Generative models in cosmology and beyond (From cosmological data analysis to fast Bayesian methods and machine learning)

Uroš Seljak

UC Berkeley/LBL
Outline

• Generative models create synthetic data
• Full N-body or hydro is a (not fast) generative model
• Generative models as an optimal parameter estimation problem in cosmology
• Physics based generative models
• Generative models for Bayesian evidence
• Generative models for machine learning
• w. B. Dai, H. Jia, C. Modi, Y. Feng, B. Yu…
Current data analysis in cosmology

• We have some data such as galaxy positions, weak lensing distortions, CMB...

• The goal of data analysis is to extract information about cosmological parameters from the probability distribution of data: data likelihood $p_\theta(x)$

• If the field is Gaussian (e.g. CMB) the likelihood depends only on correlation function or power spectrum. We have good methods (e.g. optimal quadratic estimator)

• There is a lot more information in galaxies, weak lensing, that are in higher order correlations

• How do we extract these? How do we get their covariance matrix? No obvious solution.
Alternative: “optimal” transport

• We want data likelihood $p_{\theta}(x)$
• Monge 1781: Can we transform with $y=G(x)$ a given probability distribution of the data to another, such as a simple multi-variate Gaussian?
• $p_{\theta}(x)dx=q(y)dy$, so $p_{\theta}(x)=q(y)|dy/dx|$

$$p_{\theta}(x) = N[G_{\theta}(x); 0, I]|\nabla x G_{\theta}|$$

• We need $G_{\theta}(x)$ as a function of cosmology parameters $\theta$ and Jacobian too
• Goal: finding $G(x)$ means to Gaussianize data
• In cosmology this is equivalent to reconstruction of initial density, which is Gaussian distributed
• If replace Gaussian $q(y)$ with uniform (PDF to CDF): copula
Likelihood formulation without Jacobian

\[ p_\theta(x) = N[\mathcal{G}_\theta(x); 0, I]|\nabla_x \mathcal{G}_\theta| \]

- Introduce latent space \( z = G(x) \)

\[ p_\theta(x) = \int dz \exp \left\{-\frac{1}{2} \sum_{j=1}^{N} \left[ z_j^2 + \ln 2\pi \right] \right\} \delta_D [x - \mathcal{G}^{-1}(z)] \]

\[ p_\theta(x) = \lim_{\sigma^2 \to 0} \int dz \exp \left\{-\frac{1}{2} \sum_{j=1}^{N} \left[ z_j^2 + \frac{[x_j - \mathcal{G}^{-1}_j(z)]^2}{\sigma^2} + 2 \ln 2\pi + \ln \sigma^2 \right] \right\} \]

- Introduce noise and generative (forward) model \( \mathcal{G}^{-1} \)

\[ p_\theta(x) = \int dz \exp \left\{-\frac{1}{2} \sum_{j=1}^{N} \left[ z_j^2 + \frac{[x_j - \mathcal{G}^{-1}_j(z, \sigma^2 = 0)]^2}{\sigma^2} + 2 \ln 2\pi + \ln \sigma^2 \right] \right\} \]

- We marginalize over \( z \) to get likelihood of parameters \( \theta \)
Cosmology Forward model: from initial to final dark matter to galaxies

**Initial Dark Matter Field (Gaussian) (s)**

**Particle Mesh Sims**

**Final Dark Matter Field (d)**

**Group finding: Friends of Friends**

**Dark Matter Halos (d)**

-Leapfrog evolution

\[
P_{n+1/2} = P_{n-1/2} - F(a_n) \nabla \phi_n \Delta a,
\]

\[
x_{n+1} = x_n + \frac{F(a_{n+1/2})}{a_{n+1/2}^2} P_{n+1/2} \Delta a
\]

-Poisson Equation

\[
\nabla^2 \phi = \frac{3 \Omega_0}{2} \frac{\bar{\rho} - 1}{a},
\]

For \( N \) time steps

-Initial Dark Matter
-Data

-Dark Matter Today
-

-Galaxy Light Today
FastPM performance on halos

FastPM with 5(10) steps only 3.8(7.3) times slower than initial condition generator

It enforces ZA on large scales

Comparison against very high resolution simulation: 1-2% accurate for 5 time steps using abundance matching of halos

Elena Massara, Yu Feng, US
How to find the initial density field?

• maximize a posterior (MAP) of $z$, ie solve the optimization problem in $10^{6+}$ dimensions

• To solve this we need a gradient of data $x$ with respect to initial density $z$: this is $10^{6+} \times 10^{6+}$ matrix, fortunately only its product with a vector is needed

• Get the gradient using backpropagation through FastPM kick/drift operations

• Replace FoF with differentiable operation (we use neural networks)

• $O(100)$ iterations are used in optimization
Initial density reconstruction

We replace dark matter galaxy connection physical modeling with neural network trained on simulations: differentiable and fast

Approach to Reconstruction

Forward model is not enough, we need a likelihood for the data

\[ \chi^2(s) = s^\dagger S^{-1} s + [d - F(s)]^\dagger N^{-1} [d - F(s)]. \]
Example of MAP reconstruction

We use optimization that finds the best solution in terms of final data (optimal filter). This 3-d example optimizes in 2 million dimensions. Galaxy are sparse tracers, so we loose small scale info.
Incomplete data: dark matter example
From MAP to parameter estimation

- Simple Maximum Likelihood Estimator is wrong when number of parameters $N_z$ is similar to the data size $N_x$.
- Instead we have to marginalize out latent space first.

$$p_\theta(x) = \int d\mathbf{z} \exp \left\{ -\frac{1}{2} \sum_{j=1}^{N} \left[ z_j^2 + \frac{[x_j - G_{j\theta}^{-1}(z, \sigma^2 = 0)]^2}{\sigma_j^2} \right] + 2 \ln 2\pi + \ln \sigma_j^2 \right\}$$

The marginalization integral gives rise to Hessian determinant:

$$-\ln p(x|z) = \tilde{\mathcal{L}}_p(z, \mu_s|z, x) - \frac{1}{2} \ln \det \Sigma_{s|z} - \frac{1}{2} N_s \ln(2\pi) + \ln p(z).$$

- Only now we can maximize $p(x)$ wrt $\theta$ leads to find MLE parameters.
- We use simulation based evaluation of Hessian determinant derivative: unbiased even for non-Gaussian case, no sampling is needed, but optimality is not guaranteed.
- Covariance matrix can also be obtained using simulations.
Cosmology is all about error quantification

Reconstruction of linear cosmological power: we removed BAO smearing (perfect BAO reconstruction)

Response based inverse covariance matrix

No need to run mock simulations to get covariance matrix
Future directions

• Compare these methods to HMC sampling in terms of errors (much more expensive, but has better optimality guarantees)

• Marginalizing over astrophysics parameters means many more simulations varying these parameters will be needed

• Scale up in terms of volume and mass resolution: for DESI and LSST we will likely need to run $10^{12}$ particle simulations hundreds to thousands of times

• Payoff: optimal analysis, best BAO reconstruction, up to 2 x smaller error
Potential Gradient Descent

- How to improve resolution of FastPM? Add another displacement field that moves particles inward or outward

The PGD correction displacement:

$$S = -\alpha \nabla \hat{\Omega}_h \hat{\Omega}_i \phi$$

High pass filter $\hat{\Omega}_h$ prevents the large scale growth, low pass filter $\hat{\Omega}_l$ reduces the numerical effect

- Train on high resolution simulations (Dai et al. 2017).
- Two free parameters only (shape of small scale force)
- Cheap and fast machine learning (in ML usually many more parameters to train)
- Fast to generate (2 extra FFTs)
- For hydro feedback effects: use enthalpy (EGD)
Visual inspection

More mass at the outskirts
FastPM with PGD power spectrum

Weak lensing maps
Generative models of all observables

• We have many different data in cosmology: stellar mass, gas information (X-rays, tSZ, kSZ), dark matter, HI...
• Many of these come from expensive hydro simulations
• We need a fast way to generate forward models
• We need it to be differentiable so we can take a gradient of the data with respect to initial density modes
• PGD+EGD trained on Illustris TNG-300 hydro outputs: 7 parameter (Dai et 2019, in prep) model
• These are all differentiable, so easy to do gradient backpropagation
Future directions in generative models

- Train on low resolution DM on high resolution hydro
- We need to get prior distributions of parameters from different hydro sims: astrophysics prior
- We can also create generative models from data (e.g. CMB foregrounds)
- Can we make even cheaper generative models? (Zeldovich, 1-d, 2-d)
Next step: posterior analysis

• So far we have obtained data likelihood or its summary statistic (e.g. optimal power spectrum), we need posterior of cosmological parameters marginalized over nuisance parameters (astrophysics)

• MCMC is probably out of the question, since we would need a full simulation at every point

• We need cheaper and faster posterior analyses

• Variational methods (Variational Inference): based on stochastic minimization of KL divergence: ADVI

• This is Monte Carlo integration, suffers from sampling noise: slow $N^{-1/2}$ convergence
Our proposal: EL$_2$O f-divergence

With Byeonghee Yu

\[ \mathcal{L}_q = -\ln q(z), \quad q(z) = N(z; \mu, \Sigma) \quad \mathcal{L}_p = -\ln p(z|\mathbf{x}) \]

- We propose to minimize $L_2$ norm between $L_p$ and $L_q$. It needs to be sampled from some fiducial probability distr, which can be $q$

- **EL$_2$O: expectation with $L_2$ optimization**

\[ \text{EL}_2\text{O} = \langle (\mathcal{L}_q - \mathcal{L}_p - c)^2 \rangle_{\tilde{p}} \]

  f-divergence
c is approx. log evidence

- if $q$ covers $p$ it is noiseless, if not it finds the closest solution to it

- No noise because both log $p$ and log $q$ are evaluated at the same position, $L_2$ is positive definite: solving linear least square (convex)

- **No integration: no sampling noise**

- Our proposal: replace noisy KLD with noiseless EL$_2$O
BOSS RSD analysis

• Take summary statistics of galaxy clustering $P_l(k)$, where $l = 0, 2, 4$ are the multipoles of the power spectrum and $k$ is the wavevector.

• Data: Measured $P_l(k)$ of the BOSS DR12 galaxies (LOWZ+CMASS)
• Covariance: nearly diagonal, but model dependent (sampling variance component), plus trispectrum component
• Model: Perturbation theory predicted $P_l(k)$ which depends on 13 parameters, presented in Hand et al

$$P_{gg}^S(k) = (1 - f_s)^2 P_{cc}^S(k) + 2 f_s (1 - f_s) P_{cs}^S(k) + f_s^2 P_{ss}^S(k)$$

<table>
<thead>
<tr>
<th>Sample</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>type A centrals</td>
<td>isolated centrals (no satellites in the same halo)</td>
</tr>
<tr>
<td>type B centrals</td>
<td>non-isolated centrals (at least one satellite in same halo)</td>
</tr>
<tr>
<td>type A satellites</td>
<td>isolated satellites (no other satellites in same halo)</td>
</tr>
<tr>
<td>type B satellites</td>
<td>non-isolated satellites (at least one other satellite)</td>
</tr>
</tbody>
</table>
BOSS RSD analysis with analytic PT model

Our model fits the data very well, accurately modeling $P_0$, $P_2$, and $P_4$ to $k = 0.4 \ h\text{Mpc}^{-1}$.
Near perfect agreement of $\text{EL}_2\text{O}$ posterior with MCMC with 125 $\text{EL}_2\text{O}$ evaluations vs $10^5$ for MCMC
BOSS RSD analysis cosmological constraints

EL_2O

This is the most accurate RSD analysis to date, about 30% improvement over previous BOSS analyses (including recent EFT papers).

Competitive with weak lensing

Combined fσ_8 error of 3%: smallest error to date

Consistent with standard cosmology
BayesFast Planck analysis

• Towards a general fast Bayesian posterior method
• Planck 8 dim with w: EL$_2$O (250 CAMB calls) vs MCMC (10$^6$ CAMB calls)

Code release: work in progress with He Jia
Dark energy Survey (DES) analysis

- 27 correlated dimensions, no gradients available
- Need about 300 CAMB calls, versus $10^6$ for MultiNEST (but more accurate than MultiNEST)
- Here we combined EL$_2$O with quadratic surrogate HMC
Bayesian evidence

\[
p_\theta(x) = \int dz p(z)p_{n\theta}(x|z)
\]

- This is an integral of likelihood over the prior, extremely expensive with MCMC (nested sampling, annealed importance sampling)
- Generative models are normalized, MCMC samples are not
- We can obtain it by finding a bijective generative model that reproduces the distribution of MCMC samples
- We can model very complex distributions by transporting the samples to a Gaussian (optimal transport, Gaussianization)

\[
p_\theta(x) = N[G_\theta(x); 0, I]|\nabla x G_\theta|
\]

- Need to keep track of Jacobian
- Can be improved by importance or bridge sampling
Information theory: each bijective transformation reduces multi information and increases entropy towards maximum entropy solution: Gaussian $N(0,I)$

Iterative transformation of samples into a Gaussian

One can draw samples from it using inverse of bijective transformation

This generative model gives good samples after 10-20 transformations
Bayesian evidence

- Hard example: 32-dimensional thin rotated banana
- A lot faster and more accurate than AIS or nested sampling
- 22s (our method) versus 30 min for dynesty (nested sampler)
Summary

• In cosmology we have good generative models (simulations), but we need them to be fast and we need their gradient with respect to $10^{6++}$ initial density parameters: FastPM trained on hydro sims

• Reconstruction of initial density is inverse problem: if we can solve it we can optimally extract cosmological information. We now have all the tools, we just need to scale it to the datasizes we have

• Similar generative model ideas can also be applied to Bayesian posterior and evidence calculations: potential for very large reduction in CPU relative to MCMC methods
Future of supervised ML: generative learning

• Learn $p_\theta(x)$ from labeled data or simulations
• for different hypotheses $\theta$, use likelihood ratio to classify or regress
• Supervised ML is dominated by discriminative learning (for a good reason)

• Example: 30 dimensional Atlas Higgs data, background versus signal