



Universiteit
Leiden
The Netherlands

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

Giuffrida, F.

Citation

Giuffrida, F. (2026, September 10). *Learning under constraints: inference, selection and validation with maximum entropy models*. Retrieved from <https://hdl.handle.net/1887/4309386>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4309386>

Note: To cite this publication please use the final published version (if applicable).

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 canoneke en microcanoneke modellen brengt een systematische trade-off aan het licht: microcanoneke 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 canoneke en de microcanoneke 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 canoneke of microcanoneke 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 microcanoneke geval wordt een exacte GRO e-variable in gesloten vorm afgeleid. Voor canoneke modellen, waar exacte oplossingen doorgaans niet beschikbaar zijn, wordt een microcanoneke 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 microcanoneke 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.