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

Giuffrida, F.

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.