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

Giuffrida, F.

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.