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

Giuffrida, F.

Learning Under Constraints

Inference, Selection and Validation with Maximum Entropy Models

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr. S. de Rijcke,
volgens besluit van het college voor promoties
te verdedigen op donderdag 10 september 2026
klokke 11:30 uur

door

Francesca Giuffrida
geboren te Rome
in 1995

Promotores:

Prof. dr. D. Garlaschelli
Prof. dr. P. D. Grünwald

Promotiecommissie:

Prof. dr. ir. S.J. van der Molen
Prof. dr. K.E. Schalm
Prof. dr. W.T.F. den Hollander
Dr. M. Sales-Pardo (Rovira i Virgili University)
Prof. dr. T. P. Peixoto (Interdisciplinary Transformation University Austria)

Contents

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.

CHAPTER 2

Description length of maximum entropy models

This chapter is based on:
F. Giuffrida, T. Squartini, P. Grünwald, and D. Garlaschelli. Description length of canonical and microcanonical Models. *Phys. Rev. Research 7, 043057* (2025).

Abstract

The (non)equivalence of canonical and microcanonical ensembles is a fundamental question in statistical physics, concerning whether the use of soft and hard constraints in the maximum-entropy construction leads to the same description of a system. Despite the fact that maximum-entropy models are also commonly used in statistical inference, pattern detection, and hypothesis testing, a complete understanding of the effects of ensemble nonequivalence on statistical modeling is still missing. Here, we study this problem from a rigorous model selection perspective by comparing canonical and microcanonical models via the minimum description length principle, which yields a trade-off between likelihood, measuring model accuracy, and complexity, measuring model flexibility and its potential to overfit data. We compute the normalized maximum likelihood (NML) of both formulations and find that (1) microcanonical models always achieve higher likelihood but are always more complex; (2) the optimal model choice depends on the empirical values of the constraints—the canonical model performs best when its fit to the observed data exceeds its uniform average fit across all realizations; (3) in the thermodynamic limit, the difference in description length per node vanishes when ensemble equivalence holds but persists otherwise, showing that nonequivalence implies extensive differences between large canonical and microcanonical models. Finally, we compare the NML approach to Bayesian methods, showing that (4) the choice of priors, practically irrelevant in equivalent models, becomes crucial when an extensive number of constraints are enforced, possibly leading to very different outcomes.

2.1 Introduction

Entropy maximization provides a principled approach to inference, resulting in maximally random models under specified constraints. These models have found applications across a wide range of scientific disciplines, such as network science [31, 6, 36] and time-series analysis [37, 38].

Two distinct formulations of maximum-entropy models are possible depending on how constraints are enforced. Canonical models are defined by requiring the value of the constraints to be satisfied *on average*. Microcanonical models are defined by requiring the value of the constraints to be satisfied *exactly*. These two formulations do not necessarily lead to the same description of a given system, i.e., they are, in general, *non-equivalent*.

The concept of ensemble equivalence, introduced by Gibbs [2], plays an important role in statistical mechanics. Traditional studies have shown that canonical and microcanonical ensembles tend to become equivalent in the thermodynamic limit for systems with short-range interactions. Violations of this asymptotic equivalence, known as ensemble non-equivalence, have historically been linked to systems with long-range interactions or other forms of non-additivity, such as mean-field spin models or self-gravitating systems [39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 7].

More recently, it has become clear that ensemble non-equivalence can also arise due to an additional mechanism: the presence of an *extensive number of constraints* [12, 13, 14, 15]. This has been demonstrated, for example, in network models with local constraints, such as the Configuration Model. In these cases – unlike the classical examples from statistical physics – non-equivalence is not necessarily associated with long-range interactions or phase transitions, and is not confined to specific parameter regimes, but is present across the entire parameter space. Moreover, this phenomenon can have practical implications for the expected macroscopic properties of the system under study (see [50] for an example concerning bipartite networks).

Whenever the two approaches are equivalent, they identify the same set of typical configurations. In these cases, preferring one formulation over the other becomes a matter of convenience (e.g., computational efficiency). Whenever the two approaches are non-equivalent, the choice depends on the specific characteristics of the system under study and the nature of the constraints involved. For example, in the context of choosing a null model to be compared with real data, for e.g., pattern detection, it has been argued that, when non-equivalence holds, the ensemble choice should be based on a theoretical guiding principle [31, 14, 51, 15, 52]. In this case, if the data are expected to be noisy, meaning that the measured values of the constraints generally do not match the true (unobservable) ones, one should prefer the canonical version of the null model because it allows for fluctuations in the constraint quantities, while the microcanonical one would paradoxically assign zero probability to the true configuration of the system. Conversely, if, by hypothesis, the measured values of the constraints are not affected by error or noise, the microcanonical model should be preferred.

When there is no prior expectation available about the presence or absence of noise in the empirical values of the constraints, the decision between hard and soft

constraints should be based on posterior evidence, i.e., which of the two null models best describes the data. This immediately leads to a problem of model selection between canonical and microcanonical ensembles.

Information theory provides a powerful framework for model selection by offering criteria that capture the trade-off between a model's "goodness of fit," quantified by its log-likelihood, and its "complexity." In general, model complexity measures the flexibility of a model in fitting different datasets. In other words, it measures how easily the model can overfit the data. A common way to quantify model complexity is by counting the number k of the model's free parameters, which, in the case of maximum-entropy models, corresponds to the number of constraints. This is the case of the two historically most popular model selection criteria: the *Akaike Information Criterion* (AIC) [10] and the *Bayesian Information Criterion* (BIC) [11], according to which one should pick the model minimizing

$$\mathrm{AIC} = -\ln \mathcal{L} + k \tag{2.1}$$

and

$$\mathrm{BIC} = -\ln \mathcal{L} + \frac{k}{2} \ln V, \tag{2.2}$$

respectively, where $\mathcal{L}$ is the model likelihood and V is the sample size. Canonical and microcanonical models defined by the same set of constraints share the same number of parameters. Therefore, according to both AIC and BIC, their complexity is the same, and a comparison based on these criteria would reduce to a mere comparison between their log-likelihood functions. This would naively lead to the conclusion that the microcanonical model is always the best-scoring one since it has the highest likelihood. However, neither AIC nor BIC can be computed for microcanonical models since they are derived under regularity assumptions that are not even remotely met by the latter. Moreover, both criteria are asymptotic results derived under the assumption of a finite number of parameters as the system size approaches infinity [9]. This assumption is not met by models whose number of parameters scales with the size of the system.

To overcome these issues, here we consider the Minimum Description Length (MDL) principle [16, 17, 18], a family of information-theoretic approaches to model selection seeking the model that provides the shortest description of data. Specifically, we focus on the approach based on the Normalized Maximum Likelihood (NML), or Shtarkov, distribution [53], with the aim of investigating the trade-off between the difference of log-likelihoods and that of complexities in light of ensemble non-equivalence.

A particularly useful definition of non-equivalence, naturally connected to information theory, is the *measure-level* one, which involves the Kullback-Leibler (KL) divergence between the microcanonical and canonical probability distributions. Under mild conditions, this approach is equivalent to other definitions commonly used in statistical physics [7]. According to this definition, (non-)equivalence is characterized by a (non-)vanishing relative entropy density as the system size approaches infinity. Interestingly, the KL divergence between canonical and microcanonical distributions coin-

cides with the difference between the corresponding log-likelihood functions [12, 13]. Thus, measure-level (non-)equivalence can be inspected by simply considering the asymptotic behavior of this log-likelihood difference [12], providing a natural connection to statistical inference.

While the impact of non-equivalence on the difference between log-likelihoods is relatively well understood [12, 14], its effects on the difference between complexities remain largely unexplored. To encompass a broad spectrum of applications, our study focuses on maximum-entropy models suitable for studying systems represented by binary rectangular matrices (e.g., bipartite networks, time series). More precisely, we aim at determining i) whether the influence of non-equivalence persists even when the complexity term is accounted for, and ii) whether non-equivalent canonical and microcanonical models should, in fact, be regarded as statistically distinct models. We conclude that the answer to both questions is yes, raising the question of whether non-equivalence can be defined at the model selection level. Furthermore, we show that in the presence of extensively many constraints, the choice of prior in a Bayesian framework can critically affect the resulting description length, underscoring the importance of this choice in non-equivalent settings.

The rest of the paper is organized as follows. Section 2.2 introduces MDL for discrete data and the Normalized Maximum Likelihood (NML) approach to define description lengths. In Section 2.3, we provide a general expression for the difference between the description lengths of canonical and microcanonical models as a function of the Kullback-Leibler divergence between the two distributions. In Section 2.4, we explicitly derive and compare the description lengths of both the canonical and the microcanonical formulations of models defined by one global constraint (i.e., the sum of the entries of a matrix) as well as one-sided local constraints (i.e., the row-specific sums of a matrix). While the first constraint leads to ensemble equivalence, the second set of constraints leads to ensemble non-equivalence. Finally, Section 2.5 reviews the relationship between the NML-based and Bayesian approaches when non-equivalence holds.

2.2 The Minimum Description Length principle

The MDL principle embodies the idea that the best model to describe a given dataset is the one providing the shortest description. We now introduce the basic formalism by focusing on the discrete data $\boldsymbol{x} \in \mathcal{X}$, where $\mathcal{X}$ represents the space of all possible samples of fixed size V. A parametric statistical model $\mathcal{M}$ consists of a family of probability distributions sharing the same functional form, i.e.,

$$\mathcal{M} = \{P_{\mathcal{M}}(\boldsymbol{x}; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta} \tag{2.3}$$

where $\boldsymbol{\theta}$ is the vector of parameters whose values lie in the set Θ.

In order to apply the MDL principle, one needs to compute the description length of the data $\boldsymbol{x}$ provided by a model $\mathcal{M}$. By Kraft's inequality [16, 54], the optimal number of natural digits (nats) needed to describe the outcome $\boldsymbol{x}$ of a probability distribution $P(\boldsymbol{x})$ is given by $-\ln P(\boldsymbol{x})$. Since, however, $\mathcal{M}$ encompasses many distributions corresponding to different parameter values, a natural choice is that of

considering the description length provided by the distribution minimizing the quantity $-\ln P_{\mathcal{M}}(\boldsymbol{x};\boldsymbol{\theta})$, reading

$$\mathcal{L}_{\mathcal{M}}(\boldsymbol{x}) \equiv P_{\mathcal{M}}(\boldsymbol{x};\hat{\boldsymbol{\theta}}(\boldsymbol{x})), \tag{2.4}$$

with $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ being the maximum-likelihood (ML) estimator of $\boldsymbol{\theta}$, i.e.

$$\hat{\boldsymbol{\theta}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}\in\Theta} P_{\mathcal{M}}(\boldsymbol{x};\boldsymbol{\theta}). \tag{2.5}$$

However, the distribution $\mathcal{L}_{\mathcal{M}}(\boldsymbol{x})$ is affected by (at least) two problems having the same origin: the parameters appearing in its definition are evaluated for each individual data in $\mathcal{X}$. As a consequence, i) $\mathcal{L}_{\mathcal{M}}(\boldsymbol{x})$ cannot be a probability distribution with support in $\mathcal{X}$ as it is not properly normalized, and ii) adopting $\mathcal{L}_{\mathcal{M}}(\boldsymbol{x})$ may lead to severe overfitting since it is defined *a posteriori* relative to the observations.

To solve these problems, one can build a so-called *universal distribution* relative to $\mathcal{M}$, i.e., a unique, representative distribution $\bar{P}_{\mathcal{M}}(\boldsymbol{x})$ of the statistical model $\mathcal{M}$. Once equipped with a universal distribution, the description length of $\boldsymbol{x}$ through is $\mathcal{M}$ is defined as

$$\mathrm{DL}^{\bar{P}}_{\mathcal{M}}(\boldsymbol{x}) \equiv -\ln \bar{P}_{\mathcal{M}}(\boldsymbol{x}). \tag{2.6}$$

Universal distributions (which don't necessarily belong to $\mathcal{M}$) can be defined in several ways. In what follows, we will consider the MDL recipe based upon the NML universal distribution (for a more formal introduction to universal coding, we redirect the interested reader to [16]).

2.2.1 The Normalized Maximum Likelihood universal distribution

The NML distribution corresponds to the distribution attaining the minimum pointwise *redundancy* (or *regret*) in the worst-case scenario [16], i.e.,

$$\bar{P}^{\mathrm{NML}}_{\mathcal{M}} \equiv \arg\min_{\bar{P}} \max_{\boldsymbol{x}\in\mathcal{X}} \left\{ -\ln \bar{P}(\boldsymbol{x}) - [-\ln P_{\mathcal{M}}(\boldsymbol{x};\hat{\boldsymbol{\theta}}(\boldsymbol{x}))] \right\}. \tag{2.7}$$

In words, it is the distribution that *a priori* requires the minimum number of nats when compared to the shortest description length *a posteriori*, that is $-\ln P_{\mathcal{M}}(\boldsymbol{x};\hat{\boldsymbol{\theta}}(\boldsymbol{x}))$. The minimum is found in the worst-case scenario, i.e., for the data $\boldsymbol{x}$ for which this difference is maximal. The unique solution to this *minimax* problem reads

$$\bar{P}^{\mathrm{NML}}_{\mathcal{M}}(\boldsymbol{x}) = \frac{P_{\mathcal{M}}(\boldsymbol{x};\hat{\boldsymbol{\theta}}(\boldsymbol{x}))}{\sum_{\boldsymbol{y}\in\mathcal{X}} P_{\mathcal{M}}(\boldsymbol{y};\hat{\boldsymbol{\theta}}(\boldsymbol{y}))}, \tag{2.8}$$

which is a normalized maximum likelihood. Following the universal coding approach, the description length of $\boldsymbol{x}$ through model $\mathcal{M}$ is computed as

$$\begin{aligned} \mathrm{DL}^{\mathrm{NML}}_{\mathcal{M}}(\boldsymbol{x}) &\equiv -\ln \bar{P}^{\mathrm{NML}}_{\mathcal{M}}(\boldsymbol{x}) \\ &= -\ln P_{\mathcal{M}}(\boldsymbol{x};\hat{\boldsymbol{\theta}}(\boldsymbol{x})) + \ln \sum_{\boldsymbol{y}\in\mathcal{X}} P_{\mathcal{M}}(\boldsymbol{y};\hat{\boldsymbol{\theta}}(\boldsymbol{y})). \end{aligned} \tag{2.9}$$

While the first term is (minus) a model maximum log-likelihood, the second term

$$\mathrm{COMP}_{\mathcal{M}} \equiv \ln \sum_{\boldsymbol{y}\in\mathcal{X}} P_{\mathcal{M}}(\boldsymbol{y};\hat{\boldsymbol{\theta}}(\boldsymbol{y})) = \ln \sum_{\boldsymbol{y}\in\mathcal{X}} \mathcal{L}_{\mathcal{M}}(\boldsymbol{y}) \tag{2.10}$$

is named *parametric complexity* of $\mathcal{M}$. Notice that $\mathcal{X}$ represents the set of all possible data sets of fixed size V, i.e., all the observable samples of size V. Thus, $\mathrm{COMP}_{\mathcal{M}}$ depends on both the model and $\mathcal{X}$. Finally, we can rewrite equation (2.9) as

$$\mathrm{DL}^{\mathrm{NML}}_{\mathcal{M}}(\boldsymbol{x}) = -\ln \mathcal{L}_{\mathcal{M}}(\boldsymbol{x}) + \mathrm{COMP}_{\mathcal{M}}. \tag{2.11}$$

Similarly to other criteria rooted in information theory, this description length consists of two terms: the log-likelihood term, evaluated on the data and favoring models with a high goodness-of-fit, and the complexity term, depending only on the chosen model and penalizing the ones that are too complex.

Once we have evaluated the description length of our data according to a basket of competing models, the MDL principle prescribes selecting the one providing the minimum DL.

2.3 Description length of canonical and microcanonical models

Although computing the complexity term within the NML-based description length can be challenging, models defined by a sufficient statistic admit a simpler expression for this term. This section introduces an important class of models of this kind, the maximum-entropy models (MEMs), both in their canonical and microcanonical formulations.

Specifically, we will focus on discrete data that can be represented as a binary $n \times m$ matrix $\mathbf{G}$, i.e.

$$\mathbf{G} = \{g_{ij}\}_{\substack{i=\{1\ldots n\}\\ j=\{1\ldots m\}}} \tag{2.12}$$

and denote by $\mathcal{G}$ the set of all such matrices that, when endowed with a probability distribution $P(\mathbf{G})$, constitute a proper ensemble of matrices. The matrix representation is often used to describe systems composed of n elements to be modeled (corresponding to the number of rows), each characterized by m state variables or degrees of freedom (corresponding to the number of columns). The size of the data set is fixed and supposed to represent the number of independent entries of the matrix, i.e., nm.

In the most common scenario, we can access a vector of quantities $\mathbf{c}^* = \mathbf{c}(\mathbf{G}^*)$, evaluated on the only available matrix $\mathbf{G}^*$, and representing the sufficient statistic of the model. Maximum-entropy models are, then, obtained by looking for the probability distribution that maximizes Shannon entropy

$$\mathcal{S}[P] = -\sum_{\mathbf{G}\in\mathcal{G}} P(\mathbf{G}) \ln P(\mathbf{G}) \tag{2.13}$$

while constraining the sufficient statistics. If the model is required to reproduce the value of the constraints exactly, i.e., $\mathbf{c}(\mathbf{G}) = \mathbf{c}^*$, one obtains a *microcanonical* model.

If, instead, the model is required to reproduce the value of the constraints on average, i.e., $\langle \mathbf{c}(\mathbf{G}) \rangle_P = \mathbf{c}^*$ with $\langle \cdot \rangle_P$ being the ensemble average induced by distribution P, one obtains the corresponding *canonical* formulation. This approach, introduced by Gibbs [2] and reprised by Jaynes [1], leads to ensembles of matrices that reproduce (either exactly or on average) the constrained quantities while randomizing everything else.

In what follows, we will adopt the NML-based approach to compute the description lengths of microcanonical and canonical MEMs, which is optimal (in a minimax sense) when assuming no prior knowledge on the parameters. Thus, the description length is taken as the one defined in equation (2.11). For models admitting a sufficient statistic, the complexity term takes the convenient form

$$\mathrm{COMP} = \ln \sum_{\mathbf{c} \in \mathcal{C}} \Omega(\mathbf{c}) \mathcal{L}(\mathbf{c}), \tag{2.14}$$

where $\mathcal{L}(\mathbf{c})$ denotes the value of the maximum likelihood in correspondence with a sufficient statistic reading $\mathbf{c}$ and $\Omega(\mathbf{c})$ denotes the number of configurations with $\mathbf{c}(\mathbf{G}) = \mathbf{c}$, i.e.,

$$\Omega(\mathbf{c}) \equiv \sum_{\mathbf{G} \in \mathcal{G}} \delta_{\mathbf{c}(\mathbf{G}), \mathbf{c}} \tag{2.15}$$

with $\delta_{i,j}$ being the Kronecker delta. As this set may be empty for some values of $\mathbf{c}$, we will solely focus on the *graphical values*, i.e., the set $\mathcal{C} = \{\mathbf{c} : \Omega(\mathbf{c}) \neq \emptyset\}$ of values that are verified by at least one configuration, and denote it with $\mathcal{N}_{\mathcal{C}} = |\mathcal{C}|$ its cardinality. In other words, the sum in equation (2.14) runs over the graphical values of the sufficient statistics rather than over all possible data.

2.3.1 Description length of microcanonical models

When considering microcanonical models, entropy maximization yields the following functional form

$$P_{\mathrm{mic}}(\mathbf{G}; \mathbf{c}) = \begin{cases} \frac{1}{\Omega(\mathbf{c})} & \text{if } \mathbf{c}(\mathbf{G}) = \mathbf{c} \\ 0 & \text{else,} \end{cases} \tag{2.16}$$

where $\mathbf{c}$ is the vector of parameters whose values lie in the discrete set $\mathcal{C}$ - for microcanonical models, in fact, the vector of parameters corresponds to the sufficient statistics itself - and $P_{\mathrm{mic}}(\mathbf{G}; \mathbf{c})$ is a uniform probability whose mass differs from zero only in correspondence of those configurations that verify the condition $\mathbf{c}(\mathbf{G}) = \mathbf{c}$. The derivation of this probability functional form as the solution of entropy maximization under hard constraints is given in more detail in Section 2.A of the Appendix.

We will now evaluate the description length of a microcanonical model. Its maximum log-likelihood simply reads

$$\ln \mathcal{L}_{\mathrm{mic}}(\mathbf{c}) = \ln \frac{1}{\Omega(\mathbf{c})} = -\ln \Omega(\mathbf{c}). \tag{2.17}$$

While combining this equation with equation (2.14), one obtains the following expression for the related complexity:

$$\mathrm{COMP}_{\mathrm{mic}} = \ln \sum_{\mathbf{c} \in \mathcal{C}} \Omega(\mathbf{c}) \cdot \frac{1}{\Omega(\mathbf{c})} = \ln \sum_{\mathbf{c} \in \mathcal{C}} 1 = \ln \mathcal{N}_{\mathcal{C}}. \tag{2.18}$$

Hence, the description length of a microcanonical maximum-entropy model is

$$\mathrm{DL}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G}^*) = \ln \Omega(\mathbf{c}^*) + \ln \mathcal{N}_{\mathcal{C}}. \tag{2.19}$$

2.3.2 Description length of canonical models

When considering canonical models, entropy maximization yields the following well-known exponential form

$$P_{\mathrm{can}}(\mathbf{G};\boldsymbol{\theta}) = \frac{e^{-\boldsymbol{\theta}\cdot\mathbf{c}(\mathbf{G})}}{Z(\boldsymbol{\theta})} \tag{2.20}$$

with $Z(\boldsymbol{\theta}) = \sum_{\mathbf{G}\in\mathcal{G}} e^{-\boldsymbol{\theta}\cdot\mathbf{c}(\mathbf{G})}$ being the *partition function*. In the context of network theory, the maximum-entropy canonical models are also known as *Exponential Random Graph Models* [4, 5, 6]. In this case, the vector of parameters $\boldsymbol{\theta}$ corresponds to the vector of Lagrange multipliers needed to carry out a constrained entropy maximization. The ML estimators $\hat{\boldsymbol{\theta}}(\mathbf{G})$ are found by solving the system of equations

$$\langle \mathbf{c} \rangle_{\hat{\boldsymbol{\theta}}(\mathbf{G})} = \mathbf{c}(\mathbf{G}) \tag{2.21}$$

with $\langle\cdot\rangle_{\boldsymbol{\theta}}$ representing the ensemble average induced by $P_{\mathrm{can}}(\mathbf{G};\boldsymbol{\theta})$.

The maximum log-likelihood of a canonical model is linked to the microcanonical one through the Kullback-Leibler divergence:

$$S(P_{\mathrm{mic}}||P_{\mathrm{can}})|_{\mathbf{c},\boldsymbol{\theta}} = \sum_{\mathbf{G}\in\mathcal{G}} P_{\mathrm{mic}}(\mathbf{G};\mathbf{c}) \ln \frac{P_{\mathrm{mic}}(\mathbf{G};\mathbf{c})}{P_{\mathrm{can}}(\mathbf{G};\boldsymbol{\theta})}. \tag{2.22}$$

In fact, it can be shown [13, 12] that

$$\begin{aligned} D_{\mathrm{KL}}(\mathbf{c}^*) &\equiv S(P_{\mathrm{mic}}||P_{\mathrm{can}})|_{\mathbf{c}^*,\boldsymbol{\theta}^*} \\ &= \ln \frac{P_{\mathrm{mic}}(\mathbf{G}^*;\mathbf{c}^*)}{P_{\mathrm{can}}(\mathbf{G}^*;\boldsymbol{\theta}^*)} \\ &= \ln \mathcal{L}_{\mathrm{mic}}(\mathbf{c}^*) - \ln \mathcal{L}_{\mathrm{can}}(\mathbf{c}^*) \end{aligned} \tag{2.23}$$

where $\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}}(\mathbf{c}^*)$. Hence,

$$\mathcal{L}_{\mathrm{can}}(\mathbf{c}) = \mathcal{L}_{\mathrm{mic}}(\mathbf{c}) \cdot e^{-D_{\mathrm{KL}}(\mathbf{c})} = \frac{1}{\Omega(\mathbf{c})} \cdot e^{-D_{\mathrm{KL}}(\mathbf{c})} \tag{2.24}$$

for any value of the sufficient statistic, and

$$\begin{aligned} \mathrm{COMP}_{\mathrm{can}}^{\mathrm{NML}} &= \ln \sum_{\mathbf{c}\in\mathcal{C}} \Omega(\mathbf{c}) \cdot \frac{1}{\Omega(\mathbf{c})} \cdot e^{-D_{\mathrm{KL}}(\mathbf{c})} \\ &= \ln \sum_{\mathbf{c}\in\mathcal{C}} e^{-D_{\mathrm{KL}}(\mathbf{c})}. \end{aligned} \tag{2.25}$$

Consequently, the description length of a canonical model can be expressed as

$$\begin{aligned} \mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}(\mathbf{G}^*) &= -\ln \mathcal{L}_{\mathrm{can}}(\mathbf{c}^*) + \ln \sum_{\mathbf{c}\in\mathcal{C}} e^{-D_{\mathrm{KL}}(\mathbf{c})} \\ &= \ln \Omega(\mathbf{c}^*) + D_{\mathrm{KL}}(\mathbf{c}^*) + \ln \sum_{\mathbf{c}\in\mathcal{C}} e^{-D_{\mathrm{KL}}(\mathbf{c})}. \end{aligned} \tag{2.26}$$

2.3.3 Comparing canonical and microcanonical models

We can now compare the description length of canonical and microcanonical models. Given that D_{KL} is guaranteed to be non-negative, we have

$$\begin{aligned}\Delta \ln \mathcal{L}(\mathbf{G}^*) &\equiv \ln \mathcal{L}_{\mathrm{can}}(\mathbf{G}^*) - \ln \mathcal{L}_{\mathrm{mic}}(\mathbf{G}^*) \\ &= -D_{\mathrm{KL}}(\mathbf{c}^*) \leqslant 0 \end{aligned} \tag{2.27}$$

i.e., the canonical likelihood is always smaller than or equal to the microcanonical one. For the same reason, it is straightforward to see that the canonical complexity is always smaller than, or equal to, the microcanonical one:

$$\begin{aligned}\Delta\mathrm{COMP} &\equiv \mathrm{COMP}_{\mathrm{can}} - \mathrm{COMP}_{\mathrm{mic}} \\ &= \ln \frac{\sum_{\mathbf{c}\in\mathcal{C}} e^{-D_{KL}(\mathbf{c})}}{\mathcal{N}_{\mathcal{C}}} \leqslant 0. \end{aligned} \tag{2.28}$$

Thus, microcanonical models achieve a higher goodness-of-fit but, at the same time, are more complex. The interplay between the two differences determines the difference between the canonical and the microcanonical description length:

$$\begin{aligned}\Delta\mathrm{DL}(\mathbf{G}^*) &\equiv \mathrm{DL}^{\mathrm{NML}}_{\mathrm{can}}(\mathbf{G}^*) - \mathrm{DL}^{\mathrm{NML}}_{\mathrm{mic}}(\mathbf{G}^*) \\ &= -\Delta \ln \mathcal{L}(\mathbf{G}^*) + \Delta\mathrm{COMP} \\ &= D_{\mathrm{KL}}(\mathbf{c}^*) + \ln \frac{\sum_{\mathbf{c}\in\mathcal{C}} e^{-D_{\mathrm{KL}}(\mathbf{c})}}{\mathcal{N}_{\mathcal{C}}}. \end{aligned} \tag{2.29}$$

The following reasoning provides a different way of looking at the expression above. The canonical distribution $P_{\mathrm{can}}(\mathbf{G};\boldsymbol{\theta})$ induces the following distribution on the values of the sufficient statistic

$$Q_{\mathrm{can}}(\mathbf{c};\boldsymbol{\theta}) = \Omega(\mathbf{c}) P_{\mathrm{can}}(\mathbf{c};\boldsymbol{\theta}) \tag{2.30}$$

where $P_{\mathrm{can}}(\mathbf{c};\boldsymbol{\theta})$ is the canonical probability of a matrix $\mathbf{G}$, such that $\mathbf{c}(\mathbf{G}) = \mathbf{c}$, for a given value of the vector of parameters $\boldsymbol{\theta}$. The maximum likelihood assigned to $\mathbf{c}$ by the correspondent model $\{Q_{\mathrm{can}}(\mathbf{c},\boldsymbol{\theta})\}_{\theta\in\Theta}$ is, thus,

$$\begin{aligned} Q_{\mathrm{can}}(\mathbf{c};\hat{\boldsymbol{\theta}}(\mathbf{c})) &= \Omega(\mathbf{c}) P_{\mathrm{can}}(\mathbf{c};\hat{\boldsymbol{\theta}}(\boldsymbol{c})) \\ &= \frac{\mathcal{L}_{\mathrm{can}}(\mathbf{c})}{\mathcal{L}_{\mathrm{mic}}(\mathbf{c})} = e^{-D_{\mathrm{KL}}(\mathbf{c})} \end{aligned} \tag{2.31}$$

and

$$\Delta\mathrm{DL}(\mathbf{G}^*) = -\ln Q_{\mathrm{can}}(\mathbf{c}^*;\boldsymbol{\theta}^*) + \ln \frac{\sum_{\mathbf{c}\in\mathcal{C}} Q_{\mathrm{can}}(\mathbf{c};\hat{\boldsymbol{\theta}}(\mathbf{c}))}{\mathcal{N}_{\mathcal{C}}}; \tag{2.32}$$

upon defining

$$\bar{Q}_{\mathrm{can}} \equiv \frac{\sum_{\mathbf{c}\in\mathcal{C}} Q_{\mathrm{can}}(\mathbf{c},\hat{\boldsymbol{\theta}}(\mathbf{c}))}{\mathcal{N}_{\mathcal{C}}} \tag{2.33}$$

as the uniform average of $Q_{\mathrm{can}}(\mathbf{c};\hat{\boldsymbol{\theta}}(\mathbf{c}))$ over all possible values of $\mathbf{c}\in\mathcal{C}$, we find that the difference between description lengths can be expressed as

$$\Delta\mathrm{DL}(\mathbf{G}^*) = \ln \frac{\bar{Q}_{\mathrm{can}}}{Q^*_{\mathrm{can}}}, \tag{2.34}$$

where $Q^*_{\text{can}} \equiv Q_{\text{can}}(\mathbf{c}^*, \boldsymbol{\theta}^*)$. This expression depends only on the canonical model, providing a new interpretation of the comparison between the canonical and the microcanonical description length: the canonical model provides a shorter description length whenever $\Delta\text{DL}(\mathbf{G}^*) < 0$, i.e., whenever $Q^*_{\text{can}} > \bar{Q}_{\text{can}}$. In other words, the canonical model is the best-scoring model whenever the maximum likelihood it assigns to the observed value $\mathbf{c}^*$ is higher than the maximum likelihood averaged over all possible values of the sufficient statistic (or, more intuitively, if it performs 'better than on average' on the observed sufficient statistic). Remarkably, this is in open contrast with a comparison based only on the likelihoods, according to which the microcanonical formulation should always be preferred, independently of the observed data.

Expression (2.29) shows that the entity of the difference between the two description lengths crucially depends on the value of the KL divergence. As said in the introduction, the canonical and the microcanonical ensembles are (non-)equivalent if the relative D_{KL} is (non-)vanishing as the system size approaches infinity. When studying matrices, identifying the system size with the number of entries, i.e., nm, may be inappropriate, as adding (only) one entry to a matrix without altering its structure is, in fact, impossible. In these cases, it may be much more reasonable to identify the system size with the number n of items: increasing the system size from n to $n+1$ would, thus, correspond to adding a node to a network or a time-point to a time series.

In this scenario, the canonical and the microcanonical ensembles are equivalent in the measure sense [7] if

$$\lim_{n\to\infty} \frac{D_{\text{KL}}(\mathbf{c}^*)}{n} = \lim_{n\to\infty} \frac{|\Delta \ln \mathcal{L}(\mathbf{c}^*)|}{n} = 0 \tag{2.35}$$

or, equivalently, if

$$D_{\text{KL}}(\mathbf{c}^*) = |\Delta \ln \mathcal{L}(\mathbf{c}^*)| = o(n). \tag{2.36}$$

where $o(x)$ stands for a quantity that goes to 0 when divided by x as $n \to +\infty$.

This definition of non-equivalence relies on comparing the likelihood functions of the two models. We aim to extend it in order to take into account the models' complexities. As a first step, we consider the order of ΔDL. Notice that

$$D_{\text{KL}}(\mathbf{c}) = o(n) \quad \forall \mathbf{c} \in \mathcal{C} \quad \Rightarrow \quad \Delta\text{DL}(\mathbf{c}) = o(n) \quad \forall \mathbf{c} \in \mathcal{C} \tag{2.37}$$

as it follows directly from (2.29). Thus, if the two ensembles are equivalent, the difference in description lengths between the corresponding models will be $o(n)$. However, a similar conclusion cannot be immediately derived when non-equivalence holds and D_{KL} grows at least linearly with n. Indeed, the likelihood and complexity terms in (2.29) are of the same order and have opposite signs; therefore, evaluating the order of ΔDL is a non-trivial task in this case. The following section addresses this problem by explicitly evaluating ΔDL in two notable examples of maximum-entropy models.

2.4 Applications to maximum-entropy models

The simplest constraint one may imagine to enforce is the sum of the elements of a matrix, i.e.

$$l(\mathbf{G}) \equiv \sum_{i=1}^{n}\sum_{j=1}^{m} g_{ij}. \tag{2.38}$$

In the case of bipartite binary networks, this quantity corresponds to the total number of links and induces the well-known Erdös-Rényi model. While this model is one of the most popular ones, it is also widely recognized to fall short in capturing the topological heterogeneity that characterizes most real-world systems. An alternative approach to introduce heterogeneity is to constrain the sum of elements of each row, i.e.

$$r_i(\mathbf{G}) \equiv \sum_{j=1}^{m} g_{ij} \quad i = 1 \dots n. \tag{2.39}$$

In the case of bipartite binary networks, the sequence $r_i(\mathbf{G})$, $i = 1 \dots n$ coincides with the degree sequence of the nodes belonging to just one layer and induces the canonical MEM known as Bipartite Partial Configuration Model [55]. As we will employ asymptotic results to compare the canonical and the microcanonical description lengths, we need to specify the scaling of the constraints $\mathbf{c}^*$ as $n \to \infty$. In what follows, we will focus on *dense matrices*, i.e., we will assume that $g^*_{ij} = \Theta(1)\ \forall (i,j)$, where $\Theta(x)$ indicates a quantity of the same order of x as $n \to \infty$. This choice leads to $r^*_i = \Theta(m)$ and $l^* = \Theta(nm)$, the 'actual' order of these quantities depending on the growth rate of m. We will consider two cases: (i) m finite in the limit, i.e., $m = \Theta(1)$ and (ii) m growing linearly with n, i.e., $m = \Theta(n)$. As a consequence of our assumption, the matrix density

$$p^* = \frac{l^*}{nm} \tag{2.40}$$

and the row densities

$$p^*_i = \frac{r^*_i}{m} \quad i = 1 \dots n \tag{2.41}$$

will remain finite in the asymptotic limit, i.e., $p^* = \Theta(1)$ and $p^*_i = \Theta(1)$, $\forall\, i$.

2.4.1 Enforcing one global constraint

Since we are considering binary matrices, constraining the total number of 1s in a microcanonical fashion leads to a number of configurations equal to $\Omega(l) = \binom{nm}{l}$. The corresponding microcanonical distribution, thus, becomes

$$P_{\text{mic}}(\mathbf{G}; l) = \begin{cases} \frac{1}{\binom{nm}{l}} & \text{if } l(\mathbf{G}) = l \\ 0 & \text{else,} \end{cases} \tag{2.42}$$

inducing the maximum log-likelihood

$$\ln \mathcal{L}_{\text{mic}}(l) = -\ln \binom{nm}{l}. \tag{2.43}$$

The complexity equals the logarithm of the number of graphical values, i.e.

$$\mathrm{COMP}_{\mathrm{mic}}^{\mathrm{NML}} = \sum_{l=0}^{nm} 1 = \ln[nm+1]. \tag{2.44}$$

The microcanonical description length, thus, reads

$$\mathrm{DL}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G}^*) = \ln \binom{nm}{l^*} + \ln[nm+1]. \tag{2.45}$$

The canonical probability can be computed explicitly, as this model represents one of the few canonical MEMs whose partition function can be computed analytically. Specifically, it reads $Z(\theta) = (1+e^{\theta})^{nm}$ and induces the expression

$$P_{\mathrm{can}}(\mathbf{G};\theta) = \frac{e^{-\theta l(\mathbf{G})}}{(1+e^{-\theta})^{nm}}. \tag{2.46}$$

Upon defining

$$p \equiv \frac{e^{-\theta}}{1+e^{-\theta}}, \tag{2.47}$$

we can re-write the canonical probability via the so-called *mean-value parametrization* leading to the expression

$$P_{\mathrm{can}}(\mathbf{G};p) = p^{l(\mathbf{G})}(1-p)^{nm-l(\mathbf{G})} \tag{2.48}$$

i.e., the distribution of a collection of nm i.i.d. Bernoulli variables, indicating that each matrix element is either 1, with probability p, or 0, with probability $1-p$. The ML estimator of p is given by the matrix density $\hat{p}(\mathbf{G}) = l(\mathbf{G})/nm$ and the maximum log-likelihood, obtained by inserting $\hat{p}$ into (2.48), reads

$$\ln \mathcal{L}_{\mathrm{can}}(l) = -nm \cdot h\left(\frac{l}{nm}\right) \tag{2.49}$$

where $h(p) = -p\ln(p) - (1-p)\ln(1-p)$ is the entropy of a Bernoulli distribution with a parameter equal to p. An exact expression for the canonical complexity has been derived in [56] and reads

$$\mathrm{COMP}_{\mathrm{can}}^{\mathrm{NML}} = \ln\left[\frac{e^{nm}\Gamma(nm,nm)}{nm^{nm-1}} + 1\right] \tag{2.50}$$

where $\Gamma(s,x)$ is the upper incomplete gamma function. In case s is a positive integer (as it happens to be our case), it can be expressed as

$$\Gamma(s,x) = e^{-x}(s-1)!\sum_{k=0}^{s-1}\frac{x^k}{k!}. \tag{2.51}$$

In conclusion, the canonical description length reads

$$\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}(\mathbf{G}^*) = nm \cdot h\left(\frac{l^*}{nm}\right) + \ln\left[\frac{e^{nm}\Gamma(nm,nm)}{nm^{nm-1}} + 1\right]. \tag{2.52}$$

Notably, the details of the matrix structure play no role as $\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}$ solely depend on global quantities, namely the number of 1s and the sample size - in our case, nm.

So far, all the reported results are exact. Nevertheless, as we are interested in the order of ΔDL when $n \to \infty$, we consider the following asymptotic results, derived from the exact ones (see section 2.B.1 of the Appendix):

$$\Delta \ln \mathcal{L}(\mathbf{G}^*) = -\frac{1}{2} \ln[nm] - \frac{1}{2} \ln[2\pi p^*(1-p^*)] + o(1), \tag{2.53}$$

$$\Delta\mathrm{COMP} = -\frac{1}{2} \ln[nm] + \frac{1}{2} \ln\left[\frac{\pi}{2}\right] + o(1) \tag{2.54}$$

and, upon combining them, the (asymptotic) difference between description lengths becomes

$$\Delta\mathrm{DL}(\mathbf{G}^*) = \frac{1}{2} \ln[\pi^2 p^*(1-p^*)] + o(1). \tag{2.55}$$

The expressions above provide two significant insights. First, the order of the difference between log-likelihood functions is $o(n)$, regardless of the growth rate of m, a result implying that condition (2.36) is met and ensemble equivalence holds. Second, the leading terms of expressions (2.53) and (2.54) are identical, thus canceling out when compared. Consequently, the (asymptotic) difference between description lengths depends on the size of the system only through the matrix density p^*, assumed to be finite in the dense regime. Thus, such a difference is finite in the same regime, in agreement with (2.37).

2.4.2 Enforcing n local constraints

Let us now consider canonical and microcanonical models induced by an extensive number of parameters, each one corresponding to the sum of the elements of a different row. With this choice, the matrix can be viewed as a collection of n independent strips of size $1 \times m$. Since we impose a single constraint per row, all the prior results hold row-wise, with m replacing nm and r_i replacing l. Specifically, the microcanonical probability distribution reads

$$P_{\mathrm{mic}}(\mathbf{G};\mathbf{r}) = \begin{cases} \prod_{i=1}^{n} \frac{1}{\binom{m}{r_i}} & \text{if } \mathbf{r}(\mathbf{G}) = \mathbf{r} \\ 0 & \text{else,} \end{cases} \tag{2.56}$$

and the related maximum log-likelihood is the sum of the individual ones, i.e.,

$$\ln \mathcal{L}_{\mathrm{mic}}(\mathbf{G}^*) = -\sum_{i=1}^{n} \ln \binom{m}{r_i^*}. \tag{2.57}$$

Analogously, the microcanonical complexity reads

$$\mathrm{COMP}_{\mathrm{mic}} = n \ln[m+1], \tag{2.58}$$

inducing the expression

$$\mathrm{DL}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G}^*) = \sum_{i=1}^{n} \ln \binom{m}{r_i^*} + n \ln[m+1]. \tag{2.59}$$

Similarly, the canonical probability is the product of row-specific canonical probabilities

$$P_{\text{can}}(\mathbf{G};\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{e^{-\theta_i r_i(\mathbf{G})}}{(1+e^{-\theta_i})^m}, \tag{2.60}$$

and can be rewritten as

$$P_{\text{can}}(\mathbf{G};\mathbf{p}) = \prod_{i=1}^{n} p_i^{r_i(\mathbf{G})}(1-p_i)^{m-r_i(\mathbf{G})} \tag{2.61}$$

upon considering the mean-value parametrization

$$p_i \equiv \frac{e^{-\theta_i}}{(1+e^{-\theta_i})^m} \quad i = 1 \dots n. \tag{2.62}$$

Since for each row the ML estimator of the parameter p_i is given by the row density $\hat{p}_i(\mathbf{G}) = \frac{r_i(\mathbf{G})}{m}$, the maximum log-likelihood reads

$$\ln \mathcal{L}_{\text{can}}(\mathbf{r}) = -m \sum_{i=1}^{n} h\left(\frac{r_i}{m}\right) \tag{2.63}$$

and the complexity reads

$$\text{COMP}_{\text{can}} = n \ln\left[\frac{e^m \Gamma(m,m)}{m^{m-1}} + 1\right]. \tag{2.64}$$

In conclusion, the canonical description length reads

$$\text{DL}_{\text{can}}(\mathbf{G}^*) = m \sum_{i=1}^{n} h\left(\frac{r_i}{m}\right) + n \ln\left[\frac{e^m \Gamma(m,m)}{m^{m-1}} + 1\right]. \tag{2.65}$$

In the hypothesis of m growing linearly with n, i.e., $m = \Theta(n)$, we consider the following asymptotic expressions (for a derivation, see section 2.B.2 of the Appendix)

$$\begin{aligned} \Delta \ln \mathcal{L}(\mathbf{G}^*) = &-\frac{n}{2}\ln[m] - \frac{1}{2}\sum_{i=1}^{n} \ln[2\pi p_i^*(1-p_i^*)] \\ &+ \frac{n}{12m} - \frac{1}{12m}\sum_{i=1}^{n} \frac{1}{p_i^*(1-p_i^*)} + o(1), \end{aligned} \tag{2.66}$$

$$\begin{aligned} \Delta\text{COMP} = &-\frac{n}{2}\ln[m] + \frac{n}{2}\ln\left[\frac{\pi}{2}\right] \\ &+ a \cdot \frac{n}{\sqrt{m}} + b \cdot \frac{n}{m} + o(1) \end{aligned} \tag{2.67}$$

with

$$\begin{aligned} a &= \frac{2}{3}\sqrt{\frac{2}{\pi}} \simeq 0.53, \\ b &= -\frac{11}{12} - \frac{4}{9\pi} \simeq -1.06. \end{aligned}$$

By combining these expressions, we obtain the asymptotic expression

$$\Delta \mathrm{DL}(\mathbf{G}^*) = \frac{1}{2}\sum_{i=1}^{n} \ln[\pi^2 p_i^*(1-p_i^*)] + a \cdot \frac{n}{\sqrt{m}} + c \cdot \frac{n}{m} + \frac{1}{12m}\sum_{i=1}^{n} \frac{1}{p_i^*(1-p_i^*)} + o(1)$$

where $c = -1 - \frac{4}{9\pi} \simeq -1.14$.

As for the case of one global constraint, we focus on two aspects of these results. First, the difference between log-likelihood functions is either $\Theta(n)$ or $\Theta(n \ln n)$ depending on the order of m, i.e., $m = \Theta(1)$ in the first case and $m = \Theta(n)$ in the second case. In any case, according to (2.36), the ensemble equivalence is broken. In particular, if $m = \Theta(n)$, the leading terms of expressions (2.66) and (2.67) are both of order $\Theta(n \ln n)$, thus canceling out when compared, precisely as in the equivalent case. Nonetheless, the remaining terms are still of order $\Theta(n)$. Hence, the (asymptotic) difference between description lengths increases linearly with the system size, regardless of the growth rate of m.

2.4.3 Towards model-level equivalence

At this point, it is important to emphasize that comparing canonical and microcanonical *ensembles* is not the same as comparing the corresponding *models*. The former involves comparing two probability distributions, each obtained for specific parameter values, whereas the latter involves comparing two sets of distributions (i.e., two models), with each set sharing a common functional form within itself. While in the context of the measure-level definition of ensemble equivalence, the comparison between ensembles can be reduced to comparing their likelihood terms, the comparison between models must take into account both their likelihoods and complexities.

In this section, we have shown two examples where the leading-order terms of the differences in likelihood and complexity are identical. This confirms that the complexities play as crucial a role in model comparison as the likelihoods. Furthermore, we validated relation (2.37) in the case of a single global constraint: when ensemble equivalence holds, the difference in description lengths between the canonical and microcanonical ensembles turns out to be $o(n)$. At the same time, our result that ΔDL grows with n in the case of n local constraints shows that the signature of broken ensemble equivalence remains visible when model complexities are included in the comparison.

These findings lead to the following question: can we generalize the definition of ensemble equivalence at the model level? To do so, we need a quantity that plays a role analogous to the relative entropy in comparing ensembles, but that applies to comparing models.

ΔDL might seem like a good candidate, but it is not suitable for this purpose: while it does take into account the model complexities, it still depends on the observed values. To be more general, we propose a definition involving the universal NML distributions representing the two models. Inspired by the measure level definition of

non-equivalence [7] of equation (2.35), we consider the following KL divergence

$$\begin{aligned} D_{\mathrm{KL}}^{\mathrm{NML}} &\equiv S(P_{\mathrm{mic}}^{\mathrm{NML}} || P_{\mathrm{can}}^{\mathrm{NML}}) \\ &= \sum_{\mathbf{G}\in\mathcal{G}} \bar{P}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G}) \ln \frac{\bar{P}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G})}{\bar{P}_{\mathrm{can}}^{\mathrm{NML}}(\mathbf{G})} \\ &= \sum_{\mathbf{G}\in\mathcal{G}} \bar{P}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G}) \Delta\mathrm{DL}(\mathbf{G}) \\ &= \sum_{\mathbf{c}\in\mathcal{C}} \Omega(\mathbf{c}) \frac{1}{\Omega(\mathbf{c})\mathcal{N}_{\mathcal{C}}} \Delta\mathrm{DL}(\mathbf{c}) \\ &= \overline{\Delta\mathrm{DL}(\mathbf{c})}, \end{aligned} \tag{2.68}$$

where $\overline{\Delta\mathrm{DL}(\mathbf{c})}$ represents the uniform average of $\Delta\mathrm{DL}(\mathbf{c})$ over all possible values of $\mathbf{c} \in \mathcal{C}$. Following the standard definition and assuming that the more relevant scale to consider is still the system size n, we say that the canonical and microcanonical models are equivalent *on the model level* if

$$\lim_{n\to\infty} \frac{D_{\mathrm{KL}}^{\mathrm{NML}}}{n} = \lim_{n\to\infty} \frac{\overline{\Delta\mathrm{DL}(\mathbf{c})}}{n} = 0. \tag{2.69}$$

Notice that, according to (2.37)

$$D_{\mathrm{KL}}(\mathbf{c}) = o(n) \quad \forall \mathbf{c} \in \mathcal{C} \quad \Rightarrow \quad D_{\mathrm{KL}}^{\mathrm{NML}} = o(n). \tag{2.70}$$

i.e., whenever ensemble equivalence holds, the corresponding microcanonical and canonical models are equivalent on the model level. The reverse implication, verified here in the specific case of one-sided local constraints, will be investigated in future work.

2.5 NML-based VS Bayesian approach to model selection

In the previous sections, we implicitly assumed that no prior knowledge about the parameters was available. Under this assumption, the NML-based universal distribution is the best choice in a precise minimax sense, as shown in section 2.2. Here, we consider the Bayesian approach to model selection [16, 17] by introducing the Bayesian universal distribution

$$\bar{P}_{\mathcal{M}}^{\mathrm{Bayes}}(\boldsymbol{x}) = \int_{\Theta} P_{\mathcal{M}}(\boldsymbol{x};\boldsymbol{\theta}) w(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2.71}$$

(see [57, 58] for applications in the context of network models). The main difference with respect to the NML description length lies in the presence of a prior on the parameters $w(\boldsymbol{\theta})$ to be supplied by the user. This implementation, which is formally equivalent to the *Bayes factor method* [59], provides a description length for $\boldsymbol{x}$ through model $\mathcal{M}$ that is defined as

$$\mathrm{DL}_{\mathcal{M}}^{\mathrm{Bayes}}(\boldsymbol{x}) \equiv -\ln \bar{P}_{\mathcal{M}}^{\mathrm{Bayes}}(\boldsymbol{x}). \tag{2.72}$$

In some cases, the NML-based approach can be retrieved in a Bayesian context once the Bayesian distribution is equipped with a prior such that the two description lengths are, at least asymptotically, the same. We will refer to these priors as *NML-optimal priors*. In the microcanonical case, the uniform prior is an NML-optimal prior. Indeed, when the parameter space is a discrete set, the integral in (2.72) is replaced by a sum, and the microcanonical Bayesian description length reads

$$\begin{aligned} \mathrm{DL}_{\mathrm{mic}}^{\mathrm{Bayes}}(\mathbf{G}^*) &= -\ln \mathcal{L}_{\mathrm{mic}}(\mathbf{G}^*) - \ln w(\mathbf{c}^*) \\ &= \ln \Omega(\mathbf{c}^*) - \ln w(\mathbf{c}^*), \end{aligned} \tag{2.73}$$

which coincides with the NML-based description length provided by equation (2.19), upon identifying $w(\mathbf{c})$ with a uniform prior over $\mathcal{C}$:

$$w^{\mathrm{Uniform}}(\mathbf{c}) = \frac{1}{\mathcal{N}_{\mathcal{C}}} \quad \forall\, \mathbf{c} \in \mathcal{C}. \tag{2.74}$$

In the canonical case, the situation is quite different. First, consider V i.i.d. outcomes $\boldsymbol{x} = (x_1, \dots x_V)$ from a general k-dimensional exponential model. When $\boldsymbol{x}$ is such that the ML estimators of $\boldsymbol{\theta}$ are bounded away from the boundaries of Θ as the sample size goes to infinity, and under some regularity conditions on $\mathcal{M}$ which hold in our setting, the following asymptotic results hold true:

$$\mathrm{DL}_{\mathcal{M}}^{\mathrm{NML}}(\boldsymbol{x}) = -\ln \mathcal{L}_{\mathcal{M}}(\boldsymbol{x}) + \frac{k}{2} \ln \frac{V}{2\pi} + \ln \int_{\Theta} \sqrt{\det \mathbf{I}(\boldsymbol{\theta})} d\boldsymbol{\theta} + o(1), \tag{2.75}$$

$$\mathrm{DL}_{\mathcal{M}}^{\mathrm{Bayes}}(\boldsymbol{x}) = -\ln \mathcal{L}_{\mathcal{M}}(\boldsymbol{x}) + \frac{k}{2} \ln \frac{V}{2\pi} - \ln w(\hat{\boldsymbol{\theta}}(\boldsymbol{x})) + \ln \sqrt{\det \mathbf{I}(\hat{\boldsymbol{\theta}}(\boldsymbol{x}))} + o(1), \tag{2.76}$$

where $\det \mathbf{I}(\theta)$ is the determinant of the $k \times k$ *normalized Fisher information matrix*.

If we equip the Bayesian universal distribution with the Jeffreys prior [60]

$$J(\boldsymbol{\theta}) = \frac{\sqrt{\det \mathbf{I}(\boldsymbol{\theta})}}{\int_{\Theta} \sqrt{\det \mathbf{I}(\boldsymbol{\theta})} d\boldsymbol{\theta}} \tag{2.77}$$

by posing $w(\boldsymbol{\theta}) \equiv J(\boldsymbol{\theta})$, the two asymptotic expressions coincide (for a formal derivation of these results, we redirect the interested reader to [53] and [16]). Thus, whenever the asymptotic formulas (2.75) and (2.76) hold true, the NML-based description length is asymptotically equivalent to the one provided by the Bayesian distribution equipped with Jeffreys prior (hereby, Bayes-Jeffreys description), which is NML-optimal.

This is verified in the case of one global constraint, where the resulting canonical model turns out to be an exponential i.i.d. model. The asymptotic formulas do not provide the rate of convergence between the two description lengths, which can be derived by directly comparing the exact expressions of the NML description length $\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}$ (2.65) and the Bayes-Jeffreys description length

$$\mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Jeffreys}}(\mathbf{G}^*) = \ln[\pi(nm)!] - \ln\left[\Gamma(nm - l^* + 1/2)\Gamma(l^* + 1/2)\right] \tag{2.78}$$

(see section 2.C.1 of the Appendix for a derivation). In the limit $n \to \infty$ one finds

$$\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}(\mathbf{G}^*) - \mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Jeffreys}}(\mathbf{G}^*) = \frac{2}{3}\sqrt{\frac{2}{\pi nm}} + \Theta\left(\frac{1}{nm}\right). \tag{2.79}$$

The same comparison can be carried out in the case of n local constraints by applying the results above to each row. In this case, the Bayes-Jeffreys description length reads

$$\mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Jeffreys}}(\mathbf{G}^*) = n\ln[\pi m!] - \sum_{i=1}^{n} \ln\left[\Gamma(m - r_i^* + 1/2)\Gamma(r_i^* + 1/2)\right] \tag{2.80}$$

and in the limit $n \to \infty$ one finds

$$\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}(\mathbf{G}^*) - \mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Jeffreys}}(\mathbf{G}^*) = \frac{2}{3}\frac{\sqrt{2}n}{\sqrt{\pi m}} + \Theta\left(\frac{n}{m}\right). \tag{2.81}$$

As $n \to \infty$, the last difference diverges both in case m stays finite, i.e., $m = \Theta(1)$, and in case m diverges, upon assuming that the rate at which m diverges does not exceed that of n. Consequently, in this case, it is no longer true that the NML approach can be asymptotically retrieved from the Bayesian one equipped with Jeffreys prior. Although the existence of a modified Jeffreys prior extending the equivalence of the two approaches to the case of n local constraints can indeed be hypothesized, for the time being, we merely suggest not to assume this to be true when ensemble non-equivalence holds.

In the remainder of this section, we show that the Bayesian description length becomes more sensitive to the choice of prior whenever local constraints are enforced. Let us start by considering the relationship

$$\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}} = \mathrm{DL}_{\mathrm{mic}}^{\text{Bayes-Rissanen}} \tag{2.82}$$

stating that the NML-based description length of a canonical model coincides with the Bayesian description length of its microcanonical variant, equipped with the so-called *canonical prior*, introduced by Rissanen in [61] (see the definition in section 2.C.2 of the Appendix). Moreover,

$$\mathrm{DL}_{\mathrm{mic}}^{\mathrm{NML}} = \mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Uniform}} \tag{2.83}$$

i.e., the NML-based description length of the microcanonical model coincides with the Bayesian description length of its canonical variant equipped with a uniform prior on the mean-value parameters.

The two identities, derived in 2.C.2, show that there exists a prior guaranteeing that the Bayesian description length of the canonical (microcanonical) model coincides with the NML-based description length of the microcanonical (canonical) model. This might lead to the conclusion that, in the framework of model selection, canonical and microcanonical models coincide even when non-equivalence holds if the right priors are chosen. Nevertheless, these priors are not NML-optimal. While this non-optimality is negligible in the case of one global constraint, it could play a crucial role when

local constraints are enforced. In other words, in this case, different choices of prior can result in very different Bayesian description lengths for microcanonical as for canonical models.

The n local constraints case examined in the previous sections provides an instructive example. In this case, we showed that the difference between the canonical and microcanonical NML description lengths is $\Theta(n)$. Combining this result with (2.82), we can conclude that when it comes to the microcanonical model, choosing between the uniform prior (which is NML-optimal) and the canonical prior can amount to a difference in description length of $\Theta(n)$ nats. Similarly, when it comes to the canonical model, it can be easily shown that, based on (2.81, 2.83) when $m = \Theta(n)$, choosing between the uniform and the Jeffreys prior can lead to a difference in description length of $\Theta(\sqrt{n})$ nats. In contrast, the same differences are $\Theta(1)$ when one global constraint is involved.

This simple example shows that when non-equivalence holds, the selection of different priors can result in increasingly larger differences in DL when the system size grows. In other words, fine-tuning the prior can result in a significant gain - or, conversely, a loss - in terms of compression, particularly for large-scale systems.

2.6 Conclusion

This study explored the implications of ensemble non-equivalence in the context of MDL-based model selection. We specifically compared the NML-based description lengths of canonical and microcanonical maximum-entropy models. The NML approach enables a comparison of model description lengths derived from the same underlying principle without requiring prior assumptions about the parameters.

We found that, although microcanonical models fit the data better than their canonical counterparts, they are always more complex. Determining the best-scoring model - i.e., the one that minimizes the description length - is not straightforward, as it depends on the observed values of the sufficient statistics. Specifically, we found that the canonical model performs best when its fit to the observed data exceeds its average fit across all possible values of the sufficient statistics.

Additionally, we explicitly compared the description lengths of canonical and microcanonical models for dense $n \times m$ binary matrices in two cases of interest: one with a global constraint, such as the sum of the matrix entries, and another with n local constraints, such as the sum of the entries in each row. In the former case, the model is homogeneous, where the probability of an entry being 1 is the same for all entries, while in the latter, the model becomes heterogeneous, with the probability of an entry being 1 depending on the specific row. In this second scenario, the canonical and microcanonical ensembles are no longer equivalent.

By comparing the description lengths of the canonical and microcanonical formulations of these models, we found that the difference remains finite when a global constraint is enforced but grows linearly with system size when local constraints are imposed. Our findings highlight the impact of choosing hard or soft constraints on data compression, leading us to conclude that canonical and microcanonical models should be treated as distinct from a model selection perspective in the presence

of ensemble non-equivalence. Furthermore, they suggest a new definition of model equivalence based on MDL, which requires further investigation.

In addition, we investigated the relationship between NML-based and Bayesian description lengths. We defined NML-optimal priors as those that, when applied in a Bayesian framework, result in description lengths that asymptotically match the NML-based approach. While the Jeffreys prior is well known to be NML-optimal for canonical models under a single global constraint, we show that this optimality does not hold when local constraints are imposed. Indeed, in such cases, the difference between the corresponding description lengths increases with system size. This observation serves as a caution to practitioners, suggesting that the two approaches (NML and Jeffreys priors) should not be treated as interchangeable in the presence of ensemble non-equivalence. It also raises the question of whether an NML-optimal prior exists when local constraints are enforced.

Lastly, we showed that different priors can result in largely different description lengths when local constraints are involved. This finding underscores the importance for practitioners of carefully selecting priors, since the choice of prior could greatly impact the accuracy and effectiveness of model selection.

Appendix 2.A Derivation of the microcanonical distribution from entropy maximization

It is well known that canonical maximum-entropy models can be derived by maximizing the Shannon entropy subject to soft constraints [1]. Here, we show that microcanonical models can also be seen as the solution to an entropy maximization problem under hard constraints, as a direct consequence of the Shannon-Khinchin axioms that uniquely identify Shannon entropy [62, 3].

Let $\mathcal{G}$ denote the set of all configurations, and let $\mathcal{G}_{\text{valid}} \subset \mathcal{G}$ denote the subset of configurations that exactly satisfy the hard constraints. Consider the problem of maximizing the Shannon entropy

$$S[P] = -\sum_{\mathbf{G}\in\mathcal{G}} P(\mathbf{G}) \log P(\mathbf{G}) \tag{2.84}$$

over probability distributions $P(\mathbf{G})$ supported on $\mathcal{G}$, under the requirement that $P(\mathbf{G}) = 0$ for all $\mathbf{G} \notin \mathcal{G}_{\text{valid}}$. That is, configurations violating the constraints are assigned zero probability.

The third Shannon-Khinchin axiom states that $S[P]$ is expansible, i.e., it does not change if an outcome with zero probability is added. This implies that, in this case, maximizing entropy over the full space $\mathcal{G}$ with support restricted to $\mathcal{G}_{\text{valid}}$ is equivalent to maximizing entropy over $\mathcal{G}_{\text{valid}}$ only. Then, according to the second Shannon-Khinchin axiom, which states that entropy is maximized by the uniform distribution in the absence of further constraints, the solution is the uniform distribution over $\mathcal{G}_{\text{valid}}$:

$$P_{\text{mic}}(\boldsymbol{x}) = \begin{cases} \frac{1}{|\mathcal{G}_{\text{valid}}|} & \text{if } \mathbf{G} \in \mathcal{G}_{\text{valid}}, \\ 0 & \text{else.} \end{cases} \tag{2.85}$$

With this derivation, canonical and microcanonical ensembles emerge as distinct solutions to the same entropy maximization principle, differing only in how the constraints are imposed.

Appendix 2.B Asymptotic expansions

Here we derive the asymptotic expansions used to obtain the asymptotic results in section 2.4.

2.B.1 One global constraint

We start by deriving an asymptotic expansion for the microcanonical log-likelihood:

$$\begin{aligned}\ln \mathcal{L}_{\text{mic}}(l) &= -\ln \binom{nm}{l} \\ &= -\ln(nm)! + \ln(nm-l)! + \ln l!. \end{aligned} \tag{2.86}$$

In the dense regime, l is of order $\Theta(n)$ and Stirling's approximation

$$\ln x! = x \ln x - x + \frac{1}{2}\ln(2\pi x) + \Theta(1/x) \tag{2.87}$$

can be used to evaluate all the factorials above, yielding

$$\ln \mathcal{L}_{\text{mic}}(l) = -nm \ln(nm) + (nm-l)\ln(nm-l) + l \ln l - \frac{1}{2}\ln\left(\frac{nm}{2\pi l(nm-l)}\right) + o(1), \tag{2.88}$$

which can be further expressed as a function of the density $p = l/nm$:

$$\ln \mathcal{L}_{\text{mic}}(p) = -nm \cdot h(p) + \frac{1}{2}\ln(nm) + \frac{1}{2}\ln(2\pi p(1-p)) + o(1). \tag{2.89}$$

By combining this expression with the canonical log-likelihood

$$\ln \mathcal{L}_{\text{can}}(p) = -nm \cdot h(p), \tag{2.90}$$

we obtain the asymptotic log-likelihood difference of equation (2.53).

Similarly, we combine the asymptotic expansion of the microcanonical complexity

$$\begin{aligned}\text{COMP}_{\text{mic}} &= \ln(nm+1) = \ln(nm) + \ln\left(1 + \frac{1}{nm}\right) \\ &= \ln(nm) + o(1), \end{aligned} \tag{2.91}$$

which follows from $\ln\left(1+\frac{1}{x}\right) = \Theta(1/x)$, and the following asymptotic expansion of the canonical complexity

$$\text{COMP}_{\text{can}} = \frac{1}{2}\ln(nm) + \frac{1}{2}\ln\left(\frac{\pi}{2}\right) + o(1). \tag{2.92}$$

The latter is easily derived by asymptotically expanding the following approximation

$$\mathrm{COMP}_{\mathrm{can}}^{\mathrm{NML}} \simeq \ln\left[\sqrt{\frac{\pi nm}{2}} + \frac{2}{3} + \frac{\sqrt{2\pi}}{24\sqrt{nm}} - \frac{4}{135nm} + \frac{\sqrt{2\pi}}{576\sqrt{(nm)^3}} + \frac{8}{2835(nm)^2}\right], \tag{2.93}$$

provided in [56], where the authors show it to be very precise already for small values of the total number of entries nm. Finally, equation (2.54) is obtained by comparing the two asymptotic complexities. Notice that the asymptotic formula (2.92) for the canonical complexity represents a well-known result, as it can be derived directly by applying the asymptotic formula (2.75) to the case of i.i.d. Bernoulli random variables.

2.B.2 One-sided local constraints

Here, we consider the regime in which $m = \Theta(n)$. Similarly to the previous case, we need to expand the microcanonical likelihood

$$\begin{aligned}\ln \mathcal{L}_{\mathrm{mic}}(\mathbf{r}) &= -\sum_{i=1}^{n} \ln \binom{m}{r_i} \\ &= -n\ln m! + \sum_{i=1}^{n} \ln(m-r_i)! + \ln r_i!. \end{aligned} \tag{2.94}$$

In the dense regime, all r_i's are of order $\Theta(n)$, and we could again apply Stirling's formula to the factorials above. Nevertheless, because of the factor n ahead of everything, Stirling's formula is not enough, and we turn to Stirling's series [63]

$$\ln x! = x\ln x - x + \frac{1}{2}\ln(2\pi x) + \frac{1}{12x} + \Theta(1/x^3). \tag{2.95}$$

The resulting asymptotic expression can be expressed as a function of the row densities $p_i = r_i/m$:

$$\begin{aligned}\ln \mathcal{L}_{\mathrm{mic}}(\mathbf{r}) = -m\sum_{i=1}^{n} h(p_i) + \frac{n}{2}\ln m + \frac{1}{2}\sum_{i=1}^{n}[\ln(2\pi p_i(1-p_i))] \\ - \frac{n}{12m} + \frac{1}{12m}\sum_{i=1}^{n}\left[\frac{1}{p_i(1-p_i)}\right] + o(1). \end{aligned} \tag{2.96}$$

By combining this expression with the canonical log-likelihood

$$\ln \mathcal{L}_{\mathrm{can}}(\mathbf{p}) = -m\sum_{i=1}^{n} h(p_i), \tag{2.97}$$

one obtains the asymptotic log-likelihood difference of equation (2.66).

Similarly, the asymptotic expansion of the microcanonical complexity

$$\begin{aligned}\mathrm{COMP}_{\mathrm{mic}} &= n\ln(m+1) = \ln m + n\ln\left(1+\frac{1}{m}\right) \\ &= n\ln m + \frac{n}{m} + o(1), \end{aligned} \tag{2.98}$$

which follows from $\ln\left(1+\frac{1}{x}\right)=\frac{1}{x}+\Theta(1/x^2)$, is compared to the following asymptotic expansion of the canonical complexity, based on (2.93)

$$\mathrm{COMP}_{\mathrm{can}}=\frac{n}{2}\ln m+\frac{n}{2}\ln\frac{\pi}{2}+\frac{2}{3}\sqrt{\frac{2}{\pi}}\cdot\frac{n}{\sqrt{m}}+\left(\frac{1}{12}-\frac{4}{9\pi}\right)\cdot\frac{n}{m}+o(1) \tag{2.99}$$

to express (2.64). The asymptotic expression (2.67) is obtained as the difference between the two asymptotic complexities.

Appendix 2.C NML and Bayes

2.C.1 Bayesian-Jeffreys description lengths

In what follows, we compute the Bayesian description length $\mathrm{DL}_{\mathrm{can}}^{\mathrm{B,\ Jeffreys}}$ of the canonical model obtained by constraining the sum l over the matrix. As already stated, this model is equivalent to modeling nm i.i.d. Bernoulli variables, and this result can be found in Example 8.3 of [16]. In the paper, we extend this result to the case of n one-sided local constraints.

First, we compute Jeffreys prior. For a Bernoulli variable, the Fisher information for the parameter p is $\mathbf{I}(p)=p^{-1}(1-p)^{-1}$ and $\int_0^1\sqrt{\mathbf{I}(p)}=\pi$. Thus, the Jeffreys prior reads

$$J(p)=\frac{1}{\pi\sqrt{p(1-p)}} \tag{2.100}$$

and the Bayesian-Jeffreys DL is

$$\begin{aligned}\mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Jeffreys}}(\mathbf{G}^*)&=-\ln\int_0^1 P_{\mathrm{can}}(\mathbf{G}^*;p)J(p)dp\\&=-\ln\frac{1}{\pi}\int_0^1 p^{l^*-\frac{1}{2}}(1-p)^{nm-l^*-\frac{1}{2}}dp.\end{aligned} \tag{2.101}$$

The integral above is the Beta function

$$B(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}dt=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \tag{2.102}$$

computed for $x=l^*+\frac{1}{2}$ and $y=nm-l^*+\frac{1}{2}$, where $\Gamma(x)$ is the Gamma function

$$\Gamma(x)=\int_0^{+\infty}t^{x-1}e^{-t}dt. \tag{2.103}$$

Thus

$$\mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Jeffreys}}(\mathbf{G}^*)=-\ln\frac{\Gamma\left(l^*+\frac{1}{2}\right)\Gamma\left(nm-l^*+\frac{1}{2}\right)}{\pi(nm)!}, \tag{2.104}$$

which corresponds to Equation (2.78). By employing the Stirling series, this expression can be expanded asymptotically as

$$\mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Jeffreys}}(\mathbf{G}^*)=nm\cdot h\left(\frac{l^*}{nm}\right)+\frac{1}{2}\ln\frac{\pi nm}{2}+\Theta\left(1/nm\right). \tag{2.105}$$

The last expression is compared with the asymptotic expansion of the canonical NML description length $\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}$, obtained as the difference between the asymptotic expansion of (2.93) and the canonical log-likelihood:

$$\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}(\mathbf{G}^*) = nm \cdot h\left(\frac{l^*}{nm}\right) + \frac{1}{2}\ln\frac{\pi nm}{2} + \frac{2}{3}\sqrt{\frac{2}{\pi nm}} + o(1). \tag{2.106}$$

Finally, Equation (2.79) is obtained as the difference between (2.106) and (2.105)

2.C.2 Proving identities

In what follows, we prove identities (2.82) and (2.83) of section 2.5.

Proving Identity (2.82). We begin with identity (2.82), namely:

$$\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}} = \mathrm{DL}_{\mathrm{mic}}^{\text{Bayes-Rissanen}}, \tag{2.107}$$

where the canonical prior $\hat{W}$ inducing the Bayes-Rissanen description length is expressed as

$$\begin{aligned}\hat{W}(\mathbf{c}) &= \frac{\sum_{\mathbf{G}:\mathbf{c}(\mathbf{G})=\mathbf{c}} P_{\mathrm{can}}(\mathbf{G};\hat{\boldsymbol{\theta}}(\mathbf{G})}{\sum_{\mathbf{G}} P_{\mathrm{can}}(\mathbf{G};\hat{\boldsymbol{\theta}}(\mathbf{G}))} \\ &= \frac{\Omega(\mathbf{c})\mathcal{L}_{\mathrm{can}}(\mathbf{c})}{\sum_{\mathbf{G}} \mathcal{L}_{\mathrm{can}}(\mathbf{G})}.\end{aligned} \tag{2.108}$$

We recognize the canonical complexity in the denominator of $\hat{W}$. Thus, putting this expression in the Bayesian microcanonical description length, we get

$$\begin{aligned}\mathrm{DL}_{\mathrm{mic}}^{\text{Bayes-Rissanen}}(\mathbf{G}^*) &= \ln\Omega(\mathbf{c}^*) - \ln\hat{W}(\mathbf{c}^*) \\ &= -\ln\mathcal{L}_{\mathrm{can}}(\mathbf{c}) + \mathrm{COMP}_{\mathrm{can}} \\ &= \mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}(\mathbf{G}^*),\end{aligned} \tag{2.109}$$

which proves the identity.

Proving Identity (2.83). We will now prove identity (2.83), namely:

$$\mathrm{DL}_{\mathrm{mic}}^{\mathrm{NML}} = \mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Uniform}}, \tag{2.110}$$

for the specific maximum-entropy models considered in this work, starting from the case of one global constraint. The uniform prior on the mean-value parameter p reads

$$w^{\mathrm{Uniform}}(p) = 1 \quad \text{for } p \in [0,1]. \tag{2.111}$$

Putting this prior in the Bayesian description length yields:

$$\begin{aligned}
\mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Uniform}}(\mathbf{G}^*) &= -\ln \int_0^1 P_{\mathrm{can}}(\mathbf{G}^*; p) dp \\
&= -\ln \int_0^1 p^{l^*} (1-p)^{nm-l^*} dp \\
&= \ln \binom{nm}{l^*} + \ln(nm+1) \\
&= \mathrm{DL}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G}^*),
\end{aligned} \tag{2.112}$$

where the integral is computed by integrating by parts l^* times. This proves the identity for the case of one global constraint.

The identity holds as well for the case of n one-sided local constraints, with the uniform prior reading

$$w^{\mathrm{Uniform}}(\mathbf{p}) = 1 \quad \text{for } \mathbf{p} \in [0,1]^n. \tag{2.113}$$

Indeed, we have that

$$\begin{aligned}
\mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Uniform}}(\mathbf{G}^*) &= -\ln \int_0^1 P_{\mathrm{can}}(\mathbf{G}^*; p) dp \\
&= -\sum_{i=1}^{n} \ln \int_0^1 p_i^{r_i^*} (1-p)^{m-r_i^*} dp_i \\
&= \sum_{i=1}^{n} \ln \binom{m}{r_i^*} + n \ln(m+1) \\
&= \mathrm{DL}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G}^*),
\end{aligned} \tag{2.114}$$

which proves the identity for the case of n one-sided local constraints.

CHAPTER 3

Testing maximum entropy models with e-values

This chapter is based on:
F.Giuffrida, D. Garlaschelli and P. Grünwald. Testing maximum entropy models with e-values. *Phys. Rev. E 113, 054306* (2026).

Abstract

E-values have recently emerged as a robust and flexible alternative to p-values for hypothesis testing, especially under optional continuation, i.e., when additional data from further experiments are collected. In this work, we define optimal e-values for testing between maximum entropy models, both in the microcanonical (hard constraints) and canonical (soft constraints) settings. We show that, when testing between two hypotheses that are both microcanonical, the so-called growth-rate optimal e-variable admits an exact analytical expression, which also serves as a valid e-variable in the canonical case. For canonical tests, where exact solutions are typically unavailable, we introduce a microcanonical approximation and verify its excellent performance via both theoretical arguments and numerical simulations. We then consider constrained binary models, focusing on $2 \times k$ contingency tables – an essential framework in statistics and a natural representation for various models of complex systems. Our microcanonical optimal e-variable performs well in both settings, constituting a new tool that remains effective even in the challenging case where the number k of groups grows with the sample size, as in models with growing features used for the analysis of real-world heterogeneous networks and time series.

3.1 Introduction

In recent years, scientific interest in complex data modeling has surged, due to the increasing availability of both global-scale structured data and computational power. At the same time, rising concerns about the misuse of p-values and significance testing [19, 20, 21] underscore the need for reliable statistical methods to extract knowledge from data. As a robust and flexible alternative to p-values for hypothesis testing, *e-values* [22, 23] have recently gained considerable attention. Having been independently (re)discovered several times in different contexts (including by physicists [14] — see [22] for early history) over the past decades, interest suddenly exploded in 2019 when the first versions of several breakthrough papers [24, 25, 26, 27] appeared on arXiv.

An *e-variable* is a nonnegative random variable whose expected value under the null hypothesis is at most one. The value it takes on the given sample is called the e-value. This simple definition yields several desirable properties: e-values provide rigorous control of the Type I error, retain it under optional continuation (i.e., when data from additional experiments become available), and can be interpreted as a measure of evidence against the null hypothesis. However, not all e-variables are equally useful as test statistics. To address this, a notion of *optimality* is introduced. An *optimal* e-variable is one that grows quickly under the alternative hypothesis, accumulating strong evidence against the null when the latter is false. In this paper, we focus specifically on *growth-rate optimal* (GRO) e-variables [24]. The results in [24], later extended in [64, 65], provide a general theoretical framework for constructing GRO e-variables in broad testing scenarios.

The aim of this work is to develop optimal e-variables for hypothesis testing between maximum entropy models (MEMs). These models are derived by considering each possible realization $\mathbf{x} \in \mathcal{X}$ of the data (where $\mathcal{X}$ is the set of allowed realizations) and looking for the probability distribution $P(\mathbf{x})$ that maximizes Shannon entropy

$$\mathcal{S}[P] = -\sum_{\mathbf{x}\in\mathcal{X}} P(\mathbf{x}) \log P(\mathbf{x}) \tag{3.1}$$

under a set of constraints, typically defined through a vector of observables $\mathbf{c}(\mathbf{x})$ over the data. This approach, due to Gibbs [2] and Jaynes [1], outputs ensembles of data reproducing the constrained quantities and randomizing everything else maximally.

Two main formulations of MEMs exist, depending on how the constraints are enforced. If the constraints are imposed as exact values, i.e., $\mathbf{c}(\mathbf{x}) = \mathbf{c}^*$ on each realizable $\mathbf{x}$, one obtains a *microcanonical model*, where only configurations satisfying the constraints are assigned nonzero probability. If, instead, the constraints are satisfied only on average, i.e., $\mathbb{E}_P[\mathbf{c}(\mathbf{x})] = \mathbf{c}^*$, one obtains a *canonical model*, where fluctuations are allowed and the probability distribution has exponential form. In statistical terminology, by varying $\mathbf{c}^*$ one obtains an *exponential family with discrete outcome space and uniform carrier* [66]. In both cases, the probability of $\mathbf{x}$ is entirely determined by the value of $\mathbf{c}(\mathbf{x})$, which plays the role of sufficient statistic.

MEMs are commonly used to model complex systems that give rise to structured data, e.g., in network science [31, 36, 6] and time-series analysis [38, 37], where

they capture structural properties such as (heterogeneous) node degrees in networks or empirical trends in (non-stationary) temporal data, respectively. However, while statistical tests for exponential family models are well established, testing procedures specifically tailored to maximum entropy models remain far less developed, especially when applied in the microcanonical setting. In fact, e-variables for testing between general MEMs have so far not been developed at all: the only related works we are aware of are [67, 30] and [29, 28]. The former concentrates on the very specific subcase of 2×2 tables (to re-appear as Example A–C in our paper later on), but uses e-variables which are designed for purely sequential purposes, and are therefore not optimal in the sense we define below, neither in the canonical nor in the microcanonical setting. The latter works, [29, 28], studied e-variables for testing between two exponential families with the same sufficient statistic but different carriers. By contrast, testing between MEMs amounts to testing exponential families with different sufficient statistics but the same (uniform) carrier. This is exactly the aim of this work.

This paper is organized as follows. In section 3.2, we introduce e-variables and growth-rate optimality. In section 3.3, we address the problem of finding optimal e-variables for testing between two maximum entropy models, either microcanonical (3.3.1) or canonical (3.3.2), that differ in their sufficient statistics. We introduce a method to construct optimal e-variables. We show that the microcanonical GRO e-variable is also a valid canonical e-variable, and that in some cases, it asymptotically coincides with the optimal one. This is particularly relevant: while canonical models are way more commonly used in the literature, calculating the optimal canonical e-variable is usually analytically impossible and computationally infeasible. Here, we provide a method to compute the optimal microcanonical e-variable explicitly and to further verify how well it approximates the optimal canonical e-variable. In section 3.4, we explicitly apply these results to contingency tables, highlighting their connections to important problems in network science. We first analyze the case of 2×2 contingency tables (3.4.1) and then generalize to $2 \times k$ (3.4.2). In both cases, we provide fully worked-out examples, allowing for exact calculations of the corresponding e-variables and a detailed comparison between microcanonical and canonical constructions. We show that, in all scenarios considered, the GRO microcanonical e-variable is not only a valid canonical e-variable but also an excellent approximation of the optimal canonical one.

3.2 Introduction to E-variables

Consider the typical hypothesis testing scenario, where the goal is to test a *null hypothesis* $\mathcal{M}_0$ against an *alternative hypothesis* $\mathcal{M}_1$. Both $\mathcal{M}_0$ and $\mathcal{M}_1$ are assumed to be parametric statistical models, i.e., families of distributions sharing the same functional form:

$$\mathcal{M}_j = \{P_j(\mathbf{x}; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_j}, \quad j \in \{0, 1\}, \tag{3.2}$$

where $\boldsymbol{\theta}$ represents the vector of model parameters and $\boldsymbol{\Theta}_j$ the corresponding parameter space for model $\mathcal{M}_j$.

An *e-variable* E is a non-negative random variable that satisfies the following condition under all distributions in the null hypothesis:

$$\mathbb{E}_0[\,E\,] \leqslant 1 \quad \forall P_0 \in \mathcal{M}_0. \tag{3.3}$$

The realized value of E evaluated on data $\mathbf{x}$ is called an *e-value*. Unlike p-values, *larger* e-values indicate stronger evidence against the null. This follows directly from their defining property that, under the null, their expectation is bounded by one. This simple yet powerful definition has several important implications [22, 23]:

- **Type I error control**: The condition $\mathbb{E}_0[\,E\,] \leqslant 1$ ensures that a test based on e-values controls the Type I error, that is, the probability of rejecting the null hypothesis when it is actually true. Given a significance level $0 \leqslant \alpha \leqslant 1$, by Markov's inequality, we have

$$P_0(E \geqslant 1/\alpha) \leqslant \alpha, \tag{3.4}$$

  for all $P_0 \in \mathcal{M}_0$. This guarantees that the probability of wrongly rejecting the null hypothesis does not exceed the significance level α, regardless of the true parameter value within the null model.

- **Post-hoc error control:** E-values allow a variation of valid Type-I error control even when the significance level is chosen *after* observing the data [68]. Specifically, if e is the observed e-value, then rejecting the null hypothesis at level $1/e$ preserves a Type I risk bound despite this level being data-dependent. This contrasts with traditional p-values, which only guarantee valid inference when the significance level is fixed in advance.

- **Optional continuation:** E-values support valid testing under *optional continuation*, making them well-suited for sequential analyses and meta-analyses across independent studies. If $e_{(1)}, e_{(2)}, \ldots$ are e-values computed on independent data batches (e.g., studies), their product remains a valid e-value — even if the decision to analyze further batches, to perform tests on them, or to incorporate specific prior knowledge into the e-values is guided by the outcomes of earlier batches. In this way, Type I error control is preserved, enabling flexible and robust hypothesis testing across repeated or cumulative experimental settings.

We emphasize that post-hoc error control and Type-I error control under optional continuation *cannot* be achieved with p-value-based and other classical hypothesis testing methods — in particular, classical meta-analyses in the medical and the social sciences come without any error guarantees. Regarding the latter, [69] contains an e-value-based re-analysis of 2 of 28 classic psychological findings considered in the *ManyLabs2 Project* [70]. The goal of this project was to asses the replicability of the original results. Each of the 28 psychological studies was repeated in many labs across the world, each repetition (called *replication attempt*) involving newly acquired, expensive experimental data. [69] found that, for the two findings considered, if e-values had been used, correct and statistically valid conclusions could have been

drawn based on a much smaller number of replication attempts than had actually been done, saving enormous amounts of time and money. For an overview of other practical applications of e-values (there are many, besides meta-analysis), we refer to [22, 23].

While all random variables satisfying condition (3.3) qualify as e-variables, not all of them are informative. For instance, the constant random variable $E(\mathbf{x}) \equiv 1$ satisfies the definition but provides no information. To address this, a notion of *optimal* e-variables was introduced in [24]. In particular, the authors define the *Growth Rate Optimal* (GRO) e-variable as the unique solution to a specific optimization problem based on a growth criterion, which we present below.

As a first step toward understanding GRO e-variables, we introduce the concept of *Bayesian evidence* (also known as the *Bayesian marginal likelihood*) of model j with prior density w_j, defined as:

$$P_j^{w_j}(\mathbf{x}) = \int_{\boldsymbol{\Theta}_j} P_j(\mathbf{x}; \boldsymbol{\theta}) w_j(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{3.5}$$

This quantity reflects the overall support the data provide for model j, by averaging the likelihood over the prior; it is widely used in Bayesian model selection, where models with higher evidence are preferred. We shall mostly work with prior densities w_j defined on convex parameter spaces $\boldsymbol{\Theta}_j \subset \mathbb{R}^d$ ($d > 0$), assuming they are continuous and strictly positive for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_j$. We refer to such priors as *regular priors*.

Given a fixed prior w_1 (regular or not) on the alternative hypothesis, the GRO e-variable S^{GRO} is the unique solution to the following optimization problem:

$$S^{\mathrm{GRO}} = \arg \max_{E \in \mathcal{E}_0} \mathbb{E}_{P_1^{w_1}}[\log E], \tag{3.6}$$

where $\mathcal{E}_0$ denotes the set of all e-variables relative to the null model $\mathcal{M}_0$, i.e., the set of all random variables satisfying (3.3).

This optimization can be interpreted as a growth criterion: while the expected value of any e-variable is bounded under the null, a well-designed e-variable should grow rapidly assuming the alternative is true, when the prior w_1 is correctly specified. The use of the logarithmic growth in this criterion is motivated and discussed in more detail in [24]. The quantity $\mathbb{E}_{P_1^{w_1}}[\log E]$, known as the *e-power* [71, 72, 73] of E, has become a standard measure for evaluating the performance of an e-variable.

In the most common case, a GRO e-variable solving the aforementioned optimization problem takes the form of a *Bayes factor* [74], i.e., the ratio between two Bayesian evidences:

$$S(\mathbf{x}) = \frac{P_1^{w_1}(\mathbf{x})}{P_0^{w_0}(\mathbf{x})}. \tag{3.7}$$

The fact that the solution to (3.6 looks like (3.7) is not at all obvious — it is the central result of two breakthrough papers, [24, 65]. Equation (3.7) represents the Bayes factor comparing models $\mathcal{M}_1$ and $\mathcal{M}_0$. It measures the relative support that the data provide for one model over the other. However, not all Bayes factors qualify as e-variables; to ensure the e-variable property (3.3), while w_1 may be chosen freely,

a specific prior w_0^*, depending on w_1, must then be chosen for the null hypothesis. Specifically, w_0^* is the solution to the following optimization problem:

$$w_0^* = \arg \min_{w \in \mathcal{W}_{\boldsymbol{\theta}_0}} D_{\mathrm{KL}}(P_1^{w_1} \| P_0^{w_0}) \tag{3.8}$$

where $\mathcal{W}_{\boldsymbol{\theta}_0}$ is the space of all priors on $\boldsymbol{\theta}_0$, and $D_{\mathrm{KL}}(P_1^{w_1} \| P_0^{w_0})$ denotes the *Kullback-Leibler divergence*:

$$\begin{aligned} D_{\mathrm{KL}}(P_1^{w_1} \| P_0^{w_0}) &= \sum_{\mathbf{x} \in \mathcal{X}} P_1^{w_1}(\mathbf{x}) \log \frac{P_1^{w_1}(\mathbf{x})}{P_0^{w_0}(\mathbf{x})} \\ &= \mathbb{E}_{P_1^{w_1}}[\log S(\mathbf{x})]. \end{aligned} \tag{3.9}$$

Theorem 1 in [24] proves that given w_1, among all Bayes factors of the form (3.7), the random variable

$$S^{\mathrm{GRO}}(\mathbf{x}) = \frac{P_1^{w_1}(\mathbf{x})}{P_0^{w_0^*}(\mathbf{x})}, \tag{3.10}$$

is the *only* e-variable, assuming that a w_0^* achieving the minimum in (3.8) exists [1].

To sum up, the GRO e-variable is the unique solution of two different optimization problems, defined on two different sets: for a given $P_1^{w_1}$ and null model $\mathcal{M}_0$, S^{GRO} is the only e-variable among Bayes factors of the form 3.7; at the same time it is the only e-variable maximizing the e-power (see Figure 3.1).

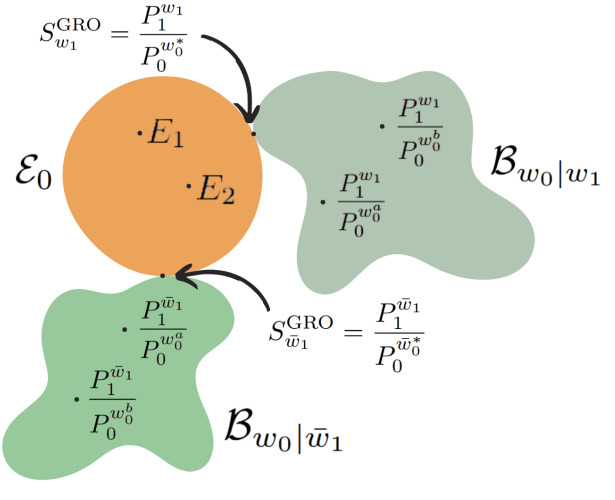


Figure 3.1: The GRO e-variable S^{GRO} is the unique intersection between the set $\mathcal{B}_{w_0|w_1}$ of all Bayes factors for a given $P_1^{w_1}$ and varying $P_0^{w_0}$, and the set $\mathcal{E}_0$ of all e-variables relative to model $\mathcal{M}_0$. At the same time, it is the unique e-variable maximizing the e-power relative to $P_1^{w_1}$. This schematic representation considers two possible alternative priors, w_1 and $\bar{w}_1$.

[1] As shown in [24], multiple distinct minimizers w_0^* may exist, but they yield the same $P_0^{w_0^*}$. Even when a minimizer does not exist, $P_0^{w_0^*}$ can be defined as a limit along a minimizing sequence w_j, ensuring that (3.10) remains a valid e-variable.

The reader may have noticed a seeming asymmetry: while e-values are defined in a frequentist sense — requiring Type I error control for *all* $P_0 \in \mathcal{M}_0$ — our optimality criterion for GRO relies on a prior w_1 over the alternative model $\mathcal{M}_1$, and thus relies on a Bayesian formulation.

It would be conceptually appealing to define an optimality criterion that, like the e-variable condition, provides performance guarantees over *all* $P_1 \in \mathcal{M}_1$ rather than *on average according to a prior* w_1. As it turns out, this is indeed possible by drawing on ideas from information theory.

To move in that direction, we first note that the result of [24] is not limited to Bayes factors. It applies to more general e-variables of the form

$$S(\mathbf{x}) = \frac{\bar{P}_1(\mathbf{x})}{P_0^{w_0^*}(\mathbf{x})}, \tag{3.11}$$

where $\bar{P}_1$ is any probability distribution over the data space. For such constructions to be useful, however, the choice of $\bar{P}_1$ must be guided by an appropriate extension of the GRO criterion.

This leads us to the concept of *regret*, also referred to as *relative growth* (*regrow*) in [24], which we adopt here using more common terminology. Regret quantifies the power loss incurred when using a candidate e-variable instead of the ideal one, which is designed for the true data-generating distribution.

To define it precisely, suppose that the data are generated according to a fixed but unknown distribution $P_1(\mathbf{x}; \boldsymbol{\theta}_1) \in \mathcal{M}_1$. If we knew $\boldsymbol{\theta}_1$, we could construct the GRO e-variable $S^{\mathrm{GRO}(\boldsymbol{\theta}_1)}$ optimal for that specific alternative:

$$S^{\mathrm{GRO}(\boldsymbol{\theta}_1)} = \frac{P_1(\mathbf{x}; \boldsymbol{\theta}_1)}{P_0^{\widetilde{w}_0^*}(\mathbf{x})} \tag{3.12}$$

where

$$\widetilde{w}_0^* = \arg \min_{w_0 \in \mathcal{W}_{\boldsymbol{\theta}_0}} \mathbb{E}_{\boldsymbol{\theta}_1} \left[\log \frac{P_1(\mathbf{x}; \boldsymbol{\theta}_1)}{P_0^{w_0}} \right] \tag{3.13}$$

and $\mathbb{E}_{\boldsymbol{\theta}_j}$ denotes the expected value under $P_j(\,\cdot\,, \boldsymbol{\theta}_j)$. Here, the alternative hypothesis reduces to a *singleton* — a statistical model containing only one distribution — and in some cases, such as the $2 \times k$ contingency tables considered later in this paper, $S^{\mathrm{GRO}(\boldsymbol{\theta}_1)}$ can be computed exactly. The regret of a candidate e-variable S_{cand} is then given by:

$$\mathrm{REG}_1(\boldsymbol{\theta}_1; S_{\mathrm{cand}}) := \mathbb{E}_{\boldsymbol{\theta}_1} \left[\log S^{\mathrm{GRO}(\boldsymbol{\theta}_1)} - \log S_{\mathrm{cand}} \right], \tag{3.14}$$

which quantifies the expected loss in log-growth due to not knowing the true parameter.

Since $\boldsymbol{\theta}_1$ is unknown, a natural robustness criterion is to consider the *worst-case regret* across the entire alternative:

$$\mathrm{REG}_1(\boldsymbol{\Theta}_1; S_{\mathrm{cand}}) := \max_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1} \mathrm{REG}_1(\boldsymbol{\theta}_1; S_{\mathrm{cand}}). \tag{3.15}$$

This leads to a new optimality principle: among all e-variables, one would seek the *minimax optimal* e-variable — i.e., the e-variable that minimizes $\mathrm{REG}_1(\boldsymbol{\Theta}_1; S_{\mathrm{cand}})$

over all valid choices of S_{cand}. However, computing this minimax-optimal e-variable is generally infeasible in practice, as we currently lack efficient algorithms to solve the corresponding optimization problem. Nevertheless, when the models $\mathcal{M}_0$ and $\mathcal{M}_1$ exhibit sufficient regularity — as is the case for Maximum Entropy models, discussed in the next section — GRO e-variables constructed from (3.11) with appropriately chosen $\bar{P}_1$ can closely approximate the minimax optimal solution. In particular, one can consider e-variables of the form (3.11), where the numerator $\bar{P}_1$ is set to a *universal distribution* relative to the alternative model $\mathcal{M}_1$. Universal distributions, which include Bayesian mixtures $P_1^{w_1}$ as special cases, arise naturally in the theory of the *Minimum Description Length (MDL) Principle* [16, 17, 18]. Such choices of $\bar{P}_1$ lead to e-variables that, while not exactly minimax-optimal, are typically close to optimal in terms of regret minimization, and therefore provide a practical and principled strategy for robust hypothesis testing. To clarify this connection, we take a brief detour to explain how e-value-based methods relate to the MDL Principle and its central concept, the universal distribution.

3.2.1 GRO e-values and description lengths

The Minimum Description Length Principle provides a general framework for model selection: from a set of candidate models, it chooses the one that yields the shortest encoding of the observed data. In this approach, each model is represented by a single probability distribution, and models are compared via their *description length*. The preferred model is the one with the smallest description length.

When comparing two models $\mathcal{M}_0$ and $\mathcal{M}_1$, the difference in description lengths is

$$\begin{aligned}\Delta\mathrm{DL}(\mathbf{x}) &= \mathrm{DL}_1(\mathbf{x}) - \mathrm{DL}_0(\mathbf{x}) \\ &= -\log \bar{P}_1(\mathbf{x}) + \log \bar{P}_0(\mathbf{x}),\end{aligned} \tag{3.16}$$

where $\bar{P}_1$ and $\bar{P}_0$ are the representative distributions for $\mathcal{M}_1$ and $\mathcal{M}_0$. By Kraft's inequality [16, 54], the code length to describe $\mathbf{x}$, using a code that compresses optimally in expectation under Q, is (up to rounding) $-\log Q(\mathbf{x})$ nats (log denoting natural logarithm here); thus, $-\log \bar{P}_j(\mathbf{x})$ is the code length implied by $\bar{P}_j$.

Universal distributions and worst-case redundancy

The key point is how to determine a single probability distribution $\bar{P}_j$ representing $\mathcal{M}_j$: it should perform well regardless of which specific distribution within $\mathcal{M}_j$ generated the data. In other words, if a distribution $P \in \mathcal{M}_j$ achieves a short expected code length $\mathbb{E}_P[-\log P(\mathbf{x})]$, then $\bar{P}_j$ should yield a similarly short one. Such $\bar{P}_j$ are called *universal distributions* for the model $\mathcal{M}_j$ [16].

To be more precise, we can define the *redundancy* of $\bar{P}_j$ relative to a parameter $\boldsymbol{\theta}_j$ as

$$\mathrm{RED}_j(\boldsymbol{\theta}_j; \bar{P}_j) := \mathbb{E}_{\boldsymbol{\theta}_j}\big[-\log \bar{P}_j(\mathbf{x}) + \log P_j(\mathbf{x}; \boldsymbol{\theta}_j)\big]. \tag{3.17}$$

This quantity measures the expected extra bits needed when using $\bar{P}_j$ instead of the expected optimal code for $P_j(\cdot; \boldsymbol{\theta}_j)$. The latter is not available in practice, since the

true $\boldsymbol{\theta}_j$ is typically unknown. Thus, it is useful to define the *worst-case redundancy*:

$$\mathrm{RED}_j(\boldsymbol{\Theta}_j; \bar{P}_j) := \max_{\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j} \mathrm{RED}_j(\boldsymbol{\theta}_j; \bar{P}_j). \tag{3.18}$$

A distribution is universal if this quantity is small.

Ideally, one would like to find a $\bar{P}_j$ that minimizes the worst-case redundancy, but this is generally infeasible. However, for a d_j-dimensional parametric model under standard regularity conditions (satisfied by the Maximum Entropy models considered later), the Bayesian choice $\bar{P}_j = P_j^{w_j}$ with a regular prior w_j attains near-optimal performance: its redundancy is within a constant of the optimal value as the sample size grows. This is formalized in the following result.

Definition 3.2.1 (INECCSI sets [16]).
Let

$$\mathcal{M}_j = \{P_j(\mathbf{x}; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_j}, \quad j \in \{0, 1\}.$$

A subset $\boldsymbol{\Theta}'_j \subset \boldsymbol{\Theta}_j$ is an *INECCSI subset* if its interior is a non-empty, convex, compact subset of the interior of $\boldsymbol{\Theta}_j$.

INECCSI subsets exclude boundary effects and ensure regular asymptotics. For instance, in the Bernoulli model with $\Theta = [0, 1]$, any $[\epsilon, 1 - \epsilon]$ with $0 < \epsilon < 1/2$ is INECCSI.

Let $\mathcal{M}_j^{(m)}$ be the i.i.d. extension of $\mathcal{M}_j$ to m observations, i.e., a model over $\mathbf{y}^m = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ where each $\mathbf{x}_i \in \mathcal{X}$ is independently sampled from $P_j(\cdot; \boldsymbol{\theta})$. Let $P_j^{(m)}$ be the i.i.d. extension of P_j and $\bar{P}_j^{(m)}$ be a distribution on $\mathbf{y}^m$. Let $\mathrm{RED}^m(\boldsymbol{\Theta}_j; \bar{P}_j^{(m)})$ be the worst-case redundancy attained by $\bar{P}_j^{(m)}$. Since $\mathbb{E}_{\boldsymbol{\theta}_j}[-\log P_j^{(m)}(\mathbf{y}^m; \boldsymbol{\theta}_j)]$ grows linearly in m, universality of $\bar{P}_j^{(m)}$ requires RED^m to grow sub-linearly in m.

A standard result [16] states that for every INECCSI subset $\boldsymbol{\Theta}'_j$ and regular prior w_j, there exists $C > 0$ such that for all m:

$$\begin{aligned} \frac{d_j}{2} \log m - C \;\leq\; \inf_{\bar{P}_j^{(m)}} \mathrm{RED}^m(\boldsymbol{\Theta}'_j; \bar{P}_j^{(m)}) \\ \leq\; \mathrm{RED}^m(\boldsymbol{\Theta}'_j; P_j^{w_j(m)}) \;\leq\; \frac{d_j}{2} \log m + C, \end{aligned} \tag{3.19}$$

where the infimum is over all distributions on $\mathcal{X}^{(m)}$. The key implications of these results are:

- the minimum achievable worst-case redundancy grows as $(d_j/2) \log m$;
- Bayesian marginal likelihoods are universal distributions, and their redundancy exceeds the minimum attainable by at most a constant — in this sense, they are asymptotically almost optimal.

Universal distributions guarantee low-regret e-variables

We can now formalize the connection between e-values and MDL: we show that using a universal distribution $\bar{P}_1$ as the numerator in the e-variable construction (3.11) leads to small regret. This provides a principled justification for the use of Bayesian mixtures in e-value methods.

To see this, notice that (minus) the log-ratio of any variable of the form

$$S(\mathbf{x}) = \frac{\bar{P}_1(\mathbf{x})}{P_0^{w_0}(\mathbf{x})}$$

induces a difference in description lengths between models $\mathcal{M}_1$ and $\mathcal{M}_0$:

$$-\log S(\mathbf{x}) = -\log \bar{P}_1(\mathbf{x}) + [\log P_0^{w_0}(\mathbf{x})]. \tag{3.20}$$

This mirrors expression (3.16), but with one crucial difference: for S to be an e-variable, the denominator $P_0^{w_0}$ cannot be chosen freely, as it must be the prior w_0^* that ensures that S qualifies as a GRO e-variable (i.e., satisfies the e-variable condition).

This formulation, however, brings a clear interpretative advantage: the description length difference expressed in (3.20) now has a direct statistical interpretation. Indeed, if S is an e-variable, the corresponding code-length difference can be mapped to a statistical significance measure, since Type I error control is guaranteed. This grounds the MDL code-length difference within a frequentist hypothesis testing framework. In particular, smaller values of $-\log S(\mathbf{x})$ correspond to larger e-values and hence stronger evidence against the null model $\mathcal{M}_0$. This observation addresses a long-standing issue in MDL: although it provides a principled model comparison method, it lacks explicit statistical guarantees such as Type I error control [16, Open Problem No. 9, page 413]. Restricting attention to code-length differences that admit an e-value interpretation not only provides such guarantees but also makes it possible to assign a well-defined evidential value to differences in description lengths. This can be seen as the natural solution to the problem — at least for the two-model comparison case [17]. Extending this insight to multiple models remains an important open challenge.

The connection between e-values and MDL becomes even more compelling when considering the regret of e-variables. Indeed, we now show that the worst-case regret of an e-variable using numerator $\bar{P}_1$ is never larger than the worst-case redundancy of $\bar{P}_1$. Let us restrict attention to an INECCSI subset $\boldsymbol{\Theta}_1' \subset \boldsymbol{\Theta}_1$. Moreover, for clarity, let's denote the regret relative to the GRO e-variable associated to $\bar{P}_1$, i.e. the regret obtained by putting $S_{\text{cand}} = \bar{P}_1(\mathbf{x})/P_0^{w_0^*}(\mathbf{x})$ in definition 3.15, as $\mathrm{REG}_1(\boldsymbol{\Theta}_1'; \bar{P}_1)$. Then, for any distribution $\bar{P}_1$, according to definitions (3.15) and (3.12) (see section 3.A) it holds:

$$\mathrm{REG}_1(\boldsymbol{\Theta}_1'; \bar{P}_1) \leqslant \mathrm{RED}_1(\boldsymbol{\Theta}_1'; \bar{P}_1). \tag{3.21}$$

This shows that, in the worst-case over $\boldsymbol{\theta}_1$, the regret of an e-variable built with numerator $\bar{P}_1$ is upper bounded by the redundancy of $\bar{P}_1$. Consequently, choosing a

universal distribution $\bar{P}_1$, which by definition provides small redundancy, guarantees small regret.

This motivates our choices in the next sections: we will construct e-variables by setting $\bar{P}_1$ to the Bayesian mixture $P_1^{w_1}$ (with a regular prior), which yields small regret of order $(d_1/2)\log m + O(1)$. This choice is also common in the literature, as Bayesian mixtures are widely adopted in the construction of e-variables [22]. In Section 3.F of the Supplementary Material, we show the same results for another universal distribution, the Normalized Maximum Likelihood $\bar{P}_1^{\mathrm{NML}}$, which yields regret of the same order, and has already been used (implicitly) in [75] as an e-variable numerator.

3.3 Application to maximum entropy models

Here, we provide explicit formulas for hypothesis tests that involve either microcanonical or canonical maximum entropy models. We focus on the case of discrete data. For a given choice of sufficient statistics $\mathbf{c}(\mathbf{x})$, we denote by $\mathcal{C}$ the discrete set of values of $\mathbf{c}(\mathbf{x})$ that are realizable by at least one $\mathbf{x} \in \mathcal{X}$; for mathematical convenience, we assume that $\mathcal{C}$ is a (finite or countable) subset of $\mathbb{R}^d$ for some $d \in \mathbb{N}$. Moreover, for any given value $\mathbf{c} \in \mathcal{C}$, let $\Omega(\mathbf{c})$ represent the number of configurations satisfying the constraint $\mathbf{c}(\mathbf{x}) = \mathbf{c}$, formally defined as:

$$\Omega(\mathbf{c}) := \sum_{\mathbf{x}\,:\,\mathbf{c}(\mathbf{x})=\mathbf{c}} 1. \tag{3.22}$$

Entropy maximization, when the hard constraints $\mathbf{c}(\mathbf{x}) = \mathbf{c}$ are enforced on each realizable configuration $\mathbf{x}$, yields a *microcanonical* model whose functional form is a uniform distribution over data satisfying the constraints:

$$P_{\mathrm{mic}}(\mathbf{x};\mathbf{c}) = \begin{cases} \frac{1}{\Omega(\mathbf{c})}, & \text{if } \mathbf{c}(\mathbf{x}) = \mathbf{c}; \\ 0, & \text{else.} \end{cases} \tag{3.23}$$

The parameters of a microcanonical model correspond to the sufficient statistics themselves, with values in the discrete parameter space $\boldsymbol{\Theta}_{\mathrm{mic}} = \mathcal{C}$.

When soft constraints $\mathbb{E}[\mathbf{c}(\mathbf{x})] = \mathbf{c}$ are enforced (that is, the value $\mathbf{c}$ of the sufficient statistic is to be met only as an ensemble average), the maximization of the entropy returns a *canonical* model where the resulting functional form of the probability distribution is, instead, exponential, with positive probability for all possible data:

$$P_{\mathrm{can}}(\mathbf{x};\boldsymbol{\theta}) = \frac{e^{-\boldsymbol{\theta}\cdot\mathbf{c}(\mathbf{x})}}{Z(\boldsymbol{\theta})} \tag{3.24}$$

where $Z(\boldsymbol{\theta}) \equiv \sum_{\mathbf{x}\in\mathcal{X}} e^{-\boldsymbol{\theta}\cdot\mathbf{c}(\mathbf{x})}$ is a normalization term known as *partition function*. Canonical models coincide with what is called *exponential families with uniform carrier function* in the statistics literature [66], and the formula above is generally referred to as canonical parametrization, where the parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\mathrm{can}}$ may be viewed as the Lagrange multipliers resulting from the entropy maximization. For each value $\mathbf{c}$

defining the microcanonical model in Eq. (3.23), there is a corresponding value $\boldsymbol{\theta}$ such that $\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{c}(\mathbf{x})] = \mathbf{c}$ under the canonical distribution in Eq. (3.24).

The above 'duality' between canonical and microcanonical models implies that, alternatively, canonical models can also be parameterized using the expected value of the sufficient statistics. Given parameters $\boldsymbol{\theta}$, define the mean value vector:

$$\boldsymbol{\mu}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{c}(\mathbf{x})]. \tag{3.25}$$

This defines a smooth, one-to-one mapping between the canonical parameter space $\boldsymbol{\Theta}_{\text{can}}$ and the set of realizable mean values, denoted by $\mathtt{M}$. In exponential family theory, $\boldsymbol{\mu}$ is known as the *mean value parameter*. For future reference, we refer to $P_{\boldsymbol{\mu}} = P_{\text{can}}(\cdot; \boldsymbol{\theta}(\boldsymbol{\mu}))$ as the canonical distribution defined in its mean value parametrization, where $\boldsymbol{\theta}(\boldsymbol{\mu})$ is the mapping from mean-value parameters to corresponding canonical parameters, i.e. the inverse of $\boldsymbol{\mu}(\boldsymbol{\theta})$.

A well-known result in this setting is that, if the set of possible constraint values $\mathcal{C}$ is finite, then:

- the canonical parameter space is the full space $\boldsymbol{\Theta}_{\text{can}} = \mathbb{R}^d$;
- the corresponding space of mean values $\mathtt{M}$ coincides with the interior of the convex hull of $\mathcal{C}$.

This result ensures that the mapping $\boldsymbol{\theta} \mapsto \boldsymbol{\mu}$ is not only bijective but also covers all "physically meaningful" expected constraint values[2]. In the rest of the paper, we will make use of this bijection and employ whichever parameterization is most convenient. In particular, when dealing with Bayesian marginal likelihoods and their priors, we will typically work in the mean-value space, bearing in mind that all results can be equivalently expressed in the canonical parameter space via the mapping $\boldsymbol{\theta} \mapsto \boldsymbol{\mu}(\boldsymbol{\theta})$.

In what follows, we define GRO e-variables for tests where both the null and the alternative hypotheses are two microcanonical or two canonical MEMs that differ in the choice of constraints. Our main theoretical results are presented in a general form, but to guide the reader through the derivations, we will use a running example throughout (Examples A, B, and C): a simple 2×2 contingency table, representing two groups of binary data.

3.3.1 Microcanonical test

Consider a test where the null $\mathcal{M}_{\text{mic},0}$ is a microcanonical model with sufficient statistics $\mathbf{c}_0$ taking values in set $\mathcal{C}_0$ and the alternative $\mathcal{M}_{\text{mic},1}$ is a microcanonical model with sufficient statistics $\mathbf{c}_1$ taking values in set $\mathcal{C}_1 \neq \mathcal{C}_0$. The parameters of microcanonical models are discrete and correspond to their sufficient statistics. In this section, we restrict our analysis to the case of microcanonical Bayesian universal distributions for the alternative hypothesis, denoted by $P_{\text{mic},1}^{W_1}$, where W_1 is a probability mass function defined on $\mathcal{C}_1$. Thus, the microcanonical GRO e-variable reads

$$S_{\text{mic}}^{\text{GRO}} = \frac{P_{\text{mic},1}^{W_1}(\mathbf{x})}{P_{\text{mic},0}^{W_0^*}(\mathbf{x})} \tag{3.26}$$

[2]It follows from the general theory of exponential families with finite support [76, Theorem 9.2], under a technical condition known as *steepness*, which holds when $\mathcal{C}$ is finite.

and it solves the discrete version of the GRO optimization problem:

$$W_0^* = \arg\min_{W \in \mathcal{W}_{\mathbf{c}_0}} D_{\mathrm{KL}}(P_{\mathrm{mic},1}^{W_1} \| P_{\mathrm{mic},0}^{W_0}) \tag{3.27}$$

where instead of prior densities w_0, we need to consider prior probability mass functions W_0, and $\mathcal{W}_{\mathcal{C}_0}$ is the set of all such distributions on the parameter space $\mathcal{C}_0$. We solve the microcanonical GRO optimization problem explicitly and exactly (full derivation in Section 3.104 of the Supplementary Material; here we report only the main results) and find the optimal prior distribution on the null:

$$W_0^*(\mathbf{c}_0) = \sum_{\mathbf{x}\,:\,\mathbf{c}_0(\mathbf{x})=\mathbf{c}_0} P_{\mathrm{mic},1}^{W_1}(\mathbf{x}), \tag{3.28}$$

i.e., $W_0^*(\mathbf{c}_0)$ is the marginal distribution of the null sufficient statistic $\mathbf{c}_0(\mathbf{x})$ induced by $P_{\mathrm{mic},1}^{W_1}$. In the special case where the alternative sufficient statistics completely determine the value of the null, we say that *Condition A* holds:

Condition A: (3.29)
there exists a function $f : \mathcal{C}_1 \to \mathcal{C}_0$ s.t. $\mathbf{c}_0(\mathbf{x}) = f(\mathbf{c}_1(\mathbf{x}))$.

Under Condition A, one can write:

$$W_0^*(\mathbf{c}_0) = \sum_{\mathbf{c}_1\,:\,f(\mathbf{c}_1)=\mathbf{c}_0} W_1(\mathbf{c}_1), \tag{3.30}$$

i.e., the GRO-optimal prior on the null is the distribution induced on the null sufficient statistics by the alternative prior, or equivalently, the marginal distribution of $\mathbf{c}_0$ induced by W_1.

Once that W_0^* is computed, the microcanonical GRO-optimal e-variable can always be expressed as

$$S_{\mathrm{mic}}^{\mathrm{GRO}}(\mathbf{x}) = \frac{\Omega_0(\mathbf{c}_0(\mathbf{x}))}{\Omega_1(\mathbf{c}_1(\mathbf{x}))} \frac{W_1(\mathbf{c}_1(\mathbf{x}))}{W_0^*(\mathbf{c}_0(\mathbf{x}))}. \tag{3.31}$$

Finally, although the fact that $S_{\mathrm{mic}}^{\mathrm{GRO}}$ is an e-variable follows from a general theorem (Theorem 1 in [24], as mentioned above Equation (3.10)), we additionally provide a further, direct proof showing that its expected value under the null is exactly one:

$$\mathbb{E}_0[S_{\mathrm{mic}}^{\mathrm{GRO}}] = 1 \quad \forall P_{\mathrm{mic},0} \in \mathcal{M}_{\mathrm{mic},0}. \tag{3.32}$$

This direct proof, as well as the section's other detailed calculations and proofs, can be found in section 3.B. For clarity, we now provide a first example application.

Example A

Let us consider the dataset $\mathbf{x} = (\mathbf{x}^a, \mathbf{x}^b)$ consisting of two groups of binary data, represented as $\mathbf{x}^a = (x_1^a, ..., x_{n^a}^a)$ and $\mathbf{x}^b = (x_1^b, ..., x_{n^b}^b)$, with n^a and n^b the respective group sizes. The total sample size is $n = n_a + n_b$. We denote by $n_1^a = \sum_{i=1}^{n_a} x_i^a$ and

$n_1^b = \sum_{i=1}^{n_b} x_i^b$ the total number of 1s in $\mathbf{x}^a$ and $\mathbf{x}^b$, and by $n_1 = n_1^a + n_1^b$ the total number of 1s in $\mathbf{x}$. The aim is to build a microcanonical test to determine whether the probability of observing $x = 1$ differs across groups. To do so, we set the alternative sufficient statistics equal to the number of 1s in each group, $\mathbf{c}_1 = (n_1^a, n_1^b)$, and the null sufficient statistic equal to the total number of 1s, $c_0 = n_1$. In the microcanonical formulation, these quantities are treated as fixed in the respective models. To find the microcanonical GRO e-variable, we apply formula (3.31), where:

- $\Omega_0(n_1) = \binom{n}{n_1}$ is the number of permutations of $\mathbf{x}$ preserving the total number of 1s;
- $\Omega_1(n_1^a, n_1^b) = \binom{n^a}{n_1^a}\binom{n^b}{n_1^b}$ is the number of permutations of $\mathbf{x}$ preserving the total number of 1s in each group.

For the sake of this example, we put independent, discrete uniform priors on the alternative parameters n_1^a and n_1^b:

$$W_1(n_1^a, n_1^b) = \mathcal{U}_a(n_1^a)\,\mathcal{U}_b(n_1^b) = \frac{1}{n^a+1}\,\frac{1}{n^b+1}. \tag{3.33}$$

In this case, Condition A (3.29) holds, as the null sufficient statistics can be written as a function of the alternative one: $n_1 = n_1^a + n_1^b$. Thus, the optimal prior on the null W_0^* is the distribution of n_1 induced by W_1. In this case, that is simply the convolution of $\mathcal{U}_a$ and $\mathcal{U}_b$, which is a triangular discrete function:

$$W_0^*(n_1) = \begin{cases} \frac{n_1+1}{(n^a+1)(n^b+1)}, & \text{if } 0 \leqslant n_1 \leqslant \min(n^a, n^b), \\ \frac{\min(n^a,n^b)+1}{(n^a+1)(n^b+1)}, & \text{if } \min(n^a,n^b) < n_1 \leqslant \max(n^a,n^b), \\ \frac{n^a+n^b+1-n_1}{(n^a+1)(n^b+1)}, & \text{if } \max(n^a,n^b) < n_1 \leqslant n^a+n^b, \\ 0, & \text{if otherwise.} \end{cases} \tag{3.34}$$

as shown in Figure 3.2. The microcanonical GRO-optimal e-value is finally obtained by substituting each term in Eq. (3.31).

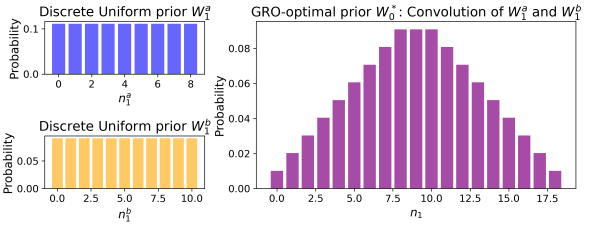


Figure 3.2: In the microcanonical Example A, when the prior on the alternative sufficient statistics n_1^a and n_1^b are uniform distributions (on the left), the resulting GRO-optimal prior on the null sufficient statistic n_1 is the convolution of the two uniform distributions, which results in a triangular distribution (on the right). In this example, $n^a = 8$ and $n^b = 10$.

3.3.2 Canonical test

Consider a test where the null $\mathcal{M}_{\text{can},0}$ is a canonical model with sufficient statistics $\mathbf{c}_0$ taking values in set $\mathcal{C}_0$, and the alternative $\mathcal{M}_{\text{can},1}$ is a canonical model with sufficient statistics $\mathbf{c}_1$ taking values in set $\mathcal{C}_1 \neq \mathcal{C}_0$. The goal is to find the canonical GRO e-variable:

$$S^{\text{GRO}}_{\text{can}} = \frac{\bar{P}_{\text{can},1}(\mathbf{x})}{P^{w_0^*}_{\text{can},0}(\mathbf{x})} \tag{3.35}$$

where w_0^* is a prior density on the mean-value parameter space $\mathtt{M}_0$ and solves the optimization problem

$$w_0^* = \arg \min_{w \in \mathcal{W}_{\boldsymbol{\mu}_0}} D_{\text{KL}}(\bar{P}_1 \| P_0^{w_0}). \tag{3.36}$$

No exact general solution is currently available for this problem. While in some cases it can be solved analytically or numerically, in the majority of cases, there is neither a known analytic solution nor a feasible numerical approach. Here, we propose two candidate approximations, the *microcanonical approximation* and the *pseudo approximation*. As we will see, the first serves as an actual approximation, and the second as a tool to assess whether the former approximation is good.

The definition of the microcanonical approximation is based on two facts, proven in section 3.C:

- A canonical universal distribution $\bar{P}_{\text{can},1}$ with sufficient statistics $\mathbf{c}_1$ can always be expressed as a microcanonical Bayesian marginal likelihood, i.e., there always exists a prior probability mass function $W_{\text{can},1}(\mathbf{c})$ such that

  $$\bar{P}_{\text{can},1} = P^{W_{\text{can},1}}_{\text{mic},1} \tag{3.37}$$

  with $W_{\text{can},1}$ obtained by setting (for $j = 1$):

  $$W_{\text{can},j}(\mathbf{c}_j) = \sum_{\mathbf{x}\,:\,\mathbf{c}_j(\mathbf{x})=\mathbf{c}_j} \bar{P}_{\text{can},j}(\mathbf{x}), \tag{3.38}$$

  i.e., $W_{\text{can},j}(\mathbf{c}_j)$ is equal to the distribution of $\mathbf{c}_j(\mathbf{x})$ induced by $\bar{P}_{\text{can},j}(\mathbf{x})$.

- Given the canonical and microcanonical models $\mathcal{M}_{\text{can}}$ and $\mathcal{M}_{\text{mic}}$ built upon the same sufficient statistic $\mathbf{c}(\mathbf{x})$, a microcanonical e-variable E is always a canonical e-variable:

  $$\mathbb{E}_P[\,E\,] \leqslant 1 \;\; \forall\, P \in \mathcal{M}_{\text{mic}} \quad \Rightarrow \quad \mathbb{E}_P[\,E\,] \leqslant 1 \;\; \forall P \in \mathcal{M}_{\text{can}}. \tag{3.39}$$

Following the first fact, given $\bar{P}_{\text{can},1}$ and using the results of the previous section, we can build the approximating microcanonical GRO e-variable $S^{\text{GRO}}_{\text{mic}}$ for the microcanonical test based on the corresponding $P^{W_{\text{can},1}}_{\text{mic},1} = \bar{P}_{\text{can},1}$. Thus (3.26) becomes

$$S^{\text{GRO}}_{\text{mic}} = \frac{\bar{P}_{\text{can},1}(\mathbf{x})}{P^{W_0^*}_{\text{mic},0}(\mathbf{x})} = \bar{P}_{\text{can},1} \frac{\Omega_0(\mathbf{c}_0(\mathbf{x}))}{W_0^*(\mathbf{c}_0(\mathbf{x}))} \tag{3.40}$$

where, readapting (3.28)

$$W_0^*(\mathbf{c}_0) = \sum_{\mathbf{x}\,:\,\mathbf{c}_0(\mathbf{x})=\mathbf{c}_0} \bar{P}_{\text{can},1}(\mathbf{x}). \tag{3.41}$$

Given the second fact (3.39), the resulting microcanonical GRO e-variable is a valid canonical e-variable, i.e., it is an e-variable for the test between two canonical models, even if, for this test, it is not the GRO-optimal one. As such, from (3.6), it will have a smaller e-power than the canonical GRO one unless the two coincide:

$$\mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{mic}}^{\text{GRO}}\right] \leqslant \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{can}}^{\text{GRO}}\right]. \tag{3.42}$$

The pseudo approximation is further built over the microcanonical one. The prior W_0^* (used to build $S_{\text{mic}}^{\text{GRO}}$), defined on $\mathcal{C}_0$, is transformed into a smooth density $w_{\text{pseudo},0}$ over the corresponding (continuous) mean-value parameter space $\mathtt{M}_0$. This is obtained through a high resolution limit, by computing $W_0^*(\mathbf{c}_0)$ for a much higher dimension and by properly rescaling and normalizing it such that it is interpreted as a Riemann approximation of a continuous density on $\boldsymbol{\mu}_0$. A practical example of this procedure, which might seem abstract at this stage, is given in Examples B and C. Moreover, a pseudo-code is provided in section 3.D. Given that, in general, $w_{\text{pseudo},0}$ is different from the GRO-optimal prior w_0^*, which in most cases remains unknown, the resulting variable

$$S_{\text{pseudo}} = \frac{\bar{P}_{\text{can},1}(\mathbf{x})}{P_{\text{can},0}^{w_{\text{pseudo},0}}(\mathbf{x})} \tag{3.43}$$

is not an e-variable, unless $w_{\text{pseudo},0} \equiv w_0^*$. Indeed, from Theorem 1 of [24], $S_{\text{can}}^{\text{GRO}}$ is the only e-variable of that form. Moreover, from (3.8), it holds:

$$D_{\text{KL}}(\bar{P}_{\text{can},1} \| P_{\text{can},0}^{w_0^*}) \leqslant D_{\text{KL}}(\bar{P}_{\text{can},1} \| P_{\text{can},0}^{w_{\text{pseudo},0}}) \tag{3.44}$$

or, equivalently:

$$\mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{can}}^{\text{GRO}}\right] \leqslant \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{pseudo}}\right]. \tag{3.45}$$

Consequently, one has

$$\mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{mic}}^{\text{GRO}}\right] \leqslant \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{can}}^{\text{GRO}}\right] \leqslant \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{pseudo}}\right],$$

i.e., the two approximations provide an upper and a lower bound for the e-power of the canonical GRO e-variable. In summary, when the canonical GRO e-variable is not available, we can follow a two-step procedure:

1. We build the corresponding microcanonical approximation, knowing that it is a valid candidate e-variable. To build it, we first transform the canonical universal distribution into a microcanonical one, by finding $W_{\text{can},1}$ as in (3.38). Then, we compute $S_{\text{mic}}^{\text{GRO}}$ according to the formulas expressed in the previous section (Equations (3.28) and (3.31)).

2. The goodness of the microcanonical approximation can be evaluated by looking at the width of the interval

$$r = \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{pseudo}}\right] - \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{mic}}^{\text{GRO}}\right] \geqslant 0 \tag{3.46}$$

where, for future reference, it is useful to note that, using definitions (3.43) and (3.40) and sufficiency, we can rewrite

$$\begin{aligned} r &= \mathbb{E}_{\bar{P}_{\mathrm{can},1}}\left[\log P^{W_0^*}_{\mathrm{mic},0}(\mathbf{x}) - \log P^{w_{\mathrm{pseudo},0}}_{\mathrm{can},0}(\mathbf{x})\right] \\ &= \mathbb{E}_{\bar{P}_{\mathrm{can},1}}\left[\log W_0^*(\mathbf{c}_0(\boldsymbol{x})) - \log W_{\mathrm{pseudo},0}(\mathbf{c}_0(\mathbf{x}))\right] \end{aligned} \tag{3.47}$$

where $W_{\mathrm{pseudo},0}(\mathbf{c}_0(\mathbf{x}))$ is defined as in (3.38).

The evaluation above is under $\bar{P}_{\mathrm{can},1}$-expectation; this makes sense if we use a Bayesian universal distribution $\bar{P}_{\mathrm{can},1} = P^{w_1}_{\mathrm{can},1}$ and the prior w_1 is a reasonable expression of our uncertainty. If we are not so sure about our priors, or if $\bar{P}_{\mathrm{can},1}$ is non-Bayesian (see 3.F), we may be interested in a more stringent, worst-case measure for evaluating the performance of the microcanonical approximation. In analogy with the worst-case REG defined in 3.15, we define an alternative version of r, denoted by r', which can be defined both relatively to a single parameter $\boldsymbol{\theta}_1$ (equivalently and more conveniently relative to $\boldsymbol{\mu}_1 = \boldsymbol{\mu}(\boldsymbol{\theta}_1)$:

$$\begin{aligned} r'(\boldsymbol{\mu}_1) &= \mathbb{E}_{\boldsymbol{\mu}_1}\left[\log S_{\mathrm{pseudo}} - \log S^{\mathrm{GRO}}_{\mathrm{mic}}\right] \\ &= \mathbb{E}_{\boldsymbol{\mu}_1}\left[\log W_0^*(\mathbf{c}_0(\boldsymbol{x})) - \log W_{\mathrm{pseudo},0}(\mathbf{c}_0(\mathbf{x}))\right], \end{aligned} \tag{3.48}$$

where $\mathbb{E}_{\boldsymbol{\mu}}$ denotes the expected value under $P_{\boldsymbol{\mu}}$, and in its worst-case version, which for clarity will be denoted by r':

$$r' := \max_{\boldsymbol{\mu}_1 \in \mathtt{M}_1} r'(\boldsymbol{\mu}_1). \tag{3.49}$$

It can be easily argued that

$$r \geqslant 0 \Rightarrow r' \geqslant 0. \tag{3.50}$$

In case r (or r') is small, we know that our easily computable microcanonical e-variable $S^{\mathrm{GRO}}_{\mathrm{mic}}$ is close to optimal according to the GRO criterion for the canonical problem, and hence can be used instead of the canonical $S^{\mathrm{GRO}}_{\mathrm{can}}$. In the following example, which is a continuation of Example A, we show a practical case where this turns out to be true.

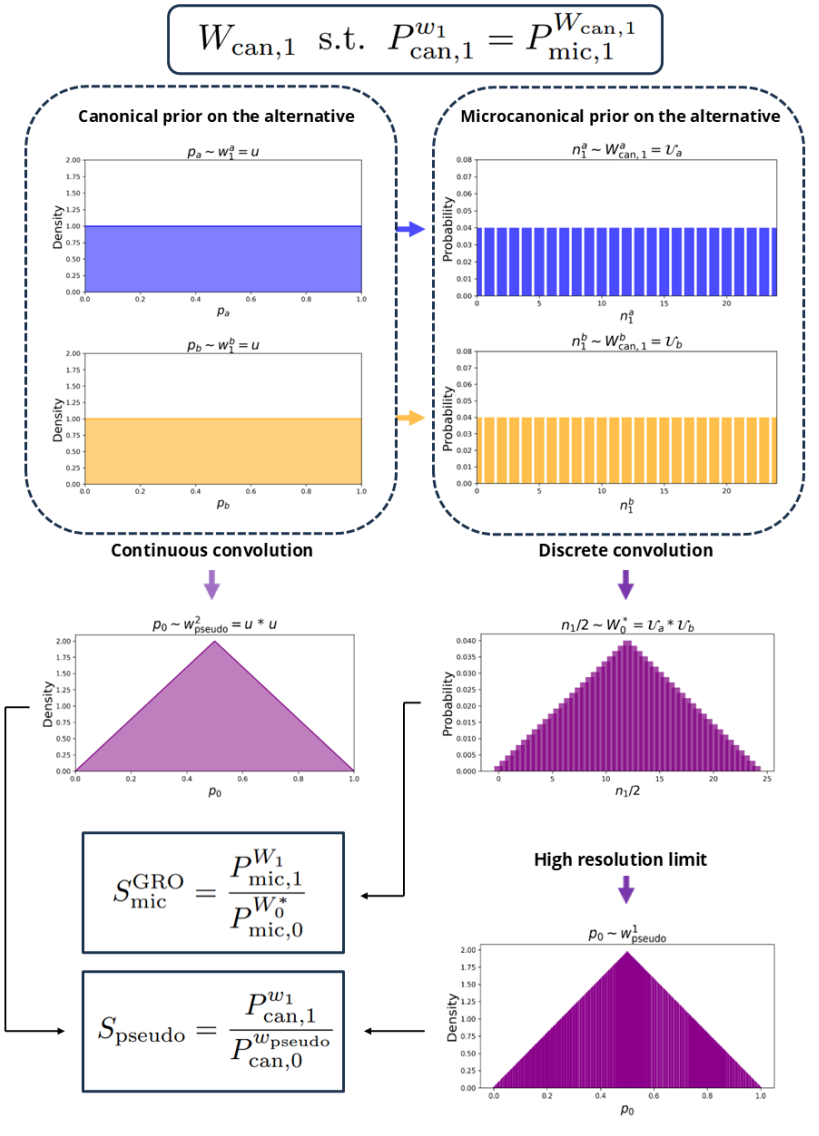


Figure 3.3: Procedures to compute the microcanonical approximation S^{GRO}_{mic} and the pseudo approximation S_{pseudo} for testing between two binary data streams, under uniform priors (as in Examples A, B, and C). Starting from two independent continuous uniform priors on the alternative (top left), we construct discrete microcanonical priors (top right) satisfying $P^{w_1}_{can,1} = P^{W_{can,1}}_{mic,1}$. In the specific case of continuous uniform priors, these discrete priors are also uniform. The optimal discrete prior W^*_0, used in S^{GRO}_{mic}, is obtained by convolving the alternative priors. The continuous prior $w_{pseudo,0}$ for S_{pseudo} is derived either from W^*_0 through a high resolution limit ($w^1_{pseudo,0}$), or by directly convolving the original continuous priors ($w^2_{pseudo,0}$).

Example B (continued from Example A)

We consider the same setting as in Example A, but in this case, we are interested in constructing a canonical test. In a canonical formulation, the observed number of 1s is fixed only in expectation. As a result, the null model is a collection of n i.i.d. Bernoulli variables, where the parameter is the probability $p_0 \in [0,1]$ of observing $x = 1$, which is the same regardless of the group. The alternative model, instead, assumes that data in the two groups are independent Bernoulli variables, where the parameters are the probabilities $(p_a, p_b) \in [0,1]^2$ of observing $x = 1$, which depend on the group. The tests aim to assess whether p_a and p_b are the same or whether they are different. Again, for the sake of this example, we put independent, continuous uniform priors on the alternative parameters p_a and p_b, $w_1(p_a, p_b) = u(p_a)u(p_b)$ where $u(p) = 1$ if $p \in [0,1]$ and $u(p) = 0$ else. In this simple case, the Bayesian marginal likelihood can be computed analytically, and it reads:

$$\begin{aligned} P_{\text{can},1}^{w_1} &= \int_0^1 p_a^{n_1^a}(1-p_a)^{n^a-n_1^a}dp_a \int_0^1 p_b^{n_1^b}(1-p_b)^{n^b-n_1^b}dp_b \\ &= \binom{n^a}{n_1^a}^{-1}\frac{1}{n^a+1}\ \binom{n^b}{n_1^b}^{-1}\frac{1}{n^b+1}. \end{aligned} \tag{3.51}$$

Following the procedure described in this section, we first build the microcanonical approximation. To do so, we need to compute the probability $W_{\text{can},1}$ induced by $P_1^{w_1}$ on the alternative sufficient statistics, such that $P_{\text{can},1}^{w_1} = P_{\text{mic},1}^{W_{\text{can},1}}$. By inspecting Eq. (3.51), it is easy to observe that $W_{\text{can},1}$ is the uniform distribution: $W_{\text{can},1} = (n_a+1)^{-1}(n_b+1)^{-1} = \mathcal{U}_a(n_1^a)\,\mathcal{U}_b(n_1^b)$. Thus, we can compute the microcanonical approximation $S_{\text{mic}}^{\text{GRO}}$ by using the results of Example A. As a second step, we check whether this microcanonical e-variable is a good approximation by studying the behavior of the interval width r as the total size n increases. To evaluate S_{pseudo}, we compute the prior $w_{\text{pseudo},0}$ as described above (denoted by $w_{\text{pseudo},0}^1$ in Figure 3.3 to be distinct from $w_{\text{pseudo},0}^2$ of the following Example C). A schematic representation of how $S_{\text{mic}}^{\text{GRO}}$ and S^{pseudo} are built is shown in Figure 3.3.

Once $S_{\text{mic}}^{\text{GRO}}$ and S^{pseudo} are computed, we can compute the gap between their e-power, r, which contains the optimal e-power: the smaller r, the better the microcanonical approximation works. In our simple example, r is already very small for small sample sizes and converges very rapidly to 0 as the sample size increases, as shown in Figure 3.4.

3.3.3 Asymptotic justification for the microcanonical approximation

We now provide a theoretical result that explains why the microcanonical approximation tends to perform very well in practice, even when the canonical GRO e-variable is not available. Specifically, it suggests that in many cases the gap r defined in Equation (3.46) converges very fast to 0 as the sample size increases.

First, let $\mathcal{M}$ be a canonical maximum entropy model with sufficient statistic $\mathbf{c}(\mathbf{x})$ taking values in a finite set $\mathcal{C} \subset \mathbb{R}^d$. The canonical distribution has an exponential

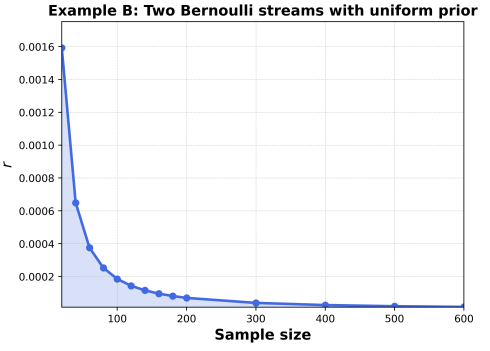


Figure 3.4: Convergence of the interval width r for a canonical test between two streams of binary data (as in Example B), for $n^a = n^b = m$, as the sample size $n = 2m$ grows.

form

$$P_{\text{can}}(\mathbf{x};\boldsymbol{\theta}) = \frac{e^{-\boldsymbol{\theta}\cdot\mathbf{c}(\mathbf{x})}}{Z(\boldsymbol{\theta})}, \tag{3.52}$$

and induces the mean-value mapping

$$\boldsymbol{\mu}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{c}(\mathbf{x})], \tag{3.53}$$

with $\boldsymbol{\mu}$ taking values in the mean-value parameter space $\mathtt{M}$.

Next, consider the i.i.d. extension $\mathcal{M}^{(m)}$ in which $\mathbf{y}^{(m)} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ are m i.i.d. samples from $P_{\text{can}}(\cdot;\boldsymbol{\theta})$. The sufficient statistic of $\mathbf{y}^{(m)}$ is the sum

$$\mathbf{s}^{(m)}(\mathbf{y}^{(m)}) = \sum_{j=1}^{m} \mathbf{c}(\mathbf{x}_j). \tag{3.54}$$

Let w be a prior density over $\mathtt{M}$, and let Q^w be the induced probability for any measurable $\mathtt{M}' \subseteq \mathtt{M}$:

$$Q^w(\boldsymbol{\mu} \in \mathtt{M}') := \int_{\mathtt{M}'} w(\boldsymbol{\mu})\, d\boldsymbol{\mu}. \tag{3.55}$$

The following theorem shows that the normalized sufficient statistic converges in distribution to Q^w. Convergence is quantified in terms of probabilities of subsets, with error decaying at rate $O(\log m/m)$.

Theorem 3.3.1.
Let w be any regular prior density on the mean value parameter space $\mathtt{M} \subset \mathbb{R}^{\mathtt{d}}$. Then, for any INECCSI (Definition 3.2.1) subset $\mathtt{M}'$ of $\mathtt{M}$, we have

$$\left| P_{\text{can}}^{w^{(m)}} \left\{ \frac{\mathbf{s}^{(m)}(\mathbf{y}^{(m)})}{m} \in \mathtt{M}' \right\} - Q^w \left\{ \boldsymbol{\mu} \in \mathtt{M}' \right\} \right| = O\left(\frac{\log m}{m} \right). \tag{3.56}$$

In words: under the Bayesian marginal likelihood $P_{\text{can}}^{w^{(m)}}$, the normalized sufficient statistic $\mathbf{s}^{(m)}/m$ becomes increasingly close to being distributed according to the prior over mean-value parameters.

Assume we can extend both canonical models $\mathcal{M}_0$ and $\mathcal{M}_1$ to i.i.d. models $\mathcal{M}_0^{(m)}$ and $\mathcal{M}_1^{(m)}$ as above, where (3.54) holds for both $\mathbf{c} = \mathbf{c}_0$ (sum denoted by $\mathbf{s}_0$) and $\mathbf{c} = \mathbf{c}_1$ (sum denoted by $\mathbf{s}_1$). In such settings, Theorem 3.3.1 supports the claim that the gap r between the microcanonical and pseudo approximations vanishes for large m. This is best explained in terms of our running example.

Example C (continued from Examples A and B)

Suppose $n^a = n^b = m$, so that data can be grouped into m i.i.d. *blocks*, each consisting of one binary outcome from group a and one from b. For each m, the sufficient statistics are:

- $\mathbf{s}_1^{(m)} = (n_1^{a\,(m)}, n_1^{b\,(m)})$: number of ones in each group under the alternative;
- $\mathrm{s}_0^{(m)} = \frac{1}{2}\big(n_1^{a\,(m)} + n_1^{b\,(m)}\big)$: average number of ones across both groups under the null. The division by 2 is required to ensure that, for a single outcome, $\mathtt{M}_0 = [0, 1]$, and can be interpreted, intuitively, as a set of probabilities.

After normalization by m, $\mathbf{s}_1^{(m)}/m \in \mathtt{M}_1 = [0, 1]^2$ and $\mathrm{s}_0^{(m)}/m \in \mathtt{M}_0 = [0, 1]$. Notice that every discrete distribution $W_j^{(m)}$ on the sufficient statistics of $\mathcal{M}_j^{(m)}$ induces a discrete distribution $V_j^{(m)}$ on the normalized sufficient statistics: $P_{\mathrm{can},1}^{w_1(m)}$ induces a probability $W_{\mathrm{can},1}^{(m)}$ on the alternative sufficient statistics $\mathbf{s}_1$, and a corresponding one, denoted here by $V_1^{(m)}$, on the normalized alternative sufficient statistics $\mathbf{s}_1/m$. Similarly, the microcanonical optimal prior $W_0^{*(m)}$ on the null sufficient statistic s_0 induces a distribution V_0^* on s_0/m.

In Example B, we used a prior w_1 under which p_a and p_b were independently and uniformly distributed, i.e., $w_1(p_a, p_b) = w_1^a(p_a) \cdot w_1^b(p_b)$ with $w_1^a = w_1^b = u$. The independent uniform prior has a remarkable property: $V_1^{(m)}$ coincides *exactly* with the product of two independent discrete uniforms, each defined on $\{0, 1/m, \ldots, 1\}$, corresponding to the components of $\mathbf{s}_1^{(m)}/m$.

Consequently, the distribution $V_0^{*(m)}$ is *exactly* equal to a triangular discrete distribution, which is the convolution of these two discrete uniforms (Figure 3.2). Theorem 3.3.1 indicates that something analogous, but now in an asymptotic sense, will happen for every regular prior $w_1(p_a, p_b) = w_1^a(p_a)w_1^b(p_b)$, as long as p_a and p_b are still independent under w_1. More in detail, even if w_1^a and/or w_1^b are not uniform, $V_1^{(m)}$ will converge to a distribution on $\mathtt{M}_1$ that is a discretized version of w_1, and $V_0^{*(m)}$ will still be the exact convolution of the two components of $V_1^{(m)}$, which are the (approximate) discretized versions of the components of w_1. To illustrate, in Example 1 below, w_1^a and w_1^b will be taken to be of general beta form rather than restricted to uniform, and then we will see Theorem 3.3.1 in action, the correspondence becoming asymptotic rather than precise at each m. Still, $V_0^{*(m)}$ will converge, as m grows, to a continuous, strictly positive density on $\mathtt{M}_0$, denoted by $w_{\mathrm{pseudo},0}^1$. This distribution can be approximated by considering a very large m and "smoothening" the corresponding $V_0^{*(m)}$ to $w_{\mathrm{pseudo},0}^1$. This limiting procedure is precisely what we referred to

earlier as the *high resolution limit*. Once $w^1_{\text{pseudo},0}$ is obtained, $P_{\text{can}}^{w^1_{\text{pseudo},0}(m)}$ induces a probability distribution $W^{1(m)}_{\text{pseudo},0}$ on the null sufficient statistics. As above, we can define

$$V^{1(m)}_{\text{pseudo},0}\left(\frac{\mathbf{s}_0^{(m)}}{m}\right) := W^{1(m)}_{\text{pseudo},0}(\mathbf{s}_0^{(m)}). \tag{3.57}$$

Now we invoke Theorem 3.3.1 again: it indicates that $V^{1(m)}_{\text{pseudo},0}$ converges to $w^1_{\text{pseudo},0}$. Thus, one may expect $V_0^{*(m)}$ and $V^{1(m)}_{\text{pseudo},0}$, and consequently $W_0^{*(m)}$ and $W^{1(m)}_{\text{pseudo},0}$, to be close and r to be small, according to Eq. (3.47). The microcanonical e-variable becomes thus an excellent approximation of the canonical one — which is what we set out to argue.

In the current example, we can go further: let $w^2_{\text{pseudo},0}$ be the continuous convolution of the independent priors w_1^a and w_1^b. Applying Theorem 3.3.1 to $\mathcal{M}_0$ with this density shows that the induced distribution $V^{2(m)}_{\text{pseudo},0}$ on $\mathbf{s}_0^{(m)}/m$ converges to a discretized version of $w^2_{\text{pseudo},0}$. At the same time, $V_0^{*(m)}$ is constructed by first discretizing the prior w_1 and then computing the discrete convolution of its two components. In the high-resolution limit $m \to \infty$, this procedure yields the discretized version of the *continuous* convolution of w_1. Thus, in the large-m limit, $w^1_{\text{pseudo},0}$ and $w^2_{\text{pseudo},0}$ become indistinguishable. In practice, one can compute S_{pseudo} either by directly convolving the continuous components of w_1, or by taking the discrete convolution $W_0^{*(m)}$ and then its high-resolution limit: both approaches yield the same result (Figure 3.3).

How precise and general is this? In the reasoning above, we invoked Theorem 3.3.1 several times to go back and forth between prior distributions on mean-value parameters and marginal distributions on sufficient statistics. Specifically: (a) at the level of $\mathcal{M}_1$ (blue and yellow arrows in Figure 3.3); and (b) at the level of $\mathcal{M}_0$, for relating $W_0^{*(m)}$ to $P_{\text{can}}^{w^1_{\text{pseudo},0}}$ (b1, bottom right arrow in Figure 3.3) and $P_{\text{can}}^{w^2_{\text{pseudo},0}}$ (b2, bottom left arrow).

In step (a), the theorem is not really needed when w_1^a and w_1^b are uniform (as in the figure). Nevertheless, as long as the priors remain independent and regular, Theorem 3.3.1 suggests that step (a) holds even if they are not uniform. More generally, moving from the binary 2-group case to a general MEM $\mathcal{M}_1$, Theorem 3.3.1 still suggests that step (a) is valid whenever w_1 factorizes into independent regular priors, making the mean-value parameters independent. We write "suggest" rather than "prove" because the convergence in (3.56) is too weak to formally imply $r \to 0$ (it concerns probabilities of sets, whereas (3.47) involves expectations of log densities). Nevertheless, it provides strong heuristic evidence, and we do observe convergence numerically (see Figure 3.4). All reasoning based on Theorem 3.3.1 should thus be understood as heuristic rather than fully formal.

Turning now to step (b) for general $\mathcal{M}_0$ and $\mathcal{M}_1$: as long as $\mathbf{s}_0^{(m)}$ is a linear function of $\mathbf{s}_1^{(m)}$, the use of Theorem 3.3.1 in steps (b1) and (b2) remains heuristically justified, provided w_1 factorizes into regular independent priors as above. This linearity condition holds in all our examples (e.g. in Example C, $\mathbf{s}_0^{(m)}$ is the average of

the components of $\mathbf{s}_1^{(m)}$). It guarantees that the limiting density $w^1_{\text{pseudo},0}$ exists, and Theorem 3.3.1 then suggests that it coincides with $w^2_{\text{pseudo},0}$.

In the more general case where $\mathbf{s}_0^{(m)}$ is a function (not necessarily linear) of $\mathbf{s}_1^{(m)}$ — i.e. Condition A holds — then it may still be true that $V_0^{*(m)}$ converges to a high-resolution limiting density $w^1_{\text{pseudo},0}$. In that case, Theorem 3.3.1 still suggests that step (b1) remains valid, so that r becomes small with growing sample size, making the microcanonical approximation effective. However, in such settings, it is less clear whether the approach based on $w^2_{\text{pseudo},0}$ still makes sense.

We stress that this asymptotic justification does not rely on $\bar{P}_1$ being a Bayesian mixture with prior w_1. The construction leading to $w^1_{\text{pseudo},0}$ applies to any universal distribution $\bar{P}_1$ on the alternative, since $\bar{P}_1$ always induces a discrete distribution on the alternative sufficient statistic; from this, one can derive $V_0^{*(m)}$ and then obtain $w^1_{\text{pseudo},0}$ via the high-resolution limit. By contrast, the alternative route based on $w^2_{\text{pseudo},0}$ explicitly requires a factorized regular prior on the alternative.

3.4 Application to contingency tables and related models

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.

Formally, a contingency table is an $l \times k$ matrix where each entry represents the count of occurrences for a particular combination of row and column categories. Such tables are widely used in fields where categorical data naturally arise, such as biostatistics, social sciences, and market analysis.

In network science, this approach plays a crucial role in link analysis, where the presence or absence of an edge ($x = 1$ or $x = 0$) in a network is studied across different subsets of nodes. For instance, in community detection, one may ask whether the probability of forming a link differs within and between predefined groups of nodes. This idea is closely related to the Stochastic Block Model (SBM), a generative model in which nodes are assigned to latent groups, and connection probabilities are determined by group memberships. Contingency tables provide a natural way to summarize and test the differences in connection probabilities across groups, helping to assess whether observed patterns deviate from a null model where edges are formed independently of group structure. See e.g., [77, 78] for connections between network modeling and contingency tables, and the discussion in 3.4.3 of this paper.

In this work, we focus on binary categorical data, which corresponds to $l = 2$ in the general $l \times k$ contingency tables setting. We first apply our results to the simple case of two groups, i.e., 2×2 contingency tables. We consider microcanonical and canonical tests. For canonical tests, our main focus will be that of finding the microcanonical approximation in practical cases; this translates into finding the induced prior on the alternative (3.37) and then applying formula (3.40). We will finally verify the approximation validity by evaluating the interval width r, and show results on the regret. Later, we extend these results to the more general case of $2 \times k$ contingency

tables.

3.4.1 2×2 contingency tables

A 2×2 contingency table is a fundamental tool to assess whether the distribution of a binary outcome differs between two groups. Given a dataset where each observation consists of a binary variable $x \in \{0, 1\}$ and a categorical label indicating group membership, the data can be summarized in the following 2×2 table:

	Group A	Group B	Total
$x = 1$	n_1^a	n_1^b	n_1
$x = 0$	n_0^a	n_0^b	n_0
Total	n^a	n^b	n

The dataset $\mathbf{x}$ consists of two groups, represented as $\mathbf{x}_a = (x_1^a, ..., x_{n^a}^a)$ and $\mathbf{x}_b = (x_1^b, ..., x_{n^b}^b)$, where n^a and n^b are the respective group sizes. The table reports the number of ones (n_1^a and n_1^b) and zeros (n_0^a and n_0^b) in each group, along with their totals, n_1 and n_0. The key question is whether the probability of observing $x = 1$ differs between the two groups. This problem translates into a hypothesis testing problem, where:

- In the alternative hypothesis, the two groups are distinct, meaning the number of ones is constrained separately in each group:

$$\mathbf{c}_1 = (n_1^a, n_1^b). \tag{3.58}$$

- In the null hypothesis, the groups are indistinguishable, so only the total number of ones is constrained:

$$\mathrm{c}_0 = n_1. \tag{3.59}$$

These constraints define the sufficient statistics under each hypothesis and form the basis for the microcanonical and canonical tests discussed next. The reader may have noticed that this is exactly the setting of Examples A, B, and C in section 3.3. Nevertheless, for the sake of clarity, in this section all quantities will be defined again and in more detail, at the cost of repeating ourselves.

2×2 microcanonical test

In the microcanonical formulation, we enforce hard constraints on the observed counts, treating them as fixed quantities. The null model with sufficient statistics n_1 reads

$$P_{\text{mic, }0}(\mathbf{x}; n_1) = \begin{cases} \frac{1}{\Omega_0(n_1)}, & \text{if } n_1(\mathbf{x}) = n_1; \\ 0, & \text{else;} \end{cases} \tag{3.60}$$

where

$$\Omega_0(n_1) = \binom{n}{n_1} \tag{3.61}$$

is the number of permutations of $\mathbf{x}$ preserving the total number of 1s. The alternative model with sufficient statistics (n_1^a, n_1^b) reads

$$P_{\text{mic, 1}}(\mathbf{x}; n_1^a, n_1^b) = \begin{cases} \frac{1}{\Omega_1(n_1^a, n_1^b)}, & \text{if } (n_1^a(\mathbf{x}), n_1^b(\mathbf{x})) \\ & \qquad = (n_1^a, n_1^b); \\ 0, & \text{else,} \end{cases} \tag{3.62}$$

where

$$\Omega_1(n_1^a, n_1^b) = \binom{n^a}{n_1^a}\binom{n^b}{n_1^b} \tag{3.63}$$

is the number of permutations of $\mathbf{x}$ preserving the total number of 1s in each group.

For any given prior W_1 on the alternative sufficient statistics, $S_{\text{mic}}^{\text{GRO}}$ is found exactly by computing W_0^* and applying (3.31). In this case, Condition A (3.29) is satisfied, as the null sufficient statistics can be written as a function of the alternative one:

$$n_1 = n_1^a + n_1^b. \tag{3.64}$$

Thus, following (3.30), the optimal prior on the null is the distribution of n_0 induced by $W_1(n_1^a, n_1^b)$. If n_1^a and n_1^b are independently distributed:

$$W_1(n_1^a, n_1^b) = W_1^a(n_1^a) \cdot W_1^b(n_1^b), \tag{3.65}$$

then W_0^* is simply the convolution of W_1^a and W_1^b:

$$W_0^* = W_1^a * W_1^b, \tag{3.66}$$

where $f * g$ represents the convolution between functions f and g.

Notice that an example application of this formula for the case of two independent uniform priors has already been carried out in Example A of section 3.3.1.

$\mathbf{2 \times 2}$ canonical test

The null canonical model obtained by constraining the average number of 1s, i.e., the expected value of n_1, is represented by the exponential distribution

$$P_{\text{can}}(\mathbf{x}; \theta_0) = \frac{e^{-\theta_0 \cdot n_1(\mathbf{x})}}{(1 + e^{-\theta_0})^n}, \tag{3.67}$$

which can be rewritten in the mean-value parametrization:

$$P_{\text{can}}(\mathbf{x}; p_0) = p_0^{n_1(\mathbf{x})}(1 - p_0)^{n - n_1(\mathbf{x})} \tag{3.68}$$

upon defining

$$p_0 = \frac{e^{-\theta_0}}{1 + e^{-\theta_0}}. \tag{3.69}$$

The null model is the distribution of a collection of n i.i.d. Bernoulli variables, where the occurrence of $x = 1$ has the same probability p_0 regardless of the group.

The alternative model, obtained by constraining the expected values of n_1^a and n_1^b, reads:

$$P_{\text{can}}(\mathbf{x}; \theta_a, \theta_b) = \frac{e^{-\theta_a \cdot n_1^a(\mathbf{x}) - \theta_b \cdot n_1^b(\mathbf{x})}}{(1 + e^{-\theta_a})^{n^a}(1 + e^{-\theta_b})^{n^b}}, \tag{3.70}$$

or, equivalently,

$$P_{\text{can}}(\mathbf{x}; p_a, p_b) = p_a^{n_1^a(\mathbf{x})}(1 - p_a)^{n^a - n_1^a(\mathbf{x})} p_b^{n_1^b(\mathbf{x})}(1 - p_b)^{n^b - n_1^b(\mathbf{x})} \tag{3.71}$$

upon defining the mean-value parameters:

$$p_a = \frac{e^{-\theta_a}}{1 + e^{-\theta_a}} \quad \text{and} \quad p_b = \frac{e^{-\theta_b}}{1 + e^{-\theta_b}}. \tag{3.72}$$

The alternative model assumes that the data in groups A and B are independent Bernoulli variables, with different probabilities of $x = 1$ across groups. In this scenario, the aim of the test is to assess whether p_a and p_b are the same or whether they are different. In what follows, we explicitly apply the procedure described in section 3.3.2 to different choices of $\bar{P}_{\text{can},1}$.

Example 1: Independent beta priors. The beta probability distribution reads:

$$\text{Beta}(y; \alpha, \beta) = \frac{y^{\alpha-1}(1 - y)^{\beta-1}}{B(\alpha, \beta)}, \quad \text{for } y \in (0, 1), \tag{3.73}$$

where $B(\alpha, \beta)$ is the beta function, defined as

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1 - t)^{\beta-1} dt. \tag{3.74}$$

The beta prior represents a popular choice because it is flexible enough to encompass several cases of interest (Table 3.1 of the Supplementary Material). Here, we put two independent beta priors w_1^a and w_1^b with parameters, respectively, (α_a, β_a) and (α_b, β_b), on p_a and p_b. The Bayesian marginal likelihood resulting from this choice can be written explicitly as

$$\bar{P}_{\text{can},1}(\mathbf{x}) = \frac{B(\bar{\alpha}^a, \bar{\beta}^a)}{B(\alpha^a, \beta^a)} \cdot \frac{B(\bar{\alpha}^b, \bar{\beta}^b)}{B(\alpha^b, \beta^b)} \tag{3.75}$$

where

$$\begin{aligned} \bar{\alpha}^a &= n_1^a + \alpha^a, & \bar{\beta}^a &= n^a - n_1^a + \beta^a, \\ \bar{\alpha}^b &= n_1^b + \alpha^b, & \bar{\beta}^b &= n^b - n_1^b + \beta^b. \end{aligned}$$

As in the previous example, to obtain the microcanonical approximation for this problem, we look for the probability mass function induced by $\bar{P}_{\text{can},1}$ on the alternative sufficient statistics, which reads:

$$W_{\text{can},1}(n_1^a, n_1^b) = W_{\text{can},1}^a(n_1^a) \cdot W_{\text{can},1}^b(n_1^b) \tag{3.76}$$

with

$$W^a_{\text{can},1}(n^a_1) = \Omega^a_1(n^a_1) \cdot \frac{B(\bar{\alpha}^a, \bar{\beta}^a)}{B(\alpha^a, \beta^a)} \tag{3.77}$$

and

$$W^b_{\text{can},1}(n^b_1) = \Omega^b_1(n^b_1) \cdot \frac{B(\bar{\alpha}^b, \bar{\beta}^b)}{B(\alpha^b, \beta^b)}. \tag{3.78}$$

With this choice, W^a_1 and W^b_1 are *beta-binomial distributions*. Given that n^a_1 and n^b_1 are independently distributed, as we expected because we put independent priors on p_a and p_b, $W^*_0(n_1)$ is the convolution of $W^a_{\text{can},1}(n^a_1)$ and $W^b_{\text{can},1}(n^b_1)$. Whether this expression can be written in closed form depends on the specific values of the beta parameters chosen. For example, if all beta parameters are equal to 1, W^*_0 reduces to the convolution between two discrete uniform distributions (3.34). When no closed form is available, the convolution can be computed numerically.

In Figure 3.7 we show W^a_1, W^b_1, W^*_0, $w^1_{\text{pseudo},0}$ and $w^2_{\text{pseudo},0}$ for different choices of the beta parameters. As expected, in all cases where w^a_1 and w^b_1 are well defined in the whole parameter space, $w^1_{\text{pseudo},0}$ and $w^2_{\text{pseudo},0}$ are almost indistinguishable. This is a consequence of Theorem 1: the distribution of the mean value parameter ($w^2_{\text{pseudo},0}$) and that of the sufficient statistic ($w^1_{\text{pseudo},0}$) resemble each other when the sample size is big (high resolution limit).

In the next section, we show results obtained by numerical simulations concerning the optimality of the microcanonical approximation and the regret, measured in the examples reported in this section. For simplicity, when necessary, we will assume that the two groups have the same sample size, i.e., $n^a = n^b = m$, and that the independent beta priors on the alternative, denoted by w^a_1 and w^b_1, have all parameters equal to a certain value $\gamma > 0$.

Evaluating the microcanonical approximation

In order to evaluate the goodness of the microcanonical approximation, we employ two approaches: a direct comparison and a comparison through r.

In the first case, we directly compare the e-power of the microcanonical approximation to the GRO-optimal canonical one, where the latter is computed by numerically solving the optimization problem (3.36). We find that the e-power of the microcanonical approximation converges to that of the canonical GRO e-variable as the total size grows (Figure 3.8). The e-power of the pseudo approximation converges as well, even though the convergence is slower compared to that of the microcanonical one. From these plots, we can already conclude that the microcanonical one works as a good approximation of $S^{\text{GRO}}_{\text{can}}$.

The numerical approach to directly compare the e-power is feasible in a few simple cases and only for relatively small sample sizes. Conversely, the value of r can be easily evaluated, even for very large system sizes. In Figure 3.9, we show the plot of r as defined earlier to evaluate the effectiveness of the microcanonical approximation in different scenarios. The results confirm those of Figure 3.8, as in all cases considered,

r converges to 0. In conclusion, we argue that the microcanonical approximation is a perfect candidate in this case.

Results on regret

Let's again consider the m-dimensional i.i.d. extension of our models. In section 3.E of the Supplementary Materials, we show that, if the error $r'(\boldsymbol{\mu}_1)$ in (3.48) vanishes as the sample size m increases, both the canonical growth-optimal e-variable $S^{\mathrm{GRO}}_{\mathrm{can}}$ and its microcanonical approximation $S^{\mathrm{GRO}}_{\mathrm{mic}}$ satisfy:

$$\mathrm{REG}_1(\boldsymbol{\mu}_1, \cdot) = \frac{d_1 - d_0}{2} \cdot \log m + O(1). \tag{3.79}$$

In the 2×2 case, where $d_1 = 2$ and $d_0 = 1$, this becomes:

$$\mathrm{REG}_1(\boldsymbol{\mu}_1; \cdot) = \frac{1}{2} \log m + O(1).$$

This result holds uniformly over all $\boldsymbol{\mu}_1 \in \mathtt{M}'_1$, provided that $\mathtt{M}'_1$ is an INECCSI set (i.e., excluding regions near the boundary of the parameter space). However, it does not extend to the full parameter space $\mathtt{M}$, where the asymptotic form (3.19) may fail to hold even in well-specified cases.

Our experiments confirm these insights. We evaluated worst-case regret in the 2×2 setting for different values of the beta prior parameter γ. Notice that in what follows we apply our reasoning to the mean value parameter spaces, $(p_a, p_b) \in \mathtt{M}_1 = [0,1]^2$ and $p_0 \in \mathtt{M}_0 = [0,1]$, and that we consider INECCSI sets with respect to $\mathtt{M}_1$. From experimental results, collected in Figure 3.5, we observe a clear dichotomy:

- For $\gamma < 1$, the convolution of the beta priors w_1^a and w_1^b is non-differentiable at $p_0 = 1/2$, as shown in Figure 3.7 (e.g., for $\gamma = 0.5$). Consequently, the convergence of $V_0^{*(m)}$ to a density over the mean-value space $\mathtt{M}_0$ (as discussed under Theorem 3.3.1) may be very slow or fail altogether. In this case, S^{pseudo} becomes incomparable to $S^{\mathrm{GRO}}_{\mathrm{can}}$ and $S^{\mathrm{GRO}}_{\mathrm{mic}}$, and (3.79) no longer holds (see Figure 3.10). Indeed, in Figure 3.5, we see that even on small INECCSI sets, the regret grows like $a \log m + b$ for some $a > 1/2$.

- For $\gamma = 1$, convergence is moderate. Although r' decays quickly (Figure 3.10), the experimental values of Figure 3.5 show areas (e.g., the yellow counter-diagonal) where regret exceeds the expected rate. These may still belong to an INECCSI set, but convergence has not yet been reached at the sample sizes considered ($m \leqslant 1800$).

- For $\gamma > 1$, the convolution is differentiable, and convergence of $V_0^{*(m)}$ is fast. The asymptotic behavior $(1/2) \log m + O(1)$ is observed on INECCSI sets (see again Figure 3.5)

These findings imply that, from a minimax perspective, using priors with $\gamma < 1$ is generally suboptimal. Such priors fail to achieve the expected regret rate of $(1/2) \log m + O(1)$ even when the true parameters lie well inside the parameter space.

This has implications for default prior choices. In both the Bayesian and MDL literatures, Jeffreys prior [16] is often recommended as a default when no prior knowledge is available, and is justified in the MDL framework because it achieves asymptotically minimax optimal redundancy (i.e., the middle inequality in (3.19) becomes an equality [79]). However, in our setting, Jeffreys prior corresponds to $\gamma = 1/2$, which, despite its MDL-optimality, is *not* optimal with respect to worst-case regret under e-values.

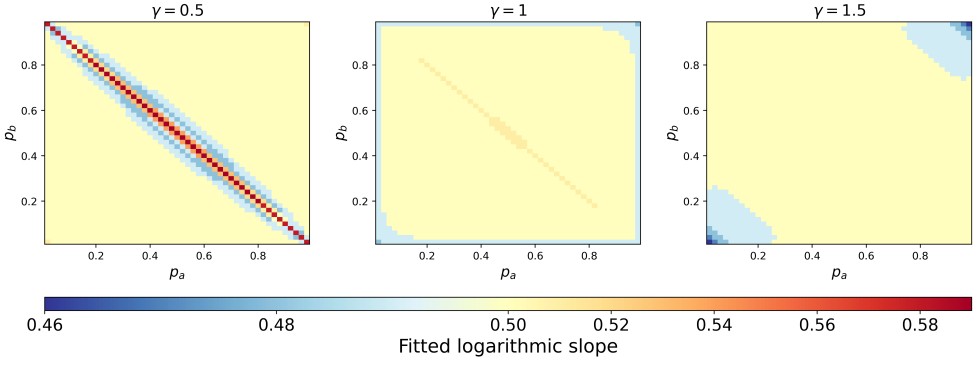


Figure 3.5: Fitted slope of the logarithmic growth $a \log m + b$ of the microcanonical approximation regret 3.14 in the 2×2 case ($n^a = n^b = m$), shown for different combinations of the alternative parameters (p_a, p_b). The expected asymptotic slope is 0.5 (yellow). Three alternative beta priors are considered: $\alpha = \beta = 0.5$, $\alpha = \beta = 1$ and $\alpha = \beta = 1.5$. The sample sizes used for fitting are $m \in \{600, 800, 1000, 1200, 1400, 1600, 1800\}$. p_a and p_b vary in the interval $[0.02, 0.98]$, with a grid step of 0.02. Values at the boundaries are excluded to improve the readability of the plots.

3.4.2 $2 \times k$ contingency tables

A $2 \times k$ contingency table is a natural extension of the 2×2 case, allowing us to assess whether the distribution of a binary outcome differs across multiple (k) groups. Given a dataset where each observation consists of a binary variable $x \in \{0, 1\}$ and a categorical label indicating group membership (among k different groups), the data can be summarized in the following $2 \times k$ table:

	Group 1	Group 2	$\cdots$	Group k	Total
$x = 1$	n_1^1	n_1^2	$\cdots$	n_1^k	n_1
$x = 0$	n_0^1	n_0^2	$\cdots$	n_0^k	n_0
Total	n^1	n^2	$\cdots$	n^k	n

The dataset consists of k groups, represented as $\mathbf{x}_i = (x_1^i, ..., x_{n^i}^i)$ for $i = 1, ..., k$,

where n^i denotes the size of group i. The table reports the number of ones (n_1^i) and zeros (n_0^i) in each group, along with their respective totals, n_1 and n_0.

The goal is to test whether all groups share the same probability of observing $x = 1$. This problem again translates into a hypothesis testing problem, where:

- Under the alternative hypothesis, the groups are allowed to have different probabilities, meaning the number of ones is constrained separately in each group:

$$\mathbf{c}_1 = (n_1^1, n_1^2, \ldots, n_1^k). \tag{3.80}$$

- Under the null hypothesis, all groups share the same probability, meaning only the total number of ones is constrained:

$$c_0 = n_1. \tag{3.81}$$

Rejection of the null, therefore, indicates a lack of homogeneity across groups, or equivalently, that at least two groups differ in their outcome distribution.

These constraints define the sufficient statistics under each hypothesis of the microcanonical and canonical tests discussed next. As the examples will illustrate, most of the results in this section naturally extend from the 2×2 case. The key distinction is that, in the latter case, the only relevant asymptotic behavior is as the total sample size n grows large. In contrast, in the present setting, both n and k can grow large, with different scenarios arising depending on the application (see 3.4.3). In all cases, the asymptotic behavior of e-variables plays a crucial role, particularly in the canonical test, where only asymptotic approximations are available, and we need to assess whether the microcanonical approximation can be used.

$2 \times k$ microcanonical test

While the null model stays the same (Eq. (3.60), the alternative model is simply the extension of (3.62) from 2 to k groups:

$$P_{\text{mic, }1}(\mathbf{x}; \{n_1^i\}) = \begin{cases} \frac{1}{\Omega_1(\{n_1^i\}),} & \text{if } (n_1^1(\mathbf{x}), ..., n_1^k(\mathbf{x})) \\ & \quad = (n_1^1, ..., n_1^k), \\ 0, & \text{else,} \end{cases} \tag{3.82}$$

where

$$\Omega_1(\{n_1^i\}) = \prod_{i=1}^{k} \binom{n^i}{n_1^i}. \tag{3.83}$$

As in the 2×2 case, Condition A (3.29) is satisfied:

$$n_1 = \sum_{i=1}^{k} n_1^i \tag{3.84}$$

and the optimal prior on the null is the marginal distribution of n_0 induced by $W_1(\{n_1^i\})$. If all n_1^i are independently distributed:

$$W_1(\{n_1^i\}) = \prod_{i=1}^{k} W_1^i(n_1^i), \tag{3.85}$$

then W_0^* is simply the convolution of the individual alternative priors:

$$W_0^* = W_1^1 * ... * W_1^k. \tag{3.86}$$

Interestingly, when the number of groups k is large, and the priors are regular enough, a Central Limit Theorem holds; thus, W_0^* is well approximated by a *discrete Gaussian distribution*, i.e., if $k \gg 1$:

$$W_0^*(n_1) \approx \frac{1}{N(\mu_k, \sigma_k)} \exp\left(-\frac{(n_1 - \mu_k)^2}{2\sigma_k^2}\right) \tag{3.87}$$

where N is the normalization constant and

$$\mu_k = \sum_{i=1}^{k} \mathbb{E}_{W_1^i}[n_1^i]$$
$$\sigma_k^2 = \sum_{i=1}^{k} \mathrm{Var}_{W_1^i}(n_1^i).$$

This result is particularly convenient: when k is big enough, the only effect of the choice of priors on the alternative, as long as they are independent and regular enough, is in determining the average and the variance of the optimal (approximated) Gaussian prior on the null.

Example 2: Independent uniform priors. Here, we extend the computations of Example A to the case of k groups. When a uniform discrete prior $\mathcal{U}$ is put on each parameter of the alternative:

$$W_1(\{n_1^i\}) = \prod_{i=1}^{k} \mathcal{U}_i(n_1^i) = \prod_{i=1}^{k} \frac{1}{n^i + 1}, \tag{3.88}$$

the GRO null prior is again the convolution of all the individual priors, i.e., the convolution of k discrete uniform distributions, which reads [80]:

$$W_0^*(n_1) = \sum_{S \subseteq \{1, \dots k\}} (-1)^{|S|} \binom{n_1 + k - 1 - \sum_{j \in S}(n^j - n)}{n_1 - 1} \left[\prod_{i=1}^{k} n^i + 1\right]^{-1}, \tag{3.89}$$

where the sum runs over all possible subsets of $\{1, \dots k\}$ and $\mid S \mid$ is the number of elements of set S. In the formula, the first factor stands for the number of ways in which a set of k non-negative numbers ($\{n_1^i\}$) can be chosen uniformly such that their sum is equal to n_1, with the constraint that for each i, n_1^i must be smaller than

or equal to n^i. The second factor represents a normalization constant. If all n^i are equal to a certain value m, the formula simplifies and reads:

$$W_0^*(n_1) = \sum_{j=1}^{\lfloor n_1/(m+1)\rfloor} (-1)^j \binom{n}{j} \binom{n_1 - j(m+1) + k - 1}{k-1} \left[\prod_{i=1}^{k} n^i + 1\right]^{-1}, \quad (3.90)$$

where $\lfloor x \rfloor$ is the floor function of x. This is the formula used to generate Figure 3.6, where we show W_0^*, along with its Gaussian approximation, for increasing values of k.

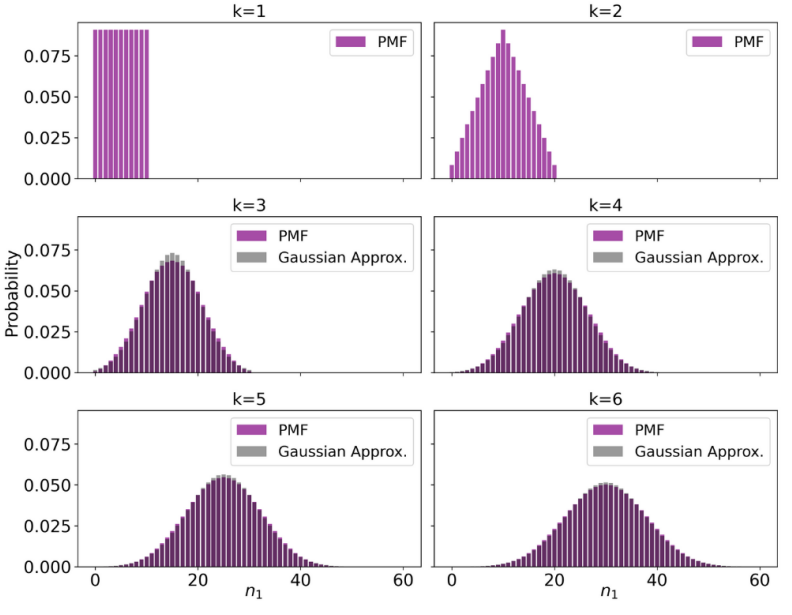


Figure 3.6: The microcanonical GRO-optimal prior on the null W_0^* for testing $2 \times k$ tables is obtained as the convolution of the k independent priors on the alternative, which are discrete uniform priors in the case shown in this picture. Each convolution, for $k > 2$, is superposed to its discrete Gaussian approximation.

$2 \times k$ canonical test

The null canonical model is the same as in the 2×2 case, Eq. (3.67), i.e., a collection of n i.i.d. Bernoulli trials, where the probability of observing $x = 1$ is the same across all groups. The alternative model extends Eq. (3.70) and (3.71) to the case of k groups, by constraining the expected values of n_1^i separately for each group i, leading to the expression:

$$P_{\text{can}}(\mathbf{x}; \boldsymbol{\theta}) = \frac{e^{-\sum_{i=1}^{k} \theta_i n_1^i(\mathbf{x})}}{\prod_{i=1}^{k} (1 + e^{-\theta_i})^{n^i}}, \quad (3.91)$$

or, equivalently, in the mean-value parametrization:

$$P_{\text{can}}(\mathbf{x};\boldsymbol{p}) = \prod_{i=1}^{k} p_i^{n_1^i(\mathbf{x})}(1-p_i)^{n^i - n_1^i(\mathbf{x})}, \tag{3.92}$$

where we define the group-specific probabilities as:

$$p_i = \frac{e^{-\theta_i}}{1+e^{-\theta_i}}, \quad \text{for each } i \in \{1,\dots,k\}. \tag{3.93}$$

In this formulation, the alternative model assumes that the data in each group are independent Bernoulli variables, with the probability of $x = 1$ depending on the group. The goal of the hypothesis test is to determine whether these probabilities are equal across all groups ($p_1 = p_2 = \cdots = p_k$) or whether they differ, indicating that the probability of observing $x = 1$ is group-dependent.

In the following sections, we apply the procedure described in section 3.3.2 to different choices of $\bar{P}_{\text{can},1}$.

Example 3: Independent beta priors. Here, we extend Example 1 to the case of k groups. We assume that each p_i is independently distributed according to a beta prior with parameters (α^i, β^i). The Bayesian marginal likelihood reads:

$$\bar{P}_{\text{can},1}(\mathbf{x}) = \prod_{i=1}^{k} \frac{B(\bar{\alpha}^i, \bar{\beta}^i)}{B(\alpha^i, \beta^i)} \tag{3.94}$$

where

$$\begin{aligned} \bar{\alpha}^i &= n_1^i + \alpha^i \\ \bar{\beta}^i &= n^i - n_1^i + \beta^i \quad \text{for each } i \in \{1,\dots,k\}. \end{aligned}$$

To derive the microcanonical approximation, we compute the probability mass function induced by $\bar{P}_{\text{can},1}(\mathbf{x})$ on $\{n_1^i\}$, which reads:

$$W_{\text{can},1}(\{n_1^i\}) = \prod_{i=1}^{k} W_{\text{can},1}^i(n_1^i) \tag{3.95}$$

with

$$W_{\text{can},1}^i(n_1^i) = \Omega_1^i(n_1^i) \cdot \frac{B(\bar{\alpha}^i, \bar{\beta}^i)}{B(\alpha^i, \beta^i)} \quad \text{for each } i \in \{1,\dots,k\}. \tag{3.96}$$

$W_0^*(n_1)$ is, then, the convolution of all $W_{\text{can},1}^i(n_1^i)$. If all beta parameters are equal to 1, W_0^* reduces to the convolution between k discrete uniform distributions (3.89). When the beta parameters are such that no closed form is available, the convolution must be computed numerically. Alternatively, if k is big enough, one can resort to the discrete Gaussian approximation (3.87). Analogously, a continuous Gaussian approximation can be used to approximate w_{pseudo}. Figure (3.11), we show $W_{\text{can},1}^i$, W_0^*, and w_{pseudo} for $k = 10$.

Evaluating the microcanonical approximation

To assess the effectiveness of the microcanonical approximation, we study the behavior of the interval width r in different cases. To simplify the problem, all beta priors considered in our results have parameters equal to the same number, γ, and all groups share the same size, i.e., $n_i = m$ for all $i \in \{1, \dots, k\}$. In this scenario, we have that $n = m \cdot k$. We consider three cases: m increases and k is fixed; n is fixed, and m and k change accordingly; finally, m and k grow together according to a certain law. We evaluate S_{pseudo}, and consequently r, by using $w^1_{\text{pseudo, }0}$, according to the procedure described in Example C, which is easily extended to the case of k groups. Our experiments (Figure 3.12) show that:

1. r converges quickly to 0 for fixed k as the m increases (or, equivalently, the total sample size increases);
2. r grows slowly for n fixed and k getting bigger;
3. r converges quickly to 0 whenever k and m grow together according to different power laws.

The only case where r does not converge to 0 corresponds to a decreasing m as $O(1/k)$. We conclude that our microcanonical approximation $S^{\text{GRO}}_{\text{mic}}$ is an optimal candidate as long as m, i.e., the data size of each group, is big enough.

3.4.3 Connection to models of networks and time series

Maximum entropy models are widely used to construct null models of complex systems that preserve specific structural or temporal features, while remaining otherwise random [1, 6, 31, 36, 81].

For instance, when applied to networks, maximum entropy models in their canonical formulations are known as *exponential random graph models* [4, 5, 6]. Examples of commonly used maximum entropy network models are the Erdős–Rényi model, Configuration Models, and Stochastic Block Models [6, 31]. The framework presented here is fully general and can be applied to build and compute e-values when testing between general maximum entropy network models with different sufficient statistics, in both their canonical and microcanonical formulations. Moreover, section 3.2.1 establishes a link between e-values and the Minimum Description Length principle — a framework increasingly used in recent years for network inference, model selection [57, 82, 58].

In particular, the hypothesis tests for contingency tables developed here have a direct correspondence with hypothesis tests between common network models. This mapping arises because the sufficient statistics in our contingency tables capture the same structural constraints as those imposed in standard binary network ensembles [31, 6]. Indeed, a binary network is represented by a binary adjacency matrix, which is a (structured) collection of 1s and 0s, corresponding to the presence or absence of a link between two nodes.

In particular, the null model considered here, in both its canonical and microcanonical formulation, corresponds to the well-known *Erdős–Rényi* model (ER), where

the sufficient statistic is the total number of links, equal to (half, if the network is undirected) the total number of 1s observed in the adjacency matrix.

In the *Stochastic Block Model* (SBM), nodes are partitioned into groups, and the adjacency matrix of a network is structured in k blocks, corresponding to the presence of inter- and intra-group links. For instance, in models of networks with community structure, intra-group link probabilities are larger than inter-group ones. The sufficient statistics are the number of links in each block. Testing an SBM against an ER model corresponds exactly to testing whether the connection probabilities are identical across all blocks (i.e., communities are absent), and this SBM vs ER problem reduces to our canonical or microcanonical $2 \times k$ contingency table test.

In the *Partial Configuration Model* (PCM) for bipartite networks [14], the degree of each node in one layer is constrained, while connections to the other layer are otherwise random. The (bi-)adjacency matrix is a $k \times m$ rectangular binary matrix, and the sufficient statistics are the number of links connected to each node in the constrained layer, i.e., the number of 1s in each row. Testing a PCM against a bipartite ER model corresponds to testing whether all nodes in the constrained layer have the same connection probability (and therefore the same expected degree), i.e., testing for homogeneity of node properties in the graph. This again maps to a $2 \times k$ contingency table, where each constrained node represents a "group" and each group size equals the number m of nodes in the unconstrained layer.

Besides network models, a final connection worth mentioning is the one between binary contingency tables and multivariate time series data describing, e.g., a system of units being active (1) or inactive (0) at discrete time steps (such as spiking neurons data). The PCM can, in this case, represent a model enforcing, for each time step, a different activation probability of the various units. Therefore, testing the PCM against a bipartite ER model amounts, in this case, to testing for non-stationarity vs stationarity of the observed process over time.

We therefore conclude that our microcanonical e-variable for contingency tables can be directly applied to a wide range of problems, both exactly in the microcanonical case and as an approximation for the canonical case. Moreover, our results on the behavior of r show that the microcanonical approximation works very well in both scenarios, as long as the size of each group is large enough. This circumstance is particularly convenient when studying models of large complex systems with a growing number of heterogeneous features, such as PCMs where the number of nodes in both layers can diverge in the "thermodynamic limit" of infinitely large graphs, SBMs used to model networks with a growing number of communities, and models of high-dimensional multivariate (nonstationary) time series. As we mentioned, the growing number of features (and parameters) in these models is generally needed to better replicate the heterogeneous properties of real-world networks and time series. At the same time, it makes the study of these models more challenging because of the breakdown of various useful approximations valid for a finite number of parameters — and even of the asymptotic equivalence between canonical and microcanonical versions of the resulting ensembles(see [12] and 2 of this work). Despite these complications, the results derived here nicely apply to those regimes.

3.5 Conclusion

In this work, we have developed a general framework for constructing optimal e-values for hypothesis testing between maximum entropy models with different constraints, in both microcanonical and canonical formulations. Our main theoretical contribution is the exact derivation of the microcanonical GRO e-variable and its use as a valid approximation to the canonical GRO e-variable when the latter is intractable. We provided analytical and numerical evidence that this approximation becomes asymptotically exact in many relevant regimes.

We illustrated our results through applications to 2×2 contingency tables, showing numerically that the microcanonical approximation provides a good proxy for the canonical solution, confirming our theoretical results. We then extended the analysis to general $2 \times k$ tables, where numerical results suggest that the microcanonical approximation works and remains asymptotically optimal for different interplays between k and the group sizes, as long as the latter are sufficiently large. Interestingly, when k becomes large, the microcanonical e-variable is itself well approximated by choosing a discrete Gaussian prior on the null. We highlighted that this framework can be naturally translated into network-science terms, where many important models can be derived as maximum entropy models.

Universal distributions play a central role in our construction. These are the same distributions that underlie the Minimum Description Length (MDL) principle, where they achieve minimax redundancy. Our results show that such universal distributions can be conveniently used to build GRO e-variables as well, thus providing a direct and convenient connection between description lengths and e-variables. A possible direction to explore in future work is to extend this connection beyond pairwise model comparisons and investigate how GRO e-variables and MDL can be combined to design tests involving multiple models at once.

Appendix 3.A Redundancy and regret

Here we show that, given the null and alternative models $\mathcal{M}_0 = \{P_0(\mathbf{x};\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0}$ and $\mathcal{M}_1 = \{P_1(\mathbf{x};\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_1}$, the regret (3.15) of the GRO e-variable (3.11), i.e.,

$$S^{\mathrm{GRO}}(\mathbf{x}) = \frac{\bar{P}_1(\mathbf{x})}{P_0^{w^*}(\mathbf{x})} \tag{3.97}$$

for a given distribution $\bar{P}_1$, is bounded by the redundancy (3.17) of $\bar{P}_1$. Indeed, for any INECCSI (Def. 3.2.1) subset $\boldsymbol{\Theta}'$ of $\boldsymbol{\Theta}$:

$$\begin{aligned}
\mathrm{REG}(\boldsymbol{\Theta}'_{\mathbf{1}}; \bar{P}_1) &= \max_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}'_1} \mathbb{E}_{\boldsymbol{\theta}_1} \left[\log S^{\mathrm{GRO}(\boldsymbol{\theta}_1)} - \log \frac{\bar{P}_1(\mathbf{x})}{P_0^{w_0^*}(\mathbf{x})} \right] \\
&= \max_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}'_1} \min_{w'_0 \in \mathcal{W}_{\boldsymbol{\theta}_0}} \mathbb{E}_{\boldsymbol{\theta}_1} \left[\log \frac{P_1(\mathbf{x}; \boldsymbol{\theta}_1)}{P_0^{w'_0}(\mathbf{x})} - \log \frac{\bar{P}_1(\mathbf{x})}{P_0^{w_0^*}(\mathbf{x})} \right] \\
&= \max_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}'_1} \left(\min_{w'_0 \in \mathcal{W}_{\boldsymbol{\theta}_0}} \mathbb{E}_{\boldsymbol{\theta}_1} \left[\log \frac{P_0^{w_0^*}(\mathbf{x})}{P_0^{w'_0}(\mathbf{x})} \right] + \mathrm{RED}_1(\boldsymbol{\theta}_1; \bar{P}_1) \right) \\
&\leqslant \mathrm{RED}_1(\boldsymbol{\Theta}'_1; \bar{P}_1).
\end{aligned} \tag{3.98}$$

Appendix 3.B Microcanonical test

In this section, the subscript "mic" is omitted for the sake of clarity, as all the models considered are microcanonical.

3.B.1 Exact solution of the optimization problem

We consider a test between a microcanonical alternative $\mathcal{M}_1$ and a microcanonical null $\mathcal{M}_0$. Given a universal microcanonical distribution $\bar{P}_1$ on the alternative, the GRO-optimal microcanonical e-variable

$$S^{\mathrm{GRO}} = \frac{\bar{P}_1}{P_0^{W_0^*}} \tag{3.99}$$

is found by solving the optimization problem

$$\begin{aligned}
W_0^* &= \arg \min_{W \in \mathcal{W}_{\mathbf{c}_0}} D_{\mathrm{KL}}(\bar{P}_1 \| P_0^W) \\
&= \arg \min_{W \in \mathcal{W}_{\mathbf{c}_0}} \sum_{\mathbf{x} \in \mathcal{X}} \bar{P}_1(\mathbf{x}) \log \frac{\bar{P}_1(\mathbf{x})}{\bar{P}_0^W(\mathbf{x})} \\
&= \arg \max_{W \in \mathcal{W}_{\mathbf{c}_0}} \sum_{\mathbf{x} \in \mathcal{X}} \bar{P}_1(\mathbf{x}) \log \bar{P}_0^W(\mathbf{x}).
\end{aligned} \tag{3.100}$$

For a microcanonical model with sufficient statistics $\mathbf{c}_i$, the Bayesian marginal likelihood reads (see Chapter 2)

$$P_i^{W_i}(\mathbf{x}) = \sum_{\mathbf{c}_i \in \mathcal{C}_i} P_i(\mathbf{x}; \mathbf{c}_i) W_i(\mathbf{c}_i) = P_i(\mathbf{x}; \mathbf{c}_i(\mathbf{x})) W_i(\mathbf{c}_i(\mathbf{x})) = \frac{W_i(\mathbf{c}_i(\mathbf{x}))}{\Omega_i(\mathbf{c}_i(\mathbf{x}))}, \tag{3.101}$$

where the latter equality is due to the definition of the microcanonical model, which assigns a positive probability only if $\mathbf{c}_i(\mathbf{x}) = \mathbf{c}_i$. By putting this result in (3.100), one gets

$$W_0^* = \arg\max_{W \in \mathcal{W}_{\mathbf{c}_0}} \sum_{\mathbf{x} \in \mathcal{X}} \bar{P}_1(\mathbf{x})[\log W_0(\mathbf{c}_0(\mathbf{x})) - \log \Omega_0(\mathbf{c}_0(\mathbf{x}))] \tag{3.102}$$
$$= \arg\max_{W \in \mathcal{W}_{\mathbf{c}_0}} \sum_{\mathbf{x} \in \mathcal{X}} \bar{P}_1(\mathbf{x}) \log W_0(\mathbf{c}_0(\mathbf{x})).$$

The latter expression can be written as a sum over the values of $\mathbf{c}_0$:

$$W_0^* = \arg\max_{W \in \mathcal{W}_{\mathbf{c}_0}} \sum_{\mathbf{c}_0 \in \mathcal{C}_0} \left[\sum_{\mathbf{x} \,:\, \mathbf{c}(\mathbf{x}) = \mathbf{c}_0} \bar{P}_1(\mathbf{x}) \right] \log W(\mathbf{c}_0). \tag{3.103}$$

According to Gibbs inequality, the distribution maximizing the quantity above is

$$W_0^*(\mathbf{c}_0) = \sum_{\mathbf{x} \,:\, \mathbf{c}_0(\mathbf{x}) = \mathbf{c}_0} \bar{P}_1(\mathbf{x}) \tag{3.104}$$

i.e., the marginal distribution of the null sufficient statistic $\mathbf{c}_0(\mathbf{x})$ induced by $\bar{P}_1$, hereby denoted, for simplicity, by $\bar{P}_1^{\mathbf{c}_0}$.

In the special case where the alternative sufficient statistics completely determine the value of the null one, we can write:

Condition A: (3.105)
there exists a function $f : \mathcal{C}_1 \to \mathcal{C}_0$ s.t.$\mathbf{c}_0(\mathbf{x}) = f(\mathbf{c_1}(\mathbf{x}))$,

and thus

$$\begin{aligned} W_0^*(\mathbf{c}_0) &= \sum_{\mathbf{x} \,:\, \mathbf{c}_0(\mathbf{x}) = \mathbf{c}_0} \bar{P}_1(\mathbf{x}) \\ &= \sum_{\mathbf{x} \,:\, \mathbf{c}_0(\mathbf{x}) = \mathbf{c}_0} \frac{W_1(\mathbf{c}_1(\mathbf{x}))}{\Omega_1(\mathbf{c}_1(\mathbf{x}))} \\ &= \sum_{\mathbf{c}_1 \,:\, f(\mathbf{c_1}) = \mathbf{c}_0} \Omega_1(\mathbf{c}_1) \frac{W_1(\mathbf{c}_1)}{\Omega_1(\mathbf{c}_1)} \\ &= \sum_{\mathbf{c}_1 \,:\, f(\mathbf{c_1}) = \mathbf{c}_0} W_1(\mathbf{c}_1). \end{aligned} \tag{3.106}$$

In other words, the GRO-optimal prior on the null is the distribution induced on the null sufficient statistics by the alternative prior, or, equivalently, the marginal distribution of $\mathbf{c}_0$ induced by W_1, hereby denoted by $W_1^{\mathbf{c}_0}$.

Once that W_0^* is computed, given that all universal microcanonical distributions considered here are, in fact, Bayesian marginal likelihoods, the microcanonical GRO-optimal e-variable can be expressed as

$$
\begin{aligned}
S^{\mathrm{GRO}}(\mathbf{x}) &= \frac{P_1^{W_1}(\mathbf{x})}{P_0^{W_0^*}(\mathbf{x})} \\
&= \frac{P_1(\mathbf{x}; \mathbf{c}_1(\mathbf{x}))}{P_0(\mathbf{x}; \mathbf{c}_0(\mathbf{x}))} \frac{W_1(\mathbf{c}_1(\mathbf{x}))}{W_0^*(\mathbf{c}_0(\mathbf{x}))} \\
&= \frac{\Omega_0(\mathbf{c}_0(\mathbf{x}))}{\Omega_1(\mathbf{c}_1(\mathbf{x}))} \frac{W_1(\mathbf{c}_1(\mathbf{x}))}{W_0^*(\mathbf{c}_0(\mathbf{x}))}.
\end{aligned} \tag{3.107}
$$

3.B.2 The average under the null of the GRO e-variable is always unitary

Here, we show that $\mathbb{E}_0[\, S^{\mathrm{GRO}} \,] = 1 \quad \forall P_0 \in \mathcal{M}_0$.

We start by picking a generic distribution inside the null model:

$$
P_0(\mathbf{x}; \bar{\mathbf{c}}_0) = \begin{cases} \dfrac{1}{\Omega_0(\bar{\mathbf{c}}_0)} & \text{if } \mathbf{c}_0(\mathbf{x}) = \bar{\mathbf{c}}_0 \\ 0 & \text{else.} \end{cases} \tag{3.108}
$$

Then, we compute the average under $P_0(\mathbf{x}; \bar{\mathbf{c}}_0)$ of S^{GRO}:

$$
\mathbb{E}_0[S^{\mathrm{GRO}}] = \mathbb{E}_0 \left[\frac{\bar{P}_1}{P_0^{W_0^*}} \right] = \sum_{\mathbf{x} \in \mathcal{X}} P_0(\mathbf{x}; \bar{\mathbf{c}}_0) \frac{\bar{P}_1(\mathbf{x})}{P_0^{W_0^*}(\mathbf{x})} \tag{3.109}
$$

$$
= \frac{1}{\Omega_0(\bar{\mathbf{c}}_0)} \sum_{\mathbf{x} \,:\, \mathbf{c}_0(\mathbf{x}) = \bar{\mathbf{c}}_0} \frac{\bar{P}_1(\mathbf{x})}{P_0^{W_0^*}(\mathbf{x})}. \tag{3.110}
$$

In what follows, we use the explicit expression of the Bayesian marginal likelihood (3.101), i.e., $P_0^{W_0}(\mathbf{x}) = \frac{W_0(\mathbf{c}_0(\mathbf{x}))}{\Omega_0(\mathbf{c}_0(\mathbf{x}))}$, and that of the GRO optimal prior (3.104)

$$
\begin{aligned}
\mathbb{E}_0[\, S^{\mathrm{GRO}} \,] &= \frac{1}{\Omega_0(\bar{\mathbf{c}}_0)} \sum_{\mathbf{x} \,:\, \mathbf{c}_0(\mathbf{x}) = \bar{\mathbf{c}}_0} \frac{\Omega_0(\mathbf{c}_0(\mathbf{x})) \bar{P}_1(\mathbf{x})}{W_0^*(\mathbf{c}_0(\mathbf{x}))} \\
&= \frac{1}{\Omega_0(\bar{\mathbf{c}}_0)} \frac{\Omega_0(\bar{\mathbf{c}}_0)}{W_0^*(\bar{\mathbf{c}}_0)} \sum_{\mathbf{x} \,:\, \mathbf{c}_0(\mathbf{x}) = \bar{\mathbf{c}}_0} \bar{P}_1(\mathbf{x}) = 1.
\end{aligned} \tag{3.111}
$$

Appendix 3.C Microcanonical approximation

3.C.1 Every microcanonical e-variable is a canonical e-variable

Given a sufficient statistic $\mathbf{c}(\mathbf{x})$, the microcanonical probability distribution can be expressed as a conditional canonical one:

$$
P_{\mathrm{mic}}(\mathbf{x}; \mathbf{c}) = P_{\mathrm{can}}(\mathbf{x}; \boldsymbol{\theta} \mid \mathbf{c}), \tag{3.112}
$$

where the probability of $\mathbf{x}$ is conditioned on a certain value of $\mathbf{c}(\mathbf{x})$. According to the law of total expectation, for every random variable $S(\mathbf{x})$ defined on $\mathcal{X}$:

$$
\mathbb{E}_{\mathrm{can}}[\, S \,] = \mathbb{E}_{\mathrm{can},}\, [\, \mathbb{E}_{\mathrm{can}}[\, S \mid \mathbf{c} \,] \,] = \mathbb{E}_{\mathrm{can}}\, [\, \mathbb{E}_{\mathrm{mic}}[\, S \,] \,]\,. \tag{3.113}
$$

It follows that if E_{mic} is a microcanonical e-variable, it is also a canonical one:

$$\mathbb{E}_{\text{mic}}[\, E_{\text{mic}} \,] \leqslant 1 \quad \forall P_{\text{mic}} \in \mathcal{M}_{\text{mic}} \quad \Rightarrow \quad \mathbb{E}_{\text{can}}[\, E_{\text{mic}} \,] \leqslant 1 \quad \forall P_{\text{can}} \in \mathcal{M}_{\text{can}}. \tag{3.114}$$

Moreover, as proven in the last section, it holds:

$$\mathbb{E}_{\text{mic}}[\, S_{\text{mic}}^{\text{GRO}} \,] = 1. \tag{3.115}$$

Thus, from (3.113), it follows that

$$\mathbb{E}_{\text{can}}[\, S_{\text{mic}}^{\text{GRO}} \,] = 1. \tag{3.116}$$

3.C.2 A canonical universal distribution can always be expressed as microcanonical Bayesian marginal likelihood

In what follows, we show that:

1. Given a canonical universal distribution $\bar{P}_{\text{can}}$ relative to a sufficient statistic $\mathbf{c}$, we can always define a prior distribution $W_{\text{can}}(\mathbf{c})$ on the sufficient statistic such that
$$\bar{P}_{\text{can}} = \bar{P}_{\text{mic}}^{W_{\text{can}}}. \tag{3.117}$$
2. The opposite is not true: for some $\bar{P}_{\text{mic}}^{W}$, there is no choice of prior density $w(\boldsymbol{\theta})$ such that $\bar{P}_{\text{mic}}^{W} = \bar{P}_{\text{can}}^{w}$.

Proof 1. By construction, a canonical universal distribution $\bar{P}_{\text{can}}(\mathbf{x})$ relative to the sufficient statistics $\mathbf{c}$ always assigns the same probability mass to configurations sharing the same value of the sufficient statistic, just as the corresponding $P_{\text{can}}(\mathbf{x}; \boldsymbol{\theta})$ does. Consequently, the following holds:

$$\bar{P}_{\text{can}}(\mathbf{x} \,|\, \mathbf{c}) = P_{\text{mic}}(\mathbf{x}; \mathbf{c}), \tag{3.118}$$

i.e., when conditioned on the sufficient statistic, the universal canonical distribution reduces to a uniform distribution over the number of configurations corresponding to the observed value, which is the microcanonical distribution. Moreover, we can always write

$$\bar{P}_{\text{can}}(\mathbf{x}) = \bar{P}_{\text{can}}(\mathbf{x} \mid \mathbf{c}(\mathbf{x}))\bar{P}_{\text{can}}^{\mathbf{c}}(\mathbf{c}(\mathbf{x})) = P_{\text{mic}}(\mathbf{x}; \mathbf{c}(\mathbf{x}))\bar{P}_{\text{can}}^{\mathbf{c}}(\mathbf{c}(\mathbf{x})), \tag{3.119}$$

where $\bar{P}_{\text{can}}^{\mathbf{c}}(\mathbf{c})$ is the distribution of $\mathbf{c}$ induced by $\bar{P}_{\text{can}}$. The proof follows by comparing the expression above with the general expression of the microcanonical Bayesian marginal likelihood (3.101) and by setting $W_{\text{can}}(\mathbf{c}) = \bar{P}_{\text{can}}^{\mathbf{c}}(\mathbf{c})$.

Proof 2. The canonical Bayesian marginal likelihood $\bar{P}_{can}^{w}$

$$P_{\text{can}}^{w}(\mathbf{x}) = \int_{\boldsymbol{\Theta}} P_{\text{can}}(\mathbf{x}; \boldsymbol{\theta}) w(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{3.120}$$

is the weighted sum of positive functions; indeed, $P_{\text{can}}(\mathbf{x};\boldsymbol{\theta})$ is an exponential function that assigns positive probability to all $\mathbf{x} \in \mathcal{X}$. As such, for all proper choices of the prior density $w(\boldsymbol{\theta})$, $\bar{P}^w_{\text{can}}$ is strictly positive everywhere:

$$\bar{P}^w_{can}(x) > 0 \quad \forall x \in \mathcal{X}. \tag{3.121}$$

Consider a microcanonical prior $W(\mathbf{c})$ such that $W(\mathbf{c}^*) = 0$ for a certain value $\mathbf{c}^*$ of the sufficient statistics. Consequently,

$$\bar{P}^W_{\text{mic}}(x^*) = 0 \quad \forall\, x : \mathbf{c}(\mathbf{x}) = \mathbf{c}^*. \tag{3.122}$$

Thus, there is no choice of canonical prior $w(\boldsymbol{\theta})$ s.t. $\bar{P}^W_{\text{mic}} = \bar{P}^w_{\text{can}}$.

3.6 Proof of Theorem 3.3.1

In the proof, we freely use well-known properties of exponential families as described by, e.g., [76]. Set $B := \mathtt{M}'$. Fix a constant a and consider the sets, for $m = 1, 2, \ldots$:

$$B^+_m = \left\{ \boldsymbol{\mu} : \inf_{\boldsymbol{\mu}' \in B} \| \boldsymbol{\mu} - \boldsymbol{\mu}' \|_2^2 \leqslant \frac{a \log m}{m} \right\}, B^-_m = \left\{ \boldsymbol{\mu} \in B : \inf_{\boldsymbol{\mu}' \notin B} \| \boldsymbol{\mu} - \boldsymbol{\mu}' \|_2^2 \geqslant \frac{a \log m}{m} \right\}.$$

B^+_m is a superset of B, including a small region (whose volume tends to 0 with sample size) just outside B's boundary; similarly B^-_m excludes a small region just inside B's boundary. To shorten notation we write $\hat{\boldsymbol{\mu}} := \mathbf{s}(\mathbf{y}^{(m)})/m$ and $P^w := P^{w\,(m)}_{\text{can}}$ and, for any measurable subset $B' \subset \mathtt{M}$, $Q^w(B') := Q^w(\boldsymbol{\mu} \in B')$, and $P_{\boldsymbol{\mu}} := P_{\text{can}}(\cdot; \boldsymbol{\theta}(\boldsymbol{\mu}))$ where $\boldsymbol{\theta}(\boldsymbol{\mu})$ is the mapping from mean-value parameters to corresponding canonical parameters, i.e. the inverse of the (1-to-1) mapping $\boldsymbol{\mu}(\boldsymbol{\theta})$ defined in the main text. We have:

$$\begin{aligned}
P^w(\hat{\boldsymbol{\mu}} \in B) &= \int_{\boldsymbol{\mu} \in \mathtt{M}} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B) w(\boldsymbol{\mu}) d\boldsymbol{\mu} && (3.123)\\
&\geqslant \int_{\boldsymbol{\mu} \in B^-_m} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B) w(\boldsymbol{\mu}) d\boldsymbol{\mu} \\
&\geqslant Q^w(B^-_m) \inf_{\boldsymbol{\mu} \in B^-_m} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B) && (3.124)\\
&\geqslant Q^w(B^-_m) \inf_{\boldsymbol{\mu} \in B^-_m} (1 - P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \notin B)) \\
&\geqslant Q^w(B^-_m) - \sup_{\boldsymbol{\mu} \in B^-_m} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \notin B) Q^w(B) + O\left(\frac{\log m}{m}\right) - \sup_{\boldsymbol{\mu} \in B^-_m} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \notin B), && (3.125)
\end{aligned}$$

and also

$$\begin{aligned}
P^w(\hat{\boldsymbol{\mu}} \in B) &= \int_{\boldsymbol{\mu} \in B^+_m} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B) w(\boldsymbol{\mu}) d\boldsymbol{\mu} + \int_{\boldsymbol{\mu} \notin B^+_m} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B) w(\boldsymbol{\mu}) d\boldsymbol{\mu} \\
&\leqslant Q^w(B^+_m) + \sup_{\boldsymbol{\mu} \in \mathtt{M}, \boldsymbol{\mu} \notin B^+_m} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B) \\
&= Q^w(B) + O\left(\frac{\log m}{m}\right) + \sup_{\boldsymbol{\mu} \in \mathtt{M} \setminus B^+_m} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B). && (3.126)
\end{aligned}$$

We will now further bound the supremum terms in the above two formulas, showing that they are both of order $O(m^{-a})$. The result then follows by plugging in $a = 1$.

Supremum in (3.126) To bound the supremum in (3.126), note first that B is a convex set. We can therefore use Csiszár's [83] multivariate generalization of Chernoff's concentration inequality (see [84] for general discussion) to get that

$$\sup_{\boldsymbol{\mu}\in\mathtt{M}\setminus B_m^+} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B) \leqslant \sup_{\boldsymbol{\mu}\in\mathtt{M}\setminus B_m^+, \boldsymbol{\mu}'\in B} e^{-mD_{\mathrm{KL}}(P_{\boldsymbol{\mu}'}\|P_{\boldsymbol{\mu}})} \tag{3.127}$$

$$= e^{-m\inf_{\boldsymbol{\mu}\in\mathtt{M}\setminus B_m^+, \boldsymbol{\mu}'\in B} D_{\mathrm{KL}}(P_{\boldsymbol{\mu}'}\|P_{\boldsymbol{\mu}})} \tag{3.128}$$

$$\leqslant e^{-m\inf_{\boldsymbol{\mu}\in\partial B_m^+, \boldsymbol{\mu}'\in B} D_{\mathrm{KL}}(P_{\boldsymbol{\mu}'}\|P_{\boldsymbol{\mu}})}, \tag{3.129}$$

where $D_{\mathrm{KL}}(P_{\boldsymbol{\mu}'}\|P_{\boldsymbol{\mu}})$ is the Kullback-Leibler divergence between $P_{\boldsymbol{\mu}'}$ and $P_{\boldsymbol{\mu}}$ defined at a single outcome, and in the final inequality we used the standard fact that the KL divergence between members of an exponential family is strictly convex in its first argument.

Now, since B is an INECCSI subset of $\mathtt{M}$, clearly there exists another INECCSI subset of $\mathtt{M}$, say $\bar{B}$, and a finite m_0 such that $B_m^+ \subset \bar{B}$ for all $m \geqslant m_0$. Since $\bar{B}$ is INECCSI, there exist c, C with $0 < c < C < \infty$ such that all eigenvalues of the Fisher information matrix $I(\boldsymbol{\mu})$ in the mean-value parameterization are in between c and C for all $\boldsymbol{\mu} \in \bar{B}$. This means that the KL divergence between any $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \bar{B}$ satisfies

$$\begin{aligned}\frac{1}{2}c^k \frac{a\log m}{m} &\leqslant \frac{1}{2}c^k\|\boldsymbol{\mu}-\boldsymbol{\mu}'\|_2^2 \\ &\leqslant D_{\mathrm{KL}}(P_{\boldsymbol{\mu}'}\|P_{\boldsymbol{\mu}}) \\ &\leqslant \frac{1}{2}C^k\|\boldsymbol{\mu}-\boldsymbol{\mu}'\|_2^2 \leqslant \frac{1}{2}C^k\frac{a\log m}{m}.\end{aligned} \tag{3.130}$$

The result now follows by plugging in the lower bound on the KL divergence implied by the above into (3.127) and then setting $a = 1$ and plugging further into (3.126).

Supremum in (3.123) To bound the supremum in (3.123), fix any $\boldsymbol{\mu} \in B_m^-$ and let $R_{m,\boldsymbol{\mu}}$ be a hyper-rectangle centered at $\boldsymbol{\mu}$ that is a subset of B and that has side-length $2\epsilon_m$. By construction there is $c' > 0$ such that, for all m, we can take $\epsilon_m = c' \cdot \sqrt{(a\log m/m)}$. Noting that we can write $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_d)$, with d the dimensionality of both the canonical space $\boldsymbol{\Theta}$ and the mean-value space $\mathtt{M}$, we set $H^{\geqslant}_{\boldsymbol{\mu},j,\epsilon} := \{\boldsymbol{\mu}' \in \mathtt{M} : \boldsymbol{\mu}'_j \geqslant \boldsymbol{\mu}_j + \epsilon\}$ and $H^{\leqslant}_{\boldsymbol{\mu},j,\epsilon} := \{\boldsymbol{\mu}' \in \mathtt{M} : \boldsymbol{\mu}'_j \leqslant \boldsymbol{\mu}_j - \epsilon\}$. We have:

$$\begin{aligned}\sup_{\boldsymbol{\mu}\in B_m^-} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \notin B) &\leqslant \sup_{\boldsymbol{\mu}\in B_m^-} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \notin R_{m,\boldsymbol{\mu}}) \\ &\leqslant \sum_{j=1}^{d}\left(P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in H^{\leqslant}_{\boldsymbol{\mu},j,\epsilon_m}) + P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in H^{\geqslant}_{\boldsymbol{\mu},j,\epsilon_m})\right) \\ &\leqslant \sum_{j=1}^{d}\left(e^{-m\inf_{\boldsymbol{\mu}'\in H^{\leqslant}_{\boldsymbol{\mu},j,\epsilon_m}} D_{\mathrm{KL}}(P_{\boldsymbol{\mu}'}\|P_{\boldsymbol{\mu}})\|)} + e^{-m\inf_{\boldsymbol{\mu}'\in H^{\geqslant}_{\boldsymbol{\mu},j,\epsilon_m}} D_{\mathrm{KL}}(P_{\boldsymbol{\mu}'}\|P_{\boldsymbol{\mu}})\|)}\right) \\ &= O\left(m^{-a}\right).\end{aligned} \tag{3.131}$$

Here the second inequality is the union bound. The third is once again Csiszár's [83] multivariate generalization of Chernoff's concentration inequality (in (3.127), we bounded $P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B)$ where B was a convex set, allowing us to use Csiszár's result directly; but in (3.131), we need to bound $P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \notin B) = P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in \mathtt{M} \setminus B)$; since $\mathtt{M} \setminus B$ is not a convex set, yet convexity is required by Csiszf's result, we now first need to cover it by $2d$ convex sets $H^{\cdot}_{\boldsymbol{\mu},\cdot,\epsilon_m}$, for each of which we then use Csiszár's result). The final inequality follows by using (3.130) again.

Appendix 3.D Pseudo approximation through the high resolution limit

Below, we provide pseudocode to compute the density $w^1_{\text{pseudo},0}$ for a canonical test on 2×2 contingency tables, under the alternative hypothesis with independent beta priors on the mean-value parameters. Notice that for $\alpha_a = \beta_a = \alpha_b = \beta_b = 1$, we retrieve the uniform prior of Example B.

The procedure starts from the induced distribution on the sufficient statistics under the alternative, denoted $W^a_{\text{can},1}$ and $W^b_{\text{can},1}$ in the main text, which follow a beta-binomial form. To approximate the continuous limit, we increase the resolution by multiplying the original group sizes by a scaling factor (Steps 1–2). The higher the scaling factor, the better the approximation — at the cost of greater computational effort.

The two high-resolution distributions are then convolved to obtain an approximation of the GRO-optimal prior on the null, W^*_0 (Step 3). Since we want a density over $p_0 \in [0, 1]$, we define a pseudo-continuous support for p_0 accordingly (Step 4). Finally, the convolved distribution is normalized to form a proper density over this support (Step 5).

```
Input:
    n_a, n_b               // group sizes
    scaling_factor         // resolution multiplier
    alpha_a, beta_a        // beta-binomial parameters for group a
    alpha_b, beta_b        // beta-binomial parameters for group b

Step 1: Define high-resolution support
    x_high_a = 0 to scaling_factor * n_a
    x_high_b = 0 to scaling_factor * n_b

Step 2: Compute high-resolution beta-binomial PMFs
    For each i in x_high_a:
        p_high_a[i] = BetaBinomialPMF(i, scaling_factor * n_a, alpha_a, beta_a)
    For each i in x_high_b:
        p_high_b[i] = BetaBinomialPMF(i, scaling_factor * n_b, alpha_b, beta_b)

Step 3: Convolve the two distributions
    conv_high = Convolve(p_high_a, p_high_b)
```

```
Step 4: Define pseudo-continuous support over [0, 1]
    p_0_fine = [0, 1, ..., len(conv_high)-1] / (scaling_factor * (n_a + n_b))

Step 5: Normalize the convolved distribution
    step = p_0_fine[1] - p_0_fine[0]
    conv_high = conv_high / (sum(conv_high) * step)

Output:
    conv_high   // approximate density over [0, 1]
    p_fine      // corresponding support
```


Appendix 3.E Bound on regret

Here we provide a bound for the regret of a maximum entropy model (3.15).

Let $\mathcal{M}_0$ be a maximum entropy model. We begin by recalling an extension of the classical $(d/2)\log m$ asymptotic redundancy result (see Equation (3.19)) that remains valid even when the model $\mathcal{M}_0$ is misspecified — i.e., it does not contain the true distribution P^*. This generalization appears in [16], and is formalized in [29, Proposition 3].

Let $\mathbf{x}^m$ be an i.i.d. sample from a distribution P^* that may lie outside $\mathcal{M}_0$. Suppose that there exists a distribution $P_{\widetilde{\boldsymbol{\theta}}_0} \in \mathcal{M}_0$ minimizing the Kullback-Leibler divergence to P^*:

$$P_{\widetilde{\boldsymbol{\theta}}_0} = \arg\min_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0} D_{\mathrm{KL}}(P^* \| P_{\boldsymbol{\theta}_0}).$$

Then, for any regular prior density w_0, we have:

$$\mathrm{RED}_0(P^*; P_0^{w_0}) := \mathbb{E}_{P^*}\left[-\log P_0^{w_0}(\mathbf{x}^m) + \log P_{\widetilde{\boldsymbol{\theta}}_0}(\mathbf{x}^m)\right] = \frac{d_0}{2}\log m + O(1). \quad (3.132)$$

In the well-specified case where $P^* \in \mathcal{M}_0$, it holds that $P^* = P_{\widetilde{\boldsymbol{\theta}}_0}$, and $\mathrm{RED}_0(P^*; P_1^{w_0})$ coincides with our earlier definition $\mathrm{RED}_0(\widetilde{\boldsymbol{\theta}}_0, P_1^{w_0})$ (up to notation), thus recovering the classical result (3.19).

We now apply this general result in the setting where $P^* = P_{\boldsymbol{\theta}_1} \in \mathcal{M}_1$, in order to bound the regret $\mathrm{REG}(\boldsymbol{\theta}_1; P_1^{w_1})$ of the Bayesian marginal $P_1^{w_1}$ for a regular prior w_1. Consider the pseudo-e-variable S_{pseudo}, used as a proxy for either $S_{\mathrm{mic}}^{\mathrm{GRO}}$ or $S_{\mathrm{can}}^{\mathrm{GRO}}$.

Using the previous result to refine equation (3.21), we obtain:

$$\begin{aligned}
\mathrm{REG}(\boldsymbol{\theta}_1; S_{\text{pseudo}}) &= \mathbb{E}_{\boldsymbol{\theta}_1}\left[\log S^{\mathrm{GRO}(\boldsymbol{\theta}_1)} - \log \frac{P_1^{w_1}(\mathbf{x})}{P_0^{w_{\text{pseudo},0}}(\mathbf{x})}\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}_1}\left[\log \frac{P_0^{w_{\text{pseudo},0}}(\mathbf{x})}{P_0^{\widetilde{w}_0'}(\mathbf{x})}\right] + \mathrm{RED}_1(\boldsymbol{\theta}_1; P_1^{w_1}) \\
&= \mathbb{E}_{\boldsymbol{\theta}_1}\left[\log \frac{P_0^{w_{\text{pseudo},0}}(\mathbf{x})}{P_{\widetilde{\boldsymbol{\theta}}_0}(\mathbf{x})}\right] + \mathrm{RED}_1(\boldsymbol{\theta}_1; P_1^{w_1}) \\
&= \mathrm{RED}_1(\boldsymbol{\theta}_1; P_1^{w_1}) - \mathrm{RED}_0(P_{\boldsymbol{\theta}_1}; P_0^{w_{\text{pseudo},\,0}}) + O(1) \\
&\overset{(a)}{=} \frac{d_1 - d_0}{2} \cdot \log m + O(1). \qquad (3.133)
\end{aligned}$$

The first two equalities are direct, and the third follows from [29, Theorem 3 (Parts 3,4)]. Equality (a) holds provided that $w_{\text{pseudo},0}$ is regular, in particular, that it has a continuous density.

Appendix 3.F Testing with the NML universal distribution

In this section, we propose an alternative to the Bayesian approach, based on the *Normalized Maximum Likelihood distribution*. In a variation of the definition of universality of 3.18, we may search for $\bar{P}_j$ such that $-\log \bar{P}_j(\mathbf{x}) + \log P_{\hat{\boldsymbol{\theta}}_j}(\mathbf{x})$ is small, in the worst-case over all possible data realizations $\mathbf{x}$; here $\hat{\boldsymbol{\theta}}_j(\mathbf{x})$ is the maximum likelihood estimator of $\boldsymbol{\theta}_j$ relative to $\mathbf{x}$. This leads to comparing $\bar{P}_j(\mathbf{x})$ to the best-fitting model *a posteriori*, that is, the model that, with hindsight, would give the shortest code for the observed data.

In the MDL literature, this more stringent criterion is often considered the ideal. The distribution achieving this is called the Normalized Maximum Likelihood (NML) distribution [16, 18, 17]. For each observed dataset $\mathbf{x}$, this distribution assigns its maximum likelihood value, normalized over all possible datasets in $\mathcal{X}$:

$$\bar{P}_j^{\mathrm{NML}}(\mathbf{x}) = \frac{P_j(\mathbf{x}; \hat{\boldsymbol{\theta}}_j(\mathbf{x}))}{\sum_{\mathbf{y}\in\mathcal{X}} P_j(\mathbf{y}; \hat{\boldsymbol{\theta}}_j(\mathbf{y}))}. \qquad (3.134)$$

Importantly, the NML distribution does not require any prior, and achieves universality both in worst-case data and in worst-case expected regret. In fact, for the MEM models introduced in this work, inequality (3.19) also holds when $P_j^{w_j(m)}$ is replaced by $\bar{P}_j^{\mathrm{NML}(m)}$.

In what follows, we explicitly compute NML-based GRO e-variables for contingency tables, introduced in 3.4 of the main text, starting from the 2×2 case and generalizing to the $2 \times k$ case. In the microcanonical case, resorting to the Normalized Maximum Likelihood approach is equivalent to putting a uniform prior on both parameters of the alternative model, as shown in Chapter 2. Consequently, this case is reduced to Example A in 3.3.1, and is not considered further. Instead, we focus on computing the microcanonical approximation of the canonical GRO e-variable.

Canonical test with NML. We first consider the case of a 2×2 contingency table, where we test between two canonical models through an NML approach, i.e., $\bar{P}_{\text{can},1} = P^{\text{NML}}_{\text{can},1}$. For a model with two independent Bernoulli distributions, the exact expression of the NML distribution reads (see Chapter 2):

$$P^{\text{NML}}_{\text{can},1}(\mathbf{x}) = P^{\text{a, NML}}_{\text{can},1}(\mathbf{x}) \cdot P^{\text{b, NML}}_{\text{can},1}(\mathbf{x}) \tag{3.135}$$

with

$$P^{\text{i, NML}}_{\text{can},1}(\mathbf{x}) = \frac{\left(\frac{n^i_1(\mathbf{x})}{n^i}\right)^{n^i_1(\mathbf{x})} \left(1 - \frac{n^i_1(\mathbf{x})}{n^i}\right)^{n^i - n^i_1(\mathbf{x})}}{\frac{e^{n^i}\Gamma(n^i, n^i)}{(n^i)^{n^i - 1}} + 1} \tag{3.136}$$

for $i \in \{a, b\}$. In the formula above, $\Gamma(s, t)$ is the upper incomplete gamma function. According to the procedure described in 3.3.2, a natural candidate e-variable is given by the microcanonical approximation, i.e., the GRO e-variable of the corresponding microcanonical test. To build it, we first need the distribution of the alternative sufficient statistics (n^a_1, n^b_1) induced by $P^{\text{i, NML}}_{\text{can},1}$, which is

$$W_{\text{can},1}(n^a_1, n^b_1) = W^a_{\text{can},1}(n^a_1) \cdot W^b_{\text{can},1}(n^b_1) \tag{3.137}$$

with

$$W^a_{\text{can},1}(n^a_1) = \Omega^a_1(n^a_1) \cdot \frac{\left(\frac{n^a_1(\mathbf{x})}{n^a}\right)^{n^a_1(\mathbf{x})} \left(1 - \frac{n^a_1(\mathbf{x})}{n^a}\right)^{n^a - n^a_1(\mathbf{x})}}{\frac{e^{n^a}\Gamma(n^a, n^a)}{(n^a)^{n^a - 1}} + 1} \tag{3.138}$$

and

$$W^b_{\text{can},1}(n^b_1) = \Omega^b_1(n^b_1) \cdot \frac{\left(\frac{n^b_1(\mathbf{x})}{n^b}\right)^{n^b_1(\mathbf{x})} \left(1 - \frac{n^b_1(\mathbf{x})}{n^b}\right)^{n^b - n^b_1(\mathbf{x})}}{\frac{e^{n^b}\Gamma(n^b, n^b)}{(n^b)^{n^b - 1}} + 1}. \tag{3.139}$$

Given that n^a_1 and n^b_1 are independently distributed, $W^*_0(n_1)$ is the convolution of $W^a_{\text{can},1}(n^a_1)$ and $W^b_{\text{can},1}(n^b_1)$, which can be computed numerically. Once that $W^*_0(n_1)$ is computed, the microcanonical approximation is evaluated through Eq. (3.40). The accuracy of the microcanonical approximation is assessed by directly inspecting the difference in e-power (Figure 3.8) and through the interval width r (bottom row of Figure 3.9), which is shown to be already close to 0 for small sample sizes and to converge towards 0 as the sample size increases.

Notice that, if n^a_1 and n^b_1 are both large enough, using $P^{\text{NML}}_{\text{can},1}$ is asymptotically equivalent to using a Bayesian universal distribution with a *Jeffreys prior* [16] on the alternative parameters. The Jeffreys prior, defined as

$$w_J(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta})}, \tag{3.140}$$

where $I(\boldsymbol{\theta})$ is the Fisher information matrix, is invariant under any smooth reparametrization of $\boldsymbol{\theta}$ and is therefore commonly used as a canonical non-informative prior. For a Bernoulli model, such as ours, the Jeffreys prior is equal to a beta prior with parameters all equal to $\gamma = 0.5$:

$$w_J(p_a, p_b) = w^a_J(p_a) w^b_J(p_b) = \frac{1}{\pi^2} \frac{1}{\sqrt{p_a(1-p_a)}} \frac{1}{\sqrt{p_b(1-p_b)}}. \tag{3.141}$$

Asymptotically, and in an expected-regret sense, the NML distribution coincides with a Bayesian universal distribution based on Jeffreys prior when the true parameter lies in the interior of $\mathtt{M}_1 = [0,1]^2$.

More explicitly, let us set $n^a = n^b = m$. Then, for any INECCSI subset $\mathtt{M}_1' \subset \mathtt{M}_1$, as $m \to \infty$:

$$\sup_{\boldsymbol{\mu}_1 \in \mathtt{M}_1} \mathbb{E}_{\boldsymbol{\mu}_1} \left[-\log P_{\text{can},1}^{w_J}(\mathbf{x}^m) + \log P_{\text{can},1}^{\text{NML}}(\mathbf{x}^m) \right] = o(1), \tag{3.142}$$

where $o(1)$ denotes a quantity that goes to 0 as $m \to +\infty$. At the boundaries of the parameter space, however, the two constructions differ: Jeffreys prior, which assigns densities that diverge to infinity at the boundaries, leads to substantially different data probabilities than the NML-induced distribution. This difference becomes even more important when convoluting the independent Jeffreys priors to compute W_0^*. This difference becomes even more important when convoluting the independent Jeffreys priors to compute W_0^*. This explains why the first and third pictures in Figure 3.8 are quite different. For this reason, in all simulations we use the exact NML formula rather than its Jeffreys approximation.

We now extend the solution to $2 \times k$ contingency tables. For a model with k independent Bernoulli distributions, the NML distribution (see Chapter 2) reads:

$$P_{\text{can},1}^{\text{NML}}(\mathbf{x}) = \prod_{i=1}^{k} P_{\text{can},1}^{\text{i, NML}}(\mathbf{x}) \tag{3.143}$$

where $P_{\text{can},1}^{\text{i, NML}}$ is given by Eq. (3.136). The microcanonical approximation is obtained by defining

$$W_{\text{can},1}(\{n_1^i\}) = \prod_{i=1}^{k} W_{\text{can},1}^i(n_1^i) \tag{3.144}$$

with

$$W_{\text{can},1}^i(n_1^i) = \Omega_1^i(n_1^i) \cdot \frac{\left(\frac{n_1^i(\mathbf{x})}{n^i}\right)^{n_1^i(\mathbf{x})} \left(1 - \frac{n_1^i(\mathbf{x})}{n^i}\right)^{n^i - n_1^i(\mathbf{x})}}{\frac{e^{n^i}\Gamma(n^i, n^i)}{(n^i)^{n^i - 1}} + 1}. \tag{3.145}$$

W_0^* is then the convolution of all $W_{\text{can},1}^i(n_1^i)$, which can be computed numerically or by resorting to the Gaussian approximation (3.87). Once that $W_0^*(n_1)$ is computed, the microcanonical approximation is evaluated through Eq. (3.40).

Appendix 3.G Supplementary tables and figures

Case	Parameter Condition	Behavior
Uniform	$\alpha = \beta = 1$	Flat
Jeffreys prior	$\alpha = \beta = 0.5$	U-shaped
Bimodal	$\alpha, \beta < 1$	Peaks at 0 and 1
Left-skewed	$\alpha > 1, \beta < 1$	Peak near 1
Right-skewed	$\alpha < 1, \beta > 1$	Peak near 0
Bell-shaped	$\alpha, \beta > 1$	Gaussian-like
Highly concentrated	$\alpha = \beta \gg 1$	Sharp peak
Degenerate	$\alpha, \beta \to \infty$	Point mass at $\frac{\alpha}{\alpha+\beta}$

Table 3.1: Behaviors of the beta distribution for different parameter values.

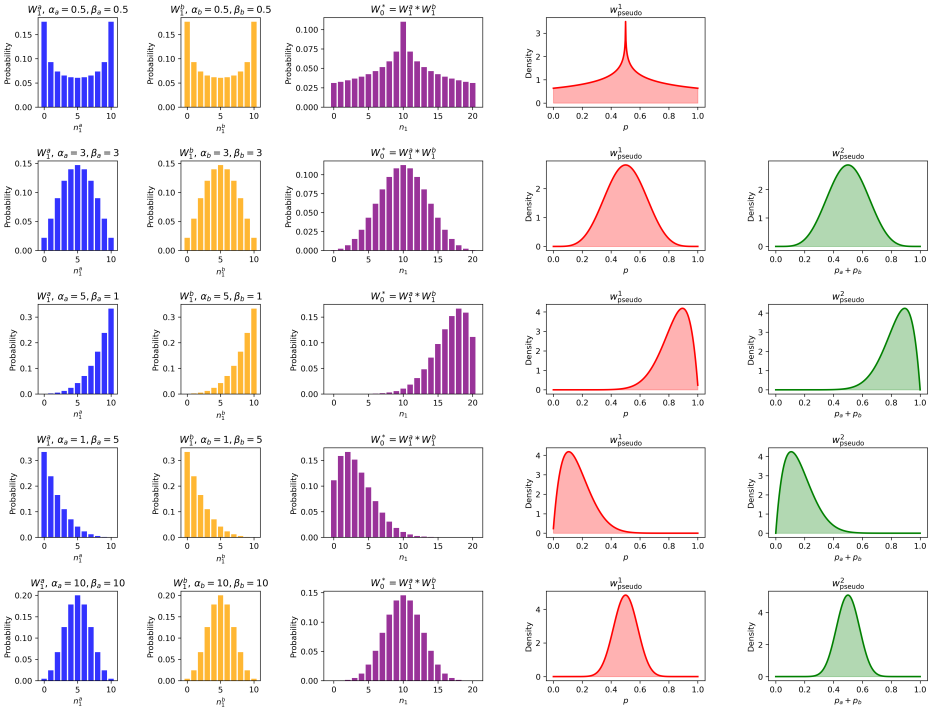


Figure 3.7: Construction of the microcanonical optimal prior W_0^* (third column) and the pseudo-prior approximations $w^1_{pseudo,0}$ (fourth column) and $w^2_{pseudo,0}$ (fifth column), for 2×2 contingency tables with independent beta priors on the alternative. Starting from the induced independent beta-binomial distribution on the alternative sufficient statistics (first and second columns), the microcanonical GRO-optimal prior W_0^* is obtained as their convolution. The pseudo prior approximation $w^1_{pseudo,0}$ is obtained starting from W_0^* through a high-resolution limit. The pseudo prior approximation $w^2_{pseudo,0}$ is obtained by directly convoluting the original continuous beta priors, when well defined on the whole parameter space (i.e., for $\gamma \geqslant 1$).

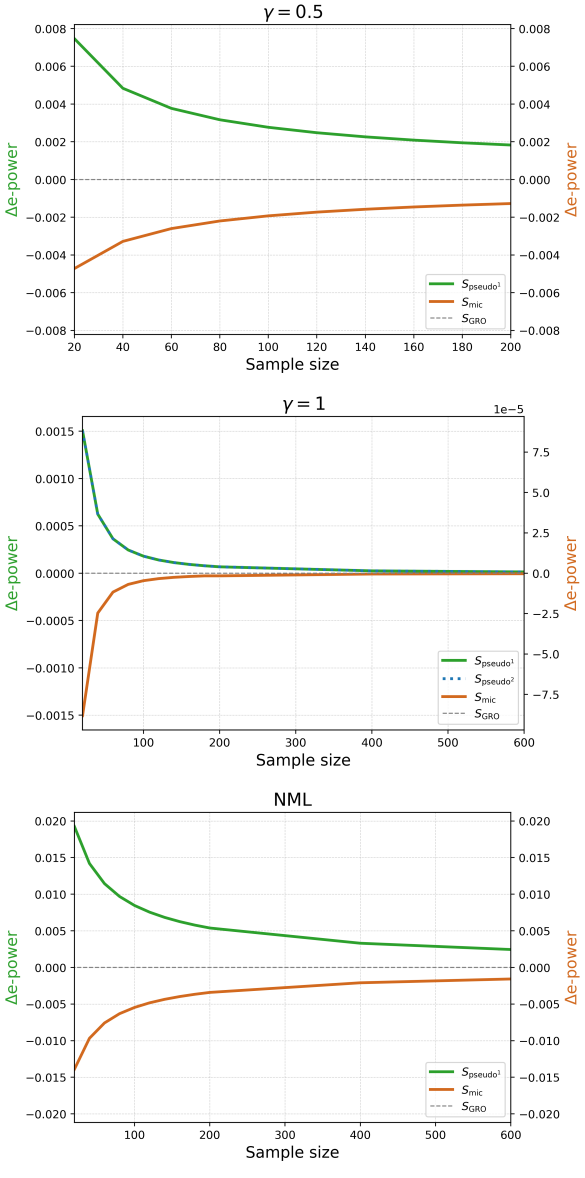


Figure 3.8: E-power difference between the canonical GRO e-variable (computed numerically), its microcanonical approximation (orange curve), and the pseudo approximation (green curve), across sample sizes and different choices on the alternative (NML and beta with all parameters equal to γ). The microcanonical and pseudo approximations provide lower and upper bounds for the canonical GRO e-power, converging to it as the sample size grows. Results are shown for the 2×2 contingency tables canonical test with $n^a = n^b = m$ and sample size equal to $2m$.

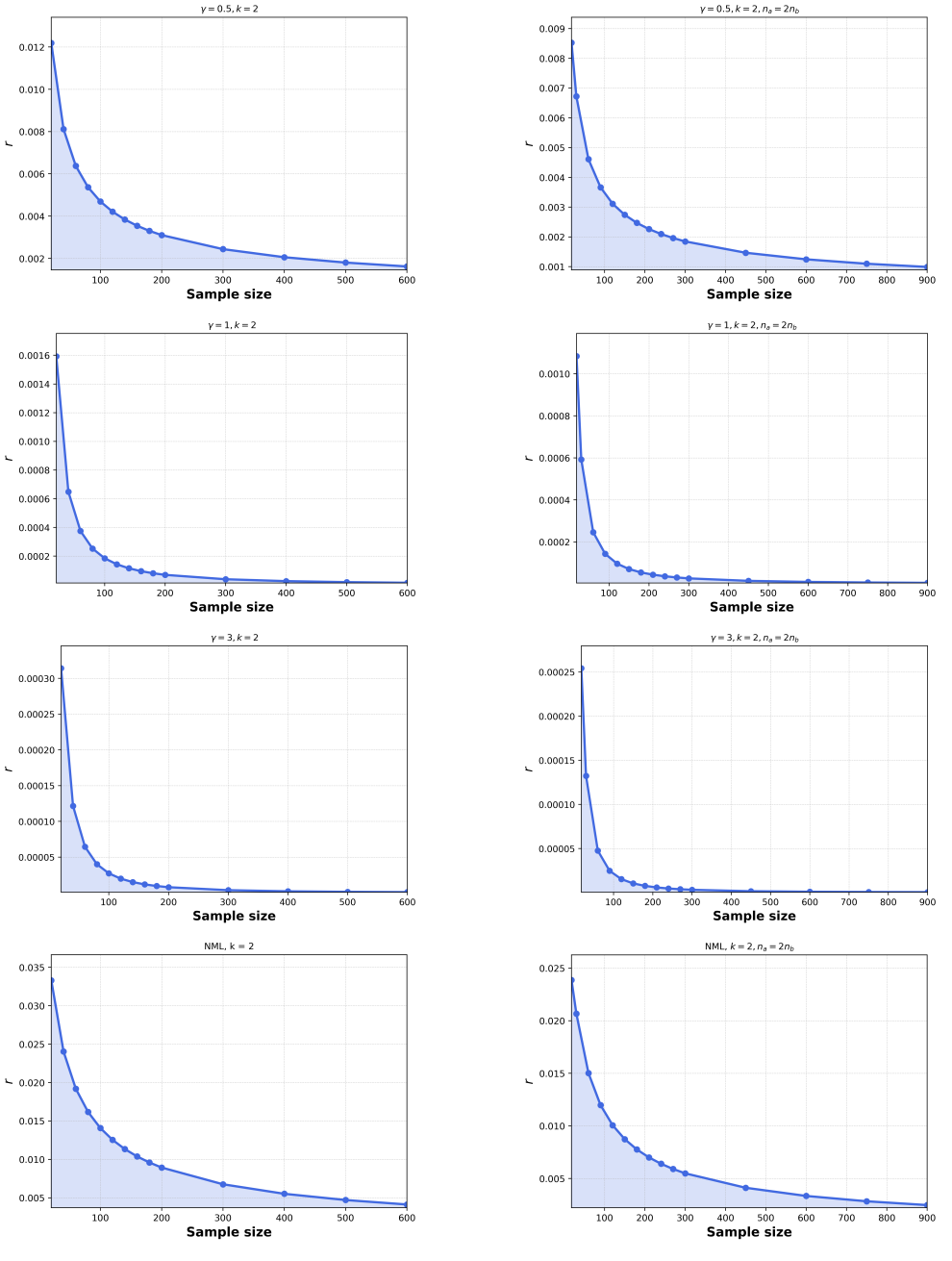


Figure 3.9: Convergence of r in 2×2 contingency tables, for $n^a = n^b$ (left column) and $n^a = 2n^b$ (right column), as the sample size $n = n^a + n^b$ grows. Results are shown for different choices of $\bar{P}_{can,1}$: beta independent priors with all parameters equal to $\gamma = 0.5$ (first row), 1 (second row), 3 (third row), and NML (fourth row).

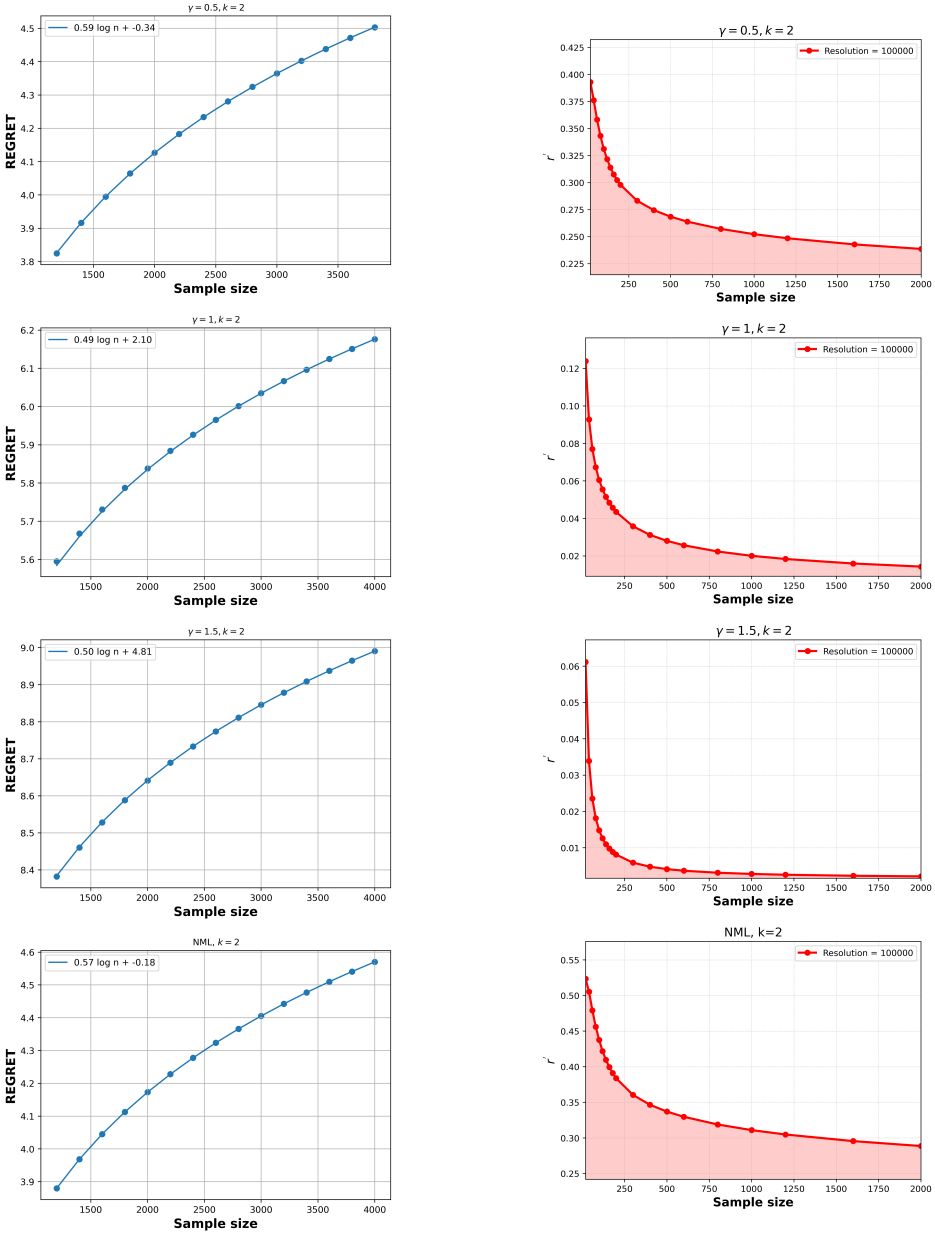


Figure 3.10: Worst case regret (left) and convergence of r' (right), for 2×2 tables with $n^a = n^b = m$ and sample size equal to $n = 2m$. Different choices of the alternative are considered: Bayesian with identical independent beta priors $B(\gamma, \gamma)$ with $\gamma = 0.5$, 1, 1.5, and NML. r' is computed by considering $w^1_{pseudo,0}$, obtained through a high resolution limit with resolution scale equal to 100000.

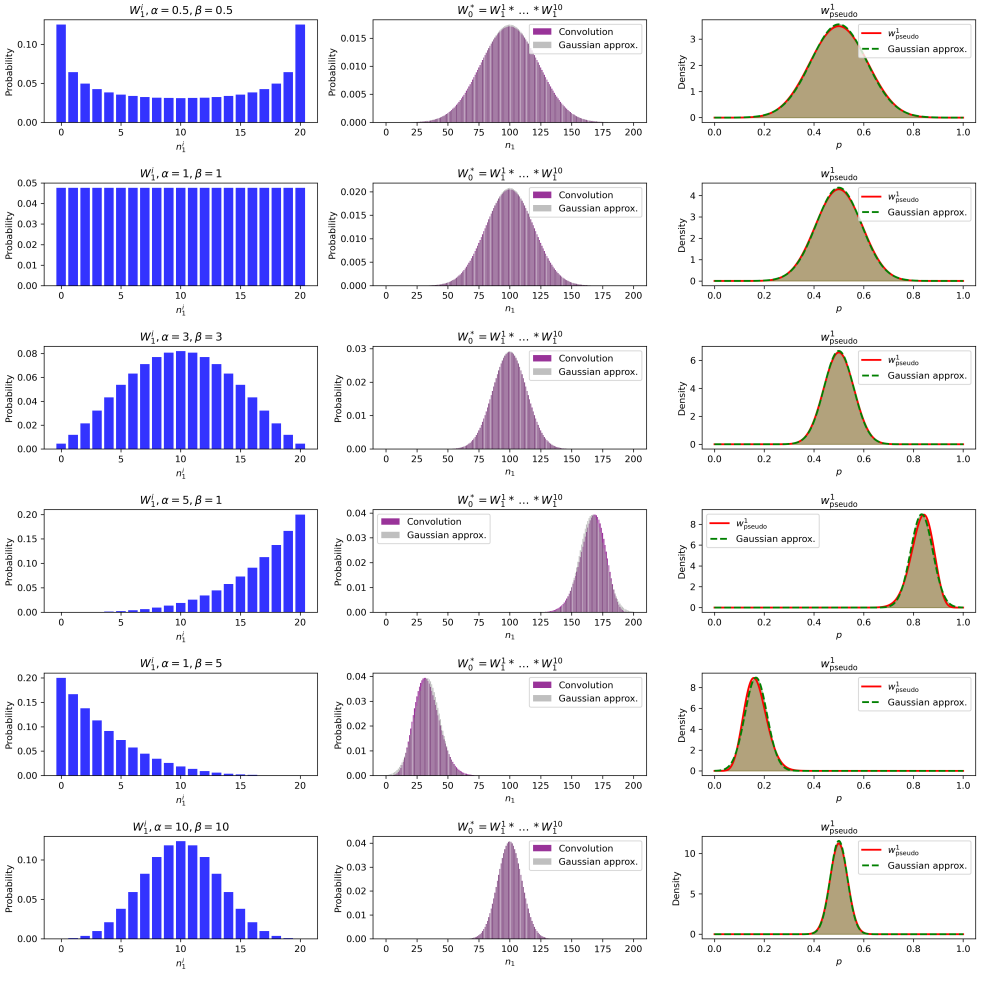


Figure 3.11: The first column shows the discrete distribution of each component of the alternative sufficient statistics, here denoted simply by W_1^i, induced by independent identical beta priors on the alternative, for different prior parameters. The GRO-optimal microcanonical prior on the null W_0^* (second column) for the $2 \times k$ test is obtained by convolving these discrete distributions k times. The pseudo prior density w^1_{pseudo} (third column) is instead obtained by directly convolving k times the corresponding beta priors. Both W_0^* and w^1_{pseudo} are shown together with their Gaussian approximations (discrete for W_0^* and continuous for w^1_{pseudo}). In this example, $k = 10$.

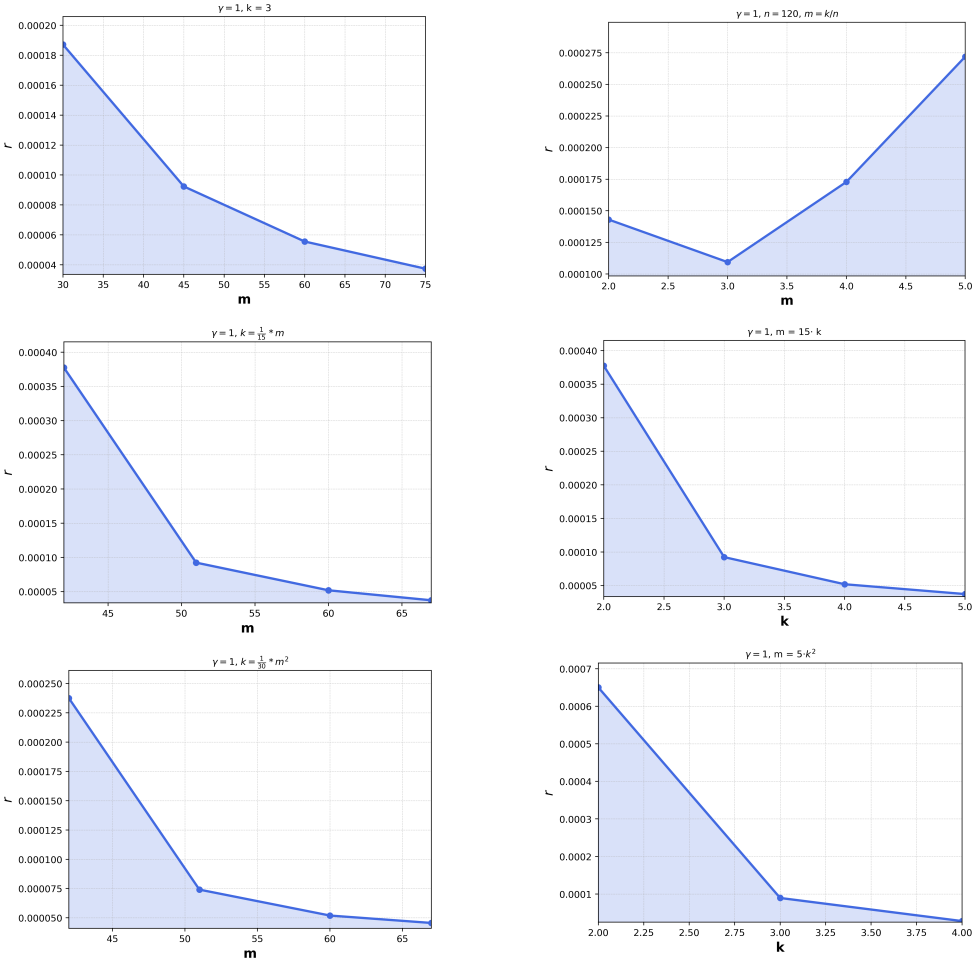


Figure 3.12: Convergence to 0 of the interval width r in the case of $2 \times k$ contingency tables, where all groups have same size m, for different interplays between the number of groups k and the size of each group m. Results are shown for identical independent beta priors on the alternative, with all parameters equal to $\gamma = 1$.

CHAPTER 4

E-values for contingency tables

This chapter is based on:
Y. Wang, S. Arnold, F. Giuffrida, P. Grünwald, and T. Dickhaus. E-values for Contingency Tables. In preparation.

Abstract

We empirically investigate several e-values and e-processes for testing in 2×2 contingency tables. While the e-process recently introduced by Turner et al. exhibits theoretical growth-optimality properties for paired data sequences, we find that, when applied to a single large table (the *batch setting*), its performance deteriorates, with the other standard e-process based on universal inference performing even worse. By contrast, *conditional* e-variables tend to perform better, both in terms of growth rate and expected stopping times, with *uniformly most powerful* e-variables showing the best overall performance. When batches arrive sequentially, the comparison becomes more nuanced: Turner's e-variable is generally optimal in the asymptotic regime, whereas the *normalized maximum likelihood* e-variable is often, though not always, preferable at practically relevant sample sizes.

4.1 Introduction

Over the last few years, e-variables have attracted increasing interest in the statistical community as new, safe methods for assessing statistical evidence; see, e.g., [24, 23, 22]. In a nutshell, an *e-variable*, also *e-statistic* or *e-value*, is a non-negative random variable S with expectation at most one under the null, and large realizations of e-variables give evidence against the null. The latter is easily seen, since, for every distribution P_0 in the null hypothesis, we have $P_0(S \geqslant 1/\alpha) \leqslant \alpha$, for any $\alpha \in (0, 1)$, as a direct consequence of Markov's inequality. E-variables allow for a monetary interpretation as bets against the null [27], and may be combined by taking averages or products. This simple behavior under combinations makes them especially attractive for real-world problems where we want to combine evidence across different studies or outcomes, and accumulate statistical evidence sequentially over time. For a comprehensive introduction to the theory of e-values, we refer to the recent monograph by [23].

In this paper, we study and propose various e-variables and e-processes for analyzing 2×2 contingency tables, a problem that dates back as far as the field of statistics itself. Such a table is defined by two group sizes, n_a and n_b, counts $Y_g \in \{0, \ldots, n_g\}$ for $g \in \{a, b\}$, where we can write $Y_g = \sum_{i=1}^{n_g} X_i^g$. According to the null hypothesis, the outcomes in both groups are i.i.d. Bernoulli with the same parameter $P(X_i^g = 1) = \theta$, where θ is an unknown nuisance parameter, making the null composite. According to the alternative, the parameter varies across the groups: outcomes within each group g are i.i.d. Bernoulli with parameter θ_g. Surprisingly, despite e-values being a rapidly growing and developing research field, much remains unknown about the design of optimal e-variables and e-processes for this simple yet fundamental problem. At first, this statement may seem to contradict the findings of [85, 30], who have derived the GRO e-variable for a given contingency table and simple given alternative (θ_a, θ_b), and then use it to derive an e-process for data arriving sequentially in small batches for composite alternatives; for the case that data arrive in pairs, they provide encouraging empirical results. Here, GRO means 'growth-optimal,' the standard and best-researched optimality criterion for e-variables. However, when one attempts to use the *Turner e-variable* for analysis of a particular table taken from [86], the results, which we reproduce in Table 4.1 below, are very disappointing. The crucial differences in the setting in which Turner performed well were that the table was *large*, with *highly different* group sizes, and that it was observed *one shot*, rather than as a sequence of small batches, such as pairs. The 43 other tables considered by [86] yielded qualitatively similar results to those in Table 4.1; for brevity, we omit them.

In this example, the p-value from Fisher's exact test is 0.00282, strongly suggesting that the outcomes depend on group membership. Ideally, we want to draw the same conclusion if we base our analysis on e-values. However, a naive calculation of Turner's e-variable for the same data gives an uninformative 1 (one), whereas the non-naive mini-batching approach suggested by [30] (and explained in detail in the next sections) yields a non-conclusive small value of 4.11 (usually we perceive e-values larger than 10 or 20 as conclusive). Upon closer analysis, we find that the problem is that the Turner approach to deal with composite alternatives, although quite a standard one,

	a	b
1	192	674
0	506	2370

Table 4.1: Data provided by [86] for $n_a = 698$ participants in group a, and $n_b = 3044$ participants in group b. The groups a and b indicate a certain difference in the genotype of the participants, and the outcomes indicate whether some studied phenotype was present (1) or absent (0).

Method	E-value
Turner (naive)	1
Turner (split)	4.11
UI (mixture)	0.0687
Fisher f-calibrated	10.087
Fisher g-calibrated	17.83
Conditional (mixture)	7.83
Conditional (UMP-2,α =0.05)	37.33
Conditional (UMP-2,α =0.01)	42.92
Conditional (NML)	2.30

Table 4.2: Different e-values for the data in Table 4.1.

does not combine well at all with the highly skewed group sizes (group b is more than four times larger than group a) of the given table, as we will explain in Section 4.3. Besides Turner's e-value, there are two further well-known available approaches to compute e-variables for the given problem: The *universal inference (UI) e-variable* proposed by [25], and the e-variable obtained by calibrating Fisher's exact p-value into an e-value by a so-called *p-to-e-calibrator*, which maps any p-value to an e-value, see [26]. The UI e-variable is easy to compute and very broadly applicable, but also known to be overly conservative, as it also turns out to be the case in our example, where we obtain a useless value of 0.0687 (see Table 4.2). There exist many valid p-to-e calibrators, and there are no guidelines as to which one to use in which situation, so we consider just two standard ones: the functions $g(p) := 1/\sqrt{p} - 1$ and $f(p) = (1 - p + p \log p)/(p(-\log p)^2$. $g(p)$ turns Fisher's exact $p = 0.00282$ into an e-value of 17.83, which seems quite reasonable. However, $g(p)$ has the property that it could vanish to 0 with non-zero probability, and as such, we do not advocate its use. To see why, recall that a central motivation for using e-values instead of p-values lies in their behaviors for sequential problems where we can multiply e-values over time to accumulate evidence from various outcomes and studies. Suppose, for example, that the study underlying Table 4.1 was carried out in several stages, each leading to a batch of data summarized by its own sub-table, and that the decision to continue with the study would have depended on the strength of evidence in the batches seen so far. Then it would not have been allowed to reuse all data at the final step to compute Fisher's exact p-value based on the entire table, since part of the data had

already been used for the intermediate inference. In contrast, providing a final e-value as the product of the e-values for all individual batches and rejecting if it exceeds the significance level α does provide a valid test. However, if one of the batches had been small, then with non-negligible probability the calibrator $g(p)$ would have given us an e-variable of 0 for that batch. Then the final product would have become 0 as well, rendering g useless (or at best: highly risky) in such a standard optional continuation scenario.

For that reason, we prefer the other standard calibrator f (which cannot reach 0), which provides an e-value of $f(0.00282) = 10.087$. The value of 10.087 serves as a benchmark in this example: for at least two reasons, we aim to find better (more powerful) e-values than the one obtained by calibration. The first reason is that the calibration operation is 'blind' and does not make use of the underlying setting — one would hope that knowledge of the setting/model should allow one to design improved e-variables. The second reason is that, while useful, we can predict it will be far from optimal in the optional continuation scenario described above. There, one can, of course, compute the p-values $p_1, \ldots, p_T$ for each batch, calibrate each into an e-value $f(p_1), \ldots, f(p_T)$, and report their product as the total summary statistic. However, such an approach is undesirable, since we do not include any potentially valuable information from the first $j-1$ studies in the computation of an e-value for the j-th study. This stands in contrast to the approach proposed by [85], which, as is possible for most e-variables designed for specific problems, incorporates implicit *learning* of the parameters, allowing us to leverage valuable information from past batches in the e-variable designed for the present one. However, as we already remarked, Turner's method does not yield satisfactory results in the table above either.

4.1.1 Novel methods considered in this paper

Here, we provide *conditional* e-values for analyzing contingency tables that are easy to implement and powerful. By 'powerful', we mean that the e-values should (with high probability) be larger than the calibrated p-value based on Fisher's exact test if the alternative hypothesis of association is true. By 'easy to compute,' we mean that resource-intensive operations, such as a loop over all contingency tables with given marginal counts, should be avoided, a requirement that applies when many e-values must be computed simultaneously, as in genome-wide association studies. Recently, [28] and [29] have studied optimal e-values in situations where the null and alternative hypotheses are given by exponential families sharing the same sufficient statistic(s). Recall that, by definition, conditioned on the sufficient statistic, the distribution of the data does not depend on the unknown parameter of interest anymore; that is, all relevant information about the parameter is contained in the sufficient statistic. The authors use this characterization of sufficiency to study *conditional e-values*, in which they condition on the sufficient statistic to convert the composite null to a simple hypothesis. Conditional e-values are attractive alternatives to GRO e-values since they are easy to compute and asymptotically optimal in the sense that the difference in expected growth in comparison to the theoretically optimal GRO e-variable grows sub-linearly (logarithmically). This paper builds on the fact that the null hypothesis

of no association may be represented as a one-dimensional exponential family with the total number of successes as a sufficient statistic, and that, conditioned on the number of ones, the null becomes simple.

Conditional e-variables are generalizations of conditional likelihood ratios; specifically, instantiated to our contingency table model, they generalize the conditional likelihood ratio for sequences of paired data as used already by Wald in his sequential probability ratio test for two proportions [87] and the conditional Bayes factors used by, for example, [88]. Indeed, what we call the conditional (mixture) e-variable in Table 4.2 is just such a conditional Bayes factor (though we extend it in a non-standard way to the sequential setting). The other versions of our conditional e-variables have, to the best of our knowledge, never been considered before. An initial study of a similar type of conditional e-variable, slightly different from the ones we consider here but that served as an inspiration for the present one, was applied, under the name *microcanonical e-variable* in Chapter 3.

Table 4.2 presents the conditional e-values in three versions, as discussed in what follows, for the data example in comparison with the previously discussed e-values. We see that one particular version, *UMP*-2, performs remarkably well on our table. UMP-2 stands for 'the 2-sided extension of the uniformly most powerful conditional Bayes factor', an e-variable construction that will be explained in the next section. We have calculated the e-variable types listed in Table 4.2 on various other tables as well, and consistently find that UMP-2 shows the best behavior, which is in accord with the heuristic analysis we give in Section 4.3. Thus, we conclude that it is the method of choice for the one-shot scenario, where we receive a single table 'thrown at us' and must analyze it.

Our findings are more intricate in the sequential setting, where batches (tables, potentially of different sizes and group size ratios) from the same source arrive sequentially. In this setting, by its very construction, UMP-2 will suffer from the same defect as the p-to-e calibrators: when data comes in batches, it does not allow one to improve the e-variable for the next batch based on previous batches. The NML conditional e-variables do allow for this behavior, and soon start to outperform UMP as the number of batches increases. Since they are harder to implement in the sequential setting, in this preliminary study, we have not experimented with the mixture conditional e-variables; we expect them to behave very similarly to the NML ones. At the same time, though, we find that the Turner e-variables generally perform better as the number of batches increases — like the NML and mixture e-variables, they adapt to the data-generating distribution as more data comes in. In a nutshell, we find that *asymptotically*, in all our experiments, the Turner e-variable outperforms the conditional NML one; but for reasonable sample sizes and reasonable targets (i.e., rejection thresholds $1/\alpha$), the NML one either is substantially better, or it is slightly (but never a lot) worse.

Thus, based on this preliminary study, our final recommendation for the case of sequentially arriving batches will be to handle the first batch (if it is not too small) using conditional UMP, and subsequent batches using the conditional NML approach, and return as evidence the product of all these e-variables. We hasten to add that the preliminary status of these findings means they should be interpreted cautiously: since

the results are not entirely clear-cut, with the Turner e-variable sometimes slightly outperforming the NML one, we aim to perform several further experiments before we submit the present manuscript to a journal.

We also want to stress that we focus on the case where (a) under the null, $\theta_a = \theta_b$, and (b) the alternative is the entire set of $(\theta_a, \theta_b) \in (0,1)^2$ with $\theta_a \neq \theta_b$ — thus there is no minimum relevant effect size, or a priori knowledge that the discrepancy between θ_a and θ_b has a certain minimum size. This is in contrast to [85], who extend the Turner e-variables to deal with nulls in which $\theta_a \neq \theta_b$ and minimum effect sizes; one can certainly extend the conditional e-variables to this setting as well, but we do not consider these extensions in this paper.

The remainder of this paper is structured as follows: in section 4.2, we introduce notation and all e-values of interest in a rigorous mathematical manner. In Section 4.3, we explain the results underlying Table 4.2, and more generally, the one-shot case, in detail. In Section 4.4, we compare the performance of the proposed e-values in simulations when batches (sub-tables) arrive sequentially. We end with a conclusion.

4.2 Preliminaries and notation

We consider binary data from two independent groups, denoted by a and b, and suppose that, in each stream, data are i.i.d. Bernoulli with parameters θ_a and θ_b, respectively, i.e.

$$X_1^a, X_2^a, \ldots \overset{\text{i.i.d.}}{\sim} \text{Ber}(\theta_a) \quad \text{and} \quad X_1^b, X_2^b, \ldots \overset{\text{i.i.d.}}{\sim} \text{Ber}(\theta_b), \tag{4.1}$$

for $(\theta_a, \theta_b) \in \Theta = (0,1)^2$. We are interested in testing the null hypothesis of no association

$$\mathcal{P} : \theta_a = \theta_b = \theta_0, \quad \theta_0 \in (0,1). \tag{4.2}$$

We abbreviate the Bernoulli distribution with parameter θ_a by P_{θ_a}, and let $P_{(\theta_a,\theta_b)}$ be the distribution of a bivariate Bernoulli vector with independent components and parameters θ_a, θ_b. We denote by p_{θ_a} and $p_{(\theta_a,\theta_b)}$ the corresponding densities, and use the same notation for multiple outcomes, naturally extended by the i.i.d. assumption. We write the null hypothesis at (4.2) equivalently as $\mathcal{P} = \{P_{(\theta_0,\theta_0)} \mid \theta_0 \in (0,1)\}$.

4.2.1 E-variables for a single batch/table

Recall that an *e-variable* for $\mathcal{P}$ is a non-negative test statistic S such that $\mathbb{E}_P(S) \leqslant 1$, for all $P \in \mathcal{P}$. In the following, we will introduce various e-variables for the testing problem at hand. For the sake of clarity, we assume, first, that we have a single batch of data available and explain in the next subsection how to derive e-processes and compute e-values if data arrives in batches over time.

Assume then that we observe a batch of data consisting of observations $X_1^k, \ldots, X_{n_k}^k \in \{0,1\}$, for $n_k \geqslant 1$ and $k = a, b$. We denote by $Y_a = \sum_{i=1}^{n_a} X_i^a$ and $Y_b = \sum_{i=1}^{n_b} X_i^b$ the number of ones (successes) in group a and b, respectively, and let $n = n_a + n_b$, and $N_1 = Y_a + Y_b$.

The Turner E-Variables For arbitrary fixed $(\theta_a, \theta_b) \in \Theta$, [30] show that

$$S_{\mathrm{GRO}(\theta_a,\theta_b)} = \frac{p_{\theta_a}((X_i^a)_{i=1}^{n_a})}{p_{\theta_0}((X_i^a)_{i=1}^{n_a})} \cdot \frac{p_{\theta_b}((X_i^b)_{i=1}^{n_b})}{p_{\theta_0}((X_i^b)_{i=1}^{n_b})}, \quad \text{for } \theta_0 = \frac{n_a}{n}\theta_a + \frac{n_b}{n}\theta_b, \tag{4.3}$$

is the GRO e-variable with respect to the simple alternative $Q = \{P_{(\theta_a,\theta_b)}\}$, see [30, Theorem 1]. Note that

$$S_{\mathrm{GRO}(\theta_a,\theta_b)} = s_{\mathrm{GRO}(\theta_a,\theta_b)}(Y_a, Y_b) \tag{4.4}$$

for some fixed function s, i.e. the e-value depends on the original outcomes only through their counts. However, often we do not want to test $\mathcal{P}$ against the simple alternative $Q = \{P_{(\theta_a,\theta_b)}\}$ but rather against the full alternative

$$\mathcal{Q} = \{P_{(\theta_a,\theta_b)} \mid (\theta_a, \theta_b) \in \Theta, \theta_a \neq \theta_b\}. \tag{4.5}$$

For this case, [30] propose to consider a prior W over Θ under which θ_a and θ_b are independent. Thus, W decomposes into sub-priors W_a, W_b such that $\theta_a \sim W_a$ and $\theta_b \sim W_b$ are independent. We can then write $P_{W_g} = \int P_{\theta'_g} \, \mathrm{d}W_g(\theta'_g)$ for $g \in \{a, b\}$. By convexity of the Bernoulli model, we have $P_{W_g} = P_{\bar{\theta}_g}$ for some $\bar{\theta}_g$, for $g \in \{a, b\}$. [30] then calculate the GRO e-variable relative to this simple alternative $(\bar{\theta}_a, \bar{\theta}_b)$. We denote the corresponding *Turner e-variable* by $S_{\mathrm{TUR}(W)}$. In case $n_a = n_b = 1$, this coincides with the GRO e-variable relative to W, but in all other cases, as long as W is non-degenerate, it does not, since if $n_g > 1$ then $P_{W_g}(X_1^g, \ldots, X_{n_g}^{n_g}) \neq P_{\bar{\theta}_g}(X_1^g, \ldots, X_{n_g}^{n_g})$; see [30] for more extensive discussion. [30] suggests using a Beta prior for W, and we use the uniform Beta prior unless otherwise stated.

Universal Inference In the given problem, a standard version of the *Universal Inference (UI) e-variable* proposed by [25] with respect to the fixed alternative $\{P_{(\theta_a,\theta_b)}\}$ is given by

$$S_{\mathrm{UI}(\theta_a,\theta_b)} = \frac{\theta_a^{Y_a}(1-\theta_a)^{n_a - Y_a}\theta_b^{Y_b}(1-\theta_b)^{n_b - Y_b}}{\hat{\theta}^{N_1}(1-\hat{\theta})^{N_1 - y}},$$

where $\hat{\theta}$ denotes the MLE for the given data with respect to the binomial distribution with sample size n_1. For the general alternative $\mathcal{Q}$ and general prior W (in this case, θ_a and θ_b do not need to be independent under W), we report the mixture e-value $S_{\mathrm{UI}(W_1)} = \int S_{\mathrm{UI}(\theta'_a,\theta'_b)} \, \mathrm{d}W(\theta'_a, \theta'_b)$.

Conditional E-Variables As mentioned in the introduction, the Turner e-variable as well as the UI e-variable may perform poorly (i.e., remain small) for some tables that intuitively provide strong evidence against the null. Recently, [28] and [29] have introduced *conditional e-values* as an attractive alternative to the theoretically optimal GRO e-value for general exponential families. Clearly, the hypothesis $\mathcal{P}$ at (4.2) forms a one-dimensional exponential family with parameter θ_0, and the hypothesis $\mathcal{Q}$ at (4.5) forms a two-dimensional exponential family with parameters (θ_a, θ_b). For the given data, the summary statistic $N_1 = Y_a + Y_b$ is sufficient for $\mathcal{P}$, and the vector

(N_1, Y_a) is sufficient for $\mathcal{Q}$. Moreover, under the null, N_1 is Binomially distributed with parameters n and θ_0, and, under the alternative, Y_k is binomially distributed with parameters n_k and θ_k, for $k = a, b$. We denote the probability mass functions of the sufficient statistic Y_a by f_{Y_a}, and refer to [66] for an extensive study of exponential families. Next, we will condition on $N_1 = n_1$, for some $n_1 \leqslant n_a + n_b$. It is well-known, and may be verified by standard calculation, that, given $N_1 = n_1$, Y_a follows a hypergeometric distribution under the null, that is

$$f_{Y_a}(y \mid N_1 = n_1) = \frac{\binom{n_a}{y}\binom{n_b}{n_1 - y}}{\binom{n}{n_1}}, \quad y \in \{\gamma^-, \dots, \gamma^+\},$$

for $\gamma^- = \max\{0, n_1 - n_b\}$ and $\gamma^+ = \min\{n_1, n_a\}$. Crucially, the density $f_{Y_a}(\cdot \mid N_1 = n_1)$ is independent of θ_0 by sufficiency. Similarly, for general $(\theta_a, \theta_b) \in \Theta$, Y_a, given $N_1 = n_1$, follows a *Fisher's noncentral hypergeometric distribution*, that is

$$f_{Y_a}(y \mid N_1 = n_1) = \frac{\binom{n_a}{y}\binom{n_b}{n_1 - y} e^{\delta y}}{\sum_{k=\gamma^-}^{\gamma^+} \binom{n_a}{k}\binom{n_b}{n_1 - k} e^{\delta k}}, \quad y \in \{\gamma^-, \dots, \gamma^+\}, \tag{4.6}$$

where δ denotes the log odds-ratio,

$$\delta = \log\left(\frac{\theta_a}{1 - \theta_a} \cdot \frac{1 - \theta_b}{\theta_b}\right) \in \mathbb{R}. \tag{4.7}$$

Note that the hypergeometric distribution results as a special case for $\delta = 0$, which holds if, and only if, $\theta_a = \theta_b$. For fixed $n_a, n_b \geqslant 1$, $n_1 \leqslant n_a + n_b$, and $\delta \in \mathbb{R}$, we denote the density of Fisher's non-central hypergeometric distribution with parameters n_a, n_b, n_1 and δ, by $f_{\mathrm{FN}(n_a, n_b, n_1, \delta)}$. Since n_a and n_b, as well as the number of ones n_1 (by conditioning) are assumed to be given, we abbreviate $f_{\mathrm{FN}(n_a, n_b, n_1, \delta)}$ simply by $f_{\mathrm{FN}(\delta)}$ whenever n_a, n_b, and n_1 are clear from the context. As seen from (4.6), for fixed n_a, n_b, n_1, the set of distributions corresponding to $f_{\mathrm{FN}(\delta)}$ themselves constitute a 1-dimensional exponential family with canonical parameter δ. The *conditional e-value* is defined as

$$S_{\mathrm{c}(\delta)}(Y_a) = \frac{f_{\mathrm{FN}(n_a, n_b, n_1, \delta)}(Y_a)}{f_{\mathrm{FN}(n_a, n_b, n_1, 0)}(Y_a)} = \frac{f_{\mathrm{FN}(\delta)}(Y_a)}{f_{\mathrm{FN}(0)}(Y_a)}, \quad \delta \in \mathbb{R}. \tag{4.8}$$

Lemma 4.2.1.
For any $\delta \in \mathbb{R}$, the conditional e-value $S_{\mathrm{c}(\delta)}$ at (4.8) is an e-value for $\mathcal{P}$.

Proof. For any $P \in \mathcal{P}$, Y_a, given $N_1 = n_1$, is hypergeometrically distributed, and thus

$$\mathbb{E}_P(S_{\mathrm{c}(\delta)} \mid N_1 = n_1) = \sum_{y=\gamma^-}^{\gamma^+} f_{\mathrm{FN}(0)}(y) S_{\mathrm{c}(\delta)}(y) = \sum_{y=\gamma^-}^{\gamma^+} f_{\mathrm{FN}(\delta)}(y) = 1.$$

The claim follows from the tower property and the fact that $S_{\mathrm{c}(\delta)}$ is non-negative.

By conditioning, we have reduced the dimensionality of the alternative parameter space, and $S_{\mathrm{c}(\delta)}$ is an e-value for any $\delta \in \mathbb{R}$. Analogously to the unconditional case, we usually do not know a priori against which particular δ we should test $\mathcal{P}$. Thus, we discuss three possible choices in the sequel.

1. *Mixture conditional e-variable.* Given a prior W on $\mathbb{R}$, it is natural to consider the *mixture conditional e-variable*

$$S_{\mathrm{c}(W)}(Y_a) = \int S_{\mathrm{c}(\delta)}(Y_a)\,\mathrm{d}W(\delta). \tag{4.9}$$

   Note that this e-variable can be thought of as a 'conditional' Bayes factor, and as such it has appeared before in the Bayesian literature: [88] and [89] give closed form expressions for the conjugate prior on δ for Fisher's non-central hypergeometric distribution. Other natural choices for W include some Gaussian prior or Jeffreys' prior.

2. *NML conditional e-variable.* Ideally, we would like to choose δ such that the enumerator in (4.8) is as large as possible, which is achieved by the MLE, denoted by $\hat{\delta}(Y_a)$, for the given data. However, $S_{\mathrm{c}(\hat{\delta}(Y_a))}$ is not an e-variable in general, since usually $\mathbb{E}_P(S_{\mathrm{c}(\hat{\delta}(Y_a))}) \gg 1$ for $P \in \mathcal{P}$. Nevertheless, by appropriately scaling, we obtain the *normalized maximum likelihood (NML)* conditional e-variable

$$S_{\mathrm{c(NML)}}(Y_a) = S_{\mathrm{c}(\hat{\delta}(Y_a))}(Y_a) \cdot \frac{1}{\sum_{k=\gamma^-}^{\gamma^+} f_{\mathrm{FN}(\hat{\delta}(k))}(k)}, \tag{4.10}$$

   where $\hat{\delta}(k)$ denotes the MLE for $f_{\mathrm{FN}(n_a,n_b,n_1,\delta)}$ with outcome $k \in \{\gamma^-, \ldots, \gamma^+\}$. The normalized maximum likelihood approach has been derived within the *Minimum Description Length (MDL)* approach to statistical model selection and prediction [18, 17] and often combined with some *luckiness function*, where we replace $\hat{\delta}(k)$ and $\hat{\delta}(Y_a)$, with the MLEs with respect to the modified data where $n'_a = n_a + k_a, n'_b = n_b + k_b$, $n'_1 = n_1 + k_1$, for some $k_a, k_b, k_1 \geqslant 0$ with the default choice $k_a = k_b = k_1 = 1$. Classic results in MDL theory for unconditional settings with simple nulls indicate that, for large enough sample sizes, and under regularity conditions, which are met in our case, the log Bayes marginal likelihood based on Jeffreys' prior will become equal, up to $o(1)$, to the log NML e-variable. Although these classic results do not cover the present conditional setting, they suggest that NML and Bayesian methods will behave similarly.

3. *UMP e-variable.* Finally, since $f_{\mathrm{FN}(\delta)}(\cdot)$ constitutes a 1-dimensional exponential family, we can use a result of [90] which states that, for every pre-specified significance level $\alpha \in (0,1)$, among all tests of our usual null $\delta = 0$ against the *one-sided* alternative $H_1 : \delta > 0$ that take the form $\mathbf{1}\{S_{\mathrm{c}(\delta)} \geqslant 1/\alpha\}$ for some $\delta \in \mathbb{R}$, the choice $\delta = \delta^+_{\mathrm{UMP}}$ is *uniformly most powerful*, where $\delta^+_{\mathrm{UMP}} > 0$ is given by

$$D_{\mathrm{KL}}\Big(f_{\mathrm{FN}(\delta^+_{\mathrm{UMP}})} \,\Big\|\, f_{\mathrm{FN}(0)}\Big) = \log(1/\alpha), \tag{4.11}$$

   and the results hold as long as $\delta^+_{\mathrm{UMP}} > 0$ satisfying the above display exists; it will then also be unique. Existence is guaranteed if $\min\{n_a, n_b\}$ is larger

than some minimal n^* [90, 91]. Even though there is no closed-form expression for δ^+_{UMP} available, we may solve (4.11) efficiently numerically by convexity in δ. Similarly, if (4.11) holds with δ^+_{UMP} replaced by some δ^-_{UMP}, then the test $\mathbf{1}\{S_{\mathrm{c}(\delta^-_{\mathrm{UMP}})} \geqslant 1/\alpha\}$ is UMP among all tests of above form against alternative $H_1 : \delta < 0$. If one is interested, like we are (since we always consider *every* distribution $(\theta_a, \theta_b) \in (0,1)^2$ with $\theta_a \neq \theta_b$ as potential alternative), in a two-sided test, this suggests to take, in analogy to the standard construction of a two-sided test, the e-variable $S_{\mathrm{UMP}-2} = (1/2)S_{\mathrm{c}(\delta^-_{\mathrm{UMP}})} + (1/2)S_{\mathrm{c}(\delta^+_{\mathrm{UMP}})}$. We call this the UMP-2 e-variable, even though it formally does not have UMP status anymore.

4.2.2 Sequential (conditional) e-variables

We now assume that data arrives in batches $(B_t)_{t\in\mathbb{N}}$ over time, and that, at each time point $t \in \mathbb{N}$, we observe batch B_t consisting of $n_{a,t}$ observations in group a and $n_{b,t}$ observations in group b. We denote by $Y_{a,t}$ and $Y_{b,t}$ the number of ones at $t \in \mathbb{N}$ in group a and b, respectively, and let $n_t = n_{a,t} + n_{b,t}$, and $N_{1,t} = Y_{a,t} + Y_{b,t}$. Similarly, we let $n^t_k = \sum_{\ell=1}^t n_{k,\ell}$, and $Y^t_k = \sum_{\ell=1}^t Y_{k,\ell}$, for $k = a, b$, and $n^t = n^t_a + n^t_b$, and $N^t_1 = Y^t_a + Y^t_b$. For $t \in \mathbb{N}$, we summarize the data B_t *at* t, and the total data $B^t = (B_1, \ldots, B_t)$ *until* t in tables as given in Tables 4.3 and 4.4.

outcome	a	b	
1	$Y_{a,t}$	$Y_{b,t}$	$n_{1,t}$
0	$n_{a,t} - Y_{a,t}$	$n_{b,t} - Y_{b,t}$	$n_{0,t}$
	$n_{a,t}$	$n_{b,t}$	n_t

Table 4.3: Data B_t at $t \in \mathbb{N}$.

outcome	a	b	
1	Y^t_a	Y^t_b	n^t_1
0	$n^t_a - Y^t_a$	$n^t_b - Y^t_b$	n^t_0
	n^t_a	n^t_b	n^t

Table 4.4: Data B^t until $t \in \mathbb{N}$.

Throughout the paper, we consider the natural filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$, where $\mathcal{F}_t = \sigma(B^t)$ for $t \in \mathbb{N}$. It was argued in the introduction that e-values are particularly useful in sequential analysis where we want to combine evidence from different studies (tables, batches) over time. We call a sequence $(S_t)_{t\in\mathbb{N}}$ of e-variables for $\mathcal{P}$ *sequential* (with respect to the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$), if $(S_t)_{t\in\mathbb{N}}$ is adapted and, for all $P \in \mathcal{P}$,

$$\mathbb{E}_P(S_t \mid \mathcal{F}_{t-1}) \leqslant 1, \quad t \in \mathbb{N}. \tag{4.12}$$

Given some sequence of sequential e-values $S_1, S_2, \ldots$, we consider their cumulative products

$$M_t = \prod_{\ell=1}^t S_\ell, \quad t \in \mathbb{N}. \tag{4.13}$$

It follows directly that $(M_t)_{t\in\mathbb{N}}$ is a nonnegative super martingale under any $P \in \mathcal{P}$, and that it satisfies thus *Ville's inequality* [92], claiming that, for any $P \in \mathcal{P}$,

$$P(\exists t : M_t \geqslant 1/\alpha) \leqslant \alpha, \quad \text{for all } \alpha \in (0,1). \tag{4.14}$$

Ville's inequality lays the foundation for safe anytime-valid inference [22]. It may be seen as a time-uniform generalization of Markov's inequality, and is stronger than $P(M_t \geqslant 1/\alpha) \leqslant \alpha$ for all $t \in \mathbb{N}$, see e.g. [93] for a discussion. Any sequence of sequential e-values thus suggests the natural sequential tests $\Psi_t = \mathbf{1}\{M_t \geqslant 1/\alpha\}$, $t \in \mathbb{N}$, where we reject the null $\mathcal{P}$ at level $\alpha \in (0,1)$ as soon as M_t exceeds the threshold $1/\alpha$, and, by (4.14), one may upper bound the probability of ever doing a type-I error by α.

Sequential Turner - Simple Alternative For the simple alternative (θ_a, θ_b), the GRO e-variable for each block is given by (4.3), which takes the form of a likelihood ratio between the alternative and a particular element of the null. It then easily follows [94] that the GRO e-variable for any sequence of blocks is given by the product of the individual e-variables (4.3) for each block, i.e. $M_t = \prod_{\ell=1}^{t} s(Y_{a,\ell}, Y_{b,\ell})$, with s given by (4.4), is the GRO e-variable for data B^t.

Sequential Turner - Composite Alternative For the full composite alternative, we may turn the Turner e-variable into a sequential sequence of e-variables as follows: we start with a prior W_1 for the first batch and calculate $S_{\mathrm{TUR}(W_1)}$. We then, for the t-th batch, use the Turner e-variable with prior W_t set equal to the posterior $W_1 \mid B_1, \ldots, B_{t-1}$ based on all previously observed data.

We note a few properties of this procedure:

1. *No Within-Batch Learning.* The posterior is updated only after a new batch arrives. Therefore, if batch B_t has large size, the Turner e-value for that batch is a likelihood ratio in which all outcomes in group a within that batch are predicted by the same Bernoulli distribution, i.e., for $g \in \{a, b\}$, every such outcome x_i^g appearing in B_t appears as a factor $p_{\bar{\theta}}(x_i^g)$ in the likelihood for the same $\bar{\theta}$ that only depends on outcomes in past batches $B_1 \ldots, B_{t-1}$.

2. *Sequential GRO in Special Cases.* If all batches B_t are pairs, $n_{a,t} = n_{b,t} = 1$, and the prior W_1 renders θ_a and θ_b independent, then so does, for each t, the posterior W_t that is used as prior in batch B_t. As a result, following the reasoning of the previous section, the t-th sequential Turner e-variable $S_{\mathrm{TUR}(W_t)}$ is the W_t-GRO e-variable and the resulting martingale $(M_t)_{t\in\mathbb{N}}$ is *sequential* W_1-*GRO*. [29] shows that, for general exponential families, the growth rate of this sequential version of GRO differs from the ('overall' optimal) GRO by $c \log n$ for some small ($< 1/2$) positive constant c.

Sequential conditional We suggest using *sequential conditional e-values* defined as

$$S_{\mathrm{c}(\delta_t),t}(Y_{a,t}) = \frac{f_{\mathrm{FN}(n_{a,t}, n_{a,t}, N_{1,t}, \delta_t)}(Y_{a,t})}{f_{\mathrm{FN}(n_{a,t}, n_{b,t}, N_{1,t}, 0)}(Y_{a,t})}, \quad t \in \mathbb{N}, \delta_t \in \mathbb{R}, \tag{4.15}$$

for some predictable sequence $(\delta_t)_{t\in\mathbb{N}} \subseteq \mathbb{R}$. Importantly, the conditional requirement at (4.12) allows us to learn the true log-odds ratio δ adaptively, and thereby increasing

power in case that the alternative is true. For the mixture approach, this suggests considering

$$S_{\mathrm{c}(W_t),t} = \int S_{\mathrm{c}(\delta),t}(Y_{a,t})\,\mathrm{d}W_t(\delta), \tag{4.16}$$

for some predictable sequence $(W_t)_{t\in\mathbb{N}}$ of priors over $\mathbb{R}$, where the natural choice is to choose W_t as the posterior for δ after having seen the data B^{t-1}. Note that, while the individual e-variables have a conditional Bayes factor interpretation, this is not so clear for their product: the conditional Bayes factor for data consisting of a series of batches $B_1, \ldots, B_T$ is not equal to the product of conditional Bayes factors for each separate batch B_t with prior W_t set to the posterior based on B^{t-1} (such a coherence results only holds for *unconditional* Bayes factors). Also, for the NML approach, we want to predictably update the MLE, which suggests using

$$S_{\mathrm{c(NML)},t} = S_{\mathrm{c}(\hat{\delta}(Y_a^t)),t}(Y_{a,t}) \cdot \frac{1}{\sum_{k=\gamma_t^-}^{\gamma_t^+} f_{\mathrm{FN}(n_{a,t},n_{b,t},N_{1,t},\hat{\delta}(k))}(k)}, \tag{4.17}$$

where $\gamma_t^- = \max\{0, N_{1,t} - n_{b,t}\}$ and $\gamma_t^+ = \min\{N_{1,t}, n_{a,t}\}$, and $\hat{\delta}(Y_a^t)$, and $\hat{\delta}(k)$ denote the MLE for the full data B^t with respect to $f_{\mathrm{FN}(n_a^t,n_b^t,N_1^t,\delta)}$ for Y_a^t and $Y_a^{t-1} + k$, respectively.

Finally, we note that the UMP-based approach presented in the previous subsection is not well-suited for a sequential application, since the optimal log-odds ratio would be computed for each batch independently, and there is no learning of the alternative between batches. That is, we actually would obtain independent e-variables rather than sequential e-variables, which are expected to perform worse in the long run.

Notation	Description
$S_{\mathrm{GRO}(\theta_a,\theta_b)}$	GRO e-variable wrt $\mathcal{Q} = \{P_{\theta_a,\theta_b}\}$
$S_{\mathrm{TUR}(W)}$	Turner e-variable with prior W
$S_{\mathrm{c}(\delta)}$, $S_{\mathrm{c}(W)}$, $S_{\mathrm{c(NML)}}$	Cond. e-value for $\delta \in \mathbb{R}$, with prior W, and cond. NML e-value
B_t (B^t)	Data entering at (until) time t
$n_{a,t}, n_{b,t}$ (n_a^t, n_b^t)	Group sizes at (until) time t
$Y_{a,t}, Y_{b,t}$ (Y_a^t, Y_b^t)	Total number of ones in group a and b at (until) time t
$N_{1,t}$ (N_1^t)	Total number of ones in both groups at (until) time t
$S_{\mathrm{c}(\delta),t}$ $(S_{\mathrm{c}(\delta)}^t)$	Conditional e-value for $\delta \in \mathbb{R}$ at (until) time t
$S_{\mathrm{c}(W),t}$ $(S_{\mathrm{c}(W)}^t)$	Conditional e-value wrt some prior W on $\delta \in \mathbb{R}$ at (until) time t
$S_{\mathrm{c(NML)},t}$ $(S_{\mathrm{c(NML)}}^t)$	NML conditional e-value at (until) time t

Table 4.5: Notation used in the paper.

4.3 Calculating and analyzing e-variables for a single table

In this section, we provide details underlying the calculations that gave rise to Table 4.2 and we analyze the dismal behavior of the Turner e-variable on the data underlying

Table 4.1. To start with the latter: if we naively apply the Turner e-variable to the whole table, considered as a single batch, we get an e-variable of 1. This is a direct consequence of the fact we alluded to before: the Turner e-variable does not facilitate parameter learning *within* a batch. In the notation of Section 4.2.1, [30] use priors $W_{1,a}$ and $W_{1,b}$ that are identical, which implies that we get $\bar{\theta} = \bar{\theta}_a = \bar{\theta}_b$ for some θ, resulting in a likelihood ratio with the alternative represented by $(\theta_a, \theta_b) = (\bar{\theta}, \bar{\theta})$. The corresponding distribution is an element of the null, so the e-variable equals 1. This was, in fact, noticed by [85], and for the case of one-shot large tables B, they implicitly suggested to decompose the table into *mini-batches*, i.e. several small tables $B_1, \ldots, B_t$ whose entries add up to those of B. We tried this for a decomposition into (approximately 700) mini-batches with $n_{a,t} = 1, n_{b,t} = 4$, and that gave rise to the value 4.11 in Table 4.1. In more detail, we computed Turner's sequential e-variable for these individual batches (with priors in each batch equal to the posteriors from the previous batch), and report their product. Additionally, we averaged the results over 5000 runs of random sample splitting (with a standard deviation of 1.90 when drawing without replacement) to obtain a more objective result, as the resulting e-value strongly depends on the particular splitting sample chosen.

The main reason why this e-variable is much inferior to the ones discussed next seems to be that this procedure forces us to throw away data points — at some point, one of the two groups is 'exhausted,' and we are left with a remaining batch of data points that all fall in the same group. These data points (252 data points in group b, for our table) cannot be used (similarly, we could have tried batches with $n_{a,t} = 1, n_{b,t} = 5$ but this forces us to throw away points in group a rather than b and the resulting e-variable still remains small). We tried several other combinations of n_a and n_b, but none of them yielded significantly better e-values. An additional issue is that, in practice, it would not be clear which decomposition to use. Another reason for the bad behavior is that, since in all reasonable decompositions $n_{a,t} \neq n_{b,t}$, the e-values we use do not have GRO status.

The *UI* e-variable behaves even worse than the Turner one. While surprising at first, this finding is consistent with the asymptotic results presented in [29]. Their Theorem 4 (parts 1 and 3) roughly states that, if we were to pick a prior W that puts all its mass on the 1-dimensional curve of all (θ_a, θ_b) that all share the same odds ratio, i.e. that satisfy (4.7) for the same δ, then the expected logarithmic difference between the UI e-variable applied to fixed sample size n and the Turner e-variable, for a total of n batches, is of order $c \log n + O(1)$ for a constant c close to $1/2$ (in the language of their Theorem 4, what we call a 'batch' is a single data point; what we call the Turner e-variable is their *sequential-RIPr* e-variable). We would then expect a ratio between both e-variables of about $\exp(c \log n + O(1)) = c_n \sqrt{n}$, where n is the number of batches, which, due to our setting $n_{a,t} = 1, n_{b,t} = 4$ (see above) was approximately 700 (accounting for a factor of $\sqrt{700} \approx 26$), and c_n tends to a positive unknown constant. The fact that this asymptotic analysis holds for *every* fixed δ, suggests that something similar still holds if, as in our implementation of UI and Turner, we use a prior W spread out over all (θ_a, θ_b) rather than just those corresponding to a particular δ. This independently suggests, but of course does not prove, that (at least this version of) UI is non-competitive.

As regards the *e-to-p calibrators*, the underlying p-value — based on Fisher's exact test — is obtained by conditioning on n_1 and modeling Y_a as a hypergeometric random variable, in close analogy with our conditional e-variables. From this perspective, some degree of similarity in their behavior is to be expected. Nevertheless, the specific choice of the calibrators f and g is not derived from, nor directly connected to, the hypergeometric model. For this reason, we found it somewhat surprising that they exhibit such good performance, at least when compared to the Turner e-variable.

Turning to the *conditional* e-variables, we observe remarkably strong performance of the UMP-2 e-variable. Importantly, although the construction requires fixing an initial significance level α to derive a (powerful) choice of δ, the resulting UMP-2 e-variable remains valid for any alternative significance level $\alpha^* \neq \alpha$. That is, under the null hypothesis, it satisfies

$$P_0\left(S_{\text{UMP-2}} > \frac{1}{\alpha^*}\right) \leqslant \alpha^* \qquad \text{for all } \alpha^* \in (0,1).$$

When $\alpha^* \neq \alpha$, the resulting test may no longer be optimal in terms of power, but it still retains the Type-I error control at level α^*. Moreover, our experiments indicate that the performance of the UMP-2 e-variable is relatively robust with respect to the choice of the initial significance level, provided that α lies within a reasonable range (e.g. $[0.01, 0.1]$).

The two remaining conditional e-variables, although grounded in standard approaches for handling composite alternatives, do not perform particularly well. This is essentially due to the fact that both are designed to achieve an expected logarithmic growth rate that, regardless of the true parameter pair (θ_a, θ_b) generating the data, remains close to that of the optimal *oracle* conditional e-variable, defined in terms of the true value of δ associated with (θ_a, θ_b) via (4.7).

This design choice entails a non-negligible overhead. In the case of the mixture conditional e-variable, the overhead arises because the prior $W(\delta)$ distributes its mass over the entire parameter space. In the conditional NML case, it is induced by the normalization factor appearing in (4.10). In the unconditional setting, the resulting loss in expected logarithmic growth relative to the oracle amounts to $(1/2)\log n + O(1)$, both for NML and mixture e-variables [17] (see also Chapter 3), and we expect a similar behavior to hold in the conditional case.

By contrast, the UMP conditional e-variables circumvent this cost by concentrating essentially all prior mass on values of δ that allow rejection with high probability for the given sample sizes and significance levels α. This advantage, however, comes at the cost of markedly poor performance when α or sample sizes differ substantially.

4.4 Comparison of Turner's and conditional e-variables in sequential settings

We assess performance in a sequential setting from three complementary perspectives: asymptotic optimality, practical evidence accumulation, and stopping-time behavior. To this end, we consider independently distributed data streams for groups a and b,

generated by two Bernoulli processes with fixed oracle parameters θ_a and θ_b, respectively. We evaluate the performance of various sequential e-variables S_t by simulation.

We present three figures, Figures 4.1–4.3, each comprising 16 panels corresponding to simulation settings that differ in the data-generating parameters (θ_a, θ_b), the total batch size $n_a + n_b$, and the group-size ratio n_a/n_b. For simplicity, we assume that all batches within a given simulation have identical sizes and ratios; that is, for each batch B_t, we have $n_{a,t} = n_a$ and $n_{b,t} = n_b$ for some fixed n_a and n_b.

We selected two configurations of the *effect size* δ for the data generating distribution: $\delta = 2\log 19$ (very large effect, rejecting the null is easy, first two rows of the figures) and $\delta = 2\log(3/2)$ (small-to-moderate effect, more difficult to discover, third and fourth row of the figures), which can be spotted by different δs' *distance* to the null, the diagonal line in the parameter space in Figure 4.5. For both these effect sizes, we chose two concrete instantiations: one on the counter diagonal, i.e. $(\theta_a, \theta_b) = (0.95, 0.05)$ corresponding to $\delta = 2\log 19$ (first row of the figures) and $(\theta_a, \theta_b) = (0.6, 0.4)$ corresponding to $\delta = 2\log(3/2)$ (third row of the figures); and one more towards the corner, i.e. $(\theta_a, \theta_b) = (0.78, 0.01)$ corresponding to $\delta = 2\log 19$ (second row of the figures) and $(\theta_a, \theta_b) = (0.16, 0.08)$ corresponding to $\delta = 2\log(3/2)$ (fourth row of the figures). Oracle parameters are visualized as individual dots in Figure 4.5. We also considered small and medium batches with group size ratio 1 (first two columns of the graphs) and with a group size ratio far from 1 (third and fourth columns of the graphs). In this way, each figure consists of 16 graphs, each depicting a separate data-generating mechanism.

Figure 4.1 measures the deviation from the (θ_a, θ_b)-GRO e-variable. This metric is defined as the expected difference

$$\mathbb{E}_{P_{(\theta_a,\theta_b)}}\left[\log \prod_{\ell=1}^{t} s_{\mathrm{GRO}(\theta_a,\theta_b)}(Y_{a,\ell}, Y_{b,\ell}) - \log \prod_{\ell=1}^{t} S_\ell\right] = \\ = \mathbb{E}_{P_{(\theta_a,\theta_b)}}\left[\sum_{\ell=1}^{t} \log \frac{s_{\mathrm{GRO}(\theta_a,\theta_b)}(Y_{a,\ell}, Y_{b,\ell})}{S_\ell}\right], \quad (4.18)$$

with the explicit formula for the GRO given by (4.3) and (4.4), and where S_ℓ denote the sequential e-values of interest. Figure 4.1 provides a comparison for S_ℓ either the sequential Turner $S_{\mathrm{TUR}(W_\ell)}$ or the sequential NML e-variable $S_{\mathrm{c(NML)},\ell}$.

Secondly, we report the expected accumulated evidence over time as a measure of performance. Figure 4.2 depicts the averaged cumulative products of the sequential Turner and sequential NML e-values with increasing number of batches. Here, the red horizontal line indicates the threshold $\log(100) \approx 4.60$, corresponding to rejecting the null at level $\alpha = 0.01$. Intuitively, we expect to reject the null after observing roughly the number of batches in which the expected evidence exceeds this threshold. However, since this reasoning can be misleading in general, we report also the stopping time distribution as a third measure of performance: In Figure 4.3, we compare the stopping times $\tau = \min\left\{t \geqslant 1 \mid \prod_{\ell=1}^{t} S_\ell \geqslant 1/\alpha\right\}$ at level $\alpha = 0.01$ for the sequential Turner and sequential NML e-values. All plots are generated by averaging over $N = 1000$ runs of each simulation.

Conclusions Figure 4.1 clearly illustrates that, asymptotically, the Turner e-variable always wins in the sense of achieving the fastest growth rate. The only panel where this behavior is not immediately apparent is the one in the third row and fourth column; however, we expect that, upon adding a sufficiently large number of additional batches, the Turner e-variable would eventually overtake the conditional NML e-variable, as observed in the other panels.

This behavior can be explained as follows. For each batch, the expected growth of the conditional NML e-variable — even if it were equipped with the oracle value of δ corresponding to the true parameter pair (θ_a, θ_b) — falls short of the expected growth of the optimal (oracle) e-variable by a small amount $\epsilon > 0$. As t increases, the maximum likelihood estimator $\hat{\delta}$ used in the NML procedure at batch t converges to this true value of δ, implying that the relative growth in (4.18) converges, to first order, to $n\epsilon$, for some small $\epsilon > 0$. As shown in [29], ϵ vanishes as the batch size increases; nevertheless, for any fixed batch size, the conditional NML e-variable remains linearly suboptimal relative to the oracle e-variable as the number of batches tends to infinity.

By contrast, the behavior of the Turner e-variable is, in a sense, dual. Consider each batch B_t itself as a sequence of observations $\{X_i^a\}$ and $\{X_i^b\}$. As already discussed in Section 4.2.2, the Turner e-process does not perform any *within-batch learning*: the approximation $(\bar{\theta}_a, \bar{\theta}_b)$ to the true, unknown parameter pair (θ_a, θ_b) used to evaluate the e-variable remains fixed throughout the batch, thereby incurring a constant approximation error for all outcomes *within* that batch.

When a new batch becomes available, however, this approximation is updated and converges to the true data-generating parameters (θ_a, θ_b). Consequently, in contrast to the conditional NML case, the approximation error — corresponding to each term in (4.18) — vanishes as the number of batches ℓ grows. Within any given batch, the approximation error remains non-zero but constant. As a result, for the Turner e-variable, smaller batch sizes are advantageous, a behavior that is consistent with what is observed in the plots.

Figure 4.2 shows that, for large effect sizes (top two rows), the conditional NML e-variable outperforms the Turner e-variable for practically relevant numbers of batches, that is, for total number of observations at which the accumulated evidence is not much larger than $\log(1/\alpha)$ for reasonable choices of $\alpha \in (0, 1)$. Note that the horizontal scale in Figure 4.2 is therefore substantially smaller than that in Figure 4.1, where our focus was on asymptotic behavior.

When the effect size is small (bottom two rows), corresponding to a more challenging testing problem, the comparison becomes less clear-cut, and the two e-variables exhibit very similar performance. Moreover, for large effect sizes, the non-centrality parameter appears to have little influence on performance, as the first two rows look essentially the same. Finally, we observe that the crossover point at which the Turner e-variable overtakes the conditional NML e-variable occurs particularly early for balanced data and small batch sizes, as illustrated by the top two rows of the first column in Figure 4.2. This observation is consistent with our earlier findings that small and balanced batch sizes are favorable for the Turner e-variable.

From Figure 4.3 we can conclude that, for large effect sizes, rejection typically

occurs slightly faster with the conditional NML e-variable than with the Turner e-variable. The difference in the required number of batches, however, is very small, since both methods lead to rejection extremely quickly: across all eight simulations in the first two rows, no more than four batches are ever needed. For smaller effect sizes (bottom two rows), the stopping times associated with the conditional NML e-variable appear to be somewhat more dispersed than those of the Turner e-variable.

Figure 4.4 zooms in on one of the sixteen cases shown in Figure 4.3 and additionally displays the individual trajectories of the conditional NML and Turner e-variables across all simulation runs. The visualization of simulated trajectories together with stopping-time distributions is inspired by the informative plots presented in [95]. From these trajectories, we can clearly observe the potential advantage of the conditional NML e-variable over the Turner e-variable already in the first batch, as well as the fact that, for certain data realizations, the evidence accumulated by the Turner method may drop to very low levels before eventually reaching significance.

4.5 Conclusion

This paper investigates e-values and e-processes for testing 2×2 contingency tables, with a particular focus on the sequential performance of Turner's e-variable and a newly proposed conditional NML e-variable. Our simulation study suggests that, while the Turner e-variable dominates asymptotically in terms of growth rate, the conditional NML e-variable performs competitively at practically relevant sample sizes and target significance levels. In these regimes, the NML-based approach is either substantially better than Turner or only slightly worse. We emphasize that these findings should be interpreted as preliminary. They provide initial evidence on the relative strengths of the methods, but do not yet offer a complete characterization of their performance. A more systematic exploration through additional experiments remains an important direction for future work.

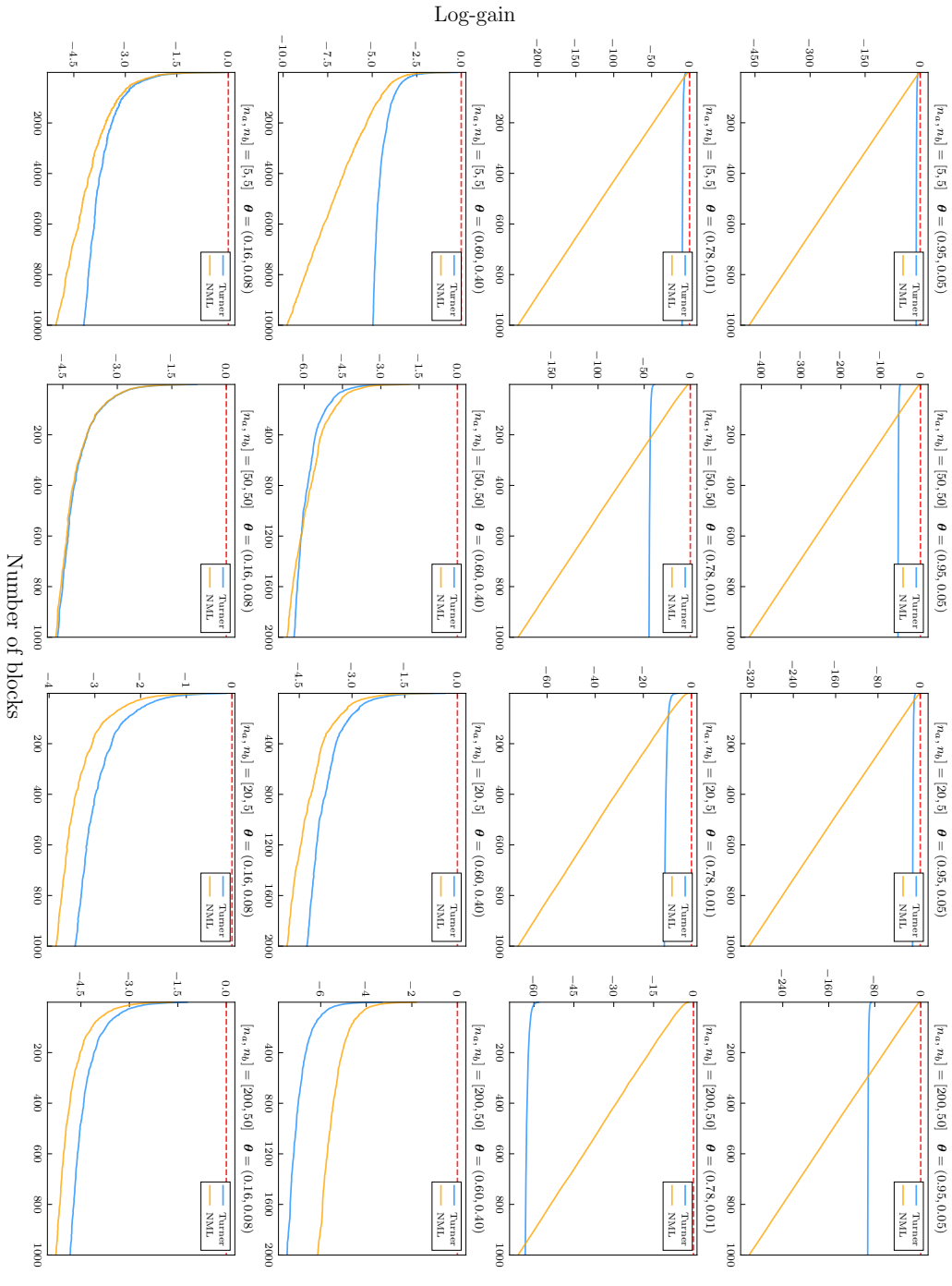


Figure 4.1: Empirical expected (log) gain (as given in (4.18)*) of theoretically optimal GRO in comparison to sequential Turner (blue) and sequential conditional NML (orange).*

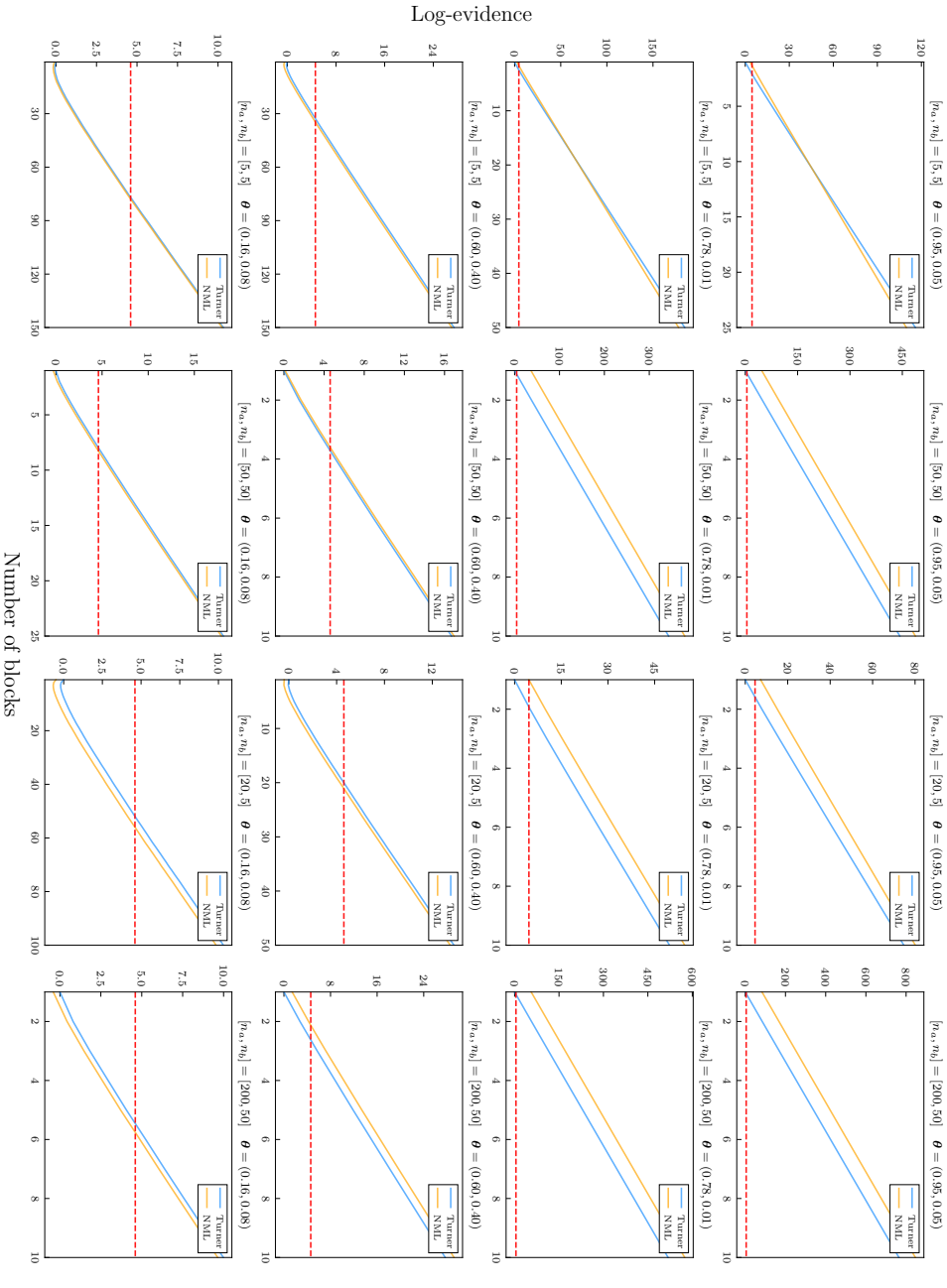


Figure 4.2: Empirical expected (log) evidence of sequential Turner (blue) and sequential conditional NML (orange). The horizontal line indicates the threshold corresponding to rejection at level $\alpha = 0.01$.

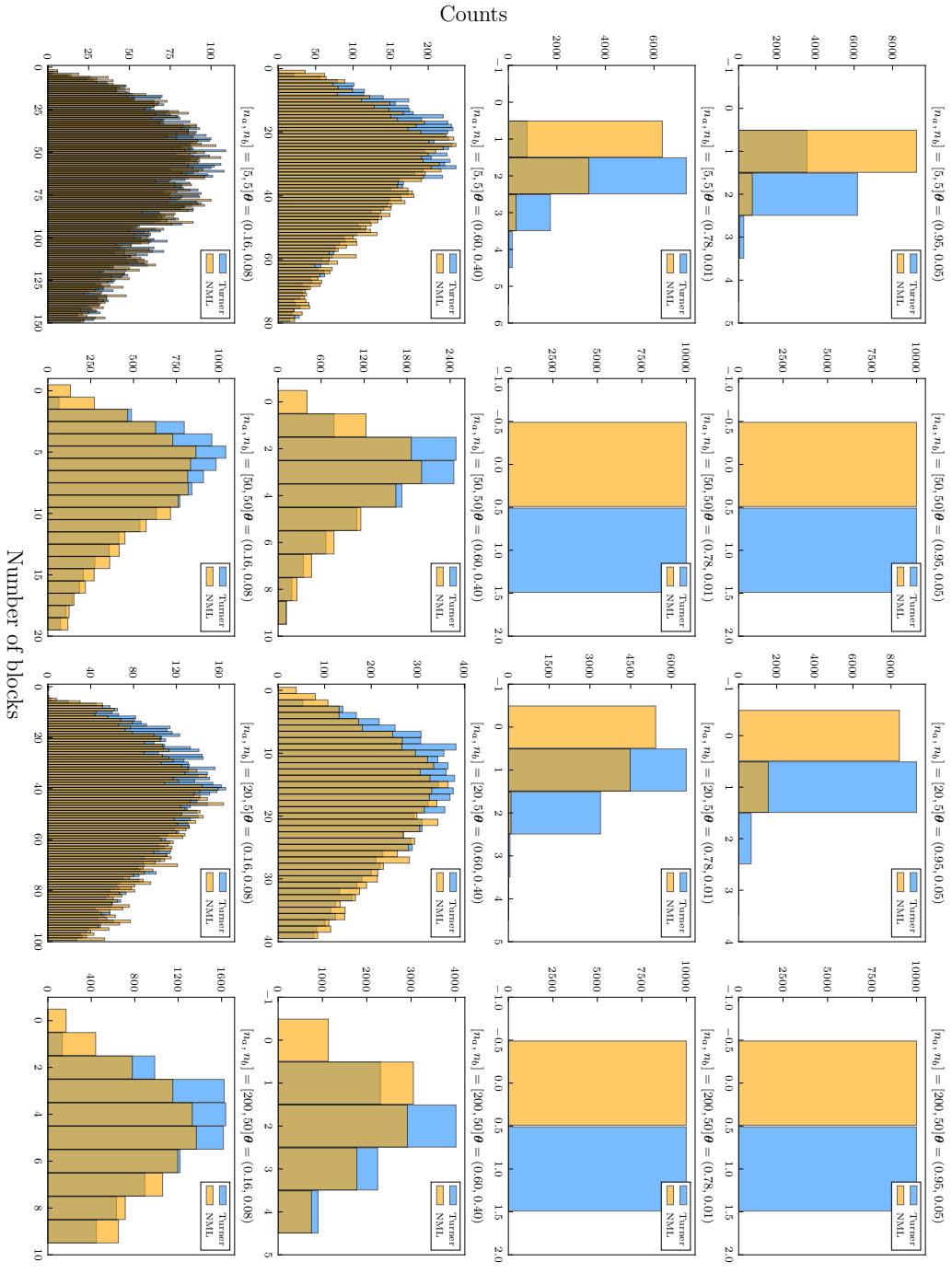


Figure 4.3: Empirical distributions of stopping times for sequential Turner (blue) and sequential conditional NML (orange) for confidence level $\alpha = 0.01$.

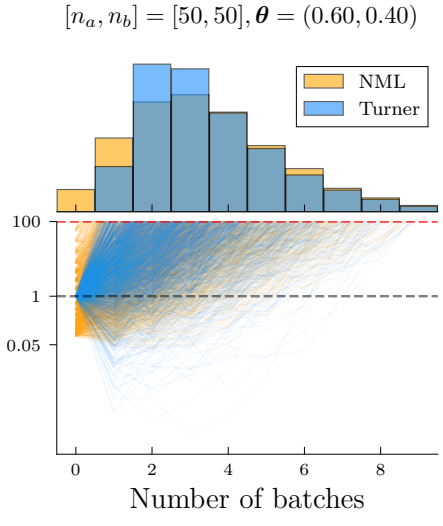


Figure 4.4: For one simulation configuration ($(n_a, n_b) = (50, 50)$, $(\theta_a, \theta_b) = (0.6, 0.4)$), the stopping time distribution for $\alpha = 0.01$ is plotted jointly with trajectories of the sequential Turner (blue) and sequential conditional NML (orange) accumulated evidence.

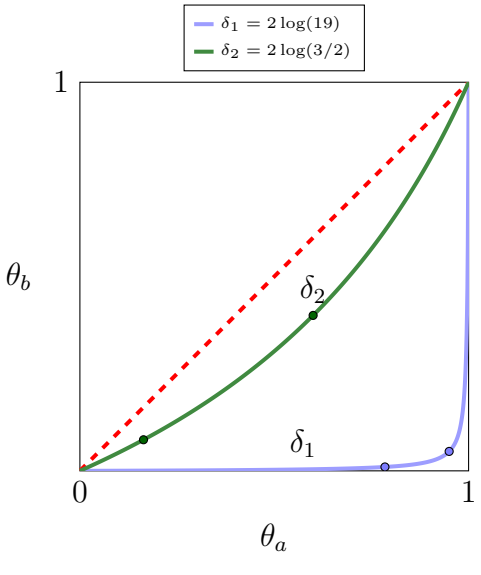


Figure 4.5: Parameter space with oracle parameters used for data generation.

CHAPTER 5

Validation of higher-order interactions: application to brain data

This chapter is based on: F. Giuffrida, E. Agrimi, M. Marzi, A. Gallo, T. Squartini and T. Gili. Signatures of higher-order coordination in the human brain. *In preparation.*

Abstract

Null models and statistical tests play a key role in validating empirical patterns in data, including relational structures that may involve interactions of arbitrary order. This perspective is particularly relevant in integrative neuroscience, where cognitive processes are understood to emerge from the coordinated activity of multiple brain regions. Here, we propose an algorithm to detect triadic hyperedges of cooperating brain regions, rooted in entropy-maximization. Within such a statistical framework, it is possible to constrain the number of lower-order co-activation patterns, thus enabling the identification of higher-order, over-expressed interactions. Our method generalizes the dyadic validation scheme by returning a matrix of hyperedge-specific p-values, from which a validated hypergraph of brain interactions can be constructed after correcting for multiple hypothesis testing. When applied to human resting-state fMRI time series, our approach suggests that subcortical, limbic, and transmodal regions are more frequently involved in triadic interactions that lack full dyadic support, whereas somatomotor and visual regions are comparatively more associated with triples consistent with dyadic closure (i.e., 2-simplices). Moreover, these purely higher-order interactions tend to link more spatially distant regions and exhibit lower redundancy than 2-simplicial configurations. By combining an information-theoretic perspective with hypergraph modeling, this framework provides a principled way to characterize higher-order functional coordination in the brain.

5.1 Introduction

The human brain is a complex system constituted of interacting neurons whose organized activity allows properties such as behaviors and cognition to emerge [96, 97]. Functional connectivity (FC) analyses represent the most popular approach to quantifying these interactions from fMRI data, typically proxied by correlations between (the activity of) different regions of the brain [98, 99, 100]. Pairwise analyses of this kind have been instrumental in mapping the large-scale architecture of the human brain, revealing a small-world, modular, and hierarchical organization [101]. Still, the neuroscientific literature has mainly focused on two-body interactions so far, thus shedding light on only a subset of the interactions that can take place in the brain, while disregarding those involving more than two regions, i.e., the so-called *higher-order interactions* (HOIs). Recent literature has successfully introduced HOIs in various contexts, including biological [102], ecological [103], financial [33], and social systems [104]; recent evidence has shown that HOIs are also key ingredients for the emergence of higher-level brain functions, such as cognition and consciousness [32, 33, 34, 35].

Several measures of HOIs have been introduced in neuroscience, the definitions of which originate from multiple theoretical contexts, including statistical inference [105], topological data analysis [33], multivariate information theory and partial information decomposition [106, 107, 108, 109], and machine learning [110].

Information-theoretic studies have characterized HOIs between brain regions in terms of *redundancy* and *synergy*: while redundancy emerges when neural populations convey overlapping information, synergy dominates when the information coming from multiple regions is integrated. The contributions of redundancy and synergy have been quantified by introducing the *S-information* [111] and *O-information* [106], quantities which have been shown to provide an exhaustive description of the higher-order interactions between three bodies [112]. These two modalities differ in their spatial distribution: while redundancy dominates in regions that process information from a single, perceptual or motor, domain – like the primary sensory and the motor cortices – synergy becomes more prominent within association networks - like the frontoparietal, default-mode and dorsal-attention networks [32, 108].

At the same time, the topological data analysis of FC data [113, 114, 115] often identifies the presence of higher-order interactions together with that of all lower-order counterparts [116]. This simplicial closure condition aligns naturally with redundant interactions, but precludes the identification of collective effects that arise without lower-order support. Models based on the concept of hypergraph, instead, allow for higher-order structures that do not require the presence of the corresponding lower-order hyperedges [114]; yet, even within this framework, applications have lacked a statistically principled procedure to determine whether empirical group-wise interactions are a consequence of simpler lower-order constraints. As a result, the extent to which higher-order functional motifs rely upon lower-order structures, thus satisfying the simplicial closure condition, remains unknown: answering this question requires a statistical approach capable of testing whether higher-order behaviors are consistent with (expectations based upon) lower-order dependencies or, instead, exceed them to

a significantly large extent, thereby reflecting a genuinely collective organization.

The framework based on entropy maximization provides a rigorous solution to this problem [31, 6], as it allows one to test whether a given pattern carries information beyond that embodied by the enforced constraints. To apply such a framework to brain activity data, we represent resting-state fMRI time series as a binary, undirected, bipartite network: a link between a region of interest (ROI) and a (discrete) time point indicates that the region is *active* at that specific time; we then turn such a representation into a functional hypergraph, where nodes are connected only if they are found to be active simultaneously a significantly large amount of time. To assess the statistical significance of dyadic interactions, we employ the Bipartite Configuration Model (BiCM) [117, 55], a linear Exponential Random Graph Model (ERGM) controlling for both the regional and the temporal heterogeneity; to assess the statistical significance of triadic interactions, we introduce the Bipartite Partial Local Overlap Model (BiPLOM), a non-linear ERGM that, in addition to controlling for regional and temporal heterogeneity, explicitly accounts for each region's aggregated propensity to co-activate in pairs. Used in combination with the BiCM, the BiPLOM allows us to assess which statistically validated higher-order interactions are better represented as simplicial complexes, in which the underlying pairwise interactions are validated as well (via the BiCM) or, conversely, if they require the hypergraphs formalism, in which the underlying pairwise interactions are not necessarily statistically significant.

We show that introducing second-order constraints reshapes the landscape of validated higher-order interactions, revealing distinct regimes of coordination across brain systems. Triads supported by validated pairwise interactions are spatially compact and display positive O-information, consistent with redundancy. In contrast, triads lacking dyadic support are more spatially dispersed and exhibit reduced or null redundancy. Importantly, the distribution of these regimes is functionally heterogeneous: limbic, subcortical, and transmodal regions show a higher prevalence of triadic configurations that would remain undetected in a simplicial framework relying exclusively on dyadic validation.

5.2 Results

5.2.1 From time series to bipartite networks

FC studies are typically carried out by exploiting the statistical (ir)regularities of the BOLD fMRI time series [118]: the BOLD signal reflects local changes in blood oxygenation, and large deviations are, in fact, believed to reflect episodes of neural activation. In this respect, a foundational method is the so-called *point process analysis* (PPA) [119], which converts normalized BOLD signals into (likely sparse) sequences of discrete events by retaining only values exceeding a predefined threshold. Despite the drastic reduction of dimensionality, PPA recovers a resting-state network (RSN) structure comparable to that obtained via standard correlation-based methods. Extensions of this approach include the so-called *co-activation pattern* (CAP) analysis [120, 121, 122], which clusters high-amplitude BOLD frames into recurring

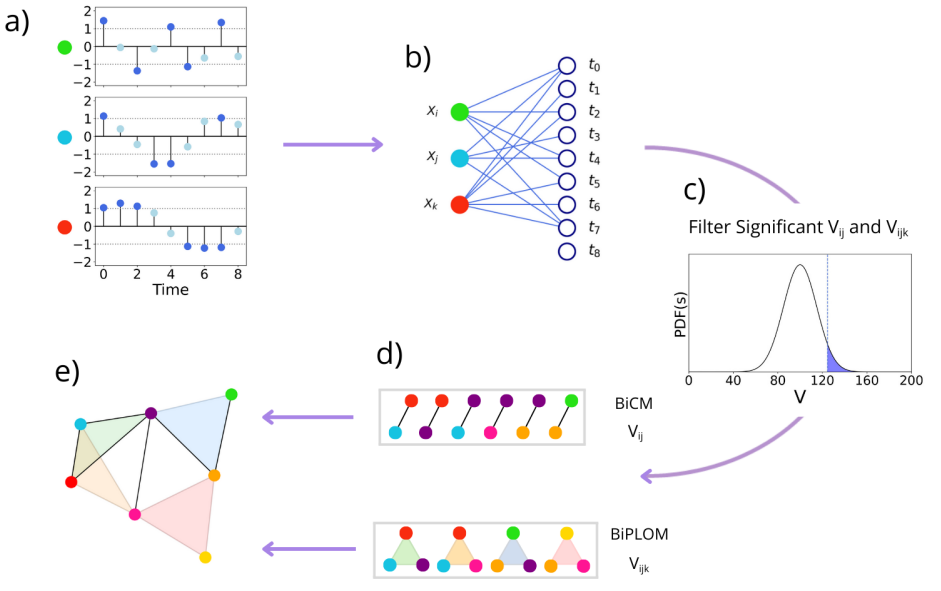


Figure 5.1: ***Infographics illustrating the pipeline of our analysis.*** *Pictorial representation of the pipeline we follow in the present contribution, from the registration of brain activity to the definition of a hypergraph the nodes of which are brain regions:* **a)** *resting-state fMRI time-series are standardized and binarized;* **b)** *time series are mapped to a binary, undirected, bipartite graph;* **c)** *the empirical abundances of dyads, V_{ij}, are validated against the benchmark named Bipartite Configuration Model (BiCM) [117, 55], while the empirical abundances of triads, V_{ijk}, are validated against the benchmark named Bipartite Partial Local Overlap Model (BiPLOM);* **d)** *the sets of validated dyads and triads are identified by the two models;* **e)** *a binary, undirected hypergraph is populated with the statistically validated interactions.*

network states, and event-based analyses in EEG data [123, 124, 125], which reveal activation patterns on sub-second timescales.

As these findings support the intuition that large fluctuations carry key information about the collective dynamics of brain, here we represent a subject's activation data as a binary, rectangular matrix $\mathbf{B} = \{b_{it}\}_{\substack{i\in\{1\dots N\}\\ t\in\{1\dots T\}}}$, where the rows correspond to the N brain regions of interest (ROIs) and the columns correspond to the T time points (see Section 5.3.1). The generic entry of such a matrix is then defined as

$$b_{it} = \begin{cases} 1, & \text{if } |z_i(t)| > \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{5.1}$$

where $z_i(t) = (x_i(t) - \mu_i)/\sigma_i$, with μ_i and σ_i being the average and standard deviation of the time series $\{x_i(t)\}_{t\in\{1\dots T\}}$ associated with region i [119, 123, 126]. In other words, region i is flagged as 'active' whenever its standardized signal deviates from the average by more than a fixed threshold τ: a typical choice is $\tau = 1$, corresponding to one standard deviation. The resulting matrix $\mathbf{B}$ can thus be interpreted as a

biadjacency matrix.

Although associations between regions and time points can be represented via a bipartite network that retains the information about the amplitude and sign of signals [127, 128], here we restrict our attention to the occurrences of extreme events, a simpler representation that offers a valid basis for statistical validation via maximum entropy null models.

5.2.2 Inferring dyadic and triadic interactions

Once a multivariate time series has been represented as a bipartite network, it is possible to leverage the information encoded in this representation to construct a network connecting brain regions that exhibit a similar behavior: what is obtained is known as *monopartite projection* [129, 55, 130, 131, 127].

To this aim, we adopt and extend the framework introduced in [55] to identify significantly similar interactions between pairs and triplets of brain regions. Specifically, two brain regions are considered (functionally) similar whenever the number of co-activations exceeds the expectation from a null model that constrains the activity of each brain region and the number of regions active at each time; analogously, three brain regions are considered as similar whenever they are joint active more frequently than expected under a null model additionally constraining the tendency of each node to co-activate in pairs.

More formally, the number of co-activations for the pair of regions (i, j) is defined as

$$V_{ij} = \sum_{t=1}^{T} b_{it} b_{jt}, \tag{5.2}$$

i.e., the number of time points at which both regions i and j are found to be active; analogously, the number of co-activations for the triplet of regions (i, j, k) is defined as

$$V_{ijk} = \sum_{t=1}^{T} b_{it} b_{jt} b_{kt}. \tag{5.3}$$

For each n-tuple of regions ($n = 2, 3$), our algorithm quantifies the statistical significance of the empirical number of co-activations. To this aim, we employ the BiCM [117, 55] to assess the statistical significance of the empirical abundances of dyadic interactions, say $\{V_{ij}\}_{i,j\in\{1\dots N\}}$, and the BiPLOM to assess the statistical significance of the empirical abundances of triadic interactions, say $\{V_{ijk}\}_{i,j,k\in\{1\dots N\}}$ (see Section 5.3.2).

Once the null model has been specified, each empirical value, say V^*_{ij} and V^*_{ijk}, is compared to the ensemble distribution of the corresponding statistics, yielding a p-value (i.e. the probability of observing a number of interactions greater than, or equal to, the empirical one) per tested pair and triplet of areas, reading

$$p\text{-value} = \Pr[V \geq V^* \mid \text{null model}] \tag{5.4}$$

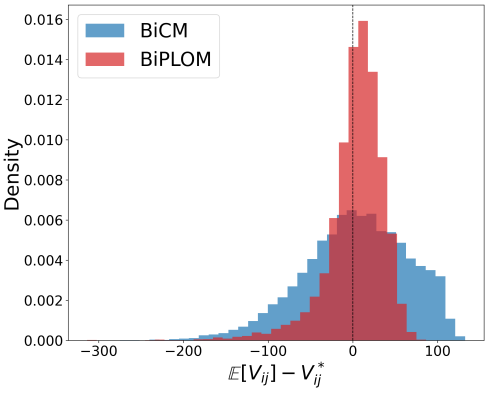


Figure 5.2: ***Co-activations residuals under BiCM and BiPLOM.*** *For a representative subject (id: 366446), histogram of residuals $\mathbb{E}[V_{ij}] - V^*_{ij}$ under BiCM (blue) and BiPLOM (red): BiPLOM yields a sharper, higher peak around zero, indicating a closer agreement between expected and empirical pairwise co-activations once pairwise tendencies are controlled for.*

(where, for the sake of generality, we have dropped the indices).

Since thousands of tests are performed across pairs of regions and millions across triplets of regions, the resulting p-values are corrected for multiple hypothesis testing. Hereby, the Benjamini-Hochberg procedure controlling for the percentage of 'false discoveries' (equivalently, 'incorrectly rejected null hypotheses'), i.e., the False Discovery Rate (FDR) [132], has been applied. The correction for dyads is performed separately from the correction for triads, thereby controlling the FDR within each order of interactions (as already done, for example, in [133]). Interactions whose p-value falls below the adjusted threshold (the single-test significance level of which has been set to $\alpha = 0.05$) are statistically validated, hence retained (see Section 5.3.3).

The complete workflow, from data pre-processing to validation and from validation to the analysis of the resulting higher-order network, is depicted in fig. 5.1.

5.2.3 Impact of second-order constraints on triplet validation

In principle, triadic interactions could be validated under both BiCM and BiPLOM. However, the two null models differ in the constraints they enforce. The BiCM imposes only first-order constraints, namely the activity of each region and the number of active regions at each time point. As such, it does not account for second-order co-activation tendencies.

The BiPLOM extends this framework by additionally constraining, for each region, its aggregated propensity to co-activate in pairs. Importantly, this does not mean that BiPLOM constrains each individual pairwise co-activations V_{ij}. Imposing such pairwise constraints would require $O(N^2)$ parameters and would render the model substantially less tractable. Instead, BiPLOM constrains node-level aggregated co-activations V_i, which encode each region's overall tendency to co-activate in pairs, without imposing constraints on each individual pairwise motif V_{ij}. The formal

definition of the model and its constraints is provided in Section 5.3.2. Notably, although V_{ij} are not directly constrained, enforcing V_i indirectly improves the reproduction of pairwise co-activations V_{ij} as well; this effect is illustrated in Figure 5.2 for a representative subject.

Because of their structural difference, the two models yield different outcomes in triplet validation. Since BiCM does not account for aggregated second-order tendencies, it is more likely to flag as significant triplets whose apparent higher-order behavior is largely driven by pairwise co-activations. In other words, part of the signal detected by BiCM may reflect dyadic structure that is not controlled for in its null model.

Empirical evidence for this interpretation comes from the analysis of closed triangles in the validated pairwise network. We found that BiCM systematically validates all such triangles also at the triadic level, indicating that, under BiCM, dyadic closure is automatically promoted to triadic significance. In contrast, BiPLOM validates only a subset of these configurations, with the retained fraction varying substantially across subjects (Figure 5.3a).

To further quantify the difference between the two models, we compared, for each subject, the sets of triplets validated under BiCM and BiPLOM, denoted by $\mathcal{T}^{\mathrm{BiCM}}$ and $\mathcal{T}^{\mathrm{BiPLOM}}$. Over their union, we computed the proportion of triplets validated exclusively by BiCM, exclusively by BiPLOM, or by both models (Figure 5.3b). Across subjects, BiPLOM validates fewer triplets overall, consistent with a stricter null model that accounts for aggregated pairwise tendencies. At the same time, the two validation sets are not simply nested: although a substantial fraction of triplets is validated by both models, BiPLOM identifies a non-negligible subset of configurations that BiCM does not recover.

5.2.4 Geometric and informational signatures of validated triplets

To characterize the triadic interaction between regions of interest, we considered three specific classes of validated triplets: *connected triplets*, in which all three constituent pairs are validated by BiCM and can therefore be modeled as simplicial complexes, *mixed triplets*, in which one or two of the constituent pairs are validated by BiCM (Fig.5.4a), and *unconnected triplets*, in which none of the constituent pairs are validated by BiCM.

To test whether these classes exhibit distinct spatial footprints, spanning across the brain's lobes or staying local, we quantified the geometric area of each validated triplet using the centroids of its constituent ROIs. To aggregate the results, we first calculated the area of each validated triplet. Then, for each individual, the areas of connected, mixed, and unconnected triplets were averaged separately, yielding a mean value per category per subject. Finally, a relative mean area was obtained for each subject by dividing its mean area by the median area of all possible triplets of regions in our atlas. This procedure yielded three sets of relative average areas, one for each subject in each set, allowing comparison of the typical spatial footprints of connected, unconnected, and mixed triplets across the entire cohort (Fig. 5.5). Unconnected and mixed triplets consistently exhibit larger spatial areas, suggesting a more dispersed

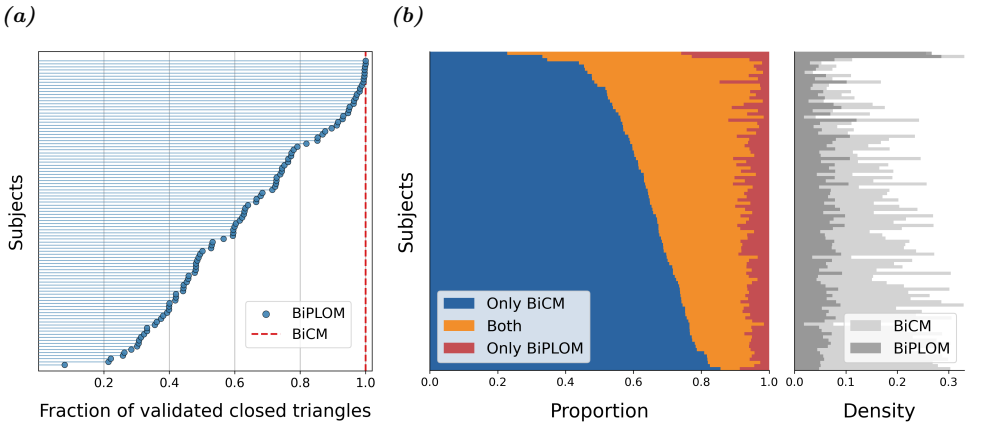


Figure 5.3: ***Comparison of BiCM and BiPLOM triplets validations.*** ***(a)*** *Fraction of closed triangles (i.e., triads of nodes whose three pairs are validated by BiCM) that are further validated as statistically significant triplets by BiPLOM (blue dots) and by BiCM (red dashed line), for all subjects ranked by their BiPLOM value. While BiCM consistently validates all these triangles across subjects, BiPLOM exhibits significant variability across subjects.* ***(b)*** *The left panel decomposes, for each subject, the union of validated triplets into three disjoint sets: triplets validated exclusively by BiCM (blue), exclusively by BiPLOM (red), and by both models (orange). Subjects are ordered according to the proportion of triplets validated exclusively by BiCM. The right panel shows the across-subject validation densities (validated triplets over the total number of possible triplets) for both models. Overall, BiPLOM yields a more conservative set of validated triplets than BiCM, while still detecting configurations that BiCM does not.*

anatomical footprint that may support integrative processes across different areas. In contrast, connected triplets tend to remain more spatially localized, supporting the coordination of activity across more homogeneous, spatially closer areas.

In addition, we investigated whether connected, mixed, and unconnected triplets also differ in their O-information values, which quantify the balance between redundancy and synergy in multivariate interactions [106]. For each subject and each validated triplet (i, j, k), we considered the corresponding time series array $\mathbf{X}_{lt}$, with $l \in \{i, j, k\}$ and $t \in \{1, \ldots, T\}$, and calculated the O-information $\Omega(\mathbf{X}_{lt})$ [134]. For each subject, the O-information values were averaged across all validated triplets. Unconnected triplets show O-information values concentrated close to zero, indicating little net redundancy. Mixed triplets display moderately positive values, while connected triplets exhibit the largest positive O-information, reflecting increasing redundancy as the number of pairwise-supported interactions within the triplet grows (Fig. 5.6).

5.2.5 Mesoscale-level characterization

Finally, we assessed mesoscale patterns by categorizing validated triplets according to the Resting State Networks (RSNs) defined by Yeo et al. [135], to which their con-

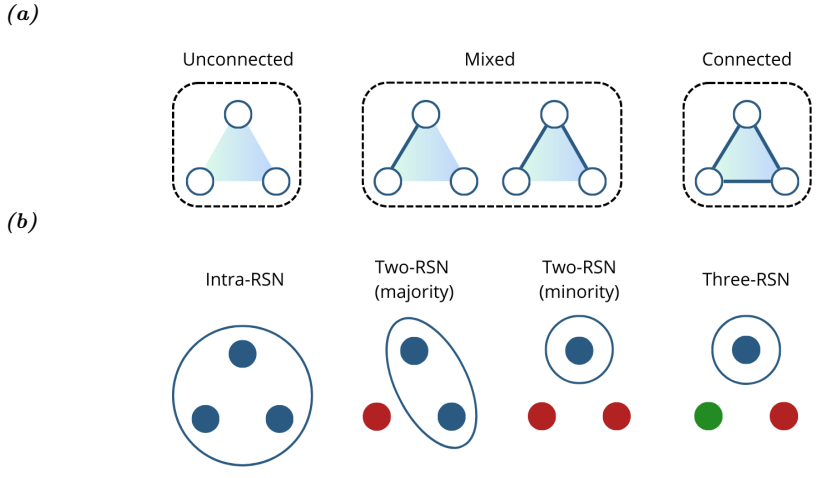


Figure 5.4: ***Triplets classification*** ***(a)*** *Validated triplets were classified according to the number of validated pairwise interactions among their underlying pairs. We distinguish between connected triplets, where all internal pairs are validated, mixed triplets, where one or two internal pairs are validated, and unconnected triplets, where no internal pair is validated.* ***(b)*** *Validated triplets were also classified based on the distribution of their nodes across resting-state networks. Specifically, for a given node in a resting-state network R, a validated triplet involving the node is said to be intra-RSN if it involves only nodes in R; two-RSN (majority) if it involves two distinct RSN and the majority of nodes belong to R; two-RSN (minority) if it involves two distinct RSN and only one node belongs to R; three RSN if it involves three distinct RSN, including R.*

stituent regions are assigned. This classification enables the analysis of co-activation motifs both within and across RSNs, thereby facilitating the interpretation of higher-order interactions in terms of known large-scale functional systems. The validated triplets were grouped into connected, mixed, and unconnected triplets. Moreover, for each RSN R, triplets involving at least one node in R were further distinguished into four categories (Fig. 5.4b): (i) intra-RSN triplets, in which all three nodes belong to R; (ii) *two-RSN (majority)* triplets, involving exactly two RSNs with two nodes in R; (iii) *two-RSN (minority)* triplets, involving exactly two RSNs with a single node in R; (iv) *three-RSN* triplets, spanning three distinct RSNs, with only one node in R.

To visualize the spatial distribution of brain regions across these motif categories, we computed regional density maps for each configuration separately. Specifically, we counted the number of validated triplets of category c that include region r in each subject s, yielding $N_s(r, c)$. These counts were normalized within each subject and configuration to obtain proportional densities expressed by the higher order regional density index $HORD_s(r, c)$. Group-level maps, one for each configuration c, were generated by averaging $HORD_s(r, c)$ across subjects.

In addition to the regional density maps, we derived regional connectedness maps to summarize whether each region tended to participate more often in connected or unconnected motifs. For each configuration type, we computed the higher-order regional flexibility index $HORF_s(r, c)$ for every subject and region, and then averaged

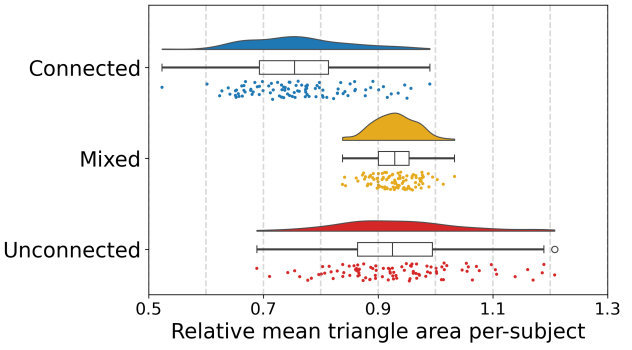


Figure 5.5: ***Comparison of relative triangle areas for validated connected, unconnected, and mixed triplets.*** *Rain plots illustrate the distribution of relative mean triangle areas across subjects (each of which is represented here by a dot). Triplets are classified according to the validation of their constituent pairs: connected triplets (all pairs validated by BiCM), unconnected triplets (no pairs validated), and mixed triplets (one or two pairs validated by BiCM). For each subject, triangle areas were averaged across all validated triplets and divided by the median of all the possible triangle areas in our atlas. Unconnected and mixed triplets consistently exhibit larger spatial areas, suggesting a more dispersed anatomical footprint that may support integrative processes across different areas. In contrast, connected triplets tend to remain more spatially localized, supporting the coordination of activity across more functionally homogeneous and spatially closer areas.*

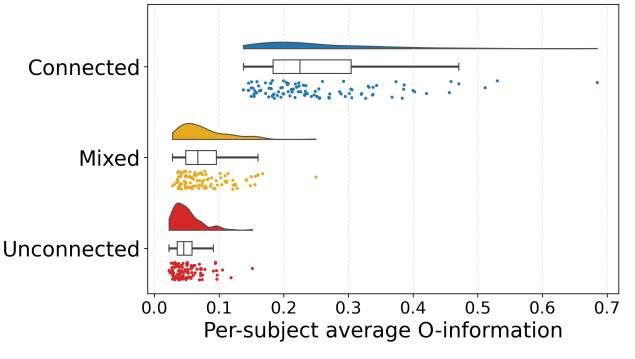


Figure 5.6: ***Comparison of O-information for validated connected, unconnected, and mixed triplets.*** *Rain plots illustrate the distribution of mean O-information values across subjects (each represented by a dot). Triplets are classified according to the validation of their constituent pairs: connected triplets (all pairs validated by BiCM), unconnected triplets (no pairs validated), and mixed triplets (one or two pairs validated by BiCM). For each subject, O-information values were averaged across all validated triplets. Connected triplets consistently show larger positive O-information, reflecting redundancy and stronger lower-order dependencies. Unconnected triplets reveal informational signatures more consistent with genuine higher-order interdependencies. Mixed triplets show an intermediate regime of coexistence between lower and higher order interactions.*

these values across subjects to obtain group-level connectedness maps. The resulting maps illustrate, for each configuration, the relative predominance of connected versus unconnected triplets across the cortex (Fig.5.7).

To aggregate these results at the RSN level, we considered, for each RSN R, all triplets including R. We then computed the proportion of connected, mixed, and unconnected triplets with a given configuration, normalizing by the total number of validated triplets. Fig. 5.8 summarizes the aggregated results, showing whether connected, mixed or unconnected triplets better describe interactions between and within specific RSNs.

5.3 Methods

5.3.1 Dataset description

We analyzed 3T resting-state fMRI data from the unrelated subjects subset of the Young Adult Human Connectome Project (HCP) [136]. The complete data set comprises approximately 1,200 participants aged 22 to 35 years, each scanned with a high-resolution multiband EPI BOLD sequence (TR = 720 ms, voxel size = 2 mm isotropic, 72 slices) optimized for whole-brain coverage and functional coherence. During acquisition, subjects were at rest with their eyes open and focused on a central cross. The imaging parameters included a repetition time (TR) of 720 ms, an echo time (TE) of 33.1 ms, a flip angle of 52°, and a spatial resolution of 2 mm isotropic across 72 slices. Each scanning session comprised two 14.5-minute runs with opposite phase-encoding directions (RL and LR), yielding 1,200 frames per run.

Data preprocessing followed the HCP minimal pipeline [137], which includes motion correction, intensity normalization, and spatial alignment to standard space. Structured noise components were further removed using the FIX denoising procedure [138], which regresses the motion-related confounding values estimated from rigid-body parameters and their temporal derivatives [139]. Temporal high-pass filtering was applied prior to regression. To maximize temporal coverage, RL and LR runs were concatenated after removing the mean signal across time.

The functional time series were divided into 116 regions of interest (ROIs), combining the 100 cortical parcels of the Schaefer atlas [140] with the 16 subcortical regions defined by the Tian atlas [141] (see Appendix 5.C). The final dataset comprised 100 multivariate time series with 116 dimensions and 2400 time points.

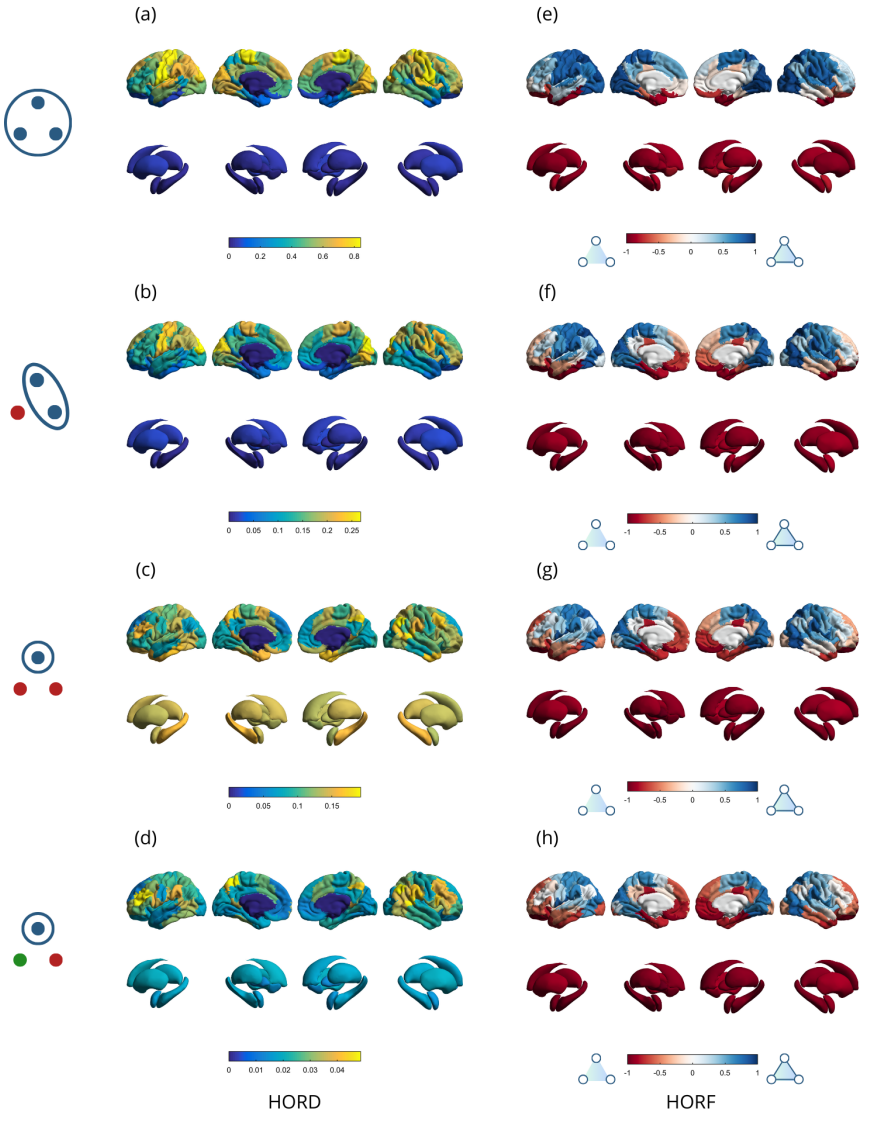


Figure 5.7: ***Regional density and connectedness of triplet configurations.*** *Panels (a–d) and (e–h) show cortical and subcortical maps of regional density and regional connectedness, respectively, for different classes of validated triplets. Each row corresponds to a distinct triplet configuration (intra-RSN, two-RSN majority, two-RSN minority, and three-RSN triplets). Panels (a–d) report regional density maps (yellow–blue scale), quantified by the higher-order regional density index (HORD), where warmer colors indicate more frequent participation in validated triplets, averaged across subjects. Panels (e–h) report regional connectedness maps (red–white–blue scale), quantified by the higher-order regional flexibility index (HORF), with blue (red) indicating a relative predominance of connected (unconnected) triplets, averaged across subjects. Notably, limbic and subcortical regions show increasing relative involvement in triplets spanning multiple RSNs and a relative predominance of unconnected configurations. In contrast, somatomotor and visual regions are more prominently associated with connected triplets.*

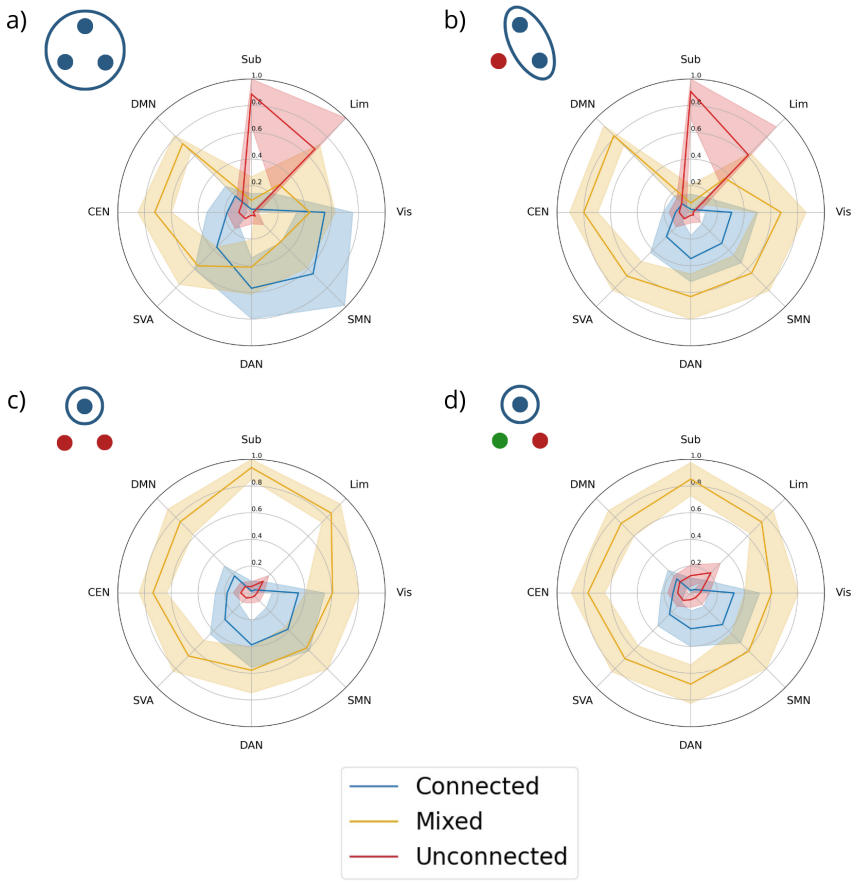


Figure 5.8: ***Mesoscale characterization of validated triplets across RSNs.*** *For each RSN R, validated triplets involving at least one node in R are classified into (a) intra-RSN triplets, (b) two-RSN (majority), (c) two-RSN triplets (minority), and (d) three-RSN triplets spanning three distinct RSNs. For each class and RSN, values are normalized to the total number of validated triplets. Shaded bands represent the standard deviation across subjects. This analysis reveals how connected, unconnected, and mixed triplets distribute across canonical functional systems, highlighting which RSNs are predominantly characterized by redundant lower-order dependencies versus genuinely higher-order interactions. A clear pattern emerges: sensorimotor and visual networks show limited involvement in unconnected triplets, whereas attentional and cognitive networks display a balance between connected and unconnected triplets, with a clear predominance of mixed triplets. At the same time, subcortical and limbic systems are often characterized by a higher prevalence of unconnected triplets than connected ones, underscoring their role as hubs of higher-order integration.*

5.3.2 Null models for validation

To assess whether the observed co-activations V_{ij}^* and V_{ijk}^* arise beyond expectations from random fluctuations, we employ canonical maximum-entropy null models, also known, in the context of network science, as *Exponential Random Graph Models*

[31, 4, 5] (see Appendix A). These models define probability ensembles of bipartite graphs that reproduce selected empirical properties of the data while randomizing all others, thereby providing rigorous statistical baselines. As anticipated, two null models are implemented in this work: the Bipartite Configuration Model for pairwise validation and the Bipartite Partial Local Overlap Model for triadic validation.

Pairwise validation: the BiCM

The BiCM [55] is a first–order[1] maximum–entropy model that preserves, in expectation, the degree sequence of both layers of the observed bipartite network $\mathbf{B}^*$. In formulas, it enforces the following constraints:

$$\mathbb{E}[k_i] = k_i^*, \qquad \forall\, i \in 1, \dots N, \tag{5.5}$$

$$\mathbb{E}[h_t] = h_t^*, \qquad \forall\, t \in 1, \dots T, \tag{5.6}$$

where $\mathbb{E}[x]$ represents the expected value of x in the chosen model, and k_i^* and h_t^* are the observed degrees of nodes i and t or, in other words, the empirical number of activations in region i and the number of active regions at time t. In this model, the entries b_{it} are *independent Bernoulli variables* by construction, with heterogeneous success probabilities p_{it} determined by imposing the above constraints. The probability of a bipartite configuration factorizes as

$$P(\mathbf{B}) = \prod_{i,t} p_{it}^{b_{it}} (1 - p_{it})^{1 - b_{it}}, \tag{5.7}$$

where

$$p_{it} = \frac{e^{-(\beta_i + \gamma_t)}}{1 + e^{-(\beta_i + \gamma_t)}}, \tag{5.8}$$

and β_i, γ_t are Lagrange multipliers estimated either by maximizing the likelihood of the model or by solving the constraint equations. This independence property implies that, for any pair of regions (i, j), the random variable $V_{ij} = \sum_t b_{it} b_{jt}$ is the sum of T independent Bernoulli trials with heterogeneous probabilities $p_{it} p_{jt}$, hence it follows a Poisson–Binomial distribution (see Appendix 5.B). Empirical values V_{ij}^* are compared with this distribution to compute p-values for pairwise validation. A detailed derivation of the Poisson-Binomial distribution and its computation can be found in [55].

Triadic validation: the BiPLOM

BiCM controls first-order heterogeneity by reproducing, in expectation, the activity level of each region and the number of active regions per time point. This allows the identification of pairwise interactions, quantities of the second order that cannot be explained solely by local differences in activation frequency. When moving to the validation of triplets, however, this model becomes insufficient: to test for genuine third–order interactions, one must account not only for local heterogeneity but also for

[1] Here, "first-order" indicates that the model constrains only quantities that are linear in the entries of $\mathbf{B}$.

the tendency of regions to co-activate at the pair level. For this reason, we introduce a second–order maximum–entropy model, the *Bipartite Partial Local Overlap Model* (BiPLOM), which constrains both first- and second-order statistics. By controlling for lower–order dependencies, BiPLOM provides a stricter statistical baseline, reducing the number of spurious triplets that would otherwise appear significant under the BiCM, as shown in the Results section. A full derivation of the BiPLOM and a detailed discussion of its properties are provided in Appendix 5.A.

For each region i, in addition to constraining its degree, as performed by the BiCM, the BiPLOM preserves the expected number of co-activations the region shares with others:

$$\langle V_i \rangle = \sum_{j \neq i} \sum_{t} p_{it} p_{jt} = V_i^* \qquad \forall i \in 1, \ldots, N. \tag{5.9}$$

This additional set of constraints ensures that the model reproduces the empirical tendency of each region to co-activate with others, effectively incorporating second–order correlations into the null ensemble. Unlike the BiCM, these additional constraints introduce dependencies among the variables b_{it}, making them non-independent Bernoulli variables. To obtain a tractable formulation, we adopt a mean–field approximation (see Appendix 5.A), which replaces the full joint distribution with a product of effective Bernoulli marginals. Under this approximation, each b_{it} becomes effectively independent, and the model regains a factorized form analogous to the BiCM, with success probabilities which read:

$$p_{it} = \frac{e^{-(\alpha_i + \beta_t + \delta_i h_t^*)}}{1 + e^{-(\alpha_i + \beta_t + \delta_i h_t^*)}}, \tag{5.10}$$

where the parameters α_i, β_t and δ_i are estimated by maximizing the likelihood. In the same mean-field approximation, the number of triplet co-activations

$$V_{ijk} = \sum_{t} b_{it} b_{jt} b_{kt} \tag{5.11}$$

is again the sum of T independent Bernoulli trials with heterogeneous probabilities $p_{it}p_{jt}p_{kt}$ and follows a Poisson–Binomial distribution. Comparison between empirical and expected values yields the p-values used for triadic validation.

Computation of p-values. Exact evaluation of Poisson–Binomial distributions is computationally demanding, especially when repeated for a large number of pairs, triplets, and subjects. For this reason, while the null distributions of V_{ij} and V_{ijk} are defined exactly by the corresponding Poisson–Binomial laws, in practice we rely on analytical approximations to compute p-values efficiently. A detailed comparison of different approximations and their performance on our dataset is provided in Appendix 5.B. Based on this analysis, we adopt the refined normal approximation throughout this work, which offers an excellent compromise between computational efficiency and accuracy.

5.3.3 Multiple hypothesis testing

The validation of dyadic and triadic interactions involves testing a large number of statistical hypotheses simultaneously. Specifically, for each subject, hypotheses are tested for all pairs and all triplets of brain regions, resulting in thousands of tests at the dyadic level and millions at the triadic level. To control for the inflation of false positives arising from multiple comparisons, we adopt a False Discovery Rate (FDR) control procedure [132].

Given a set of $|H|$ hypotheses $\{H_1, H_2, \ldots, H_{|H|}\}$, each associated with a p-value, the Benjamini–Hochberg procedure prescribes sorting the p-values in non-decreasing order,

$$\text{p-value}_1 \leq \cdots \leq \text{p-value}_{|H|}, \tag{5.12}$$

and identifying the largest index $\hat{i}$ such that

$$\text{p-value}_{\hat{i}} \leq \frac{\hat{i}}{|H|}\, t, \tag{5.13}$$

where t denotes the nominal single-test significance level, set here to $t = 0.05$.

All hypotheses H_i with $i \leq \hat{i}$ are then rejected, ensuring control of the expected proportion of false discoveries among the rejected hypotheses. In our context, each hypothesis tests whether the empirical number of co-activations for a given pair or triplet of regions is consistent with the distribution predicted by the corresponding null model (BiCM or BiPLOM).

Rejecting a dyadic hypothesis amounts to retaining the corresponding pair of regions as a statistically validated link in the monopartite projection [55]. Analogously, rejecting a triadic hypothesis results in the validation of the corresponding triplet as a higher-order interaction. FDR correction is applied separately to dyadic and triadic hypotheses, ensuring control of the FDR within each family of tests.

5.3.4 Classification and characterization of validated triplets

To characterize the validated triplets identified by the application of BiPLOM, we distinguished them based on the number of their constituent pairs validated by BiCM. Triplets in which all three pairwise interactions were validated by the BiCM were labeled as *connected triplets*, whereas triplets for which none of the three pairs were validated were labeled *unconnected triplets*. These two extreme cases provide the clearest contrast between higher-order structure strongly supported by pairwise statistics and structure that is essentially "invisible" at the pairwise level. Intermediate cases in which exactly one or two edges were validated are also possible and were labeled *mixed triplets* (see Fig. 5.4a for a schematic overview of all configurations).

Global geometric characterization of validated triplets

Having defined connected, mixed, and unconnected triplets, we next examined whether these classes differed in their spatial footprint. To this end, we associated each

ROI with a centroid in standard space, using the combined Schaefer–Tian atlas aligned to the MNI152NLin2009cAsym space, which removes inter-subject differences in overall brain size. For every validated triplet (i, j, k), with centroid coordinates $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k \in \mathbb{R}^3$, we computed the area of the triangle between the three centroids,

$$A_{ijk} = \frac{1}{2} \left\| (\mathbf{x}_j - \mathbf{x}_i) \times (\mathbf{x}_k - \mathbf{x}_i) \right\|, \tag{5.14}$$

which provides a simple Euclidean distance-based geometric measure of how spatially extended each triplet is. This also allows the use of the same measure for cortical (both ipsilateral and contralateral) and subcortical distances, whereas geodesic distances would not.

O-information analysis of higher-order dependencies

To complement the geometric characterization with an information-theoretic perspective, we quantified the nature of higher-order statistical dependencies within validated triplets using the O-information [106]. In this work, the O-information is computed on the original fMRI time series associated with each validated triplet.

Specifically, for each validated triplet (i, j, k) and for each subject, we considered the corresponding time series array

$$\mathbf{X}_{lt} = \{x_l(t)\}_{\substack{l \in \{i,j,k\} \\ t \in \{1,\dots,T\}}}, \tag{5.15}$$

where $x_l(t)$ denotes the fMRI signal of region l at time t. The time index t is treated as indexing repeated observations of the joint random variable (X_i, X_j, X_k), allowing the estimation of the associated entropic quantities from the empirical joint distribution of the signals.

The O-information $\Omega(\mathbf{X}_{lt})$ is a signed measure that quantifies whether the joint interaction among the three time series is dominated by lower-order, redundant dependencies (positive values) or by synergistic contributions (negative values). For three variables, it can be expressed as

$$\Omega(X_i, X_j, X_k) = TC(X_i, X_j, X_k) - DTC(X_i, X_j, X_k), \tag{5.16}$$

where TC and DTC denote the Total Correlation and Dual Total Correlation, respectively. The Total Correlation is defined as

$$TC(X_i, X_j, X_k) = H(X_i) + H(X_j) + H(X_k) - H(X_i, X_j, X_k), \tag{5.17}$$

while the Dual Total Correlation reads

$$\begin{aligned} DTC(X_i, X_j, X_k) = H(X_i, X_j, X_k) \; &- \big[H(X_i \mid X_j, X_k) \\ &+ H(X_j \mid X_i, X_k) + H(X_k \mid X_i, X_j)\big], \end{aligned} \tag{5.18}$$

with $H(X)$ denoting Shannon entropy of variable X and $H(X \mid Y)$ denoting the conditional entropy of X with respect to Y.

For each subject, $\Omega(\mathbf{X}_{lt})$ was computed for all validated triplets and subsequently averaged within each class (connected, mixed, and unconnected), yielding subject-level estimates of the prevailing informational regime associated with each class.

Triplet configuration types and higher-order regional density

To relate validated higher-order interactions to their spatial embedding across the cortex and subcortex, we further classified triplets according to the distribution of their constituent regions across resting-state networks (RSNs). We distinguished three main configuration types: *intra-RSN* triplets, in which all three regions belong to the same RSN; *two-RSN* triplets, involving exactly two distinct RSNs; and *three-RSN* triplets, spanning three different RSNs. For two-RSN triplets, we additionally distinguished between *majority* and *minority* configurations with respect to a given RSN R, depending on whether R contributed two or only one region to the triplet (see Fig. 5.4b for a schematic overview).

This classification allows the characterization of individual brain regions by how often they participate in higher-order motifs associated with different RSN-based configurations. Specifically, for each subject s, region r, and configuration type c (e.g., intra-RSN, two-RSN majority, two-RSN minority, three-RSN), we counted the number of validated triplets of type c that include region r, denoted by $n_s(r, c)$.

To enable comparison across subjects, these counts were normalized within each subject and configuration, defining the higher-order regional density (HORD) as

$$HORD_s(r, c) = \frac{n_s(r, c)}{\sum_{r'} n_s(r', c)}. \tag{5.19}$$

Accordingly, $HORD_s(r, c)$ represents the fraction of validated triplets of configuration c that involve region r for subject s. Group-level regional density maps were then obtained by averaging $HORD_s(r, c)$ across subjects.

Regional connectedness index

In addition to quantifying how frequently regions participate in different triplet configurations, we assessed whether regions tend to be involved preferentially in triplets supported by validated pairwise interactions or in triplets that emerge only at the higher-order level. To this end, we retained the same triplet classification described above and further distinguished, for each subject s, region r, and configuration c, how many of the contributing triplets were *connected* or *unconnected*. We denote these counts by $n_s^{\text{conn}}(r, c)$ and $n_s^{\text{unconn}}(r, c)$, respectively.

We then defined a regional connectedness index, termed higher-order regional flexibility (HORF), which contrasts these two contributions:

$$HORF_s(r, c) = \frac{n_s^{\text{conn}}(r, c) - n_s^{\text{unconn}}(r, c)}{n_s^{\text{conn}}(r, c) + n_s^{\text{unconn}}(r, c)}. \tag{5.20}$$

This index takes values in the interval $[-1, 1]$. Positive values indicate that region r participates more frequently in connected triplets of type c, whereas negative values indicate a predominance of unconnected triplets. Values close to zero reflect a balanced involvement in both types of higher-order interactions. The index is defined only for regions participating in at least one validated triplet of configuration c.

For each configuration type, $HORF_s(r, c)$ was averaged across subjects to obtain group-level maps that summarize the relative involvement of each region in lower-order-supported versus genuinely higher-order interaction patterns.

To aggregate these results at the RSN level, we considered, for each RSN R, all triplets including R. We then computed the proportion of connected, mixed, and unconnected triplets with a given configuration, normalizing by the total number of validated triplets. Fig. 5.8 summarizes the aggregated results, showing whether interactions between and within specific RSNs are better described by connected, mixed, or unconnected triplets.

5.4 Discussion and conclusion

Understanding how brain regions coordinate their activity beyond pairwise interactions is a central challenge in network neuroscience. While pairwise functional connectivity has provided valuable insights into large-scale brain organization, it is increasingly recognized that collective interactions involving more than two regions may capture additional aspects of neural coordination that are not evident at the dyadic level. In this work, we addressed this problem by introducing a statistically grounded framework to identify and characterize higher-order co-activation patterns directly from multivariate brain time series.

By representing binarized fMRI signals as a bipartite network between brain regions and time points, and by validating co-activation motifs through maximum-entropy null models, we obtained a principled mapping from temporal data to statistically validated higher-order interactions. Crucially, the use of null models that explicitly control for lower-order statistics allows us to distinguish between triadic co-activations that are consistent with pairwise structure and those that exceed it. This separation provides an operational definition of higher-order interactions in statistical terms, independent of any *a priori* functional interpretation.

The resulting validated triplets reveal a non-trivial organization of higher-order co-activation patterns in resting-state brain activity. Rather than forming a homogeneous class, triadic interactions span a spectrum of statistical configurations, reflecting different degrees of support from lower-order dependencies. This motivates a systematic comparison of triplets based on their statistical, spatial, and informational properties.

A central outcome of our analysis is the emergence of distinct regimes of triadic coordination, defined by the extent to which higher-order co-activations are supported by validated pairwise interactions. *Connected* triplets are fully supported by dyadic structure, indicating that their higher-order organization can be largely explained in terms of redundant lower-order dependencies. *Unconnected* triplets lack any statistically validated pairwise support and therefore constitute genuinely higher-order interactions in a strict statistical sense. In addition, a substantial fraction of validated triplets falls into an intermediate *mixed* category, in which only a subset of the underlying pairs is statistically supported. Importantly, these regimes are not defined heuristically, but arise directly from comparisons between null models encoding different levels of constraints. An important methodological feature of our analysis is

that unconnected and mixed triplets do not satisfy the simplicial closure condition. As a consequence, they would remain undetected in frameworks based on simplicial complexes or clique expansions of pairwise networks [113, 114].

These different regimes exhibit systematic differences in both spatial organization and informational signatures. Connected triplets tend to be spatially compact and are associated with positive, large O-information values, consistent with interaction patterns dominated by redundant, lower-order dependencies. Unconnected triplets, in contrast, are typically more spatially extended and display attenuated or near-zero O-information values. Consistently, mixed triplets occupy an intermediate regime in O-information values, reflecting the coexistence of lower-order and higher-order interdependencies within the same motif. However, the spatial extension of mixed triplets is compatible with that of unconnected ones, showing that both categories help the integration of information across the brain.

At the mesoscale, these higher-order interaction regimes exhibit structured relationships with the organization of resting-state networks (RSNs). Across RSNs, connected, unconnected, and mixed triplets are observed in all configuration types (intra-, two-, and three-RSN), but their relative prevalence varies systematically across functional systems. Unimodal networks, such as visual and somatomotor systems, are predominantly associated with connected and within-RSN triplets, consistent with their tightly coupled, locally and functionally redundant dynamics [32].

In contrast, limbic and subcortical areas display a marked predominance of unconnected and mixed configurations. These areas are primarily involved in information routing across RSN and integrating, relaying, and modulating signals between cortical areas [142, 143, 144]. Our findings suggest that these functions are predominantly supported by spatially sparse, purely higher-order interactions. Notably, these interactions favor integration between subcortical/limbic regions and functionally diverse areas, rather than forming isolated clusters within the limbic or subcortical systems themselves.

A particularly coherent mesoscale profile is observed for Control- and association-related networks, most notably the Central Executive Network (CEN), the Default Mode Network (DMN), and the Salience/Ventral Attention Network (SVA), where mixed triplets dominate in number over both connected and unconnected ones across all the types of configurations: within-RSN, two-RSN, and three-RSN. This shared signature suggests that their functional interplay relies on a balance between lower-order supported interactions and genuinely higher-order coordination.

This is consistent with the *Triple Network Model* [145, 146, 147], which conceptualizes these systems as a tightly integrated control ensemble supporting adaptive cognition. Our results extend this framework by showing that interactions involving these networks are supported by statistically validated higher-order co-activation patterns that cannot be fully reduced to dyadic structures.

More generally, the prominence of higher-order configurations, especially in cross-network triplets, highlights the role of higher-order dependencies as a flexible substrate for integration across functional systems. These interactions capture coordination patterns that cannot be reduced to purely dyadic ones, yet recur robustly across subjects and spatial scales. Our results, therefore, identify a class of statistically

validated higher-order interactions that are invisible to simplicial representations by construction, despite playing a non-marginal role in the empirical organization of brain activity. This finding suggests that restricting the analysis to cliques or simplices may overlook important dynamics such as those related to large-scale functional integration.

Beyond methodological considerations, the observed distinction establishes a statistically grounded link between topological and information-theoretic descriptions of multivariate interactions. By combining topological validation with the O-information, we show that triplets lacking full pairwise support are not only non-simplicial from a structural standpoint but are also associated with distinct informational regimes, characterized by near-zero redundancy. Consistently, connected triplets exhibit a markedly higher level of redundancy. Crucially, this correspondence does not follow from the construction of the analysis: despite being obtained through independent procedures, topological validation on binarized data and information-theoretic quantification on the full time series, the resulting distinction persists, indicating that topological and informational notions of higher-order interaction converge at the level of empirically validated brain dynamics.

Despite the novelty and robustness of the methods and the reported findings, some limitations must be acknowledged. First, the present formulation relies on binary and unsigned representations of the data obtained through thresholding. Although this choice facilitates the analysis, it inevitably removes information about the magnitude and polarity of fluctuations, which may still be relevant, especially in the context of higher-order synergistic interactions. Recent extensions of pairwise validation techniques to signed bipartite networks [127] and to signed brain time series [128] provide promising avenues for future research. Generalizing these approaches to the higher-order setting discussed here could enable the explicit inclusion of sign and weight information in the validation process.

Second, the current framework remains invariant under temporal reshuffling. Although rs-fMRI dynamics are typically modeled under stationarity assumptions, such simplifications inherently overlook the full complexity of brain activity over time. While some aspects of temporal heterogeneity are indirectly captured via degree-based constraints, the explicit temporal ordering of events is ignored. As a result, the method cannot account for time-dependent correlations or causal interactions. Future developments could address this by introducing second-order null models that retain temporal structure or constrain sequential motifs, thereby enabling a more direct investigation of co-activation dynamics.

Third, although computationally efficient formulations, such as canonical models and mean-field approximations, have been introduced, the scalability of higher-order validation remains a practical challenge, especially when testing all possible triplets or larger combinations. Optimizing sampling procedures or developing analytical approximations for motif distributions may be necessary to ensure applicability to large-scale datasets.

Despite the methodological limitations discussed above, our results demonstrate that statistically validated higher-order interactions provide a meaningful and interpretable level of description for brain activity that complements and extends pairwise

connectivity analyses. By revealing robust triadic coordination patterns that cannot be reduced to dyadic structure, this work shows that collective neural organization cannot be fully captured within purely pairwise or simplicial frameworks.

More broadly, the proposed approach illustrates how null-model–based validation can be used to define higher-order interactions in operational statistical terms, enabling principled comparisons between structural, spatial, and informational signatures of multivariate coordination. In this sense, higher-order co-activation patterns emerge not merely as descriptive constructs, but as statistically grounded objects that can inform our understanding of large-scale integration in complex systems, in the brain and beyond.

5.5 Data availability

Data concerning rs-fMRI from the HCP can be found on the Human Connectome Project website; for a detailed description, see the HCP S1200 Release Reference Manual.

Code availability

The Python package named `TRIDENT` (*TRIplets DEtectioN via staTistical validation*), implementing the algorithms described in the main text, is available on PyPI and at the URL `https://github.com/mattiamarzi/TRIDENT`.

Appendix 5.A Canonical null models for hypothesis testing

The canonical null models employed in this work belong to the general family of Exponential Random Graph Models (ERGMs) [4, 148, 5, 6]. Such models can be derived as maximum-entropy ensembles under *soft constraints*, that is, by requiring that selected topological quantities match their observed expectations while maximizing randomness in every other respect. In this work, we are interested in implementing null models for undirected, binary, and bipartite networks. As in [117], we follow the analytical approach described in [148] and further developed in [31]. This approach rests upon an inference procedure allowing us to find a probability distribution which is maximally non-committal with respect to the missing data; quantitatively, it is based on the maximization of the so called *Shannon entropy* $S[P] = -\sum_{\mathbf{B}\in\mathbb{B}} P(\mathbf{B}) \ln P(\mathbf{B})$, where $\mathbb{B}$ is the set of all undirected, binary, bipartite graphs having N nodes in one layer and T nodes in the other, and $P(\mathbf{B})$ is a probability distribution defined over $\mathbb{B}$. The entropy maximization is carried out while constraining the expected values of a chosen vector of proper topological quantities, the constraints vector $\mathbf{c}(\mathbf{B})$. Solving such a maximization problem leads us to find the exponential probability distribution:

$$P(\mathbf{B};\boldsymbol{\theta}) = \frac{e^{-H(\mathbf{B},\boldsymbol{\theta})}}{Z(\boldsymbol{\theta})} = \frac{e^{-\boldsymbol{\theta}\cdot\mathbf{c}(\mathbf{B})}}{Z(\boldsymbol{\theta})}, \tag{5.21}$$

where $\boldsymbol{\theta}$ is the vector of Lagrangian multipliers associated with the vector of constraints $\mathbf{c}(\mathbf{B})$, $H(\mathbf{B},\boldsymbol{\theta}) = \boldsymbol{\theta}\cdot\mathbf{c}(\mathbf{B})$ is the *Hamiltonian* of the model and $Z(\boldsymbol{\theta}) =$

$\sum_{\mathbf{B}\in\mathbb{B}} e^{-\boldsymbol{\theta}\cdot\mathbf{c}(\mathbf{B})}$ is the *partition function* ensuring normalization. To tune the parameters defining the entropy-based null model, we require that the expected values of the constrained quantities $\mathbb{E}[\mathbf{c}(\mathbf{B})]$ equate the corresponding values $\mathbf{c}^* = \mathbf{c}(\mathbf{B}^*)$ computed on the observed data $\mathbf{B}$. Interestingly, it can be shown that in this framework the maximum-likelihood estimation of the Lagrangian multipliers leads precisely to satisfy the condition $\mathbb{E}[\mathbf{c}(\mathbf{B})] = \mathbf{c}(\mathbf{B}^*)$ [149].

In what follows, we focus on models in which the probability of each link can be written as an independent Bernoulli random variable. This assumption holds exactly when the imposed constraints are linear functions of the adjacency matrix, as in the case of node degrees (see the BiCM below). When non-linear quantities are constrained, independence no longer holds exactly; in such cases, we adopt a mean-field approximation, as discussed later for the BiPLOM. Crucially, the independent-link formulation allows the distribution of the number of shared neighbors for any n-tuple of nodes to be expressed as a Poisson–Binomial law, providing a tractable and analytically consistent basis for hypothesis testing. From a modeling perspective, this independence structure is essential: it ensures that the significance of higher-order co-occurrences can be computed analytically rather than estimated by computationally expensive simulations. While here we apply these models to validate interactions between time series, the proposed framework is far more general: it applies to all binary, two-mode datasets that can be represented as bipartite networks. For this reason, in this section, while retaining the same notation as in the main text, we do not explicitly talk about brain or time nodes, but, respectively, the bottom and top layers.

5.A.1 Bipartite Configuration Model

The Bipartite Configuration Model [55] is a linear ERGM for binary, bipartite networks, defined by two sets of local constraints, the degree sequences $\boldsymbol{k}(\mathbf{B}) = (k_1(\mathbf{B}), \cdots, k_N(\mathbf{B}))$ and $\boldsymbol{h}(\mathbf{B}) = (h_1(\mathbf{B}), \cdots, h_T(\mathbf{B}))$. Its Hamiltonian reads

$$H(\mathbf{B};\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{i=1}^{N}\beta_i k_i(\mathbf{B}) + \sum_{t=1}^{T}\gamma_t h_t(\mathbf{B}) = \sum_{i=1}^{N}\sum_{t=1}^{T}(\beta_i+\gamma_t)b_{it}. \tag{5.22}$$

where the N-dimensional vector $\boldsymbol{\beta} = (\beta_1, \cdots \beta_N)$ and the T-dimensional vector $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_T)$ are the parameter vectors corresponding to the top and bottom degrees. The partition function can be computed analytically, and the probability distribution reads

$$P_{\text{BiCM}}(\mathbf{B};\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{e^{-\sum_{i=1}^{N}\sum_{t=1}^{T}(\beta_i+\gamma_t)b_{it}}}{\prod_{i=1}^{N}\prod_{t=1}^{T}(1+e^{-\beta_i-\gamma_t})} = \prod_{i=1}^{N}\prod_{t=1}^{T} p_{it}^{b_{it}}(1-p_{it})^{1-b_{it}}, \tag{5.23}$$

where the parameters p_{it} in the above expression are defined as

$$p_{it} \equiv \frac{e^{-\beta_i-\gamma_t}}{1+e^{-\beta_i-\gamma_t}} \qquad \forall i,t. \tag{5.24}$$

The formula above is often written in terms of the exponential parameters $\boldsymbol{x}, \boldsymbol{y}$, obtained by posing $x_i = e^{-\beta_i}$ and $y_t = e^{-\gamma_t}$ $\forall i, t$:

$$p_{it} \equiv \frac{x_i y_t}{1 + x_i y_t} \quad \forall i, t, \tag{5.25}$$

p_{it} represents the probability that nodes i and t are linked. Thus, under the BiCM, each entry b_{it} of a biadjacency matrix $\mathbf{B} \in \mathbb{B}$ is distributed as a Bernoulli random variable with probability coefficient p_{it}. In turn, this implies that $\mathbf{B}$ is a collection of $N \times T$ independent, non-identically distributed Bernoulli variables.

As mentioned, to tune the parameters defining an ERGM, we can either maximize the likelihood function or solve the constraint equations. We now explicitly show this, as it underscores the differences with the BiPLOM (defined later), where this equivalence is partially lost. The BiCM log-likelihood function reads

$$\begin{aligned}
\mathcal{L}_{\text{BiCM}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) &\equiv \ln P_{\text{BiCM}}(\mathbf{B}^*; \boldsymbol{\beta}_i, \boldsymbol{\gamma}_t) \\
&= \sum_{i=1}^{N} b_{it}^* \ln \left(\frac{e^{-\beta_i - \gamma_t}}{1 + e^{-\beta_i - \gamma_t}} \right) - \sum_{i=1}^{N} \sum_{t=1}^{T} (1 - b_{it}^*) \ln(1 + e^{-\beta_i - \gamma_t}) \\
&= \sum_{i=1}^{N} b_{it}^* \ln(e^{-\beta_i - \gamma_t}) - \sum_{i=1}^{N} \sum_{t=1}^{T} \ln(1 + e^{-\beta_i - \gamma_t}) \\
&= - \sum_{i=1}^{N} \sum_{t=1}^{T} b_{it}^* (\beta_i + \gamma_t) - \sum_{i=1}^{N} \sum_{t=1}^{T} \ln(1 + e^{-\beta_i - \gamma_t}).
\end{aligned} \tag{5.26}$$

In order to maximize it with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we need to solve the system of equations

$$\frac{\partial \mathcal{L}_{\text{BiCM}}(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \beta_i} = - \sum_{t=1}^{T} b_{it}^* + \sum_{t=1}^{T} \frac{e^{-\beta_i - \gamma_t}}{1 + e^{-\beta_i - \gamma_t}} \quad \forall i, \tag{5.27}$$

$$\frac{\partial \mathcal{L}_{\text{BiCM}}(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \gamma_t} = - \sum_{i=1}^{N} b_{it}^* + \sum_{i=1}^{N} \frac{e^{-\beta_i - \gamma_t}}{1 + e^{-\beta_i - \gamma_t}} \quad \forall t, \tag{5.28}$$

and, upon equating them to zero, we find

$$k_i^* = \sum_{t=1}^{T} \frac{e^{-\beta_i - \gamma_t}}{1 + e^{-\beta_i - \gamma_t}} = \sum_{t=1}^{T} p_{it} = \mathbb{E}[k_i] \quad \forall i, \tag{5.29}$$

$$h_t^* = \sum_{i=1}^{N} \frac{e^{-\beta_i - \gamma_t}}{1 + e^{-\beta_i - \gamma_t}} = \sum_{i=1}^{N} p_{it} = \mathbb{E}[h_t] \quad \forall t. \tag{5.30}$$

The BiCM thus provides a first-order baseline that fully accounts for heterogeneity in node activation patterns but not for correlations between nodes. As discussed in the main text, this limitation becomes crucial when validating structures of order three

or higher. Notice that the system above can be solved only numerically. A ready-to-use implementation of the Bipartite Configuration Model, along with several other Maximum Entropy models, is available in the open-source Python package NEMtropy [150].

5.A.2 From first to second-order null models

As explained in the main text, in order to validate triplets, first-order constraints might be insufficient. This motivates the introduction of second-order null models in the validation pipeline described in this work. The main limitation of introducing second-order constraints in a canonical model is the loss of link independence, and consequently of the analytical simplicity that stems from it, an aspect that is central to the validation framework introduced by Saracco et al. [55] and that we aim to extend here. In addition, second-order models are typically not analytically solvable.

To overcome these issues, a mean-field approach is employed [148, 151, 152]. The mean-field approximation builds on the concept of *separable Hamiltonians*, in which the Hamiltonian is expressed as a sum of local contributions that may depend on global or mesoscopic properties of the network. In contrast to linear ERGMs, whose Hamiltonians take the general form

$$H(\mathbf{B}, \boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{t=1}^{T} \theta_{it}\, b_{it}, \tag{5.31}$$

and yield independent link probabilities, separable Hamiltonians adopt a more general expression:

$$H(\mathbf{B}, \boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{t=1}^{T} b_{it}\, f_{it}(\mathbf{B}, \boldsymbol{\theta}), \tag{5.32}$$

where the functions f_{it} introduce non-linear dependencies (e.g., on k_i, h_t, or motif statistics), possibly involving all entries of $\mathbf{B}$.

The mean-field approximation consists of replacing the adjacency matrix $\mathbf{B}$ with its expectation $\mathbf{P} = \mathbb{E}[\mathbf{B}]$, i.e., the matrix of link probabilities, inside the functional terms f_{it}. Under this approximation, the expression becomes separable, and links become independent:

$$P(\mathbf{B}) = \prod_{i=1}^{N} \prod_{t=1}^{T} p_{it}^{b_{it}} (1 - p_{it})^{1 - b_{it}}, \tag{5.33}$$

where now

$$p_{it} = \frac{e^{-f_{it}(\mathbf{P}, \boldsymbol{\theta})}}{1 + e^{-f_{it}(\mathbf{P}, \boldsymbol{\theta})}}. \tag{5.34}$$

This approximation restores independence at the cost of only approximate constraint satisfaction, which is the price paid to maintain analytic tractability.

It is important to note that, once mean-field is applied, solving the constraint equations and maximizing the likelihood are no longer guaranteed to be equivalent. In linear ERGMs (such as the BiCM), the two procedures coincide, but in non-linear models, the mean-field approximation effectively modifies the model class. This subtlety will play a key role in the construction of the BiPLOM in the next section.

5.A.3 The Bipartite Partial Local Overlap Model

Here, we introduce the Bipartite Partial Local Overlap Model, a second-order ERGM which we solve in the mean-field approximation.

Ideally, in order to validate co-occurrences between triplets, a model designed to reproduce, beyond degree sequences, local second-order statistics is preferred. A direct way to impose second-order information would be to constrain all pairwise overlaps V_{ij}. However, since there are $O(N^2)$ of them, this approach is computationally prohibitive and would make the model intractable. For this reason, we seek summary statistics that encode local second-order structure while keeping the number of constraints linear in N. To this aim, for each node i in the bottom layer, define the aggregate overlap

$$V_i = \sum_{t=1}^{T} \sum_{\substack{j=1 \\ (j\neq i)}}^{N} b_{it} b_{jt} = \sum_{t=1}^{T} b_{it}(h_t - b_{it}) \qquad \forall i. \tag{5.35}$$

V_i counts the total number of V_{ij} motifs in which node i participates, or equivalently, the total number of bottom-layer nodes that share at least one neighbor with i. By constraining both degree sequences and the V_i patterns, the resulting null model accounts for (i) the local tendency of each node to form links (through degrees) and (ii) its local tendency to share neighbors with others. This provides a suitable baseline for statistically validating triplets whose co-occurrences cannot be explained by lower-order tendencies. Such constraints induce the Hamiltonian

$$H(\mathbf{B}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = \sum_{i=1}^{N} \beta_i k_i(\mathbf{B}) + \sum_{t=1}^{T} \gamma_t h_t(\mathbf{B}) + \sum_{i=1}^{N} \delta_i V_i(\mathbf{B}) \tag{5.36}$$

$$= \sum_{i=1}^{N} \sum_{t=1}^{T} \beta_i b_{it} + \sum_{t=1}^{T} \sum_{i=1}^{N} \gamma_t b_{it} + \sum_{i=1}^{N} \sum_{t=1}^{T} \delta_i b_{it}(h_t(\mathbf{B}) - b_{it}) \tag{5.37}$$

$$= \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\beta_i + \gamma_t + \delta_i \big(h_t(\mathbf{B}) - b_{it}\big) \right] b_{it}. \tag{5.38}$$

The mean field approximation requires replacing $\left[\beta_i + \gamma_t + \delta_i(h_t(\mathbf{B}) - b_{it})\right]$ with its expectation under the model, $\left[\beta_i + \gamma_t + \delta_i(\mathbb{E}[h_t] - p_{it})\right]$. The induced mean-field link probabilities read

$$p_{it} = \frac{e^{-\beta_i - \gamma_t - \delta_i(\mathbb{E}[h_t] - p_{it})}}{1 + e^{-\beta_i - \gamma_t - \delta_i(\mathbb{E}[h_t] - p_{it})}}. \tag{5.39}$$

Upon assuming that the constraints on the degrees are met, one gets:

$$p_{it} = \frac{e^{-\beta_i - \gamma_t - \delta_i(h_t^* - p_{it})}}{1 + e^{-\beta_i - \gamma_t - \delta_i(h_t^* - p_{it})}}. \tag{5.40}$$

However, it is easily shown that the solution to the constraint equation is degenerate

in the mean-field approximation. In fact, for each i the expectation of V_i reads:

$$\begin{aligned}\mathbb{E}[V_i] &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{\substack{j=1\\j\neq i}} b_{it}b_{jt}\right] = \sum_{t=1}^{T}\sum_{\substack{j=1\\j\neq i}} p_{it}p_{jt} \\ &= \sum_{t=1}^{T}\sum_{j=1}^{N} p_{it}p_{jt} - \sum_{t=1}^{T} p_{it}^2 \qquad (5.41)\\ &= \sum_{t=1}^{T} p_{it}(h_t^* - 1) + \sum_{t=1}^{T} p_{it}(1-p_{it}) \qquad (5.42)\\ &= \sum_{t=1}^{T} p_{it}(h_t^* - 1) + \mathrm{Var}[k_i], \qquad (5.43)\end{aligned}$$

where we assumed again that the constraints on the degrees are met. Requiring $\mathbb{E}[V_i] = V_i^*$ then implies

$$\sum_{t=1}^{T} p_{it}(h_t^* - 1) + \mathrm{Var}[k_i] = \sum_{t=1}^{T} b_{it}^*(h_t^* - 1) \qquad \forall i. \tag{5.44}$$

Summing both sides over i, one gets:

$$\sum_{t=1}^{T} h_t^*(h_t^*-1) + \sum_{i=1}^{N}\mathrm{Var}[k_i] = \sum_{t=1}^{T} h_t^*(h_t^*-1) \implies \sum_{i=1}^{N}\mathrm{Var}[k_i] = 0 \implies \mathrm{Var}[k_i] = 0\ \forall i. \tag{5.45}$$

Moreover, in the independent-links approximation,

$$\mathrm{Var}[k_i] = \sum_{t=1}^{T} p_{it}(1-p_{it}) = 0 \qquad \forall i, \tag{5.46}$$

which is obtained if and only if $p_{it} \in \{0, 1\}$. Thus, the model collapses to a deterministic configuration: no canonical ensemble exists.

Since exact constraint satisfaction is impossible, one may attempt to maximize likelihood. However, in this case, the expression for p_{it} in (5.40) does not yield a closed form depending only on model parameters and observed data.

To solve this problem, we introduce a modified version of V_i, including the *self term* $j = i$. Formally, we constrain, together with the degrees, the quantities

$$\widetilde{V}_i = \sum_{t=1}^{T}\sum_{j=1}^{N} b_{it}b_{jt} = \sum_{t=1}^{T} b_{it}\, h_t \qquad \forall i. \tag{5.47}$$

This modification does not remove the degeneracy that would arise from enforcing the constraints exactly: if one attempts to solve the constraint equations, the system

still collapses to a deterministic configuration:

$$\mathbb{E}_{\mathrm{MF}}[\widetilde{V}_i] = \mathbb{E}_{\mathrm{MF}}\left[\sum_{t=1}^{T} b_{it}h_t\right] = \sum_{t=1}^{T}\left(\sum_{\substack{j=1\\ j\neq i}}^{N} p_{it}p_{jt} + p_{it}\right)$$
$$= \sum_{t=1}^{T}\left(\sum_{j=1}^{N} p_{it}p_{jt} - p_{it}(p_{it}-1)\right) = \sum_{t=1}^{T} p_{it}h_t^* - \mathrm{Var}_{\mathrm{MF}}[k_i] \qquad \forall i, \quad (5.48)$$

where we used the fact that $b_{it}^2 = b_{it}$. Equating this expression to $\widetilde{V}_i$ and summing over i, one ends up again with condition (5.46). The role of $\widetilde{V}_i$ is, then, different and purely operational: by including the self-term, the link probabilities p_{it} acquire a closed-form expression, which makes maximum-likelihood estimation feasible even though exact constraint satisfaction remains impossible.

The modified vector of patterns $\widetilde{\boldsymbol{V}} = (\widetilde{V}_1, \cdots, \widetilde{V}_N)$, together with the degree sequences $\boldsymbol{k}$ and $\boldsymbol{h}$, define the constraints of the Bipartite Partial Local Overlap Model. The model Hamiltonian reads:

$$H(\mathbf{B}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\beta_i + \gamma_t + \delta_i\, h_t(\mathbf{B})\right] b_{it}. \quad (5.49)$$

Applying the mean-field approximation yields:

$$H_{\mathrm{MF}}(\mathbf{B}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = \sum_{i=1}^{N}\sum_{t=1}^{T} b_{it}\left[\beta_i + \gamma_t + \delta_i\, \mathbb{E}[h_t]\right] = \sum_{i=1}^{N}\sum_{t=1}^{T} b_{it}\left[\beta_i + \gamma_t + \delta_i\, h_t^*\right], \quad (5.50)$$

where we assume that the constraints on the degrees $\boldsymbol{h}$ are satisfied. Thus, in mean-field, the distribution factorizes:

$$P_{\mathrm{MF}}(\mathbf{B}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = \prod_{i=1}^{N}\prod_{t=1}^{T} p_{it}^{b_{it}}(1-p_{it})^{1-b_{it}}, \quad (5.51)$$

with link probabilities admitting the closed form

$$p_{it} = \frac{e^{-\beta_i - \gamma_t - \delta_i h_t^*}}{1 + e^{-\beta_i - \gamma_t - \delta_i h_t^*}}. \quad (5.52)$$

The log-likelihood reads:

$$\mathcal{L}_{\mathrm{MF}}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = -\sum_{i=1}^{N}\sum_{t=1}^{T} b_{it}^*(\beta_i + \gamma_t + \delta_i h_t^*) - \sum_{i=1}^{N}\sum_{t=1}^{T} \ln\left(1 + e^{-\beta_i - \gamma_t - \delta_i h_t^*}\right). \quad (5.53)$$

Setting its derivatives to zero, we get for the degree constraints:

$$\frac{\partial \mathcal{L}}{\partial \beta_i} = -\sum_{t=1}^{T} b_{it}^* + \sum_{t=1}^{T} p_{it} = 0 \quad \Longrightarrow \quad k_i^* = \sum_{t=1}^{T} p_{it} = \mathbb{E}_{\mathrm{MF}}[k_i], \quad (5.54)$$

$$\frac{\partial \mathcal{L}}{\partial \gamma_t} = -\sum_{i=1}^{N} b_{it}^* + \sum_{i=1}^{N} p_{it} = 0 \quad \Longrightarrow \quad h_t^* = \sum_{i=1}^{N} p_{it} = \mathbb{E}_{\mathrm{MF}}[h_t]. \quad (5.55)$$

For the $\widetilde{V}_i$ constraints:

$$\frac{\partial \mathcal{L}}{\partial \delta_i} = -\sum_{t=1}^{T} h_t^* b_{it}^* + \sum_{t=1}^{T} h_t^* p_{it} = 0 \qquad \Longrightarrow \qquad \widetilde{V}_i^* = \sum_{t=1}^{T} h_t^* \, p_{it} \tag{5.56}$$

while the expected value is:

$$\mathbb{E}_{\mathrm{MF}}[\widetilde{V}_i] = \sum_{t=1}^{T} p_{it} h_t^* - \mathrm{Var}_{\mathrm{MF}}[k_i]. \tag{5.57}$$

Therefore:

$$\widetilde{V}_i^* = \mathbb{E}_{\mathrm{MF}}[\widetilde{V}_i] + \mathrm{Var}_{\mathrm{MF}}[k_i] \qquad \forall i. \tag{5.58}$$

The model thus reproduces, in expected value, first-order patterns exactly, and second-order patterns as closely as possible while preserving fluctuations. By doing so, it captures each node's local tendency to co-activate with others, without collapsing onto the observed configuration.

Figures 5.9 and 5.10 summarize how BiPLOM enforces the constraints in practice. The parity plots in Figure 5.9 compare the constraint implementation of BiCM (top row) and BiPLOM (bottom row) for a representative subject. Both models accurately reproduce the degree sequences. In particular, the BiPLOM mean-field estimates $\mathbb{E}\mathrm{MF}[k_i]$ and $\mathbb{E}\mathrm{MF}[h_t]$ lie essentially on the identity line when plotted against the empirical degrees, indicating that first-order patterns are matched up to the numerical tolerance of the solver. As expected, BiCM does not reproduce the individual tendency to activate in pairs, as reflected in the aggregated overlaps V_i. By construction, BiPLOM captures these second-order aggregated tendencies much more accurately, albeit at the cost of the mean-field approximation. The bottom-right panel reports the aggregated overlaps $\widetilde{V}_i$, comparing the empirical values with both the plain mean-field prediction $\mathbb{E}_{\mathrm{MF}}[\widetilde{V}i]$ and the variance-corrected estimate $\mathbb{E}_{\mathrm{MF}}[\widetilde{V}_i] + \mathrm{Var}_{\mathrm{MF}}[k_i]$.

A complementary error diagnostic for $\widetilde{V}_i$ across subjects is shown in Figure 5.10: the histogram displays the distribution of relative errors across all nodes and subjects; the violin plot reports the per-subject median relative error; and the hexbin plot compares the empirical discrepancy $\widetilde{V}_i^* - \mathbb{E}_{\mathrm{MF}}[\widetilde{V}_i]$ with the corresponding variance-correction term. Taken together, these panels show that including the variance term systematically compensates for the bias inherent in the plain mean-field estimate of $\widetilde{V}_i$.

Although this systematic bias is intrinsic to our approximation, it is predictable and therefore controllable: a direct check on the variances provides an a posteriori diagnostic of the magnitude of the correction. In this sense, the residual discrepancy represents the price to pay for retaining a model that remains analytically tractable and computationally efficient, and that can therefore be implemented simply and efficiently.

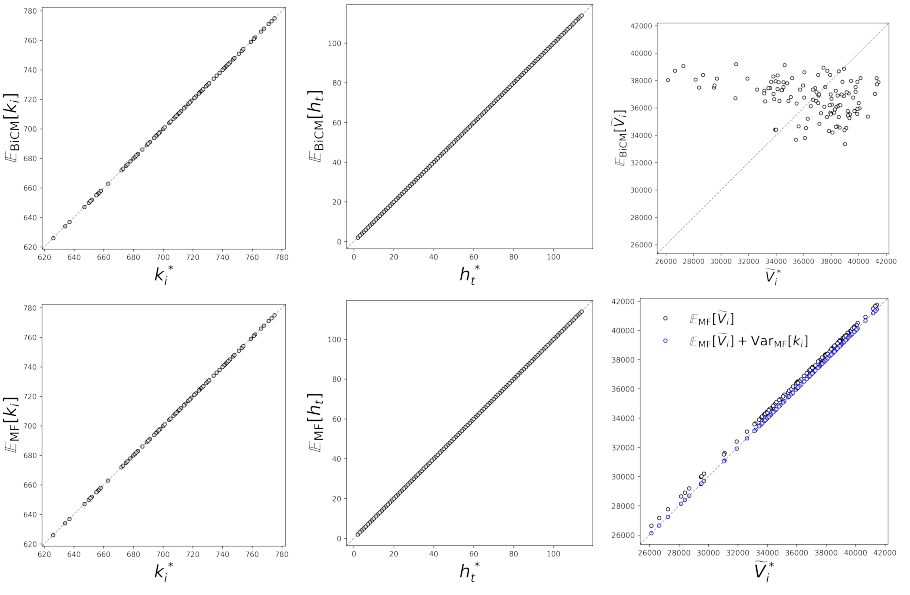


Figure 5.9: ***Constraint implementation with BiCM and BiPLOM.*** *Top row: parity plots for a random subject (id: 366466) for bottom layer degrees k_i (left), top layer degrees h_t (centre), and aggregate overlaps $\widetilde{V}_i$ (right). Each marker reports an empirical value on the horizontal axis and the corresponding mean-field prediction on the vertical axis; the dashed line represents the identity.*

Appendix 5.B Approximations of the Poisson-binomial distribution

As discussed in Section 5.2.2, for each subject and each pair or triplet of regions, we count the number of co-activations V_{ij} and V_{ijk} from the bipartite matrix $\mathbf{B}$. Under the BiCM and BiPLOM null models (Section 5.3.2), the entries b_{it} are modeled as independent Bernoulli variables with success probabilities p_{it}. For a fixed n-tuple of regions ($n = 2, 3$), the contribution of each time point t is Bernoulli with probability

$$q_t = \begin{cases} p_{it}p_{jt}, & n = 2, \\ p_{it}p_{jt}p_{kt}, & n = 3, \end{cases} \tag{5.59}$$

so that the total number of co-activations can be written in the general form

$$S = \sum_{t=1}^{T} X_t, \tag{5.60}$$

where the X_t are independent Bernoulli variables with heterogeneous success probabilities q_t. In this setting, S follows a Poisson-binomial distribution:

$$f_{\text{BP}}(S = s) = \sum_{\substack{A \subseteq \{1,\dots,T\} \\ |A| = s}} \prod_{t \in A} q_t \prod_{t \notin A} (1 - q_t), \qquad s = 0, 1, \dots, T, \tag{5.61}$$

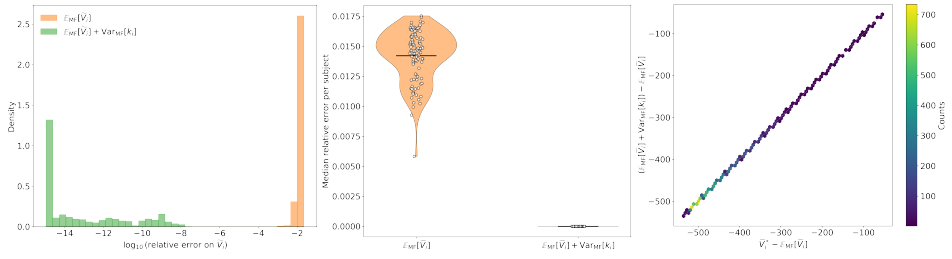


Figure 5.10: ***BIPLOM error diagnostics for the aggregate overlaps.*** *Left panel: distribution of* $\log_{10}$ *relative errors on* $\widehat{V}_i$ *across all nodes and subjects, comparing the prediction* $\mathbb{E}_{MF}[\widehat{V}_i]$ *(orange) with the variance corrected prediction* $\mathbb{E}_{MF}[\widehat{V}_i] + \mathrm{Var}_{MF}[k_i]$ *(green). Central panel: violin plot of the per-subject median relative errors for the same two predictions. Right panel: hexbin plot of* $\widehat{V}_i^* - \mathbb{E}_{MF}[\widehat{V}_i]$ *versus* $(\mathbb{E}_{MF}[\widehat{V}_i] + \mathrm{Var}_{MF}[k_i]) - \mathbb{E}_{MF}[\widehat{V}_i]$*, with color indicating the number of nodes falling in each hexagonal bin.*

whose mean, variance, and skewness are

$$\mu = \mathbb{E}[S] = \sum_{t=1}^{T} q_t, \tag{5.62}$$

$$\sigma^2 = \mathrm{Var}(S) = \sum_{t=1}^{T} q_t(1 - q_t), \tag{5.63}$$

$$\gamma = \sigma^{-3} \sum_{t=1}^{T} q_t(1 - q_t)(1 - 2q_t). \tag{5.64}$$

Given an observed co-activation count V_{obs}, the exact upper-tail p-value under the Poisson-binomial model is

$$p_{\mathrm{PB}}(V_{\mathrm{obs}}) = \mathbb{P}_{\mathrm{PB}}(S \geqslant V_{\mathrm{obs}}) = \sum_{s=V_{\mathrm{obs}}}^{T} f_{\mathrm{PB}}(s), \tag{5.65}$$

where f_{PB} is the Poisson-binomial probability mass function. In our implementation, this tail is evaluated using an FFT-based Hong-type algorithm and serves as the reference for assessing the accuracy of the approximations described below.

Exact Poisson-binomial tails provide a principled baseline but become computationally expensive when applied to all pairs and triplets of all 100 subjects. Following [55], we therefore consider three approximations that replace the Poisson-binomial distribution with simpler surrogates, all sharing the same mean μ and variance σ^2.

The Poisson approximation assumes that the success probabilities q_t are small and that $\sum_t q_t^2$ is not too large. In this regime, S is approximated by a Poisson variable with rate μ, and the upper-tail p-value is

$$p_{\mathrm{Poisson}}(V_{\mathrm{obs}}) \approx \mathbb{P}\big(\mathrm{Poisson}(\mu) \geqslant V_{\mathrm{obs}}\big) = \sum_{s=V_{\mathrm{obs}}}^{\infty} \frac{\mu^s \, \mathrm{e}^{-\mu}}{s!}. \tag{5.66}$$

This approximation is simple and efficient, but it can underestimate the tail probability when the q_t are heterogeneous or not uniformly small.

A second approximation relies on the central limit theorem. For sufficiently large T and moderate skewness, S can be approximated by a Gaussian variable with mean μ and variance σ^2. Introducing the continuity-corrected standardized statistic

$$Z = \frac{V_{\text{obs}} + 0.5 - \mu}{\sigma}, \tag{5.67}$$

the upper tail is approximated by

$$p_{\text{normal}}(V_{\text{obs}}) \approx \mathbb{P}\big(\mathcal{N}(0,1) \geqslant Z\big) = \Phi(-Z), \tag{5.68}$$

where Φ is the standard normal cumulative distribution function. The continuity correction improves accuracy in the discrete setting, particularly near the validation threshold.

The refined normal (rna) approximation incorporates the skewness γ of S into the Gaussian prediction, providing a one-parameter correction to the normal tail that remains tractable for large numbers of tests. Following [55], define

$$G(x) = \Phi(-x) - \gamma \frac{1 - x^2}{6}\, \varphi(x), \tag{5.69}$$

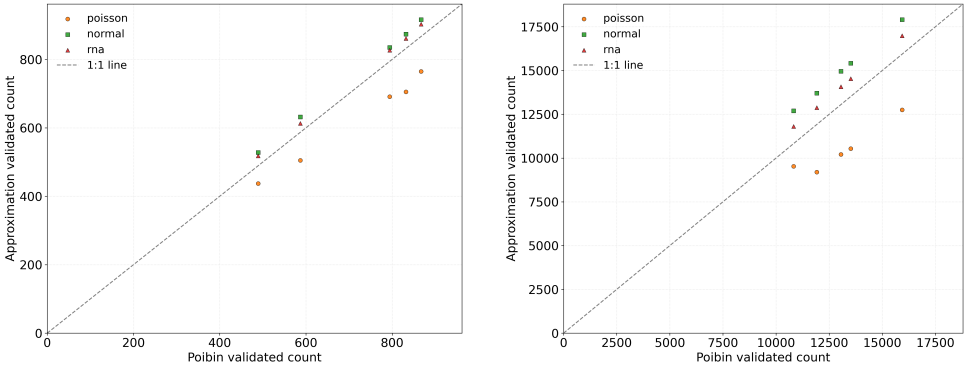
where φ is the standard normal density. The upper-tail p-value is then approximated as

$$p_{\text{rna}}(V_{\text{obs}}) \approx G\left(\frac{V_{\text{obs}} + 0.5 - \mu}{\sigma}\right). \tag{5.70}$$

For $\gamma = 0$, the refined approximation reduces to the ordinary normal approximation. For moderate skewness, the skewness correction improves the agreement with the Poisson-binomial tail without sacrificing computational efficiency. In our implementation, these three approximations are available as alternative options in the validation routines for pairs and triplets (BiPLOM).

To select a surrogate that can be safely used on the full cohort of 100 subjects, where exact Poisson-binomial calculations for all pairs and triplets would be computationally prohibitive, we benchmark Poisson, normal, and rna against the exact Poisson-binomial model on five randomly selected subjects, identified as 100307_treshold1, 125525_treshold1, 149539_treshold1, 198451_treshold1, and 899885_treshold1. For each subject and structure type (pairs for BiCM, triplets for BiPLOM), we compute, for each method, the total number of validated structures, the distribution of associated p-values, and the overlap between the corresponding sets of validated structures.

The scatter plots in Figure 5.11 compare, for each subject, the number of validated structures under Poisson, normal, and rna approximations with respect to the Poisson-binomial baseline. For BiCM pairs, the Poisson approximation validates on average about 0.87 times as many pairs as the Poisson-binomial model (mean ratio Poisson/PoiBin ≈ 0.871), whereas the normal and rna approximations are slightly more liberal, with mean ratios normal/PoiBin ≈ 1.063 and rna/PoiBin ≈ 1.044. The same pattern becomes more pronounced for BiPLOM triplets, where Poisson validates on average only about 0.80 times as many triplets as Poisson-binomial (mean ratio ≈ 0.803), while normal and rna validate about 1.15 and 1.08 times as many triplets, respectively (mean ratios ≈ 1.147 and ≈ 1.079). The points lie close to



(a) *(b)*

Figure 5.11: ***Validated counts under Poisson-binomial surrogates.*** *Scatter plots comparing the number of validated structures obtained with Poisson (orange), normal (green), and rna (red) approximations against the Poisson-binomial baseline (x-axis) across the five benchmark subjects. (a) BiCM pairs. (b) BiPLOM triplets. The dashed diagonal marks equality.*

the diagonal for all methods, indicating that the approximations track the baseline counts across subjects, but with a systematic conservative bias for Poisson and a mild anti-conservative bias for normal and rna. Across both structure types, rna remains close to the diagonal and avoids the strong underestimation of Poisson and the more pronounced over-validation of normal.

To characterize p-value distributions beyond simple counts, we analyze the transformed scores $z = -\log p$ (natural logarithm). For each approximation m and structure type, the violin plots in Figure 5.12 display the empirical distribution of z across structures and subjects. Each violin encodes a smoothed kernel density estimate

$$\hat{f}_m(z) = \frac{1}{Nh} \sum_{k=1}^{N} K\left(\frac{z - z_k}{h}\right), \tag{5.71}$$

where z_k are the observed $-\log p$ values for method m, K is a kernel function and h is the bandwidth. Higher and wider violins toward large z indicate heavier tails of small p-values and thus more aggressive validation. For both BiCM pairs and BiPLOM triplets, the rna approximation produces $-\log p$ distributions that closely follow the Poisson-binomial reference, with only a modest shift towards larger values. By contrast, the Poisson approximation concentrates more mass at smaller z, that is, larger p-values, which is consistent with its systematically smaller number of validated structures.

We also inspect the kernel density of the original p-values. For each approximation m, the curves in Figure 5.13 represent the kernel density estimate

$$\hat{g}_m(p) = \frac{1}{Nh} \sum_{k=1}^{N} K\left(\frac{p - p_k}{h}\right), \tag{5.72}$$

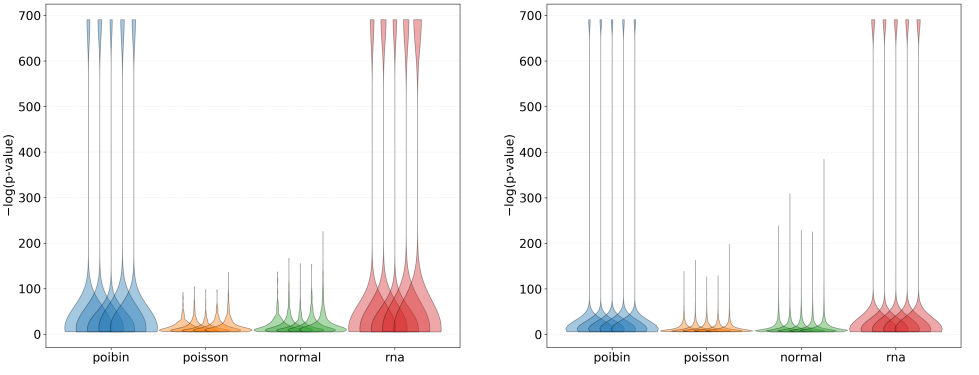
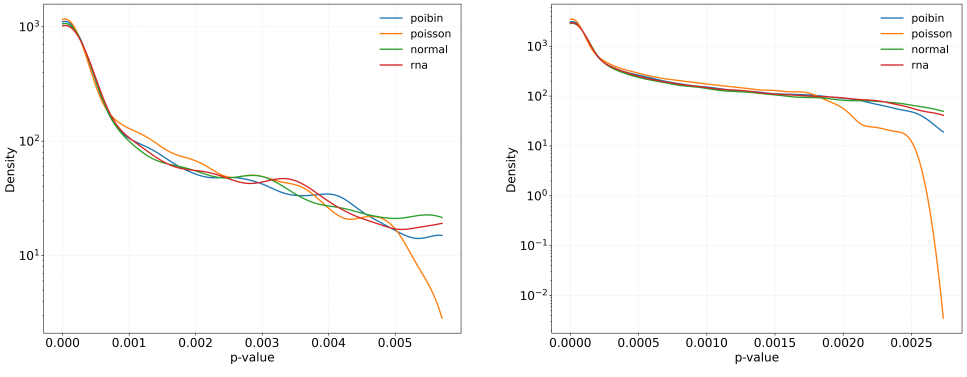


Figure 5.12: ***Distribution of significance scores across approximations.*** *Violin plots of $z = -\log p$ (natural logarithm) for Poisson, normal, and rna approximations, shown separately for each benchmark subject. (a) BiCM pairs. (b) BiPLOM triplets. Colors follow Fig. 5.11.*

constructed from all p-values p_k pooled across the five subjects. The horizontal axis reports the p-value on a linear scale, while the vertical axis reports the estimated density on a logarithmic scale, which emphasizes the behavior near $p = 0$. For both structure types, Poisson shows a marked depletion of small p-values compared to the Poisson-binomial baseline, while normal and rna closely track the Poisson-binomial density. The rna curves are slightly enhanced in the extreme tail, in line with the modestly higher number of validated structures reported in Figure 5.11, but remain very close to the Poisson-binomial reference.

The stacked overlap plots in Figure 5.14 decompose, for each benchmark subject, the union of the Poisson-binomial and approximate validated sets into structures validated only by Poisson-binomial, by both methods, or only by the approximation. For BiCM pairs, the Poisson approximation recovers on average about 87% of Poisson-binomial validated pairs, and 100% of Poisson-validated pairs are also validated by Poisson-binomial. The normal and rna approximations recover essentially all Poisson-binomial validated pairs (mean coverage close to 100%). In contrast, about 94% and 96% of normal- and rna-validated pairs, respectively, are also validated by the Poisson-binomial model. For BiPLOM triplets, the pattern is similar but more pronounced: Poisson recovers about 80% of Poisson-binomial triplets, and all Poisson-validated triplets are contained in the Poisson-binomial set, whereas normal and rna recover essentially 100% of Poisson-binomial triplets and about 87% and 93% of their own validated triplets overlap with Poisson-binomial. For the representative subject 899885, the rna approximation validates 14539 triplets at a subject-specific threshold $p \leqslant 0.00287$, corresponding to approximately $14539/253460 \approx 5.7\%$ of all possible triplets, and 92.98% of these rna-validated triplets are also validated under Poisson-binomial, while all Poisson-binomial validated triplets are recovered by rna. All proportions are computed on linear scales in both axes.



(a) *(b)*

Figure 5.13: ***p-value densities under Poisson-binomial surrogates.*** *Kernel density estimates of p-values for Poisson-binomial (blue), Poisson (orange), normal (green), and rna (red), pooled across benchmark subjects. (a) BiCM pairs. (b) BiPLOM triplets. The density axis is shown on a logarithmic scale to emphasize the small-p regime driving validation.*

To summarize the similarity between the Poisson-binomial and approximate validated sets, we use the Jaccard index. For each subject and approximation, we define

$$J\big(H_{\mathrm{PoiBin}}, H_{\mathrm{approx}}\big) = \frac{|H_{\mathrm{PoiBin}} \cap H_{\mathrm{approx}}|}{|H_{\mathrm{PoiBin}} \cup H_{\mathrm{approx}}|}, \tag{5.73}$$

where H_{PoiBin} is the set of structures validated by the Poisson-binomial model and H_{approx} is the set validated by the approximation. The Jaccard index equals 1 when the two sets coincide and 0 when they are disjoint. The bar plots in Figure 5.15 report the mean Jaccard index across the five subjects, together with one standard deviation. For BiCM pairs, the mean Jaccard index is about 0.87 for Poisson, 0.95 for normal, and 0.96 for rna. For BiPLOM triplets, the mean Jaccard index decreases slightly to about 0.78 for Poisson, 0.87 for normal, and 0.93 for rna. These values confirm that all approximations remain strongly nested within the Poisson-binomial benchmark, while highlighting systematic differences: Poisson is clearly conservative, with smaller validated sets and lower Jaccard values, whereas normal and especially rna achieve a higher recall of Poisson-binomial validated structures at the cost of a modest number of additional validations.

Taken together, the evidence from validated counts, p-value distributions, overlap decompositions, and Jaccard indices consistently points to rna as the most faithful approximation to the Poisson-binomial model. Compared with Poisson, rna avoids the strong underestimation of the number of validated structures and the depletion of small p-values. Compared with normal, rna achieves higher Jaccard indices and a closer match to the Poisson-binomial p-value density, particularly in the small p-value regime that drives validation. Based on this five-subject benchmark, we therefore adopt rna as the default surrogate for the Poisson-binomial model in the large-scale analysis of all 100 subjects.

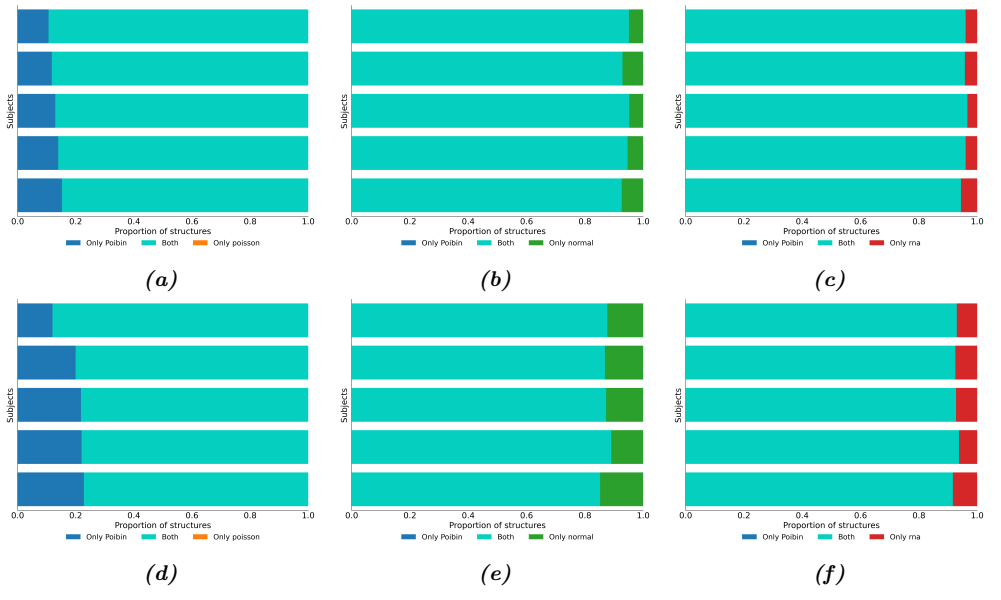


Figure 5.14: ***Overlap between Poisson-binomial and approximate validated sets.*** *Stacked decomposition of the union between Poisson-binomial validations and each approximation into structures validated only by Poisson-binomial, by both methods, or only by the approximation. Panels compare (a,d) Poisson, (b,e) normal, and (c,f) rna approximations against Poisson-binomial. Top row: BiCM pairs. Bottom row: BiPLOM triplets. For each subject, bar lengths are normalized by the union size so that each bar sums to one; subjects are ordered by decreasing fraction of Poisson-binomial-only structures.*

Appendix 5.C Atlas

The atlas considered, being the union between the Schaefer Atlas and the Tian Atlas, includes both subcortical and cortical regions. Subcortical regions are divided into left and right hippocampus, left and right amygdala, left and right posterior and anterior thalamus, left and right nucleus accumbens, left and right globus pallidus, left and right putamen, and left and right caudate nucleus. Cortical regions, instead, are grouped based on the seven canonical Resting-State Networks (RSNs) defined by Yeo et al. [135]. Specifically, these include 17 different areas corresponding to the visual system (VS), 14 accounting for the sensorimotor network (SMN), 15 for the Dorsal Attention Network (DAN), 12 for the Salience Ventral Attention Network (SVA), 5 for the limbic system, 13 for the Central Executive Network (CEN), also called Fronto-Parietal Network (FPN), and 24 for the Default Mode Network (DMN). The above partition is represented in Fig.5.16.

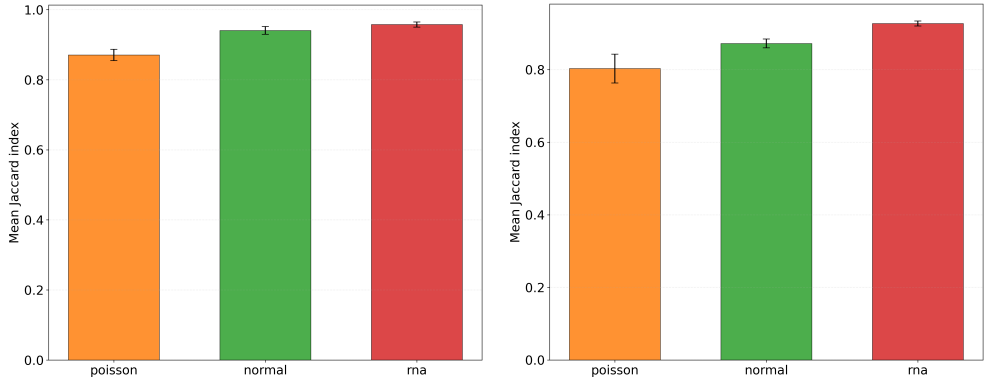


Figure 5.15: ***Agreement with Poisson-binomial validated sets.*** *Mean Jaccard index between the Poisson-binomial validated set and the sets obtained with Poisson, normal, and rna approximations, computed across the five benchmark subjects. (a) BiCM pairs. (b) BiPLOM triplets. Error bars indicate one standard deviation across subjects.*

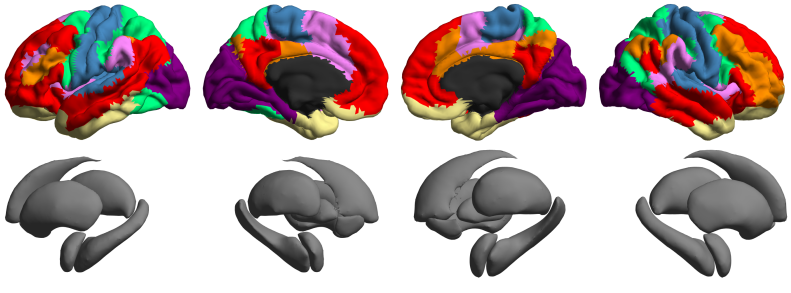


Figure 5.16: ***Yeo partition illustration on the brain.*** *Here, we show Yeo's partition through the Enigma Toolbox [153]. In gray, the subcortical structures that were not actually included in the Yeo analysis and here have been grouped together, in blue the SMN, in purple VS, in orange FPN, in green DAN, in pink VAN, in yellow the limbic system, in red the DMN.*

CHAPTER 6
Conclusion

This thesis has investigated how learning from data can be carried out when models are defined through constraints, focusing on three closely related inferential tasks: model selection, hypothesis testing, and validation. Throughout, maximum entropy models have provided a unifying formal framework for addressing these tasks in a coherent way. In what follows, we discuss the main results of this work and outline a number of open questions that naturally arise from them.

6.1 Model selection and ensemble non-equivalence

The first part of the thesis focused on model selection between canonical and microcanonical maximum entropy formulations. Chapter 2 addressed this problem within the Minimum Description Length principle, using Normalized Maximum Likelihood as a universal coding distribution. From this perspective, the distinction between canonical and microcanonical formulations becomes operational, as reflected in differences in the description length that depend on the observed data. The chapter shows that these differences remain finite when ensemble equivalence holds, but can scale with the system size whenever non-equivalence is induced by an extensive number of constraints.

While the NML approach is less widely adopted than Bayesian methods within the complex systems community, it has a clear and well-founded justification in terms of data compression and provides a principled way to construct description lengths that applies the same criterion across competing models. Chapter 2 also clarified the relationship between NML-based and Bayesian approaches. In settings where ensemble equivalence holds, the NML construction asymptotically coincides with the use of Jeffreys' prior. However, this relation breaks down in the presence of non-equivalence induced by local constraints. Whether a continuous prior exists that asymptotically reproduces the NML description length in such settings remains an open problem.

Chapter 2 also shows that, from a Bayesian perspective, canonical and microcanonical formulations can in principle be made to agree through a suitable choice of prior. This is somehow confirmed in Chapter 3 through the construction of the microcanonical approximation: a universal canonical distribution can always be rewritten as a Bayesian microcanonical universal distribution (even though the opposite is not always true). Consequently, the same holds for description lengths. Rather

than resolving the issue of non-equivalence, however, this observation highlights a more general point: different priors can induce differences in description length that themselves scale with system size. In this sense, the Bayesian formulation effectively shifts the problem from the choice of ensemble to the choice of prior, underscoring that prior assumptions can yield large gains or losses in description length and model selection outcomes, depending on how well they align with the data.

Finally, Chapter 2 opens the way to a fully model-based notion of equivalence. The standard measure-level definition of ensemble equivalence is formulated at the level of probability measures, comparing the maximum-likelihood distributions associated with different ensembles via their Kullback–Leibler divergence evaluated on the observed data. This notion, therefore, concerns ensembles, understood as individual distributions, rather than models, understood as families of distributions. Since the presence or absence of non-equivalence is determined by the choice of constraints - that is, by the choice of models themselves - it is natural to seek a definition of equivalence at the model level, taking into account both likelihoods and complexities. In this direction, we suggested a generalization based on the Kullback–Leibler divergence between the universal NML distributions of two models. As shown in Chapter 2, this quantity depends only on the chosen constraints and not on the observed data. We proved that ensemble equivalence implies model-level equivalence in this sense, whereas the converse implication remains open.

As a final remark, the analysis in Chapter 2 deliberately restricted attention to non-equivalence arising from the choice and scaling of constraints. Connections to other sources of non-equivalence, such as those associated with long-range interactions, as well as their detectability from an MDL perspective, were left open for future investigation.

6.2 E-values and MDL

In Chapter 3, growth-rate-optimal e-values were constructed for tests in which the competing hypotheses correspond to canonical or microcanonical maximum entropy models. Exact constructions were obtained for the microcanonical case, while for the canonical case, a microcanonical approximation was proposed and justified both theoretically and empirically. The resulting e-values were illustrated in examples for 2×2 and $2 \times k$ contingency tables, showing that the approximation performs well across a range of settings.

A key insight emerging from this analysis is that universal distributions, already central in MDL due to their redundancy-minimizing properties, naturally induce e-values with small regret. This observation provides a principled justification for constructing GRO e-variables from universal distributions and establishes a clear conceptual link between e-values and inference based on description lengths.

Building on this connection, an important open problem is the unification of model selection based on description lengths and hypothesis testing based on e-values within a single framework. While this thesis has clarified their connection in settings involving pairwise comparisons between models, extending this perspective to multi-model selection problems remains open. In particular, one would like to attach

interpretable measures of statistical evidence or significance (type-I error control) to differences in description length across a collection of competing models. Despite its technical challenges, such a unification is conceptually appealing, as it would bring model selection and hypothesis testing under a common notion of evidence and further strengthen the role of universal distributions and description lengths as central inferential quantities.

6.3 Conditional and microcanonical e-values

Chapter 4 focused explicitly on 2×2 contingency tables and on the practical question of which e-values are appropriate, relative to classical p-values, in different inferential scenarios. Both single-shot and sequential settings were considered, with particular attention to finite-sample behavior. Preliminary results highlighted a trade-off across inferential regimes. E-variables optimized for sequential accumulation of evidence, such as Turner's e-variable, perform well when group sizes are comparable and data arrive in many small batches, that is, when information is updated repeatedly. In contrast, conditional e-values, and especially uniformly most powerful ones, are comparatively preferable when data arrive in a small number of large batches. The analysis was intentionally exploratory and aimed at clarifying qualitative differences between approaches rather than providing a definitive prescription.

In this study, the microcanonical e-values introduced in Chapter 3 were not employed directly, even though they partially inspired the definition of conditional e-variables. The precise formal relationship between microcanonical e-variables and the conditional e-values used in Chapter 4 remains to be clarified. Microcanonical e-variables are defined for general choices of constraints under both the null and the alternative and therefore apply broadly to tests between constrained models. By contrast, the conditional e-variables considered in Chapter 4, while related to the general theory of exponential families, were specifically tailored to the 2×2 contingency table setting.

At a conceptual level, however, the two constructions are closely related. In the specific case of 2×2 contingency tables, conditional e-variables are introduced as tools for canonical models and are obtained by conditioning on the observed null sufficient statistic, thereby reducing the problem to a single continuous parameter. Microcanonical e-variables, instead, are defined directly for microcanonical models; nevertheless, such models can themselves be derived from canonical formulations by conditioning both the null and the alternative on their respective sufficient statistics, leading to a description in terms of two discrete parameters. In this sense, both approaches exploit conditioning to reduce complexity and render the resulting inference tractable, albeit through formally different constructions.

Although these differences are primarily technical, clarifying their implications – and, in particular, understanding which scientific or modeling choices they correspond to in practice – would contribute to a more unified picture of e-values for contingency tables and, more generally, for hypothesis testing with constrained models.

6.4 Refining higher-order validation under null models

The final part of the thesis shifted focus from inference between models to the use of maximum entropy models as null models for validation. Chapter 5 addressed the problem of identifying statistically validated three-wise interactions in brain data, starting from multivariate fMRI time series. By representing the data as a bipartite network between time points and brain regions and projecting it onto the layer of regions, the problem of validation was cast in terms of reproducing local constraints under suitable null models. While established approaches already exist for validating pairwise interactions using models such as the Bipartite Configuration Model, the chapter extended this framework to the validation of triplets by introducing a non-linear canonical null model (BiPLOM), solved within a mean-field approximation.

Applying this approach across multiple subjects made it possible to characterize validated triplets according to their internal pairwise structure and their spatial and functional organization. In particular, connected triplets were found to be more redundant and spatially compact, whereas unconnected triplets displayed near-zero redundancy and were more prevalent in transmodal areas. These latter structures would not be detected within a simplicial-complex framework, underscoring the importance of null-model-based validation compared to approaches that restrict attention to fully connected motifs.

From a methodological perspective, several possible improvements emerge naturally. The current analysis is restricted to binary interactions, whereas extending the null model to include signed and possibly weighted interactions would allow one to incorporate information about the direction and magnitude of co-fluctuations. This extension is not merely technical: preliminary investigations suggest that synergistic effects tend to emerge in frustrated interactions, for which knowledge of the signs of the underlying time-series fluctuations is essential. Capturing such effects would therefore require a signed and/or weighted formulation of the model.

Another methodological aspect concerns multiple-testing correction. In the present analysis, statistical validation relies on the Benjamini–Hochberg correction, which is standard in the statistical validation literature and provides a practical and well-established baseline. In settings such as the one considered here, where tested structures may share nodes or other components, alternative correction strategies that explicitly account for dependence could be explored as a methodological refinement.

In this broader context, e-values offer a potentially attractive perspective. Their use would make it possible to easily combine evidence across subjects and to apply multiple-testing corrections that remain valid under general dependence structures. At the same time, defining suitable e-variables for higher-order validation problems remains a nontrivial task. As an initial benchmark, one could consider e-values obtained by calibrating classical p-values, which, while not optimal, may serve as a useful reference against which more tailored constructions can be assessed.

Bibliography

[1] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, 1957.

[2] J. Gibbs. *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics.* Cambridge Library Collection - Mathematics. Cambridge University Press, 2010.

[3] A. Somazzi and D. Garlaschelli. Learn your entropy from informative data: an axiom ensuring the consistent identification of generalized entropies. *arXiv preprint arXiv:2301.05660*, 2023.

[4] P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

[5] D. R. Hunter and M. S. Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583, 2006.

[6] G. Cimini, T. Squartini, F. Saracco, D. Garlaschelli, A. Gabrielli, and G. Caldarelli. The statistical physics of real-world networks. *Nature Reviews Physics*, 1, 2019.

[7] H. Touchette. Equivalence and nonequivalence of ensembles: Thermodynamic, macrostate, and measure levels. *Journal of Statistical Physics*, 159(5):987–1016, 2015.

[8] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley–Interscience, 2006.

[9] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A practical information-theoretic approach.* Springer, 2010.

[10] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike.* Springer New York, 1973.

[11] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[12] T. Squartini, J. de Mol, F. den Hollander, and D. Garlaschelli. Breaking of ensemble equivalence in networks. *Phys. Rev. Lett.*, 115:268701, Dec 2015.

[13] T. Squartini and D. Garlaschelli. Reconnecting statistical physics and combinatorics beyond ensemble equivalence. *arXiv preprint arXiv:1710.11422*, 2017.

[14] Q. Zhang and D. Garlaschelli. Strong ensemble nonequivalence in systems with local constraints. *New Journal of Physics*, 24(4):043011, 2022.

[15] D. Garlaschelli, F. den Hollander, and A. Roccaverde. Covariance structure behind breaking of ensemble equivalence in random graphs. *Journal of Statistical Physics*, 173(3–4):644–662, 2018.

[16] P. Grünwald. *The minimum description length principle*. MIT press, 2007.

[17] P. Grünwald and T. Roos. Minimum description length revisited. *International journal of mathematics for industry*, 11(01):1930001, 2019.

[18] K. Yamanishi. *Learning with the Minimum Description Length Principle*. Springer Nature Singapore, 2023.

[19] J. P. A. Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[20] D. J. Benjamin et al. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10, 2017.

[21] B. B. McShane, D. Gal, A. Gelman, C. Robert, and J. L. Tackett. Abandon statistical significance. *The American Statistician*, 73(sup1):235–245, 2019.

[22] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statist. Sci.*, 38(4):576–601, 2023.

[23] Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 1, 2025. Inaugural Issue.

[24] P. Grünwald, R. de Heide, and W. Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024.

[25] L. Wasserman, A. Ramdas, and S. Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

[26] V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3), 2021.

[27] G. Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(2):407–431, 2021.

[28] Peter Grünwald, Tyron Lardy, Yunda Hao, Shaul K. Bar Lev, and Martijn de Jong. Optimal e-values for exponential families: the simple case. *arXiv: 2404.19465*, 2024.

[29] Yunda Hao and Peter Grünwald. E-values for exponential families: the general case. *arXiv: 2409.11134*, 2024.

[30] Rosanne Turner, Alexander Ly, and Peter Grünwald. Generic e-variables for exact sequential k-sample tests that allow for optional stopping. *Statistical Planning and Inference*, 230:106116, 2024.

[31] T. Squartini and D. Garlaschelli. *Maximum-Entropy Networks: Pattern Detection, Network Reconstruction and Graph Combinatorics*. Springer Cham, 2017.

[32] A. I. Luppi, P. A. M. Mediano, F. E. Rosas, and et al. A synergistic core for human brain evolution and cognition. *Nat Neurosci*, 25:771–782, 2022.

[33] Andrea Santoro, Federico Battiston, Giovanni Petri, and Enrico Amico. Higher-order organization of multivariate time series. *Nature Physics*, 19(2):221–229, 2023.

[34] Andrea I Luppi, Pedro AM Mediano, Fernando E Rosas, Judith Allanson, John Pickard, Robin L Carhart-Harris, Guy B Williams, Michael M Craig, Paola Finoia, Adrian M Owen, et al. A synergistic workspace for human consciousness revealed by integrated information decomposition. *Elife*, 12:RP88173, 2024.

[35] Andrea Santoro, Federico Battiston, Maxime Lucas, Giovanni Petri, and Enrico Amico. Higher-order connectomics of human brain function reveals local topological signatures of task decoding, individual identification, and behavior. *Nature Communications*, 15(1):10244, 2024.

[36] T. Squartini, G. Caldarelli, G. Cimini, A. Gabrielli, and D. Garlaschelli. Reconstruction methods for networks: The case of economic and financial systems. *Physics Reports*, 757, 2018.

[37] R. Marcaccioli and G. Livan. Maximum entropy approach to multivariate time series randomization. *Scientific Reports*, 10(1):10656, 2020.

[38] R. Marcaccioli and G. Livan. Correspondence between temporal correlations in time series, inverse problems, and the spherical model. *Physical Review E*, 102(1):012112, 2020.

[39] M. Blume, V. J. Emery, and R. B. Griffiths. Ising model for the λ transition and phase separation in he 3-he 4 mixtures. *Physical review A*, 4(3):1071, 1971.

[40] D. Lynden-Bell. Negative specific heat in astronomy, physics and chemistry. *Physica A: Statistical Mechanics and its Applications*, 263(1-4):293–304, 1999.

[41] R. S. Ellis, K. Haven, and B. Turkington. Large deviation principles and complete equivalence and nonequivalence results for pure and mixed ensembles. *Journal of Statistical Physics*, 101(5-6):999 – 1064, 2000.

[42] M. D'Agostino, F. Gulminelli, Ph. Chomaz, M. Bruno, F. Cannata, R. Bougault, F. Gramegna, I. Iori, N. Le Neindre, G.V. Margagliotti, et al. Negative heat capacity in the critical region of nuclear fragmentation: an experimental evidence of the liquid-gas phase transition. *Physics Letters B*, 473(3-4):219–225, 2000.

[43] J. Barré, D. Mukamel, and S. Ruffo. Inequivalence of ensembles in a system with long-range interactions. *Phys. Rev. Lett.*, 87:030601, Jun 2001.

[44] P. H. Chavanis. Gravitational instability of isothermal and polytropic spheres. *Astronomy & Astrophysics*, 401(1):15–42, 2003.

[45] R. S. Ellis, H. Touchette, and B. Turkington. Thermodynamic versus statistical nonequivalence of ensembles for the mean-field blume–emery–griffiths model. *Physica A: Statistical Mechanics and its Applications*, 335(3-4):518–538, 2004.

[46] F. Bouchet and J. Barré. Classification of phase transitions and ensemble inequivalence in systems with long range interactions. *Journal of Statistical Physics*, 118(5–6):1073–1105, 2005.

[47] J. Barré and B. Gonçalves. Ensemble inequivalence in random graphs. *Physica A: Statistical Mechanics and its Applications*, 386(1):212–218, 2007.

[48] C. Radin and L. Sadun. Phase transitions in a complex network. *Journal of Physics A: Mathematical and Theoretical*, 46(30):305002, 2013.

[49] A. Campa, T. Dauxois, D. Fanelli, and S. Ruffo. Equilibrium statistical mechanics of long-range interactions. In *Physics of Long-Range Interacting Systems*. Oxford University Press, 08 2014.

[50] M. Bruno, F. Saracco, D. Garlaschelli, C. J. Tessone, and G Caldarelli. The ambiguity of nestedness under soft and hard constraints. *Scientific Reports*, 10(1), 2020.

[51] T. Squartini, R. Mastrandrea, and D. Garlaschelli. Unbiased sampling of network ensembles. *New Journal of Physics*, 17(2):023052, 2015.

[52] D. Garlaschelli, F. den Hollander, and A. Roccaverde. Ensemble nonequivalence in random graphs with modular structure. *Journal of Physics A: Mathematical and Theoretical*, 50(1):015001, 2016.

[53] J.J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, 1996.

[54] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York, 3rd edition, 2008.

[55] Saracco. F., M. J. Straka, R. Di Clemente, A. Gabrielli, G. Caldarelli, and T. Squartini. Inferring monopartite projections of bipartite networks: an entropy-based approach. *New Journal of Physics*, 19(5):053022, 2017.

[56] P. P. A. Staniczenko, M.J. Smith, and S. Allesina. Selecting food web models using normalized maximum likelihood. *Methods in Ecology and Evolution*, 5(6):551–562, 2014.

[57] T. P. Peixoto. Nonparametric bayesian inference of the microcanonical stochastic block model. *Phys. Rev. E*, 95:012317, 2017.

[58] T. Vallès-Català, T. P. Peixoto, M. Sales-Pardo, and R. Guimerà. Consistencies and inconsistencies between model selection and link prediction in networks. *Phys. Rev. E*, 97:062316, 2018.

[59] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[60] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[61] J. Rissanen. Strong optimality of the normalized ml models as universal codes and information in data. *IEEE Transactions on Information Theory*, 47(5):1712–1717, 2001.

[62] A. I. Khinchin. *Mathematical Foundations of Information Theory*. Dover Publications, New York, 1957.

[63] V. De Angelis. Stirling's series revisited. *The American Mathematical Monthly*, 116(9):839–843, 2009.

[64] Tyron Lardy, Peter Grünwald, and Peter Harremoës. Reverse information projections and optimal e-statistics. *IEEE Transactions on Information Theory*, 70(11):7616–7631, 2024.

[65] Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire e-variable and reverse information projection. *Annals of Statistics*, 2025.

[66] L.D. Brown. *Fundamentals of Statistical Exponential Families, with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, Hayward, CA, 1986.

[67] Rosanne Turner and Peter Grünwald. Anytime-valid confidence intervals for contingency tables and beyond. *Statistics and Probability Letters*, 2023.

[68] Peter D. Grünwald. Beyond neyman–pearson: E-values enable hypothesis testing with a data-driven alpha. *Proceedings of the National Academy of Sciences*, 121(39):e2302098121, 2024.

[69] Alexander Ly, Udo Boehm, Peter Grünwald, Aaditya Ramdas, and Don van Ravenzwaaij. A tutorial on safe anytime-valid inference: Practical maximally flexible sampling designs for experiments based on e-values, 2025. PsyArXiv Preprint.

[70] Richard A Klein, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams Jr, Sinan Alper, Mark Aveyard, Jordan R Axt, Mayowa T Babalola, Štěpán Bahník, et al. Many labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018.

[71] Z. Zhang, A. Ramdas, and R. Wang. On the existence of powerful p-values and e-values for composite hypotheses. *arXiv: 2305.16539*, 2024.

[72] Q. Wang, R. Wang, and J. Ziegel. E-backtesting. *arXiv: 2209.00991*, 2024.

[73] V. Vovk and R. Wang. Efficiency of nonparametric e-tests. *arXiv: 2208.08925*, 2024.

[74] Richard D Morey, Jan-Willem Romeijn, and Jeffrey N Rouder. The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72(6-18):36, 2016.

[75] Kyoungseok Jang, Kwang-Sung Jun, Ilja Kuzborskij, and Francesco Orabona. Tighter pac-bayes bounds through coin-betting. In *Proceedings COLT 2023*, 2023.

[76] O.E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, Chichester, UK, 1978.

[77] Hélder Alves, Paula Brito, and Pedro Campos. Community detection in interval-weighted networks. *Data Mining and Knowledge Discovery*, 38(2):653–698, 2024.

[78] Max Jerdee, Alec Kirkley, and Mark E. J. Newman. Mutual information and the encoding of contingency tables. *Physical review. E*, 110 6-1:064306, 2024.

[79] B.S. Clarke and A.R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41:37–60, 1994.

[80] M. Earnest (user). Extended stars-and-bars problem(where the upper limit of the variable is bounded). Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/3182858 (version: 2019-04-14).

[81] A Golan, G Judge, and D Miller. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. John Wiley), Chichester, UK, 1996.

[82] Tiago P. Peixoto. Network reconstruction via the minimum description length principle. *Phys. Rev. X*, 15:011065, Mar 2025.

[83] I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *The Annals of Probability*, pages 768–793, 1984.

[84] Peter Grünwald, Yunda Hao, and Akshay Balsubramani. Growth-optimal e-variables and an extension to the multivariate Csiszár-Sanov-Chernoff theorem. *arXiv: 2412.17554*, 2024.

[85] Rosanne J. Turner and Peter D. Grünwald. Exact anytime-valid confidence intervals for contingency tables and beyond. *Statistics & Probability Letters*, 198:109835, 2023.

[86] Thorsten Dickhaus, Karsten Straßburger, Daniel Schunk, Carlos Morcillo-Suarez, Thomas Illig, and Alejandro Navarro. How to analyze many contingency tables simultaneously in genetic association studies. *Statistical Applications in Genetics and Molecular Biology*, 11(4), 2012.

[87] Abraham Wald. *Sequential Analysis*. Wiley, New York, 1947.

[88] María Eglée Pérez and Luis Raúl Pericchi. A case study on the Bayesian analysis of 2×2 tables with all margins fixed. *Brazilian Journal of Probability and Statistics*, 8(1):27–37, 1994.

[89] Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281, March 1979.

[90] Valen E. Johnson. Uniformly most powerful Bayesian tests. *The Annals of Statistics*, 41(4):1716–1741, August 2013.

[91] Peter Grünwald, Rianne de Heide, and Wouter Koolen. Authors' reply to the discussion of 'safe testing'. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1163–1171, November 2024.

[92] J. Ville. *Étude Critique de la Notion de Collectif*. PhD thesis, L'Université de Paris, 1939.

[93] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *Ann. Stat.*, 49(2), 2021.

[94] W. M. Koolen and P. Grünwald. Log-optimal anytime-valid e-values. *International Journal of Approximate Reasoning*, 141:69–82, 2022. Festschrift for G. Shafer's 75th Birthday.

[95] J. Ter Schure, M.F. Perez-Ortiz, A. Ly, and P. Grünwald. The anytime-valid logrank test: Error control under continuous monitoring with unlimited horizon. *New England Journal of Statistics in Data Science*, 2(2):190–214, 2024.

[96] Suzana Herculano-Houzel. The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in human neuroscience*, 3:857, 2009.

[97] Bratislav Mišić and Olaf Sporns. From regions to connections and networks: new bridges between brain and behavior. *Current opinion in neurobiology*, 40:1–7, 2016.

[98] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186–198, 2009.

[99] Karl J. Friston. Functional and effective connectivity: A review. *Brain Connectivity*, 1(1):13–36, 2011. PMID: 22432952.

[100] Danielle S Bassett and Olaf Sporns. Network neuroscience. *Nature neuroscience*, 20(3):353–364, 2017.

[101] O. Sporns. *Networks of the Brain*. MIT Press, 2016.

[102] Abel Jansma, Yuelin Yao, Jareth Wolfe, Luigi Del Debbio, Sjoerd V Beentjes, Chris P Ponting, and Ava Khamseh. High order expression dependencies finely resolve cryptic states and subtypes in single cell data. *Molecular Systems Biology*, 21(2):173–207, 2025.

[103] Jacopo Grilli, György Barabás, Matthew J Michalska-Smith, and Stefano Allesina. Higher-order interactions stabilize dynamics in competitive network models. *Nature*, 548(7666):210–213, 2017.

[104] Unai Alvarez-Rodriguez, Federico Battiston, Guilherme Ferraz de Arruda, Yamir Moreno, Matjaž Perc, and Vito Latora. Evolutionary dynamics of higher-order interactions in social networks. *Nature Human Behaviour*, 5(5):586–595, 2021.

[105] Clélia de Mulatier, Paolo P Mazza, and Matteo Marsili. Statistical inference of minimally complex models. *arXiv e-prints*, pages arXiv–2008, 2020.

[106] Fernando E. Rosas, Pedro A. M. Mediano, Michael Gastpar, and Henrik J. Jensen. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Phys. Rev. E*, 100:032305, Sep 2019.

[107] Andrea I Luppi, Fernando E Rosas, Pedro AM Mediano, David K Menon, and Emmanuel A Stamatakis. Information decomposition and the informational architecture of the brain. *Trends in Cognitive Sciences*, 28(4):352–368, 2024.

[108] T. F. Varley, M. Pope, M. G. Puxeddu, J. Faskowitz, and O. Sporns. Partial entropy decomposition reveals higher-order structures in human brain activity, 2023. arXiv preprint.

[109] Rikkert Hindriks, Tommy AA Broeders, Menno M Schoonheim, Linda Douw, Fernando Santos, Wessel van Wieringen, and Prejaas KB Tewarie. Higher-order functional connectivity analysis of resting-state functional magnetic resonance imaging data using multivariate cumulants. *Human brain mapping*, 45(5):e26663, 2024.

[110] Marta Niedostatek, Anthony Baptista, Jun Yamamoto, Jürgen Kurths, Ruben Sanchez Garcia, Ben D MacArthur, and Ginestra Bianconi. Mining higher-order triadic interactions. *Nature Communications*, 2025.

[111] R. G. James, C. J. Ellison, and J. P. Crutchfield. Anatomy of a bit: Information in a time series observation. *Chaos*, 21:037109, 2011.

[112] Fernando E Rosas, Aaron Gutknecht, Pedro A M Mediano, and Michael Gastpar. Characterising high-order interdependence via entropic conjugation. *arXiv preprint arXiv:2410.10485*, 2024.

[113] Giovanni Petri, Paul Expert, Federico Turkheimer, Robin Carhart-Harris, David Nutt, Peter J Hellyer, and Francesco Vaccarino. Homological scaffolds of brain functional networks. *Journal of The Royal Society Interface*, 11(101):20140873, 2014.

[114] Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, and Giovanni Petri. Networks beyond pairwise interactions: Structure and dynamics. *Physics reports*, 874:1–92, 2020.

[115] Jacob Billings, Manish Saggar, Jaroslav Hlinka, Shella Keilholz, and Giovanni Petri. Simplicial and topological descriptions of human brain dynamics. *Network Neuroscience*, 5(2):549–568, 2021.

[116] Allen Hatcher. *Algebraic topology*. Cambridge university press, New York, 2001.

[117] F. Saracco, R. Di Clemente, A. Gabrielli, and T. Squartini. Randomizing bipartite networks: the case of the world trade web. *Scientific Reports*, 5, 2015.

[118] Lukas Roell, Stephan Wunderlich, David Roell, Florian Raabe, Elias Wagner, Zhuanghua Shi, Andrea Schmitt, Peter Falkai, Sophia Stoecklein, and Daniel Keeser. How to measure functional connectivity using resting-state fmri? a comprehensive empirical exploration of different connectivity metrics. *NeuroImage*, 312:121195, 2025.

[119] Enzo Tagliazucchi, Pablo Balenzuela, Daniel Fraiman, and Dante R Chialvo. Criticality in large-scale brain fmri dynamics unveiled by a novel point process analysis. *Frontiers in Physiology*, 3:15, 2012.

[120] Xiao Liu and Jeff H Duyn. Time-varying functional network information extracted from brief instances of spontaneous brain activity. *Proceedings of the National Academy of Sciences*, 110(11):4392–4397, 2013.

[121] Xiahai Zhuang, Zhiguo Yang, and Dietmar Cordes. Incorporating spatial constraint in co-activation pattern analysis to explore the dynamics of resting-state networks: An application to parkinson's disease. *NeuroImage*, 172:64–84, 2018.

[122] Ivan Cifre, Mohammad Zarepour, Silvina G Horovitz, Sergio A Cannas, and Dante R Chialvo. Further results on why a point process is effective for estimating correlation between brain regions. *Papers in Physics*, 12:051, 2020.

[123] Oren Shriki, Jeff Alstott, Frederick Carver, Tom Holroyd, Richard NA Henson, Marie L Smith, Richard Coppola, Edward Bullmore, and Dietmar Plenz. Neuronal avalanches in the resting meg of the human brain. *Journal of Neuroscience*, 33(16):7079–7090, 2013.

[124] Fabrizio Lombardi, Oren Shriki, Hans J Herrmann, and Lucilla de Arcangelis. Long-range temporal correlations in the broadband resting state activity of the human brain revealed by neuronal avalanches. *Neurocomputing*, 461:657–666, 2021.

[125] KC K Nkurumeh, Han Yuan, and Lei Ding. Recurring transient brain-wide co-activation patterns from eeg spatially resembling time-averaged resting-state networks. *bioRxiv*, pages 2025–06, 2025.

[126] Fabrizio Lombardi, Hans J Herrmann, Liborio Parrino, Dietmar Plenz, Silvia Scarpetta, Anna Elisabetta Vaudano, Lucilla De Arcangelis, and Oren Shriki. Beyond pulsed inhibition: Alpha oscillations modulate attenuation and amplification of neural activity in the awake resting state. *Cell reports*, 42(10), 2023.

[127] Anna Gallo, Fabio Saracco, and Tiziano Squartini. Statistically validated projection of bipartite signed networks. *npj Complexity*, 2(1):22, Jul 2025.

[128] Marzio Di Vece, Emanuele Agrimi, Samuele Tatullo, Tommaso Gili, Miguel Ibáñez-Berganza, and Tiziano Squartini. Assessing (im)balance in signed brain networks. *arXiv: 2508.00542*, 2025.

[129] Michele Tumminello, Salvatore Miccichè, Fabrizio Lillo, Jyrki Piilo, and Rosario N. Mantegna. Statistically validated networks in bipartite complex systems. *PLoS ONE*, 6(3):e17994, March 2011.

[130] Giulio Cimini, Alessandro Carra, Luca Didomenicantonio, and Andrea Zaccaria. Meta-validation of bipartite network projections. *Communications Physics*, 5(1):76, 2022.

[131] Zachary P Neal, Annabell Cadieux, Diego Garlaschelli, Nicholas J Gotelli, Fabio Saracco, Tiziano Squartini, Shade T Shutters, Werner Ulrich, Guanyang Wang, and Giovanni Strona. Pattern detection in bipartite networks: A review of terminology, applications, and methods. *PLOS Complex Systems*, 1(2):e0000010, 2024.

[132] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 1995.

[133] Federico Musciotto, Federico Battiston, and Rosario N. Mantegna. Detecting informative higher-order interactions in statistically validated hypergraphs, 2021.

[134] Matteo Neri, Dishie Vinchhi, Christian Ferreyra, Thomas Robiglio, Onur Ates, Marlis Ontivero-Ortega, Andrea Brovelli, Daniele Marinazzo, and Etienne Combrisson. HOI: A Python toolbox for high-performance estimation of Higher-Order Interactions from multivariate data. *Journal of Open Source Software*, 9(103):7360, November 2024.

[135] B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L. Roffman, Jordan W. Smoller, Lilla Zöllei, Jonathan R. Polimeni, Bruce Fischl, Hesheng Liu, and Randy L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, 2011.

[136] David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E. Behrens, Essa Yacoub, and Kamil Ugurbil. The wu-minn human connectome project: an overview. *NeuroImage*, 80:62–79, 2013.

[137] Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013.

[138] Ludovica Griffanti, Gholamreza Salimi-Khorshidi, Christian Beckmann, Edward Auerbach, Gwenaëlle Douaud, Claire Sexton, Enikő Zsoldos, Klaus Ebmeier, Nicola Filippini, Clare Mackay, Steen Moeller, Junqian Xu, Essa Yacoub, Giuseppe Baselli, Kamil Ugurbil, Karla Miller, and Stephen Smith. Ica-based artefact removal and accelerated fmri acquisition for improved resting state network imaging. *NeuroImage*, 95, 03 2014.

[139] Theodore D Satterthwaite, Mark A Elliott, Raphael T Gerraty, Kosha Ruparel, James Loughead, Monica E Calkins, Simon B Eickhoff, Hakon Hakonarson, Ruben C Gur, Raquel E Gur, et al. An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *Neuroimage*, 64:240–256, 2013.

[140] Alexander Schaefer, Ru Kong, Evan M. Gordon, Timothy O. Laumann, Xi-Nian Zuo, Alexander J. Holmes, Simon B. Eickhoff, and B. T. Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral Cortex*, 28(9):3095–3114, 2018.

[141] Y. Tian, D. S. Margulies, M. Breakspear, and A. Zalesky. Topographic organization of the human subcortex unveiled with functional connectivity gradients. *Nature neuroscience*, 23(11):1421–1432, 2020.

[142] Kai Hwang, Maxwell A Bertolero, William B Liu, and Mark D'Esposito. The human thalamus is an integrative hub for functional brain networks. *Journal of Neuroscience*, 37(23):5594–5607, 2017.

[143] Takuma Kitanishi, Ryoko Umaba, and Kenji Mizuseki. Robust information routing by dorsal subiculum neurons. *Science advances*, 7(11):eabf1913, 2021.

[144] Peter T Bell and James M Shine. Subcortical contributions to large-scale network communication. *Neuroscience & Biobehavioral Reviews*, 71:313–322, 2016.

[145] Vinod Menon. Large-scale brain networks and psychopathology: a unifying triple network model. *Trends in Cognitive Sciences*, 15(10):483–506, 2011.

[146] Vinod Menon. Salience network. *Brain Mapping*, pages 597–611, 2015.

[147] Weidong Cai, Srikanth Ryali, Ramkrishna Pasumarthy, Viswanath Talasila, and Vinod Menon. Dynamic causal brain circuits during working memory and their functional controllability. *Nature communications*, 12(1):3314, 2021.

[148] Juyong Park and Mark E J Newman. Statistical mechanics of networks. *Phys. Rev. E*, 70(6):66117, dec 2004.

[149] Diego Garlaschelli and Maria I. Loffredo. Maximum likelihood: Extracting unbiased information from complex networks. *Physical Review E*, 78, 2008.

[150] Nicolò Vallarano, Matteo Bruno, Emiliano Marchese, Giuseppe Trapani, Fabio Saracco, Giulio Cimini, Mario Zanon, and Tiziano Squartini. Fast and scalable likelihood maximization for exponential random graph models with local constraints. *Scientific Reports*, 11:15227, 2021.

[151] Juyong Park and M E J Newman. Solution of the 2-star model of a network. *Physical Review E*, 70(6):066146, 2004.

[152] Mattia Marzi, Francesca Giuffrida, Diego Garlaschelli, and Tiziano Squartini. Reproducing the first and second moments of empirical degree distributions. *Phys. Rev. Res.*, 8(1):013047, 2026.

[153] Sara Larivière, Casey Paquola, Bo-yong Park, Jessica Royer, Yezhou Wang, Oualid Benkarim, Reinder Vos de Wael, Sofie L Valk, Sophia I Thomopoulos, Matthias Kirschner, et al. The enigma toolbox: multiscale neural contextualization of multisite neuroimaging datasets. *Nature methods*, 18(7):698–700, 2021.

List of publications

This dissertation is based on the following papers:

- Chapter 2 is based on
  F. Giuffrida, T. Squartini, P. Grünwald, and D. Garlaschelli. Description length of canonical and microcanonical Models. *Phys. Rev. Research 7, 043057* (2025).
- Chapter 3 is based on
  F.Giuffrida, D. Garlaschelli and P. Grünwald. Testing maximum entropy models with e-values. *Phys. Rev. E 113, 054306* (2026).
- Chapter 4 is based on
  Y. Wang, S. Arnold, F. Giuffrida, P. Grünwald, and T. Dickhaus. E-values for Contingency Tables. In preparation.
- Chapter 5 is based on
  F. Giuffrida, E. Agrimi, M. Marzi, A. Gallo, T. Squartini and T. Gili. Signatures of higher-order coordination in the human brain. *In preparation.*

The following paper was completed during the PhD time but is not included in this dissertation:

- M. Marzi, F. Giuffrida, D. Garlaschelli and T. Squartini. Reproducing the first and second moment of empirical degree distributions. *Phys. Rev. Research 8, 013141* (2026).

Samenvatting

Maximum-entropiemodellen zijn kansmodellen die een voorgeschreven verzameling empirische nevenvoorwaarden reproduceren, maar voor het overige maximaal willekeurig blijven. Gebaseerd op het Maximum Entropy-principe, in 1957 geïntroduceerd door Jaynes, generaliseren deze modellen de ensembles van de evenwichtsstatistische fysica naar willekeurige systemen die door nevenvoorwaarden worden gedefinieerd, dat wil zeggen elke verzameling observabelen waarvan de empirische waarden relevant worden geacht voor het probleem in kwestie. In die zin codeert een maximum-entropiemodel precies wat over een systeem bekend is en laat het al het overige aan het toeval over. Maximum-entropiemodellen vormen de rode draad van dit proefschrift, dat hun rol onderzoekt in drie fundamentele inferentietaken: modelselectie (Hoofdstuk 2), het toetsen van hypothesen (Hoofdstukken 3 en 4) en statistische validatie (Hoofdstuk 5).

Een centraal en terugkerend thema in dit hele werk is het onderscheid tussen twee formuleringen van maximum-entropiemodellen. In de microcanonieke formulering worden de nevenvoorwaarden exact opgelegd: alleen configuraties die precies aan de nevenvoorwaarden voldoen krijgen een positieve kans, en al deze configuraties worden als even waarschijnlijk beschouwd. In de canonieke formulering worden de nevenvoorwaarden slechts in verwachtingswaarde opgelegd: de kansverdeling neemt een exponentiële vorm aan, en de grootheden waarop de nevenvoorwaarden betrekking hebben, mogen fluctueren rond hun voorgeschreven gemiddelde waarden. In de evenwichtsstatistische fysica komen deze twee formuleringen overeen met verschillende experimentele omstandigheden en is bekend dat ze in typische regimes asymptotisch equivalent zijn, een fenomeen dat bekendstaat als ensemble-equivalentie. Deze equivalentie kan echter wegvallen wanneer het aantal nevenvoorwaarden meegroeit met de omvang van het systeem, een regime dat op natuurlijke wijze optreedt in complexe systemen zoals netwerken en tijdreeksen, waar lokale nevenvoorwaarden afzonderlijk aan elke systeemcomponent worden opgelegd. Om deze reden wordt het onderscheid tussen deze twee formuleringen in dit hele proefschrift als een centraal ordenend principe gehandhaafd.

Dit proefschrift is opgebouwd rond vier bijdragen, uitgewerkt in de Hoofdstukken 2 tot en met 5. Hoofdstuk 2 behandelt het probleem van modelselectie tussen de canonieke en de microcanonieke formulering van maximum-entropiemodellen die dezelfde verzameling nevenvoorwaarden delen. Deze vergelijking wordt geformaliseerd binnen het Minimum Description Length (MDL)-principe, met Normalized Maximum Likelihood (NML) als universal coding distribution. Het MDL-principe vat statistische inferentie op als een compressieprobleem, waarbij het beste model de meest beknopte beschrijving van de waargenomen data oplevert, in een afweging tussen goodness-

of-fit en modelcomplexiteit. De vergelijking tussen de NML-description lengths van canonieke en microcanonieke modellen brengt een systematische trade-off aan het licht: microcanonieke modellen behalen altijd een hogere likelihood, maar zijn ook systematisch complexer, zodat de voorkeursformulering afhangt van de waargenomen data. Bovendien wordt aangetoond dat de verschillen in description length begrensd blijven wanneer ensemble-equivalentie geldt, maar meegroeien met de systeemomvang wanneer dat niet het geval is, wat een direct en kwantitatief verband legt tussen ensemble-non-equivalentie en statistische modelvergelijking. Hoofdstuk 2 verheldert ook de relatie tussen op NML gebaseerde en Bayesiaanse benaderingen van modelselectie. Vanuit een Bayesiaans perspectief kan het onderscheid tussen de canonieke en de microcanonieke formulering in principe worden overbrugd door een geschikte keuze van prior. Deze observatie lost het onderliggende probleem van non-equivalentie echter niet op; ze verschuift het probleem veeleer naar de keuze van de prior zelf. Opmerkelijk genoeg wordt de keuze van de prior, die onder ensemble-equivalentie grotendeels onbelangrijk is, cruciaal in niet-equivalente regimes, waar verschillende priors tot sterk uiteenlopende description lengths kunnen leiden.

Hoofdstuk 3 ontwikkelt growth-rate optimal (GRO) e-values voor hypothesetoetsen tussen canonieke of microcanonieke maximum-entropiemodellen die door verschillende verzamelingen nevenvoorwaarden worden gedefinieerd. E-values zijn een recent voorgesteld alternatief voor p-values die optional continuation van dataverzameling en een principieel opeenstapelen van bewijs over experimenten heen ondersteunen, zonder de type-I-fout te vergroten. Anders dan p-values fungeren e-values als directe maten van bewijs en kunnen ze over onafhankelijke waarnemingen heen worden vermenigvuldigd, wat ze bijzonder geschikt maakt voor sequentiële settings en settings met meerdere studies. In het microcanonieke geval wordt een exacte GRO e-variable in gesloten vorm afgeleid. Voor canonieke modellen, waar exacte oplossingen doorgaans niet beschikbaar zijn, wordt een microcanonieke benadering geïntroduceerd, waarvan wordt aangetoond dat ze zowel theoretisch als via numerieke experimenten goed presteert. Deze resultaten worden toegepast op de context van $2 \times k$-contingentietabellen — een fundamenteel en alomtegenwoordig instrument dat op natuurlijke wijze opduikt in een breed scala aan associatieproblemen, van de genetica tot de sociale wetenschappen, en dat een natuurlijke representatie biedt voor diverse modellen van complexe systemen. Een conceptueel belangrijk resultaat verbindt de constructie van optimale e-variables met MDL universal coding: universele verdelingen, die al centraal stonden in de analyse van de description length in Hoofdstuk 2, induceren op natuurlijke wijze e-values met kleine regret, waarmee een principiële verbinding wordt gelegd tussen modelselectie op basis van MDL en hypothesetoetsing op basis van e-values.

Hoofdstuk 4 richt zich op e-values voor contingentietabellen en onderzoekt welke het meest geschikt zijn voor verschillende inferentietaken. Verschillende e-value-constructies worden vergeleken voor het toetsen van associatie in $2 \times k$-tabellen, waaronder nieuw geïntroduceerde conditional e-values, geïnspireerd op de microcanonieke e-values van Hoofdstuk 3. De analyse beschouwt twee inferentieregimes: een single-shot-regime, waarin de volledige dataset in één keer wordt waargenomen, en een sequentieel regime, waarin data in opeenvolgende batches binnenkomen en het bewijs in de loop van de tijd wordt bijgewerkt. De vergelijking laat zien dat uniformly most

powerful conditional e-values het best presteren wanneer de data in een klein aantal grote batches binnenkomen, terwijl sequentially designed e-variables de voorkeur verdienen in het asymptotische regime van vele kleine stappen. In plaats van één enkele aanpak te bepleiten, identificeert Hoofdstuk 4 de omstandigheden waaronder elke constructie het meest effectief is.

Hoofdstuk 5 verlegt de focus van modelvergelijking naar het gebruik van maximum-entropiemodellen als nulmodellen voor het valideren van structurele patronen. Op basis van resting-state-fMRI-tijdreeksen behandelt dit hoofdstuk het probleem van het identificeren van statistisch significante interacties tussen hersengebieden binnen een maximum-entropiekader. Na thresholding worden de data gerepresenteerd als een binair bipartiet netwerk dat hersengebieden en tijdstippen verbindt, waarbij een link aangeeft dat een gebied op een bepaald tijdstip actief is. Paarsgewijze en triadische interacties worden vervolgens gevalideerd door te toetsen of de waargenomen co-activaties de verwachtingen onder geschikte maximum-entropie-nulmodellen overtreffen. Paarsgewijze interacties worden beoordeeld met behulp van het Bipartite Configuration Model, terwijl triadische interacties worden getoetst aan een nieuw geïntroduceerd canoniek model, het Bipartite Partial Local Overlap Model, dat niet-lineaire nevenvoorwaarden oplegt en analytisch wordt opgelost onder een mean-field-benadering. De analyse identificeert twee kwalitatief verschillende regimes van triadische coördinatie: redundante, ruimtelijk compacte triaden met volledige paarsgewijze ondersteuning, geconcentreerd in unimodale corticale gebieden, en ruimtelijk verspreide triaden zonder paarsgewijze ondersteuning, die vaker voorkomen in transmodale gebieden.

In alle vier de bijdragen is het verbindende inzicht dat statistische conclusies alleen betekenisvol zijn ten opzichte van een goed gespecificeerd referentiemodel. Maximum-entropiemodellen bieden een principiële taal om deze keuzes expliciet te maken en om te begrijpen hoe ze de conclusies vormgeven die uit data worden getrokken.

Summary

Maximum entropy models are probability models that reproduce a prescribed set of empirical constraints while remaining otherwise maximally random. Based on the Maximum Entropy Principle introduced by Jaynes in 1957, these models generalize the ensembles of equilibrium statistical physics to arbitrary systems defined by constraints, i.e., any collection of observables whose empirical values are deemed relevant to the problem at hand. In this sense, a maximum entropy model encodes precisely what is known about a system and leaves everything else to chance. Maximum entropy models provide the unifying thread of this thesis, which investigates their role in three fundamental inferential tasks: model selection (Chapter 2), hypothesis testing (Chapters 3 and 4), and statistical validation (Chapter 5).

A central and recurring theme throughout this work is the distinction between two formulations of maximum entropy models. In the microcanonical formulation, constraints are imposed exactly: only configurations that satisfy the constraints precisely are assigned positive probability, and all such configurations are treated as equally likely. In the canonical formulation, constraints are enforced only in expectation: the probability distribution takes an exponential form, and the constrained quantities are free to fluctuate around their prescribed average values. In equilibrium statistical physics, these two formulations correspond to different experimental settings and are asymptotically equivalent in typical regimes, a phenomenon known as ensemble equivalence. However, equivalence can break down when the number of constraints grows with the size of the system, a regime that arises naturally in complex systems such as networks and time series, where local constraints are imposed separately on each system component. For this reason, the distinction between these two formulations is maintained throughout this thesis as a central organizing principle.

This thesis is built around four contributions, developed in Chapters 2 to 5. Chapter 2 addresses the problem of model selection between the canonical and microcanonical formulations of maximum entropy models that share the same set of constraints. This comparison is formalized within the Minimum Description Length (MDL) principle, using Normalized Maximum Likelihood (NML) as a universal coding distribution. The MDL principle frames statistical inference as a compression problem, where the best model is the one that yields the most concise description of the observed data, balancing goodness-of-fit with model complexity. The comparison between NML description lengths of canonical and microcanonical models reveals a systematic trade-off: microcanonical models always achieve higher likelihood, but are also systematically more complex, so the preferred formulation depends on the observed data. Moreover, it is shown that the description length differences remain bounded when ensemble equivalence holds, but grow with system size when it does not, providing a direct and

quantitative connection between ensemble non-equivalence and statistical model comparison. Chapter 2 also clarifies the relationship between NML-based and Bayesian approaches to model selection. From a Bayesian perspective, the distinction between canonical and microcanonical formulations can in principle be bridged through a suitable choice of prior. However, this observation does not resolve the underlying issue of non-equivalence; rather, it shifts the problem to the choice of prior itself. Remarkably, prior choice, largely inconsequential under ensemble equivalence, becomes crucial in non-equivalent regimes, where different priors can lead to very different description lengths.

Chapter 3 develops growth-rate optimal (GRO) e-values for hypothesis tests between canonical or microcanonical maximum entropy models defined by different sets of constraints. E-values are a recently proposed alternative to p-values that support optional continuation of data collection and principled accumulation of evidence across experiments, without inflating type-I error. Unlike p-values, e-values serve as direct measures of evidence and can be multiplied across independent observations, making them particularly well suited to sequential and multi-study settings. In the microcanonical case, an exact closed-form GRO e-variable is derived. For canonical models, where exact solutions are generally unavailable, a microcanonical approximation is introduced and shown to perform well both theoretically and through numerical experiments. These results are applied to the setting of $2 \times k$ contingency tables — a fundamental and ubiquitous tool arising naturally in a wide range of association problems, from genetics to the social sciences, and providing a natural representation for several models of complex systems. A conceptually important result connects the construction of optimal e-variables to MDL universal coding: universal distributions, already central to the description-length analysis of Chapter 2, naturally induce e-values with small regret, establishing a principled link between model selection based on MDL and hypothesis testing based on e-values.

Chapter 4 focuses on e-values for contingency tables, examining which are most appropriate for different inferential scenarios. Several e-value constructions are compared for testing association in $2 \times k$ tables, including newly introduced conditional e-values, inspired by the microcanonical e-values of Chapter 3. The analysis considers two inferential regimes: a single-shot setting, in which the entire dataset is observed at once, and a sequential setting, in which data arrive in successive batches and evidence is updated over time. The comparison shows that uniformly most powerful conditional e-values perform best when data arrive in a small number of large batches, whereas sequentially designed e-variables are preferable in the asymptotic regime of many small increments. Rather than advocating a single approach, Chapter 4 identifies the conditions under which each construction is most effective.

Chapter 5 shifts the focus from model comparison to the use of maximum entropy models as null models for validating structural patterns. Based on resting-state fMRI time series, it addresses the problem of identifying statistically significant interactions between brain regions within a maximum entropy framework. After thresholding, the data are represented as a binary bipartite network connecting brain regions and time points, where an edge denotes that a region is active at a given time point.

Pairwise and triadic interactions are then validated by testing whether the observed co-activations exceed those expected under suitable maximum entropy null models. Pairwise interactions are assessed using the Bipartite Configuration Model, while triadic interactions are tested against a newly introduced canonical model, the Bipartite Partial Local Overlap Model, which enforces non-linear constraints and is solved analytically under a mean-field approximation. The analysis identifies two qualitatively distinct regimes of triadic coordination: redundant, spatially compact triplets with complete pairwise support, concentrated in unimodal cortical areas, and spatially distributed triplets lacking pairwise support, which are more prevalent in transmodal regions.

Across all four contributions, the unifying insight is that statistical conclusions are meaningful only relative to a well-specified reference model. Maximum entropy models provide a principled language for making these choices explicit, and for understanding how they shape the conclusions drawn from data.

Acknowledgements

These acknowledgements mention very few names: the people who accompanied me through these years are simply too many. I hope that each of them will recognise themselves, at least in part, in the words that follow.

First of all, I wish to thank my supervisors, Diego Garlaschelli and Peter Grünwald, who both represent to me an example of curious, genuine science that holds its ground, and who have been a guide throughout these years, academically and personally. Diego encouraged me from the very beginning to look beyond and to gain experience elsewhere: without his invitation, I might never have left for Leiden, and this journey would have unfolded very differently. I am grateful for our many conversations, both scientific and personal, which carried me through some of the most difficult moments along the way. I thank Peter for his patience in building a shared vocabulary, for listening to me, and for guiding me through a field that was not my own. I am also grateful to the Machine Learning group at CWI for welcoming me on many occasions in the months following my return to Lucca.

I would like to thank my friends and colleagues in Leiden, for filling my time in the Netherlands with new food and new languages, new experiences and new ways of seeing things — and for still being there, ever since.

I thank the entire NETWORKS unit in Lucca, in its past and present composition: each of its members contributed to a different phase of my doctoral path. In particular, I thank Tommaso and Tiziano, who witnessed from the start my attempts to find my footing among the topics at the heart of this thesis and who contributed actively to its realisation.

More broadly, I am grateful to my friends and colleagues in Lucca, who have been a source of inspiration, warmth and familiarity from the first year of my PhD until today: from the microeconomics exercises of that first year to the final stretch, I am not sure I could have done this without you. A special mention goes to Alessio, with whom I shared nearly every phase of this long journey — taking apart his most outlandish ideas has been one of the most scientifically valuable exercises I have had the honour of performing.

I also wish to acknowledge the community of researchers and friends I have met over the years at conferences, from NetSci — including its Dutch edition — to CCS. These meetings offered this work, and me, essential moments of confirmation and growth, and introduced me to many curious and delightful people whom I hope to meet again along the strange paths that life takes.

To my friends Marco, Clara, Anna, Lorenzo and Lisa: thank you for managing to be there, even while scattered across Europe.

To my ever-growing family, including its youngest members: I still owe you all a proper explanation of what it is I actually do (and perhaps I always will). I am grateful for your presence and your support, which I feel strong and steady even when I forget to call (sorry, Mum).

To Agnese's family, thank you for all the birthdays celebrated "as a surprise".

Finally, to Agnese, who has been there in every moment, the happiest ones and especially the hardest. Together, there is nothing we can't face.

Curriculum Vitae

Francesca Giuffrida was born in Rome, Italy, on August 28, 1995. She attended the Liceo Scientifico Talete in Rome. In 2014, she enrolled in the Bachelor's degree programme in Physics at Sapienza University of Rome, where she graduated *cum laude* in 2017. She subsequently continued her studies at the same institution in Theoretical Physics, with a focus on Statistical Physics, obtaining her Master's degree *cum laude* in 2020. During the final months of her Master's, she became interested in Complex Systems and Network Science. Her Master's thesis, written under the supervision of Prof. Andrea Gabrielli and Prof. Giulio Cimini, focused on the introduction of link-to-link interactions in heterogeneous random graphs.

In November 2020, she began her PhD in Systems Science at the IMT School for Advanced Studies Lucca, within the ENBA track (Economics, Networks and Business Analytics), under the supervision of Prof. Diego Garlaschelli. During the first year of the PhD program, she completed coursework in econometrics, game theory, microeconomics, complex systems, and network science.

In November 2022, she became a PhD candidate at the LION Institute of Leiden University within a joint supervision agreement with IMT Lucca, under the supervision of Prof. Diego Garlaschelli and Prof. Peter Grünwald. She spent one year in Leiden and subsequently moved back to Lucca, while carrying out repeated research visits to Leiden University and to the Centrum Wiskunde & Informatica (CWI) in Amsterdam. In November 2024, she was awarded a one-year research fellowship at the University of Palermo, under the scientific supervision of Prof. Rosario N. Mantegna, which concluded in November 2025. She currently lives in Monza, Italy.