Implicit Inference

Justin Alsing (Oskar Klein Center), Benjamin D. Wandelt

Niall Jeffrey (UCL), Matt Ho (IAP), Xiaosheng Zhang (Tsinghua), Gabriel Jung (IAS, Orsay), Lucas Makinen (Imperial), Will Coulton (CCA), Stephen Feeney, ...
Cosmological data covers a hierarchy of scales on the past light cone. Smaller scales → increasing complexity
How to deal with complexity on small scales?
How to deal with complexity on small scales?

Smooth away small scales?
How to deal with complexity on small scales?

Use a computational model?
Approach 1: “Explicit” inference

1. Write down an explicit model of the full physical and stochastic model of data given parameters $\theta$. This is the **Likelihood**.
2. Get data $d$.
3. Specify **prior**
4. Write down **posterior**

\[ P(\theta|d) = \frac{P(d|\theta)P(\theta)}{P(d)} \]

5. Explore/sample posterior for fixed data as a function of parameters.
6. Done!
Approach 1: “Explicit” inference

1. Write down an explicit model of the full physical and stochastic model of data given parameters $\theta$. This is the **Likelihood**.
2. Get data $d$.
3. Specify **prior**
4. Write down **posterior**
\[
P(\theta|d) = \frac{P(d|\theta)P(\theta)}{P(d)}
\]
5. Explore/sample posterior for fixed data as a function of parameters.
6. Done!
Approach 1: “Explicit” inference

1. Work out the full physical and stochastic model of data given parameters $\theta$. This is the Likelihood.
2. Get data $d$.
3. Specify prior
4. Write down posterior
5. Explore/sample posterior for fixed data as a function of parameters.

What if $d = ?$
Approach 1: “Explicit” inference

1. Work out the full physical and stochastic model of data given parameters $\theta$. This is the likelihood.
2. Get data $d$.
3. Specify prior
4. Write down posterior
5. Explore/sample posterior for fixed data as a function of parameters.
What we want

To succeed we need more freedom than a traditional likelihood approach can provide:

• FREEDOM to make our physical model anything we want
• FREEDOM to project/summarize/cut/mask our data any way we want

Simulating data is often much easier than deriving an accurate likelihood.

*Can we analyze data if all we can do is simulate it?*
Can we analyze data if all we can do is simulate it?

Yes!

A major shift over the last 5 years.

Likelihood is represented *implicitly* through simulations

\[ d \leftrightarrow p(d|\theta) \]

Let’s do a simple example.

Challenge: Keep a running count of the number of likelihood and prior evaluations!
The easiest diagram to explain Bayesian statistics
Generate $\theta$ from the prior
Simulate/generate $d$ given $\theta$

i.e., draw from $p(d|\theta)$
This is a sample from $p(\theta, d)$
Slice through $p(\theta, d)$ at $d$ ("condition on $d$")
Slice through $p(\theta, d)$ at $d$ ("condition on $d$")
This was the true $\theta$
Posterior $p(\theta | d)$

How many likelihood evaluations?

How many prior evaluations?
This is *Implicit* Inference

• When likelihood and/or prior are not *explicitly* specified but *implicit* in...
  – simulations, generative models, labelled data.

• Various forms known as
  – Likelihood-free inference
  – Simulation-based inference
  – Approximate Bayesian Computation (ABC)
The Machine Learning revolution in computational astrophysics and cosmology

• Many problems that we considered impossible solved in the last ~5 years:
  – Automated finding of informative data summary statistics
    • computing score functions and Fisher Information for intractable models (e.g., IMNN, FI)
  – Posteriors/likelihoods/priors for intractable models
    • Implicit Inference (likelihood-free, or simulation-based): LRE, DELFI
    • Routinely used to compute posterior moments (e.g., Moment Networks)
    • Posterior samples for huge non-linear inverse problems (e.g., Initial Conditions)
  – Bayesian Evidence for intractable models
    • Simulation-based model comparison(e.g., Evidence Networks)

IMNN: Charnock, Lavaux & Wandelt, arXiv:1802.03537; LRE: Cranmer, Pavez & Louppe, 1506.02169; DELFI: Papamakarios, Murray et al., 1705.07057, 1805.07226; Alsing & Wandelt, 1712.00012; Alsing, Feeney & Wandelt, 1801.01497, 1903.01473; MN & EN: Jeffrey & Wandelt, 2011.05991, 2305.11241; FI: Coulton & Wandelt, 2305.08994, ICs: Legin et al., 2304.03788
Machine learning takes us the rest of the way

- *Recast inference problems as optimization problems.*

- Write down a loss that defines the problem
  - Parameterize the solution using a neural network
  - Minimize
  - Validate
First example: variational Bayes

- Define a parameterized family of distributions
- Minimize Kullback-Leibler loss between neural family and true likelihood

When using a neural density estimator this is DELFI, a (now) classic example of simulation-based inference.

$$D_{KL}(p^* \mid p) = \int p^*(t|\theta) \ln \left( \frac{p(t|\theta; w)}{p^*(t|\theta)} \right) dt$$

$$-\ln U(w\{\theta, t\}) = -\sum_{i=1}^{N_{\text{samples}}} \ln p(t_i|\theta_i; w)$$

Papamakarios, Murray + coauthors, arXiv:1605.06376, 1705.07057, 1805.07226
Alsing, Feeney & Wandelt, arXiv: 1801.01497, 1903.01473
Example:

Strong Gravitational Lensing

Legin et al.
arXiv:2212.00044
Simulated images
Inference

Works with multimodal posteriors!

Legin et al arXiv 2212.00044
Validation

Legend

- MDN w/ dropout
- MDN w/o dropout
- BNN w/o dropout

CDF

Observation | True Lens | True

1 | Sample 2 | Sample 3

Legin et al arXiv 2212.00044
What if the number of parameters is large or simulations are scarce?

- In general, neural density estimation becomes exponentially hard as number of dimensions increases.
- How do we handle high-dimensional problems?

- Simplify.
MOMENT AND POSTERIOR MARGINAL NETWORKS

Main idea: construct $\mathcal{F}(d), \mathcal{G}(d)$ to go directly from data to posterior.

- **Moment networks**: obtain posterior moments directly from data by training NNs to solve

$$\langle \theta \rangle_{p(\theta|d)} = \arg\min_{F(d)} \int ||\theta - F(d)||_2^2 p(d, \theta) d\theta$$

$$\text{Var} [\theta]_{p(\theta|d)} = \arg\min_{G(d)} \int ||\theta - \langle \theta \rangle_{p(\theta|d)}||_2^2 - G(d)||_2^2 p(d, \theta) d\theta$$

Moment Network Example

Cosmology and astrophysics from full hydrodynamical simulations including black holes, star formation,...
Large suites of full, cosmological hydrosimulations as a function of cosmological parameters and astrophysics models with multiple codes (AREPO/Illustris, GIZMO/SIMBA, Astrid,...).

Cosmology on small scales with baryons

15 different 2-dimensional fields:

1. Gas mass
2. Dark matter mass
3. Stellar mass
4. Gas velocity
5. Dark matter velocity
6. Neutral hydrogen mass
7. Gas temperature
8. Electron density
9. Gas metallicity
10. Gas pressure
11. Magnetic fields
12. Mg/Fe
13. Total mass
14. N-body
15. All fields except dark matter

15,000 images per field from 1,000 CAMELS-IllustrisTNG simulations.

Each image:

• 250x250 pixels
• 25x25 (Mpc/h)^2
• 100 kpc/h resolution
**SBI: COSMOLOGY FROM SMALL-SCALE HYDRO**

Computing posterior means & variances from gas temperature

\[ \mathcal{L} = \sum_{i=1}^{6} \log \left( \sum_{j \in \text{batch}} (\theta_{i,j} - \mu_{i,j})^2 \right) + \sum_{i=1}^{6} \log \left( \sum_{j \in \text{batch}} ((\theta_{i,j} - \mu_{i,j})^2 - \sigma_{i,j}^2)^2 \right) \]

Posterior means & variances computed by moment network minimizing \( \mathcal{L} \)

Villaescusa-Navarro, arXiv:2109.09747
Computing posterior means & variances from gas metallicity

\[ L = \sum_{i=1}^{6} \log \left( \sum_{j \in \text{batch}} (\theta_{i,j} - \mu_{i,j})^2 \right) + \sum_{i=1}^{6} \log \left( \sum_{j \in \text{batch}} ((\theta_{i,j} - \mu_{i,j})^2 - \sigma_{i,j}^2)^2 \right) \]

Posterior means & variances computed by moment network minimizing \( L \)

Villaescusa-Navarro et al., in prep.

Villaescusa-Navarro, arXiv:2109.09747
What the cosmological AI tells us about the CAMELS Multifield Data set

1. There is cosmological information on very small scales (100 kpc)

2. The hydro outputs contain more information than the dark matter density

3. For total matter, inferences are robust to baryonic physics (good news for weak lensing!)

Cosmology robust to baryonic physics

Villaescusa-Navarro et al., arXiv:2109.10360
Illustris TNG

SIMBA

Same initial conditions!
High dimensional application of Moment Networks (with a single training image)

Uses a generative model based on Wavelet Phase Harmonics

Moment networks:
Posterior means and variances pass quantile test
Summary Statistics and Compression
Nuisance hardened data compression for fast likelihood-free inference

Justin Alsing\textsuperscript{1,2,3}\textsuperscript{*} and Benjamin Wandelt\textsuperscript{2,4}

\textsuperscript{1}Oskar Klein Centre for Cosmoparticle Physics, Stockholm University, Stockholm SE-106 91, Sweden
\textsuperscript{2}Center for Computational Astrophysics, Flatiron Institute, 163 5th Ave, New York City, NY 10010, USA
\textsuperscript{3}Imperial Centre for Inference and Cosmology, Department of Physics, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK
\textsuperscript{4}Sorbonne Université, Institut Lagrange de Paris (ILP), 98 bis boulevard Arago, F-75014 Paris, France

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT
In this paper we show how nuisance parameter marginalized posteriors can be inferred directly from simulations in a likelihood-free setting, without having to jointly infer the higher-dimensional interesting and nuisance parameter posterior first and marginalize a posteriori. The result is that for an inference task with a given number of interesting parameters, the number of simulations required to perform likelihood-free inference can be kept (roughly) the same irrespective of the number of additional nuisance parameters to be marginalized over. To achieve this we introduce two extensions to the standard likelihood-free inference set-up. Firstly we show how nuisance parameters can be re-cast as latent variables and hence automatically marginalized over in the likelihood-free framework. Secondly, we derive an asymptotically optimal compression from $N$ data down to $n$ summaries – one per interesting parameter – such that the Fisher information is (asymptotically) preserved, but the summaries are insensitive (to leading order) to the nuisance parameters. This means that the nuisance marginalized inference task involves learning $n$ interesting parameters from $n$ “nuisance hardened” data summaries, regardless of the presence or number of additional nuisance parameters to be marginalized over. We validate our approach on two examples from cosmology: supernovae and weak lensing data analyses with nuisance parameterized systematics. For the supernova problem, high-fidelity posterior inference of $\Omega_m$ and $w_0$ (marginalized over systematics) can be obtained from just a few hundred data simulations. For the weak lensing problem, six cosmological parameters can be inferred from just $O(10^3)$ simulations, irrespective of whether ten additional nuisance parameters are included in the problem or not. If needed, an approximate posterior for the nuisance parameters can be re-constructed a posteriori as a pseudo-Blackwell-Rao estimator (without running any additional simulations).

Key words: data analysis: methods
Automatic Physical Inference with Information Maximizing Neural Networks

• Goal: obviate the need to “guess” heuristic, informative summaries of the data
• Setup: design a neural network that maps the data into a small set of informative summaries.
• The training loss is (– the Fisher information) under an assumed simple likelihood for the summaries.

• Training uses physical simulations of the model to maximize the information in the summaries about the parameters of the model.
• The achieved loss on a test set is meaningful – it’s the information content of the data.

Charnock, Lavaux, Wandelt (arXiv:1802:03537)
Information maximizing neural network

Charnock, Lavaux, Wandelt (arXiv:1802:03537)
Information maximizing neural networks: asymptotically optimal analysis, Information Matrix, score computation *if you don’t know the likelihood*

\[ \mathcal{L} = -\ln \det (F) p(d|\theta_{\text{fid}}) \]

Charnock, Lavaux, Wandelt (arXiv:1802:03537)
Example 1: inference of variance

- Perfect information gives $|F| = 5$ in this problem
- Any linear summary gives $|F| = 0.5$
The IMNN finds a minimal sufficient statistic for this inference problem.

Charnock, Lavaux, Wandelt (arXiv:1802:03537)
Example 2: Automatic physical inference with unknown noise

Charnock, Lavaux, Wandelt (arXiv:1802:03537)
Example 3: Lyman-α forest inference

• The idea is to infer the variance of the underlying density field from a non-linearly transformed, photon-noise dominated Lyman-α forest spectrum

Charnock, Lavaux, Wandelt (arXiv:1802:03537)
Example 3: Lyman-\(\alpha\) forest inference

Charnock, Lavaux, Wandelt (arXiv:1802:03537)
Example 4: excellent generalization from fiducial model

Infer frequency of LISA gravitational wave chirp

Full log-likelihood (LLH)

Gaussian LLH based on linear compression

Gaussian LLH based on IMNN compression

The Information Maximizing Network summary gives the correct unique likelihood peak.

Charnock, Lavaux, Wandelt (arXiv:1802:03537)
IMNN recovers full info directly from the field

Makinen et al., arXiv:2107.07405
The IMNN recovers the full information

Theoretical (Cramer-Rao) information bound

Realized training loss

Makinen et al., arXiv:2107.07405

11 minutes on 1 GPU
Non-Gaussian field inference with IMNN and DELFI


Makinen et al., arXiv:2107.07405
Can define both information matrix and score function on distributions of graphs

Example of using clusters of galaxies to infer cosmological parameters

Uses neurally derived Fisher score within pyDELFi.

Makinen et al. arXiv:2207.05202
Other ways to compute the Information Matrix *implicitly*

- Typically, compute at a fiducial parameter point. *Can we compute Fisher information efficiently everywhere in parameter space?*
  - Fishnets (Makinen et al, in prep)
Reasoning about models with Bayesian Machine Learning

\[ p(\theta | d) = \frac{p(d | \theta) p(\theta)}{p(d)} \]

Actually, \( p(d | M_i) \)
Bayesian model comparison

\[
\frac{p(M_i|d)}{p(M_j|d)} = \frac{p(d|M_i)}{p(d|M_j)} \frac{p(M_i)}{p(M_j)}
\]

Bayes factor $K$
Bayesian model comparison

*Even if* likelihood and posterior are explicitly given

- Likelihood can be costly to evaluate
- **Evidence can be hard to compute**

\[
P(\theta | d, M) = \frac{P(d | \theta, M)P(\theta | M)}{P(d | M)}
\]

\[
\Rightarrow P(d | M) = \int P(d | \theta, M)P(\theta | M)d\theta
\]
Sampling high-dimensional posteriors in Implicit Inference and Information-Ordered Bottlenecks

Benjamin D. Wandelt

Ronan Legin (Montreal), Niall Jeffrey (UCL), Matt Ho (IAP), Xiaosheng Zhang (Tsinghua), Gabriel Jung (IAS, Orsay), Lucas Makinen (Imperial), Will Coulton (CCA), Stephen Feeney, ...
“Score”-based Diffusion

- Consider a random walk of images
- Initialise with initial conditions
- Add Gaussian noise at every step
- Central limit theorem: this has an attractor a Gaussian noise distribution

- Then sample by solving a series of inference problems to go from Gaussian noise back to a sample of the initial conditions
- If the number of steps is large enough, each step is a Gaussian inference problem.
- Train a neural network on simulations to learn the posterior mean for each of these steps
Exploring high-dimensional posterior pdfs with “score”-based diffusion

R. Legin et al., arxiv:2304.03788
Train the “score” on QUIJOTE n-body simulations

• Largest release of N-body simulation data to date
  – 43,100 full GADGET 3 simulations (1 Gpc)$^3$, $512^3$ or $1024^3$ particles
  – ~1 PB of data
• Goal: quantify statistics information content of non-Gaussian non-linear density field about cosmological parameters
• Includes full dark matter snapshots, halo and void catalogues, and many pre-computed statistics.

First full-field inference of initial conditions from fully non-linear density field

- 1 Gpc GADGET1024$^3$ simulation at z=0
- Binned on 128$^3$ grid

R. Legin et al., arxiv:2304.03788
First full-field inference of initial conditions from fully non-linear density field

R. Legin et al., arxiv:2304.03788
First full-field inference of initial conditions from fully non-linear density field

R. Legin et al., arxiv:2304.03788
First full-field inference of initial conditions from fully non-linear density field

R. Legin et al., arxiv:2304.03788
First full-field inference of initial conditions from fully non-linear density field

R. Legin et al., arxiv:2304.03788
First full-field inference of initial conditions from fully non-linear density field

R. Legin et al., arxiv:2304.03788
Faithful reconstruction...

R. Legin et al., arxiv:2304.03788
including uncertainties (posterior variance)

R. Legin et al., arxiv:2304.03788
Accurate reconstructions

Points to note:

• full non-linear gravity
• No need for differentiability

R. Legin et al., arxiv:2304.03788
The principal impact of ML on cosmology so far

An extremely powerful tool to answer the question

“Does A contain information about B?”
Theoretical discovery example: breaking degeneracy

- Scientific results now motivated by machine learning tools

We study signatures of non-linear scales, using a combination of the power spectrum, and bispectrum, and their responses of our summary statistics constructed from the Q2Pk dataset: local, equilateral and orthogonal power spectrum and bispectrum products. Roughly a factor 2, while second, it helps break in $b_{\phi}$, without relying on $\Delta f_{\text{NL}}^{\text{local}} = 40, \Delta f_{\text{NL}}^{\text{equiv}} = 90$, for a cubic volume of $r^{3}$, with a halo number density of $n \sim 5.1 \times 10^{-5} h^{3} \text{Mpc}^{-3}$, at $z = 1$, and considering scales up to $k_{\text{max}} = 0.5 h \text{Mpc}^{-1}$. 

HMF + P + B ($k_{\text{max}} = 0.2 h \text{Mpc}^{-1}$)

HMF + P + B ($k_{\text{max}} = 0.5 h \text{Mpc}^{-1}$)

Same, assuming $b_{\phi}$ fixed
From information to insight

Can we use machine learning to discover the most important degrees of freedom in data?
The Information-Ordered Bottleneck

\[ \hat{\theta} = \arg \max_{\theta \in \Theta} \sum_{i=1}^{N} \sum_{k=0}^{k_{\text{max}}} \rho_k \ell \left[ f_{\theta}^{(k)}(x_i), y_i \right] \]

Ho, Zhao & Wandelt arXiv:2305.11213
The IOB orders latents by information

Ho, Zhao & Wandelt arXiv:2305.11213
The IOB discovers the number of degrees of freedom in data

Ho, Zhao & Wandelt arXiv:2305.11213
## Information Ordered Bottleneck:
### SOTA discovery of global intrinsic dimensionality

<table>
<thead>
<tr>
<th>ID Estimator</th>
<th>S-curve</th>
<th>1-Ball</th>
<th>2-Ball</th>
<th>3-Ball</th>
<th>4-Ball</th>
<th>MS-COCO CLIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA [23]</td>
<td>3</td>
<td>33</td>
<td>37</td>
<td>39</td>
<td>38</td>
<td>106</td>
</tr>
<tr>
<td>MADA [25]</td>
<td>2.5</td>
<td>inf</td>
<td>13.2</td>
<td>16.9</td>
<td>19.5</td>
<td>22.7</td>
</tr>
<tr>
<td>TwoNN [24]</td>
<td>2.9</td>
<td>5.3</td>
<td>13.6</td>
<td>16.3</td>
<td>21.4</td>
<td>21.4</td>
</tr>
<tr>
<td>Linear IOB*</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>10</td>
<td>14</td>
<td>322</td>
</tr>
<tr>
<td>Geometric IOB*</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>10</td>
<td>12</td>
<td>196</td>
</tr>
</tbody>
</table>

| Data Dimensionality | 3       | 1024   | 1024   | 1024   | 1024   | 768          |
| True Dimensionality | 2       | 3      | 6      | 9      | 12     | ≤ 768        |

Ho, Zhao & Wandelt arXiv:2305.11213
Training on heterogenous data sets can classify individual instances by complexity.
Adaptive IOB Compression in Semantic Latent Space

Information-Ordered Bottleneck applied to CLIP embeddings.

Dataset: MS-COCO

Check out Ho, Zhao & Wandelt arXiv:2305.11213 for more examples!
Information-Ordered Bottlenecks for insight in a world exploding with data

What is the (relative) complexity of Galaxies and Active Supermassive Black Holes spectra?

What parameter combinations does the CMB really constrain?

What are natural coordinates for the parameter space of cosmological hydro-simulations?