximizing the likelihood of

[2]The notion of "randomness" depends on the chosen entropy functional. Modifying either the entropy functional or the imposed constraints leads to a different reference notion of randomness. While Shannon entropy is a standard and well-founded choice, alternative entropy functionals can also be considered [3], resulting in different classes of maximum entropy models.

[3]We restrict attention to the *graphical values* of $\mathbf{c}$, namely the set $\mathcal{C} = \{\mathbf{c} : \Omega(\mathbf{c}) \neq 0\}$ of values realized by at least one configuration.

the model given the data. To this end, notice that the microcanonical likelihood can take only two values, either 0 or $\frac{1}{\Omega(\mathbf{c})}$. As a result, likelihood maximization is a rather degenerate problem in the microcanonical setting. Nevertheless, the likelihood-based perspective remains useful, as it also applies to the canonical case and provides a unified way to frame the problem.

The *canonical distribution* is obtained by maximizing the entropy under the soft constraint $\mathbb{E}_P[\mathbf{c}(X)] = \mathbf{c}$, i.e., by requiring the probability distribution to reproduce a prescribed value of the observables only in expectation. The resulting distribution takes the exponential form

$$P_{\text{can}}(X;\boldsymbol{\theta}) = \frac{e^{-\boldsymbol{\theta}\cdot\mathbf{c}(X)}}{Z(\boldsymbol{\theta})}, \tag{1.4}$$

where the normalization factor

$$Z(\boldsymbol{\theta}) = \sum_{X\in\mathcal{X}} e^{-\boldsymbol{\theta}\cdot\mathbf{c}(X)}$$

is known as the *partition function*. In network science, the canonical maximum entropy formulation gives rise to what are commonly referred to as Exponential Random Graph Models (ERGMs) [4, 5, 6]. In this setting, the parameter vector $\boldsymbol{\theta}$ emerges naturally from the entropy maximization procedure as the set of Lagrange multipliers associated with the imposed constraints. The corresponding *canonical model* is defined as the family of distributions sharing the functional form of Eq. (1.4) and parameters ranging over the parameter set $\boldsymbol{\Theta} = \mathbb{R}^K$, where K is the number of components of $\mathbf{c}$, or in other words, the number of constrained observables:

$$\mathcal{M}_{\text{can}} = \{P_{\text{can}}(X;\boldsymbol{\theta}) \,:\, \boldsymbol{\theta}\in\boldsymbol{\Theta}\}. \tag{1.5}$$

Returning once more to the classroom example, a canonical model would enforce the average number of students wearing glasses, allowing fluctuations across the ensemble of classrooms. As in the microcanonical case, the distribution used to represent an observed system is typically selected by choosing the parameters $\boldsymbol{\theta}$ so that the expected value of the sufficient statistics matches the empirical one, which is equivalent to solving the constraints equation $\mathbb{E}_P[\mathbf{c}(X)] = \mathbf{c}$. This choice is again equivalent to maximizing the likelihood of the model with respect to the observed data.

Notice that the microcanonical distribution can always be obtained from the corresponding canonical distribution by conditioning on the observed value of the sufficient statistics. From a statistical perspective, microcanonical models can therefore be interpreted as conditional models associated with the canonical formulation.

Relaxing hard constraints and enforcing them only in expectation, as in the canonical formulation, is often convenient in practice. Besides allowing for fluctuations that may be genuinely present in empirical systems, soft constraints can lead, in many relevant settings, to models with simpler analytical structure or more efficient numerical implementations. In such cases, canonical models may admit closed-form expressions and smoother parameter spaces, making them a natural choice for inference.

At the same time, these advantages are not general across all settings. The canonical formulation can be less transparent from an intuitive standpoint, as it describes ensembles in which the constrained quantities fluctuate rather than remain fixed. Moreover, in discrete settings, there are important cases in which the microcanonical formulation is technically simpler to implement, as in the case of Bayesian inference, where counting configurations is often easier than computing integrals.

For this reason, the choice between microcanonical and canonical formulations is inherently context-dependent. This distinction becomes particularly relevant when the two formulations are not asymptotically equivalent. Several formal definitions of ensemble equivalence have been introduced in the literature and shown to be equivalent under mild assumptions [7]. Throughout this work, we adopt the so-called *measure-level* definition of equivalence. In the case of maximum entropy models, under this definition, (non-)equivalence can be established in terms of the Kullback-Leibler divergence [8] between the two distributions that maximize the likelihood under the respective formulations. When ensemble equivalence does not hold, selecting between hard and soft constraints extends beyond modeling convenience and acquires an inferential meaning. One natural way to formalize this comparison is to frame it as a model selection problem, which is the perspective adopted in the next section.

1.2 Model selection

In general terms, model selection aims to choose, given observed data, one model from a set of competing candidates [9], where a model is defined as a parametric family of probability distributions sharing the same functional form. Many popular model selection approaches are grounded in information theory; in most cases, these criteria assign a score to each model and output a model ranking. Among the best-known approaches of this kind are the Akaike Information Criterion (AIC) [10] and the Bayesian Information Criterion [11]. According to these criteria, the best model is the one that minimizes

$$\mathrm{AIC} = -\ln \mathcal{L} + K \tag{1.6}$$

and

$$\mathrm{BIC} = -\ln \mathcal{L} + \frac{K}{2} \ln V. \tag{1.7}$$

Note that both criteria are often defined with an additional factor of 2, which does not change the relative ranking of models. Both criteria incorporate a trade-off between the maximized model likelihood (denoted by $\mathcal{L}$), which measures the model goodness-of-fit relative to the observed data, and the model complexity, which measures the tendency of a model to overfit. The complexity term, for both AIC and BIC, depends on the number K of free parameters in the model, which, in the case of MEMs, is equal to the number of constraints (with BIC additionally depending on the system volume V).

Chapter 2 of this thesis is devoted to the comparison of microcanonical and canonical formulations defined by the same set of constraints. While these two formulations

encode the same type of information, they may lead to different statistical descriptions, especially in regimes where ensemble equivalence does not hold [12, 13, 14, 15]. Model selection provides a natural framework for formalizing this comparison within the context of statistical inference. According to AIC and BIC, the two formulations share the same complexity; nevertheless, there are several reasons why neither AIC nor BIC are appropriate for comparing canonical and microcanonical MEMs (see the Introduction of Chapter 2 for more details). To overcome this problem, one can turn to a more refined information-theoretic approach based on the Minimum Description Length (MDL) principle [16, 17, 18]. The MDL principle frames statistical inference as the task of identifying structure or regularities in data. This idea can be interpreted in terms of compressibility: the more regular the data, the more concisely it can be described. The Minimum Description Length principle brings these ideas together by framing learning as a problem of data compression and by stating that the best model is the one that outputs the shortest description of the data.

There are several ways of computing this description. Nowadays, the standard approach is the *universal coding* approach, which encompasses several methods, among them the Normalized Maximum Likelihood approach and the Bayesian approach. In this work, we focus on the first one but provide a thorough analysis of the interplay between the two methods when comparing microcanonical and canonical MEMs.

Outline of Chapter 2

Chapter 2 examines, from a model-selection perspective, the inferential consequences of choosing hard (microcanonical) versus soft (canonical) constraints in maximum entropy modeling. Using the Minimum Description Length principle and the Normalized Maximum Likelihood (NML) formulation, it compares canonical and microcanonical models defined by the same sufficient statistics, providing a principled framework for comparing discrete maximum entropy models, including regimes in which ensemble equivalence does not hold.
Closed-form expressions for the NML description lengths are derived, revealing a trade-off between likelihood and complexity: microcanonical models achieve higher likelihood but are systematically more complex, so the preferred formulation depends on the observed constraints. These results are illustrated on binary rectangular matrices, showing finite description-length differences for global constraints and extensive differences (i.e., that scale at least linearly with the system size) for local constraints. The chapter also connects the NML analysis to Bayesian universal coding, highlighting how prior choice can have non-negligible inferential consequences in non-equivalent settings.

So far, we have focused on comparing microcanonical and canonical formulations defined by the same set of constraints. A different inferential question arises when alternative sets of constraints are considered. We tackle this problem using hypothesis testing with e-values, which we show in Chapter 3 to be deeply connected to MDL universal coding.

1.3 Hypothesis testing

When the models to be compared are defined by *different* sets of constraints, the comparison can be framed as a problem of hypothesis testing. In the classical Neyman–Pearson framework, one specifies a *null hypothesis* and an *alternative hypothesis*. A test statistic, defined as a function of the data, is then chosen, and its distribution under the null hypothesis is used to compute the probability of observing a value at least as extreme as the one obtained. This probability is known as the *p-value*. The p-value is compared to a predefined significance level α, and the null hypothesis is rejected if the p-value falls below this threshold.

P-values are widely used across scientific disciplines. However, increasing attention has been drawn to limitations in their interpretation and use [19, 20, 21], particularly in settings involving multiple testing, sequential data collection and analyses where methodological decisions are informed by the observed data[4]. These concerns have motivated the development of alternative measures, designed to provide more robust and transparent inferential procedures. Recently, *e-values* [22, 23, 24, 25, 26, 27] have emerged as a compelling alternative to p-values, designed to address long-standing issues in traditional hypothesis testing. Unlike p-values, e-values serve as direct measures of evidence that support the combination of experimental results, allowing for optional continuation of data sampling. Despite their relatively recent formalization, e-values have quickly gained attention, appearing frequently in high-impact statistical literature as a promising framework for valid and transparent inference.

While e-values are rapidly attracting interest as they offer desirable properties compared with p-values, this comes at the price of increased technical difficulties in finding simple, scalable, and ready-to-use implementations for practical purposes – even in the light of their comparatively more recent history in the statistical literature. Chapter 3 focuses on constructing *optimal* e-values [24] for hypothesis tests between canonical or microcanonical MEMs with different sufficient statistics. In light of ensemble non-equivalence and its non-trivial consequences on model selection, such tools should, in principle, be tailored to both formulations, although the resulting e-values remain closely related, as shown in Chapter 3. While e-values are rapidly attracting interest as they offer desirable properties compared with p-values, this comes at the price of increased technical difficulties in constructing simple, scalable, and ready-to-use procedures, especially given their comparatively recent history in the statistical literature. Chapter 3 focuses on the construction of *optimal* e-values [24] for hypothesis tests between canonical or microcanonical maximum entropy models with different sufficient statistics. In light of ensemble non-equivalence and its non-trivial consequences for inference, such tools should, in principle, be tailored separately to the two formulations, although the resulting e-values turn out to be closely related, as shown in Chapter 3.

Moreover, the construction of optimal e-values is closely connected to that of model description lengths. In particular, the *universal distributions* underlying MDL-based description lengths – including Bayesian and Normalized Maximum Likelihood

[4]These concerns have entered both the methodological and the broader scientific discourse under labels such as *p-hacking*, and have even become the subject of popular satire and online memes.

distributions – can also conveniently be used to construct GRO-optimal e-variables, establishing a direct link between hypothesis testing with e-values and model selection through universal coding.

Outline of Chapter 3

Chapter 3 develops growth-rate optimal (GRO) e-variables for hypothesis tests between maximum entropy models (MEMs) that differ in their sufficient statistics, treating separately the cases in which both null and alternative hypotheses are formulated as microcanonical (hard-constraint) models or as canonical (soft-constraint) models. The chapter also connects the construction of optimal GRO e-variables to the MDL principle, showing that *universal distributions* can be used to build model description lengths as well as GRO-optimal e-variables.
In the microcanonical setting, the chapter derives an explicit closed-form expression for the *growth rate optimal* (GRO) e-variable. For canonical models, the optimal GRO construction is characterized by an optimization problem over priors on the parameter space, which is typically intractable. To address this difficulty, the chapter introduces a microcanonical approximation: a canonical universal distribution under the alternative is rewritten as a microcanonical mixture by inducing a prior on the sufficient statistic, yielding a (microcanonical) computationally tractable candidate for the canonical test. A key result is that any microcanonical e-variable is automatically a valid canonical e-variable when the two models share the same sufficient statistic.
The quality of this approximation is analyzed both theoretically and empirically, showing that the gap between the canonical GRO e-variable and its microcanonical approximation decreases rapidly with sample size in relevant regimes. These results are illustrated through applications to contingency tables, including explicit calculations for 2×2 tables and extensions to general $2 \times k$ tables under different sampling regimes. Across these settings, the microcanonical GRO construction is shown to provide an accurate and practically effective approximation to the canonical GRO e-variable, even when the number of groups grows with the sample size.

A simple and yet powerful application of e-values for MEMs is found in contingency tables. Contingency tables are a fundamental tool in statistical analysis for examining the relationship between categorical variables. Given a dataset where observations are classified according to categorical factors, a contingency table provides a structured way to summarize the frequencies of different category combinations. In this work, we focus on binary categorical data with k categories or groups, which are represented as $2 \times k$ contingency tables. In fact, our simple classroom example can be phrased as a 2×2 contingency table. We have two categories: wearing glasses and not wearing glasses, and two groups: people sitting in the front sector and people sitting in the back one. What we observe can be summarized as the following contingency tables:

Glasses	Front	Back	Total
Yes	$n_{\mathrm{g}}^{\mathrm{f}}$	n_{g}^{b}	n_{g}
No	$n_{\mathrm{ng}}^{\mathrm{f}}$	n_{ng}^{b}	n_{ng}
Total	n^{f}	n^{b}	n

The question whether the probability of sitting in the front or the back depends on wearing glasses is the standard test for contingency tables of this kind, and can be interpreted as a test between canonical maximum entropy models, differing in the structure of their constraints (as explained in Chapter 3). The same holds for the more general case of $2 \times k$ contingency tables.

Despite the simplicity and ubiquity of this setting, a systematic comparison between classical p-value-based methods and e-values tailored to contingency tables is still lacking. This gap is particularly relevant in applied contexts in which many tables must be analyzed under heterogeneous conditions, such as studies relating genotypic and phenotypic information, where statistical evidence needs to be assessed in a reliable yet flexible manner.

A natural bridge between classical p-value and e-values is provided by p-to-e calibrators [26], which allow one to construct valid e-values starting from p-values while retaining the main advantages of the e-value framework, such as optional continuation and evidence accumulation. Since a wide variety of p-values are available for testing association in contingency tables, calibrated e-values provide an immediate and broadly applicable solution in this setting.

However, calibration alone does not guarantee optimality: the resulting e-values are not, in general, tailored to maximize evidence against specific alternatives. This raises the question of whether e-values can be constructed directly in a way that improves upon calibrated approaches.

Rather than advocating a single universal approach, Chapter 4 addresses this problem by developing and comparing different e-value constructions for contingency tables in a concrete and practically motivated setting. The chapter examines how different e-values behave in practice and how their properties depend on the way data are collected. Particular attention is devoted to defining new *conditional e-values* [28, 29], which exploit the exponential-family structure of the null model and simplify hypothesis testing by conditioning on its sufficient statistics. Although formally different, this construction is closely related to the microcanonical approach developed in Chapter 3, as both rely on conditioning on sufficient statistics.

Importantly, the conclusions also depend on how the data are collected. In some applications, evidence is assessed based on a single contingency table, observed at once. In other applications, contingency tables arise as data are collected over time, for example, across different studies, experimental batches, or successive observation windows, and inferential conclusions can be updated as new information becomes available. Although these situations correspond to the same underlying question,

they impose different requirements on statistical evidence. This distinction is particularly relevant for e-values, whose defining properties are inherently linked to the way evidence is accumulated. Indeed, some e-value constructions for contingency tables have been explicitly designed to reach asymptotic optimality in a sequential setting [30].

Understanding how different e-value constructions behave under these different modes of data acquisition is therefore essential for assessing their practical usefulness in applied settings, and it is the main focus of Chapter 4.

Outline of Chapter 4

This chapter presents a comparative study of different e-values for testing association in 2×2 contingency tables, as they arise in applications such as genetic association analysis. Several e-value constructions are compared, including existing proposals and newly introduced conditional e-values that reduce composite testing problems by conditioning on the sufficient statistics of the null model. The analysis considers both a one-shot setting, where a single table is observed, and a sequential setting, where tables arrive in batches over time. In the one-shot case, conditional e-values – in particular, the *uniformly most powerful* constructions – tend to perform best, whereas in the sequential setting the comparison becomes more nuanced: Turner's e-variable, designed for sequential settings, is asymptotically optimal, while the normalized maximum likelihood conditional e-variable is often preferable at practically relevant sample sizes.

Contingency tables are a simple yet powerful example. Nevertheless, there are more complex settings where MEMs can be defined and applied to obtain statistically sound results. A particularly representative example is that of pattern validation.

1.4 Pattern validation

Maximum entropy models are often employed not as competing hypotheses, but as stand-alone *null models* used to assess the statistical relevance of observed structures. This leads to the problem of *pattern validation*, where the inferential goal is to determine whether a structured feature observed in data can be accounted for solely on the basis of the information encoded in a reference model, or whether it signals the presence of additional organization not captured by the imposed constraints. In this setting, inference does not aim to select one model over another, but rather to identify which aspects of the data are incompatible with a given notion of randomness.

From a statistical perspective, pattern validation can be viewed as a collection of hypothesis tests carried out against a common null ensemble. One first specifies a maximum entropy model by fixing a set of constraints that encode the information deemed relevant. This model defines a reference distribution over admissible configurations. Observed patterns are then evaluated by comparing their empirical realization to what is typical under this distribution. Patterns that deviate sufficiently from the null ensemble are considered statistically validated, in the sense that they cannot be

plausibly explained by the constraints alone. Maximum entropy models thus serve as principled statistical baselines, against which deviations acquire inferential significance.

A paradigmatic application of this framework arises in the analysis of bipartite networks. Bipartite representations are a natural way of encoding relational data involving two distinct classes (or *layers*) of entities, such as countries and products, authors and publications, or users and items in recommender systems. In such datasets, relations are observed only between nodes belonging to different layers, reflecting the structure of the data-generating process. In many applications, however, the scientific question of interest concerns relationships *within* one of the two layers. For instance, in a user–item dataset, one may wish to identify connections between users who exhibit similar patterns in the items they select, even though no direct user–user interactions are observed. A common approach to address this problem is to project the bipartite network onto a monopartite one, where nodes represent entities from a single layer, and links encode some notion of co-occurrence or similarity inferred from the shared connections in the original bipartite data.

Entropy-based null models have been successfully introduced to statistically validate projected links [31, 6]. In this framework, one defines a bipartite maximum entropy model, either microcanonical or canonical, that reproduces selected local properties of the observed bipartite network, such as degree sequences on both layers. The number of co-occurrences between two nodes in the projected layer is then compared to its distribution under the null ensemble. Only co-occurrences that are unlikely to arise under the reference model are retained, yielding a statistically validated projection.

This approach highlights a key aspect of pattern validation: the role of the null model is not to reproduce the observed structure in detail, but to encode a minimal and interpretable set of assumptions about the system. Statistical significance is therefore inherently model-dependent. What counts as a validated pattern depends explicitly on which information is treated as relevant and which is left to chance.

While pairwise validation in projected networks has proven effective in many empirical settings, it becomes insufficient when the phenomena of interest involve interactions that are intrinsically higher-order. In several complex systems, the collective behavior of groups comprising three or more components cannot be reduced to the sum of pairwise effects alone. This is particularly evident in neuroscience, where increasing empirical evidence suggests that coordinated activity among multiple brain regions plays a fundamental role in information processing and in the emergence of higher-level functions, such as cognition and consciousness [32, 33, 34, 35]. In such cases, restricting attention to pairwise interactions may obscure essential aspects of the underlying organization.

Extending the pattern-validation framework to higher-order structures requires null models capable of reproducing relevant lower-order properties. From a statistical standpoint, this corresponds to fixing a suitable set of lower-order constraints and testing whether higher-order patterns observed in the data are compatible with the resulting ensemble. Maximum entropy models provide a natural language for formalizing this procedure, as they allow one to control which features are enforced into the model and which are left available for validation.

As shown in Chapter 5, the statistical validation of higher-order structures may require null models that enforce constraints that are non-linear functions of the system variables. Incorporating such constraints into maximum entropy models poses substantial methodological challenges, both computational and analytical, even in the canonical framework, which is typically more tractable: non-linear constraints are known to induce phase transitions and degeneracy in the corresponding ensembles. Chapter 5 of this thesis addresses these challenges by developing and analyzing a second-order maximum entropy model that retains non-linear constraints while remaining computationally manageable. Combined with suitable statistical testing procedures, this enables the validation of higher-order interactions that cannot be solely explained by the chosen null ensemble. These methods are developed and applied in the context of brain data, with a focus on resting-state fMRI time series. In this context, the approach enables the identification and characterization of statistically validated higher-order topological interactions between brain areas, providing insight into forms of functional coordination that cannot be reduced to pairwise relationships.

Outline of Chapter 5

Chapter 5 investigates the statistical validation of higher-order interaction patterns in resting-state fMRI time series within a maximum-entropy null-model framework. Pairwise interactions are validated using the Bipartite Configuration Model (BiCM), while triplet interactions are assessed using the Bipartite Partial Local Overlap Model (BiPLOM), a canonical maximum entropy model introduced in this chapter to include non-linear constraints, and solved in the mean-field approximation.
Validated triplets are later classified into connected triplets, supported by validated pairwise interactions, and non-connected triplets, which lack pairwise support. Their statistical and spatial properties are analyzed across multiple scales, including the whole brain, the seven resting-state networks of Yeo, and individual brain regions.
The results reveal two distinct regimes of higher-order functional coordination. Connected triplets form spatially compact and redundant structures, reflecting strong lower-order dependencies, whereas non-connected triplets are more spatially distributed and exhibit weak or absent redundancy, consistent with genuinely higher-order coordination. These regimes display distinct anatomical distributions, with unimodal regions predominantly involved in connected triplets and transmodal regions more frequently associated with non-connected ones.

From the simple classroom example to complex empirical systems, the unifying theme of this thesis is that statistical inference is meaningful only relative to a well-defined notion of chance. Maximum entropy models provide a principled way to formalize this notion, and to study how different inferential tasks depend on the assumptions one chooses to encode.