

Flow-based likelihoods for non-Gaussian inference



BCCP

Ana Díaz Rivero
November 10, 2020



Outline

1. Tensions in cosmology
2. Gaussian likelihoods
3. Data-driven likelihoods:
 - a. Gaussian mixture models
 - b. Independent component analysis
 - c. Flow-based likelihoods
4. Quantifying non-Gaussianity in a dataset
5. Application to the weak lensing convergence power spectrum
6. Summary and conclusions

Tensions in cosmology

With the avalanche of data over the past 10-20 years we are in the era of precision cosmology, where parameters are measured at the percent level.

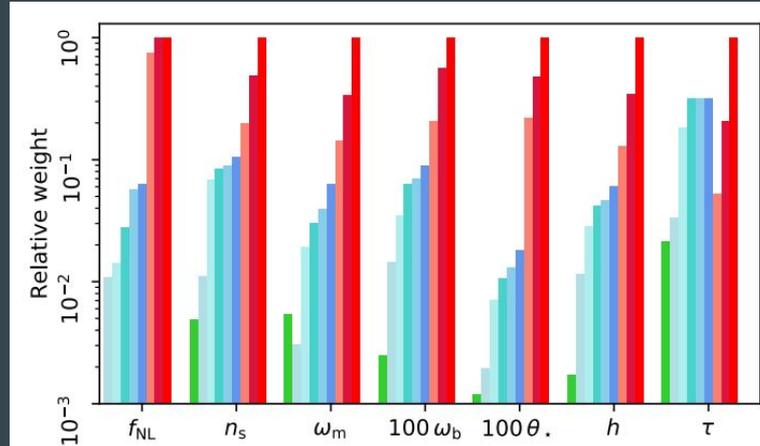
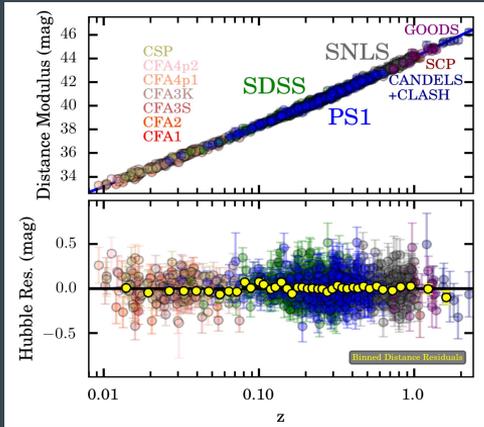


Fig. 13. Increase in the “statistical weight” (i.e., $1/\sigma^2$, where σ for each parameter comes from marginalising over the rest of the set) for a selection of Λ CDM parameters as a function of time. The bars represent the same divisions as in Figs. 11 and 12: pre-WMAP (green); WMAP1, WMAP3, WMAP5, WMAP7, and WMAP9 (blue shades); and Planck13, Planck15, and Planck18 (red shades).

Tensions in cosmology

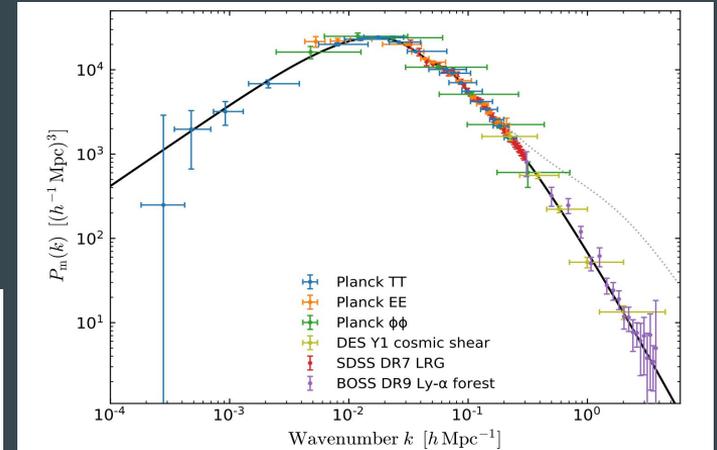
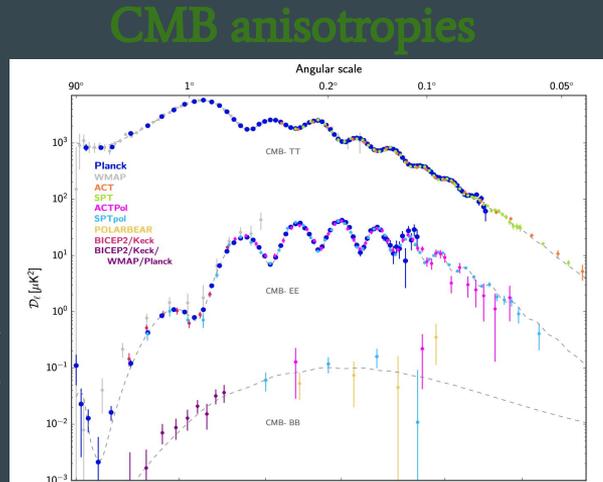
Λ CDM parameters are not **predicted** but rather **inferred** from observations*.

We can obtain **independent estimates** of the parameters from **different observables** in our single universe to check for consistency between our universe and the Λ CDM model.



Expansion history

Planck (2020)



Clustering of matter

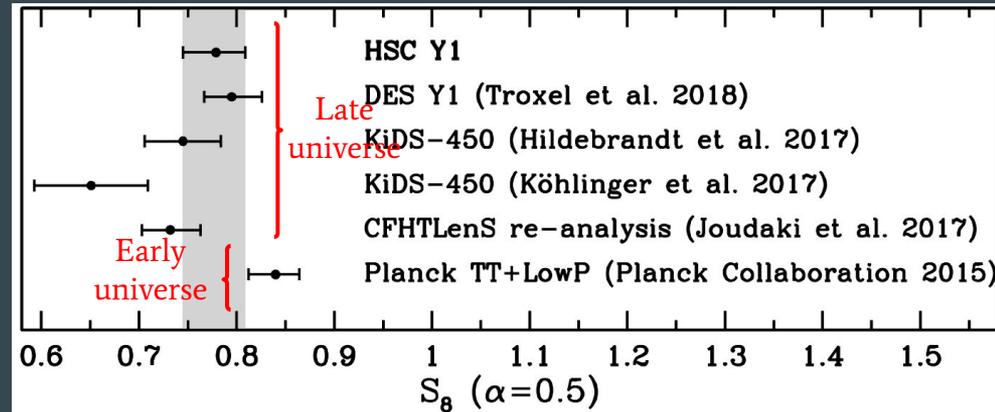
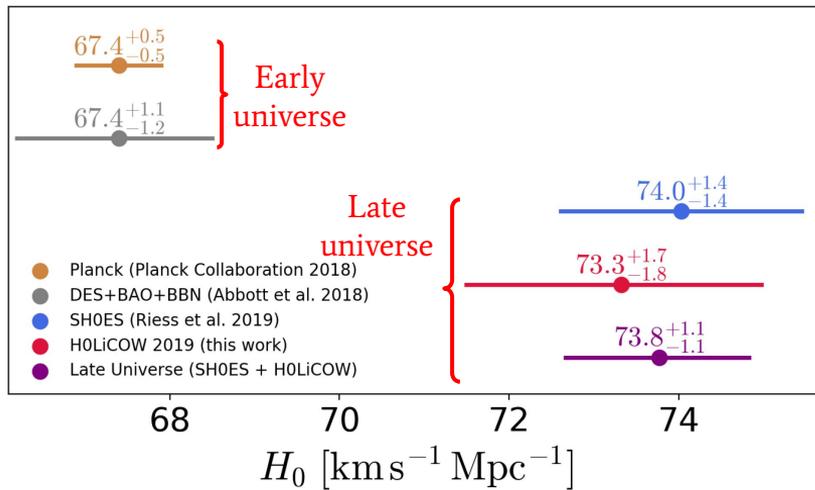
Planck (2020)

Scolnic+ (2018)

Tensions in cosmology

While individual datasets do not favor extensions to the base model, we see significant **tensions** between some of the Λ CDM (derived) parameters from different observables.

flat Λ CDM



Growth of structure

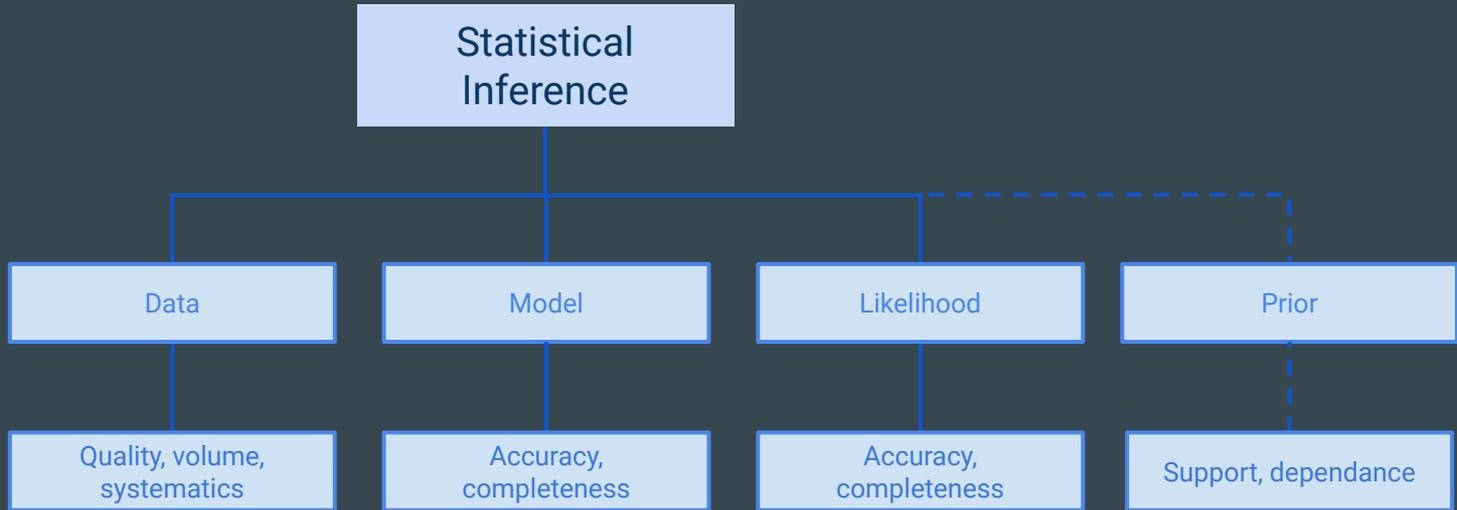
Hikage+ (2019)

Wong+ (2019) Current expansion rate of the universe

Tensions in cosmology

Are cosmological parameters as **accurate** as they are **precise**?

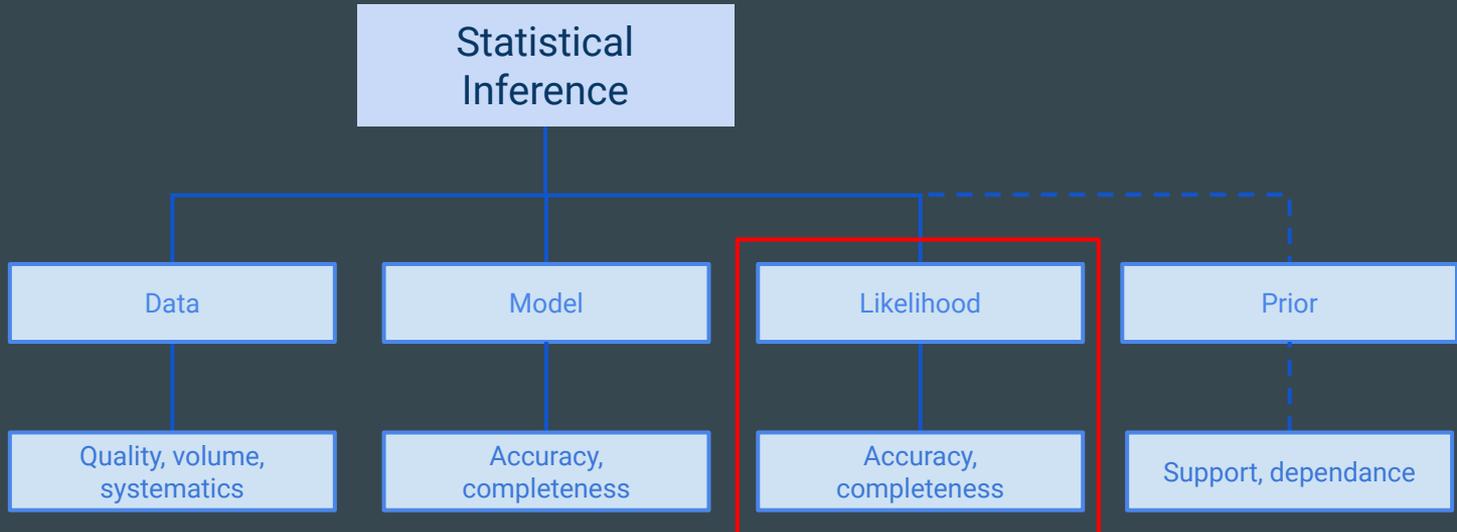
Without answering this question we can't know if tensions point to new physics.



Tensions in cosmology

Are cosmological parameters as **accurate** as they are **precise**?

Without answering this question we can't know if tensions point to new physics.



The Likelihood

The likelihood measures the extent to which a sample provides support for particular values in a statistical model.

Much of statistical inference is predicated on the likelihood:

- Maximum likelihood estimates
- likelihood ratio
- posteriors
- Bayes factor
- ...

Gaussian likelihoods

Gaussian likelihoods are very widespread: well understood, only need a covariance matrix, CLT...

CMB power spectra (*Planck* 2018)

maps. Specifically, the low- ℓ temperature (TT) likelihood is constructed by approximating the marginal distribution of the temperature angular power spectrum derived from Gibbs sampling-based component separation. The low- ℓ polarization (EE) likelihood is built by comparing a cross-frequency power spectrum of two foreground-corrected maps to a set of simulations. The temperature and polarization high- ℓ likelihoods (TT, TE, and EE) uses multiple cross-frequency spectra estimates, assuming smooth foreground and nuisance spectra templates and a Gaussian likelihood approximation.

Shear 2pt function (HSC)

$$-2\ln \mathcal{L}(\mathbf{p}) = \sum_{i,j} (d_i - m_i(\mathbf{p})) \text{Cov}_{ij}^{-1} (d_j - m_j(\mathbf{p}))$$

$$p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp \left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right]$$

Galaxy power spectrum (SDSS-III BOSS)

0.655, and 0.750 ± 0.040 and the DAO scale parameter is $D_V r_s^{\text{fid}}/r_s = 1493 \pm 28, 1913 \pm 35, \text{ and } 2133 \pm 36$ Mpc. Assuming Gaussian likelihood, we provide a covariance matrix which contains the parameter constraints as well as their correlations (see appendix B).

Galaxy clustering + weak lensing (KiDS-1000)

6.2. Gaussian likelihood assumption

Along with the vast majority of large-scale structure cosmological analyses, we adopt a multivariate Gaussian likelihood. This is expected to be a generally excellent approximation if the summary statistics entering the likelihood have been averaged over many modes in the underlying fields. Exact likelihood expres-

Gaussian likelihoods

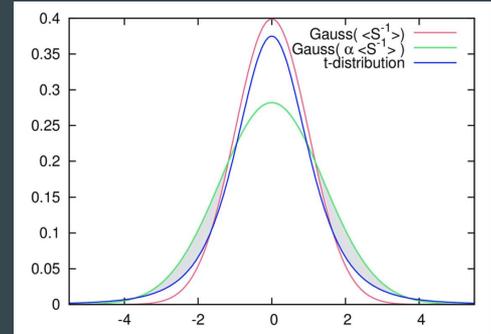
Gaussian likelihoods are very widespread: well understood, only need a covariance matrix, CLT...

$$p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp \left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right]$$

However,

- CLT isn't always applicable (e.g. power spectra at small wavenumbers)
- For estimated covariance, marginalize over the true covariance (Gaussian \rightarrow t -distribution)
- Systematic effects can introduce non-Gaussian correlations
- Physics giving rise to an observable: a nonlinear function of Gaussian RVs is not Gaussian distributed (CMB vs. galaxy distributions)

$$p(\mathbf{x}|\boldsymbol{\mu}, \hat{\Sigma}, N) = \int d\Sigma \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) p(\Sigma|\hat{\Sigma}, N)$$



Sellentin+ (2016)

Gaussian likelihoods

Gaussian likelihoods are very widespread: well understood, only need a covariance matrix, CLT...

$$p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp \left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right]$$

Using a wrong likelihood introduces a source of **systematic uncertainty**:

can bias parameters inferred from some data → are tensions created/amplified by the use of wrong likelihoods?

Gaussian likelihoods

There isn't always a clear alternative/better likelihood:

ACT Thermal SZ one-point PDF (Hill+, 2015)

Gaussianizing
data →

parameters to be broken. In this analysis, the data are not quite at the level needed to strongly break the cosmology-ICM degeneracy. The problem is made more challenging by the highly correlated, non-Gaussian nature of the PDF likelihood function (see Section V below), which we simplify by combining many of the bins in the tail of the tSZ PDF. With a more sophisticated approach to the likelihood function and wider, deeper maps, future measurements of the tSZ PDF should allow for a stronger breaking of the cosmology-ICM degeneracy.

CFHTLenS shear correlation (Sellentin+, 2018)

Removing NG
bins →

As demonstrated in the previous section, the correlations between various data points of CFHTLenS give rise to non-Gaussianities at a 30% level according to our definition. Here, we present a preliminary study of how these non-Gaussianities might impact parameter constraints, by excluding the most contaminated data points from the likelihood. However, as essentially the entire CFHTLenS dataset is contaminated (see Fig. 3), such exclusions are clearly a suboptimal strategy. We nonetheless report our findings as intermediate results and postpone an update to a non-Gaussian likelihood to future work.

An alternative: data-driven likelihoods (DDLs)

Data-driven likelihoods are learned from data:

- We can think of (mock) data as independent draws from the underlying true likelihood function.
- We can estimate the data's PDF with sufficient samples from it.

The hope is that DDLs can accurately capture non-Gaussianities in the data.

An alternative: data-driven likelihoods (DDLs)

Gaussian Mixture Models (GMM)

$$\hat{p}_{\text{GMM}}(\mathbf{x}) = \sum_{i=1}^K \phi_i \mathcal{N}(\mathbf{x} | \mu_i, \Sigma_i)$$

weights unknown
 parameters

Use expectation maximization to find parameters, BIC to determine K .

Independent Component Analysis (ICA)

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

Rotate principal components to maximize statistical independence.

$$\hat{\mathbf{S}} \equiv \mathbf{x}_{\text{ICA}} = \mathbf{W}\mathbf{x} = \{\mathbf{x}_{1,\text{ICA}}, \dots, \mathbf{x}_{N,\text{ICA}}\}$$

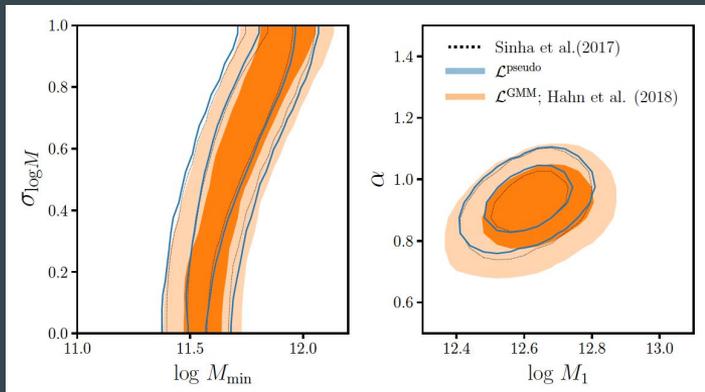
KDE

$$\hat{p}_{\text{ICA}}(\mathbf{x}) = \prod_{n=1}^N \hat{p}_n(\mathbf{x})$$

An alternative: data-driven likelihoods (DDLs)

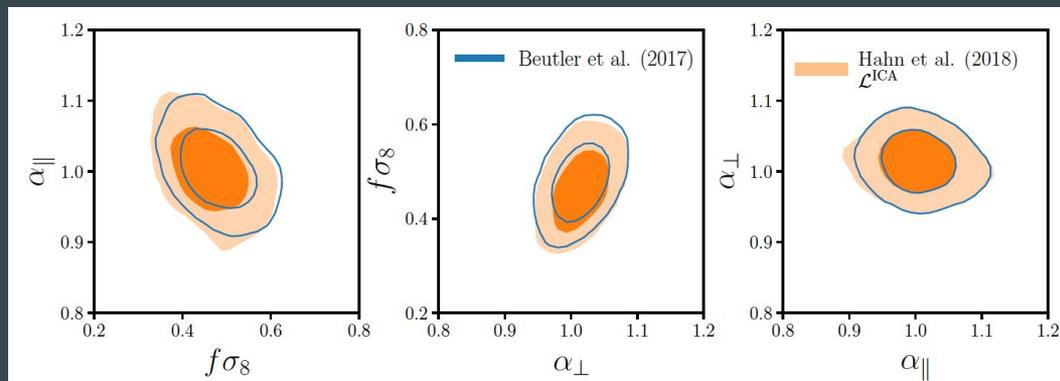
Hahn+ (2018): Large-scale structure with non-Gaussian likelihoods ($<0.5\sigma$ shifts)

Group multiplicity function (with GMM)



20,000 mocks

Galaxy power spectrum (with ICA)



2,048 mocks

An alternative: data-driven likelihoods (DDLs)

Flow-based Likelihoods (FBLs, Diaz Rivero & Dvorkin 2020)

→ I will introduce flow-based generative models

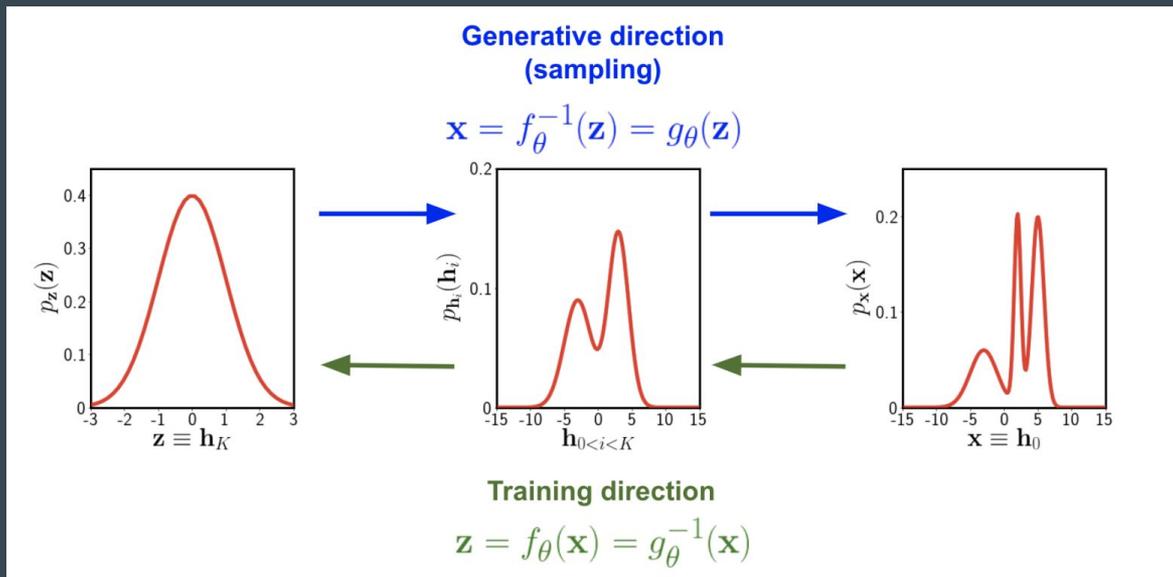
→ Their minimization objective is what we will call a flow-based likelihood

See also literature on **simulation-based inference** and **likelihood-free inference** (e.g. DELFI), which have used flows for density estimation as well!

Flow-based generative models

Generative models aim to learn the probability distribution that gave rise to data \mathbf{x} , such that new samples can be drawn.

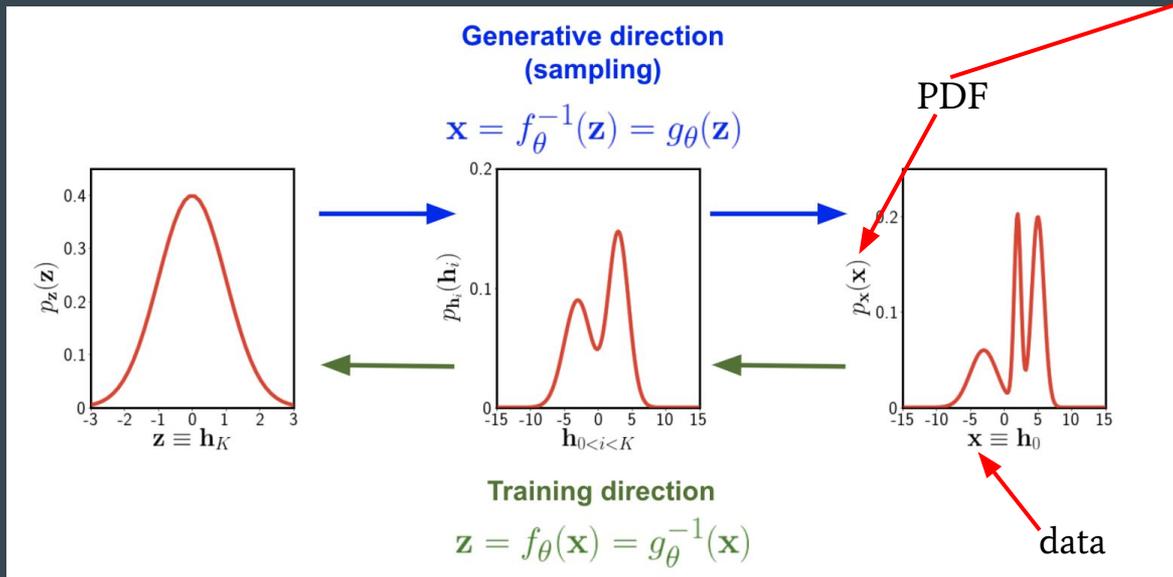
In *flow-based models*, a simple distribution is repeatedly transformed to match $p(\mathbf{x})$.



Flow-based generative models

Generative models aim to learn the probability distribution that gave rise to data \mathbf{x} , such that new samples can be drawn.

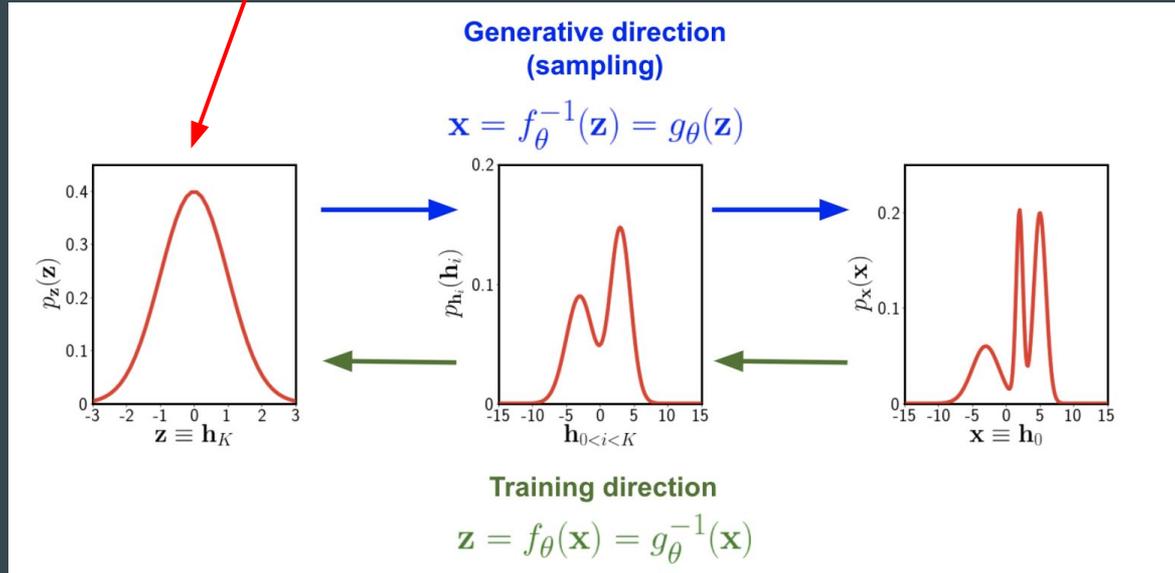
In *flow-based models*, a simple distribution is repeatedly transformed to match $p(\mathbf{x})$.



Flow-based generative models

Generative models aim to learn the probability distribution that gave rise to data \mathbf{x} , such that new samples can be drawn.

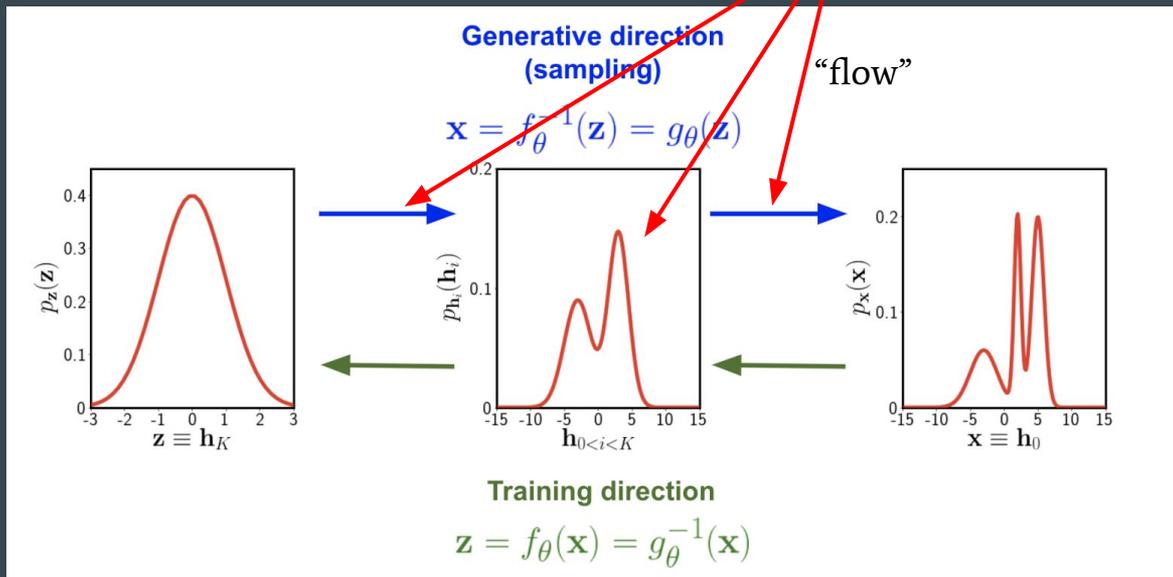
In *flow-based models*, a simple distribution is repeatedly transformed to match $p(\mathbf{x})$.



Flow-based generative models

Generative models aim to learn the probability distribution that gave rise to data \mathbf{x} , such that new samples can be drawn.

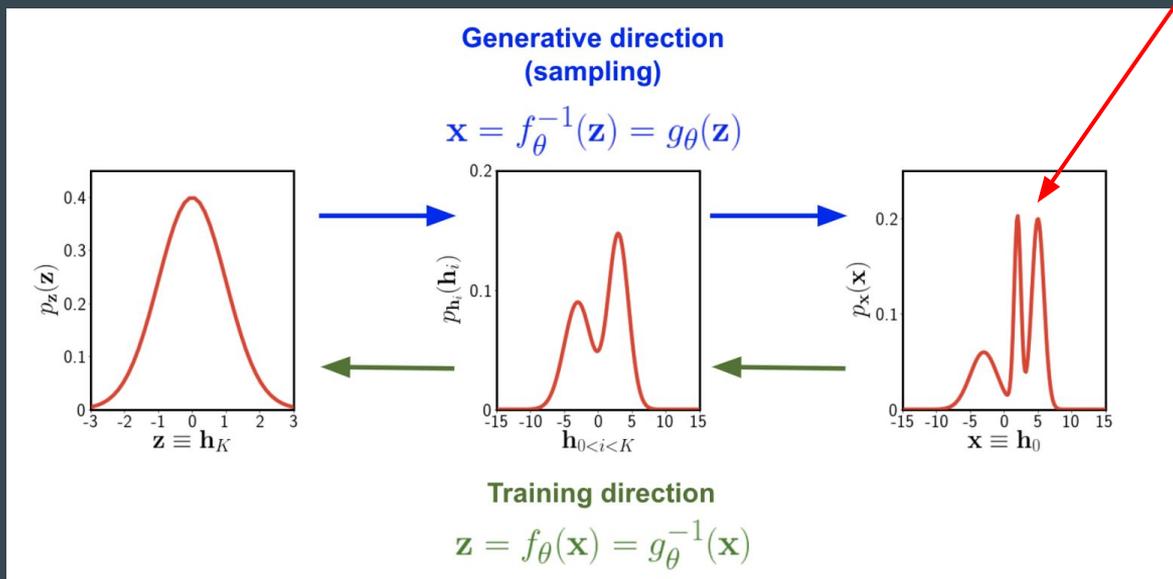
In *flow-based models*, a simple distribution is repeatedly transformed to match $p(\mathbf{x})$.



Flow-based generative models

Generative models aim to learn the probability distribution that gave rise to data \mathbf{x} , such that new samples can be drawn.

In *flow-based models*, a simple distribution is repeatedly transformed to match $p(\mathbf{x})$.



Flow-based generative models

Generative models aim to learn the probability distribution that gave rise to data \mathbf{x} , such that new samples can be drawn.

In *flow-based models*, a simple distribution is repeatedly transformed to match $p(\mathbf{x})$.

$$\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})$$

$$\mathbf{x} \equiv \mathbf{h}_0 \begin{array}{c} \xleftarrow{f_1} \\ \xrightarrow{g_K} \end{array} \mathbf{h}_1 \begin{array}{c} \xleftarrow{f_2} \\ \xrightarrow{g_{K-1}} \end{array} \dots \begin{array}{c} \xleftarrow{f_{K-1}} \\ \xrightarrow{g_2} \end{array} \mathbf{h}_{K-1} \begin{array}{c} \xleftarrow{f_K} \\ \xrightarrow{g_1} \end{array} \mathbf{h}_K \equiv \mathbf{z}$$

$$f = f_1 \circ f_2 \circ \dots \circ f_K$$

$$g = g_1 \circ g_2 \circ \dots \circ g_K$$

$$\mathbf{x} = g_{\theta}(\mathbf{z}) = f_{\theta}^{-1}(\mathbf{z})$$

$$\mathbf{z} = g_{\theta}^{-1}(\mathbf{x}) = f_{\theta}(\mathbf{x})$$

Flow-based generative models

Generative models aim to learn the probability distribution that gave rise to data \mathbf{x} , such that new samples can be drawn.

In *flow-based models*, a simple distribution is repeatedly transformed to match $p(\mathbf{x})$.


$$\begin{aligned}\log p_{\mathbf{x}}(\mathbf{x}) &= \log p_{\mathbf{z}}(\mathbf{z}) + \log \left| \det \left(\frac{d\mathbf{z}}{d\mathbf{x}} \right) \right| \\ &= \log p_{\mathbf{z}}(\mathbf{z}) + \sum_{i=1}^K \log \left| \det \left(\frac{d\mathbf{h}_i}{d\mathbf{h}_{i-1}} \right) \right|.\end{aligned}$$

Flow-based generative models

The goal is to train a model to learn these transformations.

- Transformations can involve (invertible) neural networks to make them very expressive.
- The loss is the negative log-likelihood over the training set.

$$\mathcal{L} = -\frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \log p_{\theta}(\mathbf{x})$$

If training is successful, the learned likelihood == the true data likelihood == a DDL.

Flow-based generative models

BUT, transformations must

- be easily invertible,
- have an easy-to-compute Jacobian determinant (scales as N^3),

which limits their expressivity.

Different tricks in the literature:

- Restrict the form of the transformation to exploit identities
- Make Jacobian triangular by making transformations auto-regressive or splitting up dimensions and applying affine transformations

Ideally also want quick density estimation **and** sampling.

Flow-based generative models

Uncertainty Quantification with Generative Models

Vanessa Böhm, François Lanusse, Uroš Seljak

Berkeley Center for Cosmological Physics, Department of Physics
University of California, Berkeley, California, USA

Lawrence Berkeley National Laboratory, Cyclotron Rd, Berkeley, California, USA

Abstract

We develop a generative model-based approach to Bayesian inverse problems, such as image reconstruction from noisy and incomplete images. Our framework addresses two common challenges of Bayesian reconstructions: 1) It makes use of complex, data-driven priors that comprise all available information about the uncorrupted data distribution. 2) It enables computationally tractable uncertainty quantification in the form of posterior analysis in latent and data space. The method is very efficient in that the generative model only has to be trained once on an uncorrupted data set, after that, the procedure can be used for arbitrary corruption types.

arXiv:1910.10046

Fast likelihood-free cosmology with neural density estimators and active learning

Justin Alsing,^{1,2,3*} Tom Charnock⁴, Stephen Feeney² and Benjamin Wandelt^{2,5}

¹Oskar Klein Centre for Cosmoparticle Physics, Stockholm University, Stockholm SE-106 91, Sweden

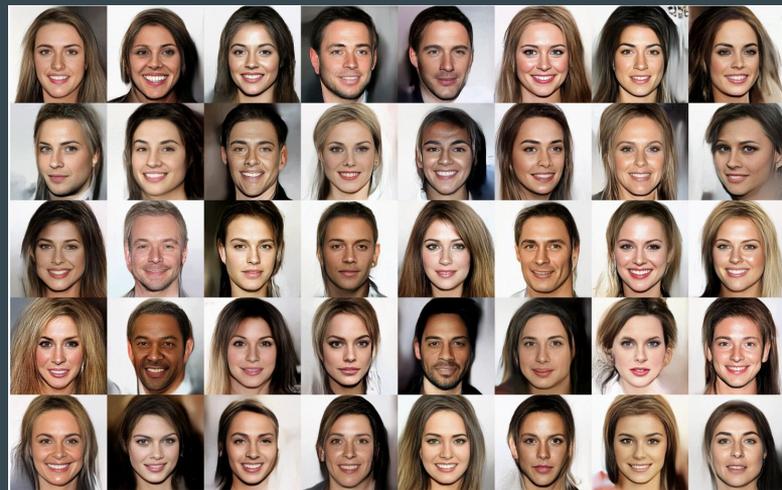
²Center for Computational Astrophysics, Flatiron Institute, 162 5th Ave, New York City, NY 10010, USA

³Imperial Centre for Inference and Cosmology, Department of Physics, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK

⁴Sorbonne Université, CNRS, UMR 7095, Institut d'Astrophysique de Paris, 98 bis bd Arago, 75014 Paris, France

⁵Sorbonne Université, Institut Lagrange de Paris (ILP), 98bis boulevard Arago, F-75014 Paris, France

arXiv:1903.00007



Glow (Kingma & Dhariwal 2018)

Constraining the Reionization History using Bayesian Normalizing Flows

Héctor J. Hortúa[Ⓧ], Luigi Malagò, Riccardo Volpi[Ⓧ]

Machine Learning and Optimization Group,
Romanian Institute of Science and Technology (RIST),
Cluj-Napoca, Romania

arXiv:2005.07694

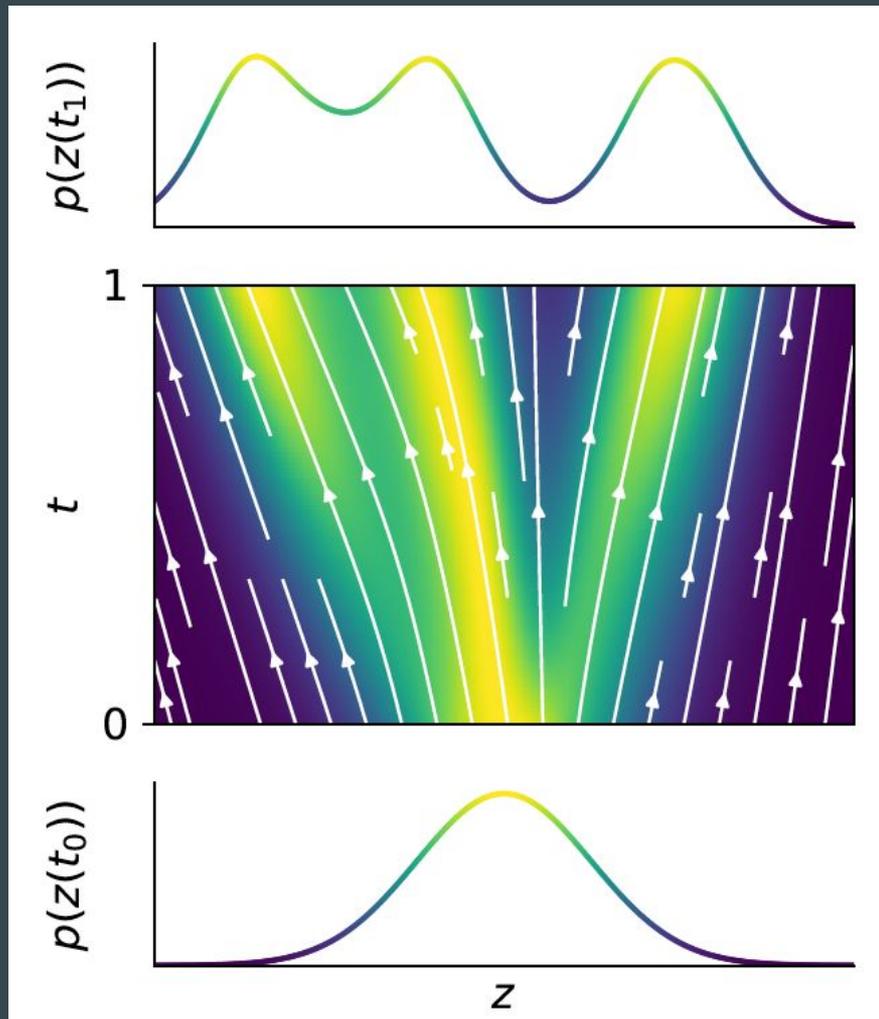
FFJORD (Grathwohl+ 2018)

Transformation from prior to data is seen as evolution in time.

$$\frac{\partial \mathbf{z}(t)}{\partial t} = f(\mathbf{z}(t), t; \theta)$$

$$\frac{\partial \log p(\mathbf{z}(t))}{\partial t} = -\text{Tr} \left(\frac{\partial f}{\partial \mathbf{z}(t)} \right)$$

$$\log p(\mathbf{z}(t_1)) = \log p(\mathbf{z}(t_0)) - \int_{t_0}^{t_1} \text{Tr} \left(\frac{\partial f}{\partial \mathbf{z}(t)} \right) dt.$$



FFJORD (Grathwohl+ 2018)

Transformation from prior to data is seen as evolution in time.

neural
network

$$\frac{\partial \mathbf{z}(t)}{\partial t} = f(\mathbf{z}(t), t; \theta)$$

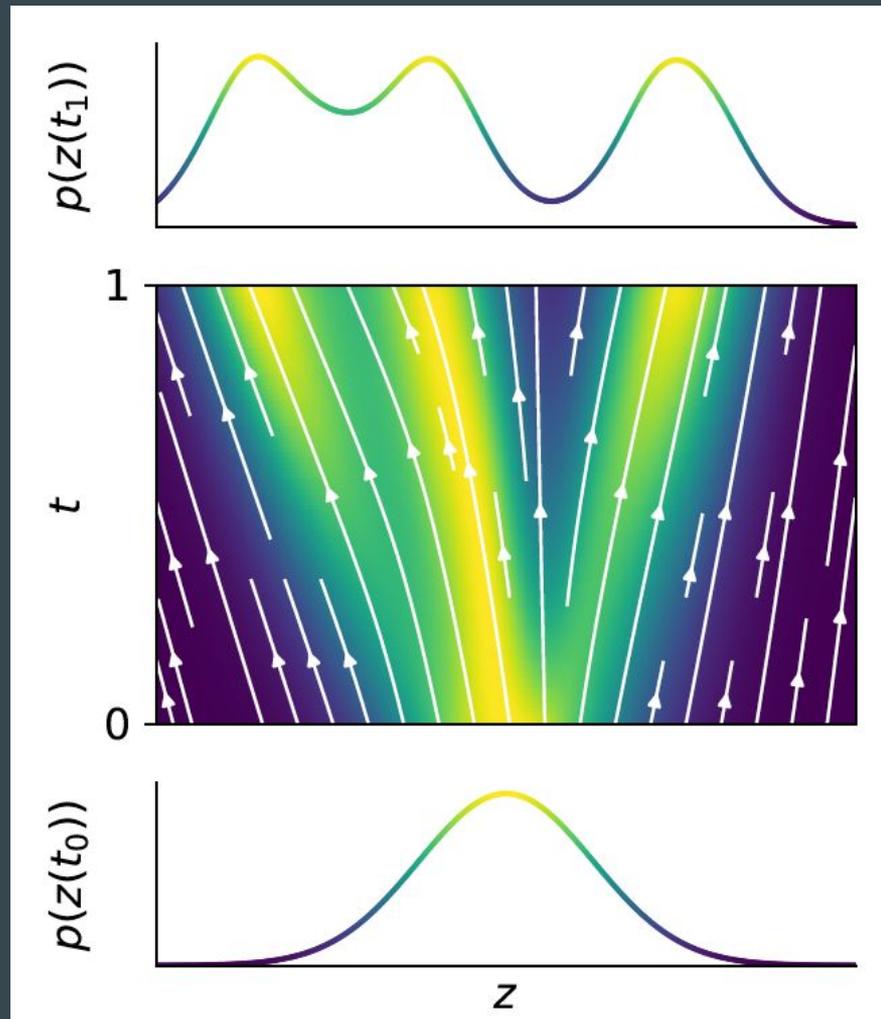
trace instead of
logdet

$$\frac{\partial \log p(\mathbf{z}(t))}{\partial t} = -\text{Tr} \left(\frac{\partial f}{\partial \mathbf{z}(t)} \right)$$

$$\log p(\mathbf{z}(t_1)) = \log p(\mathbf{z}(t_0)) - \int_{t_0}^{t_1} \text{Tr} \left(\frac{\partial f}{\partial \mathbf{z}(t)} \right) dt.$$

data

prior (IC)



FFJORD (Grathwohl+ 2018)

Transformation from prior to data is seen as evolution in time.

neural
network

$$\frac{\partial \mathbf{z}(t)}{\partial t} = f(\mathbf{z}(t), t; \theta)$$

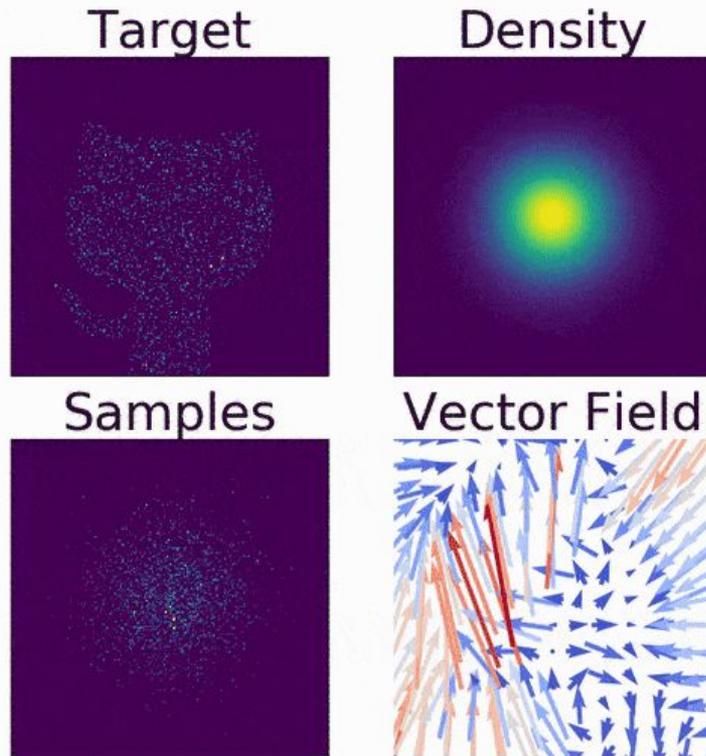
trace instead of
logdet

$$\frac{\partial \log p(\mathbf{z}(t))}{\partial t} = -\text{Tr} \left(\frac{\partial f}{\partial \mathbf{z}(t)} \right)$$

$$\log p(\mathbf{z}(t_1)) = \log p(\mathbf{z}(t_0)) - \int_{t_0}^{t_1} \text{Tr} \left(\frac{\partial f}{\partial \mathbf{z}(t)} \right) dt.$$

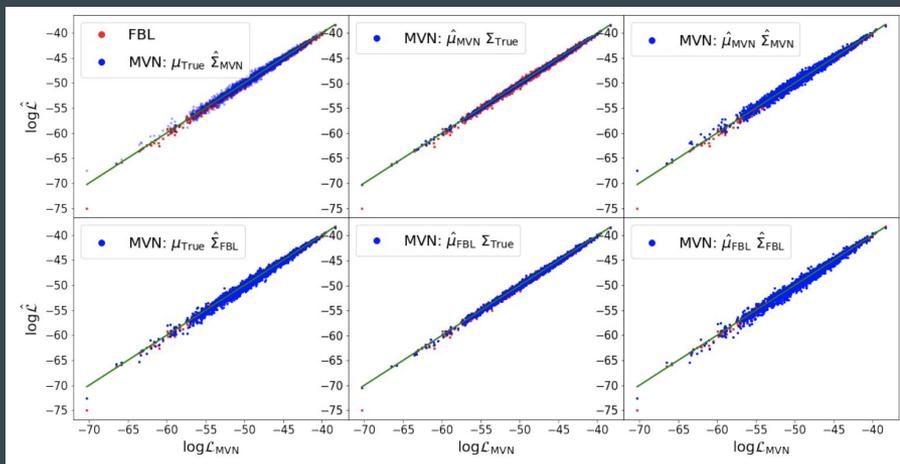
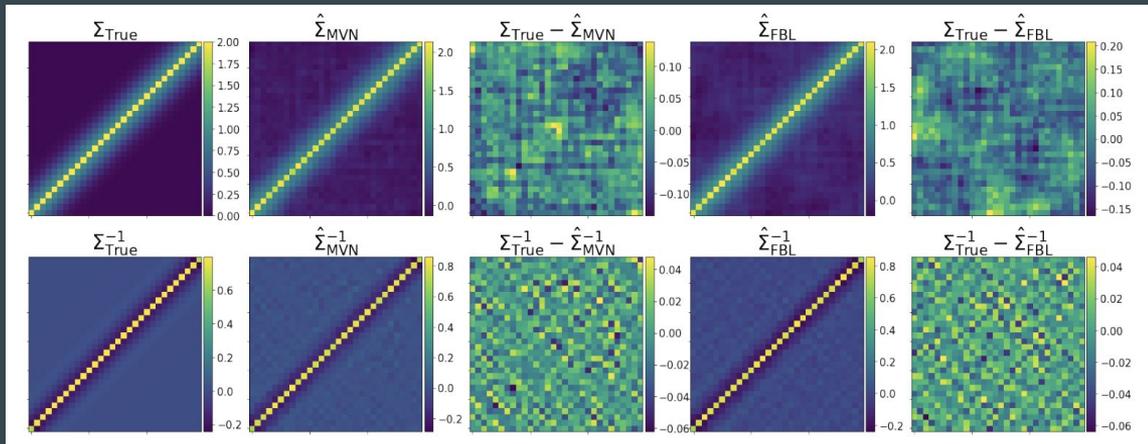
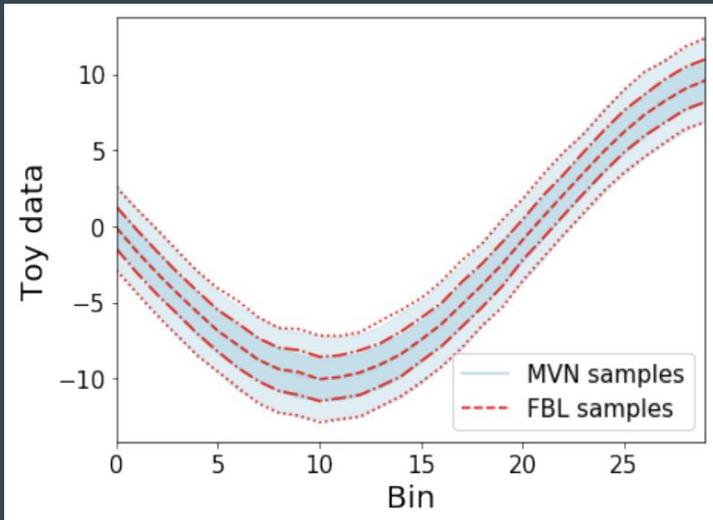
data

prior (IC)



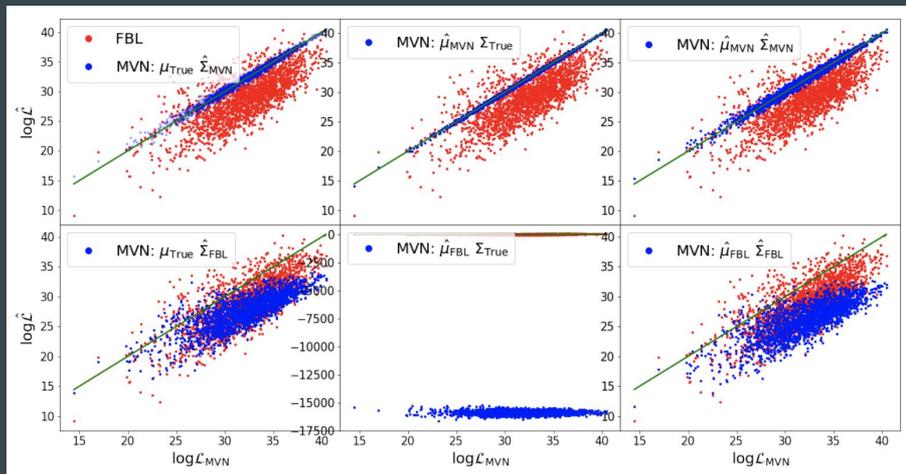
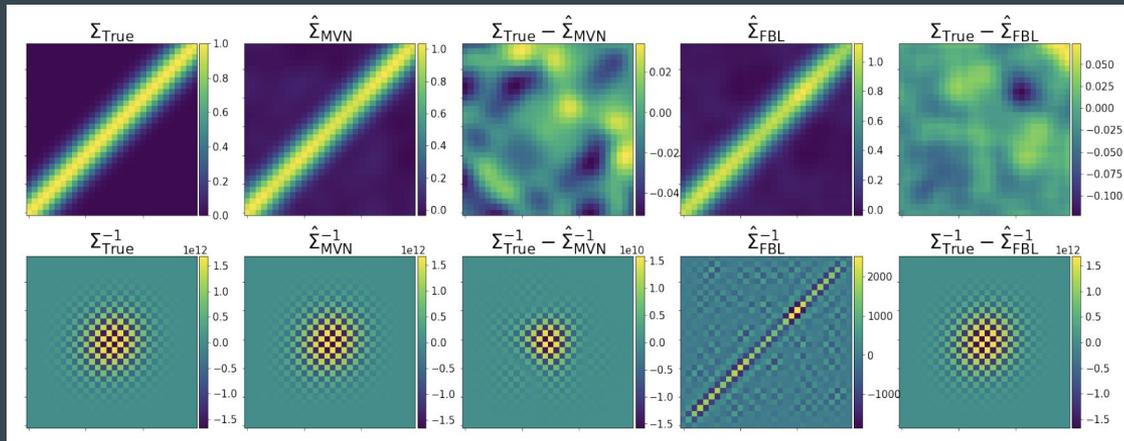
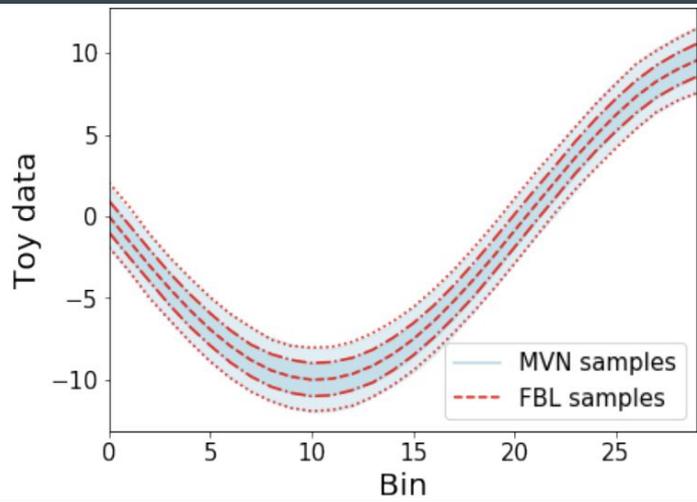
Samples vs likelihood quality

Non-singular covariance



Samples vs likelihood quality

Singular covariance



Quantifying non-Gaussianity in a dataset

We propose identifying non-Gaussianities (NG) in three ways:

1. *t*-statistic of skewness and excess kurtosis for every bin in the data

$$t = \frac{\hat{\beta} - \beta_{\text{null}}}{\text{SE}(\hat{\beta})}$$

0 for a Gaussian



Quantifying non-Gaussianity in a dataset

We propose identifying non-Gaussianities (NG) in three ways:

2. **Transcovariance matrix (Sellentin+ 2018)**, which considers the non-Gaussianity of all pairs of data points

$$s_i^{u,v} = x_i^u + x_i^v$$

Should be equal for whitened Gaussian data

$$\frac{1}{b} \sum_{a=1}^b [\mathcal{K}(s_a^{u,v}) - \mathcal{N}(0, 2)]^2 \equiv S_{u,v}^+$$

Total non-Gaussian contamination for each bin

$$\epsilon_u^+ = \sum_{v \neq u} S_{u,v}^+$$

Quantifying non-Gaussianity in a dataset

We propose identifying non-Gaussianities (NG) in three ways:

3. **KL divergence of (the data w.r.t. a MVN) vs (MVN with itself)** (Hahn+ 2018)

$$D_{n,m}(p||q) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}$$


$$\hat{D}_{n,m}(p||q) = \frac{d}{n} \sum_{i=1}^n \log \frac{\nu_k(i)}{\rho_k(i)} + \log \frac{m}{n-1}$$

Unbiased kNN estimator (Wang+ 2009)

Quantifying non-Gaussianity in a dataset

Going forward we will:

1. Apply these three tests to a mock dataset to look at the different ways in which NG can manifest themselves.
2. Generate samples from the three DDLs to assess whether each likelihood has successfully captured the NGs.

Weak gravitational lensing

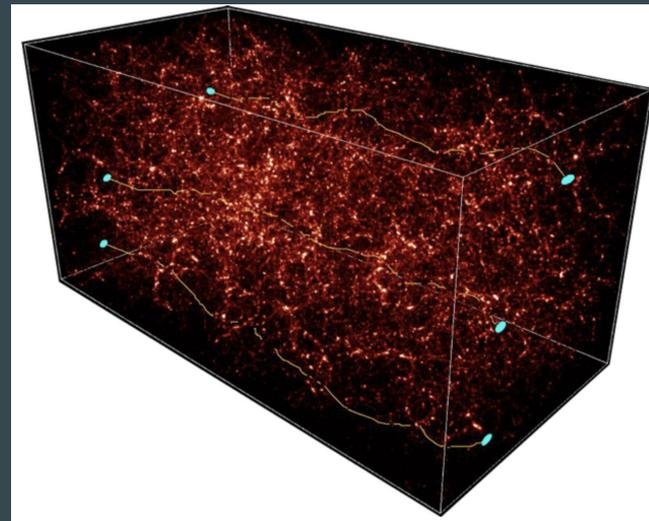
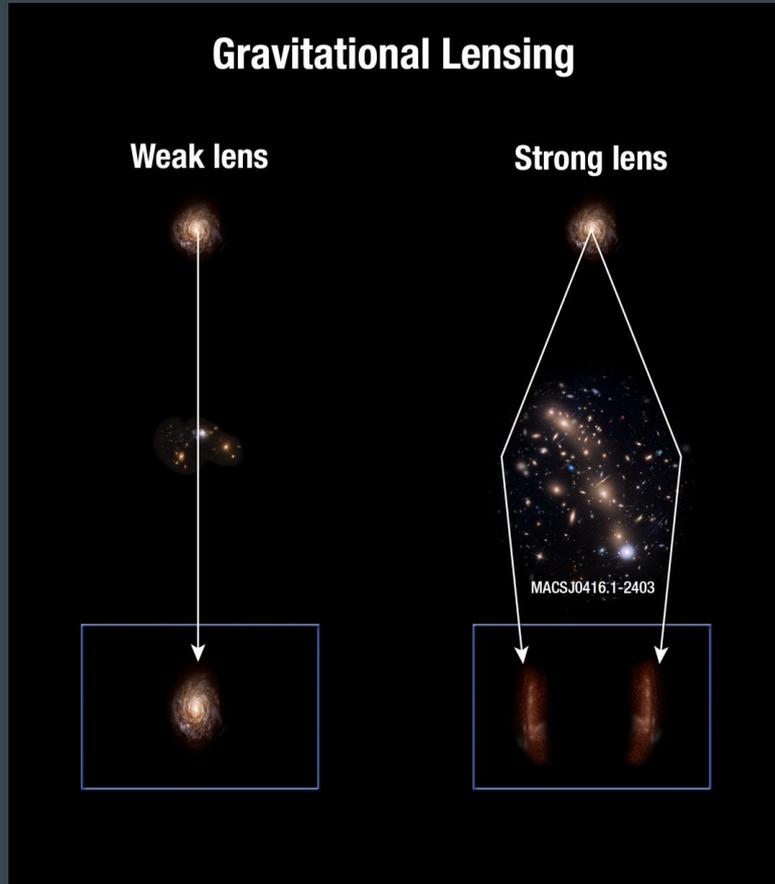
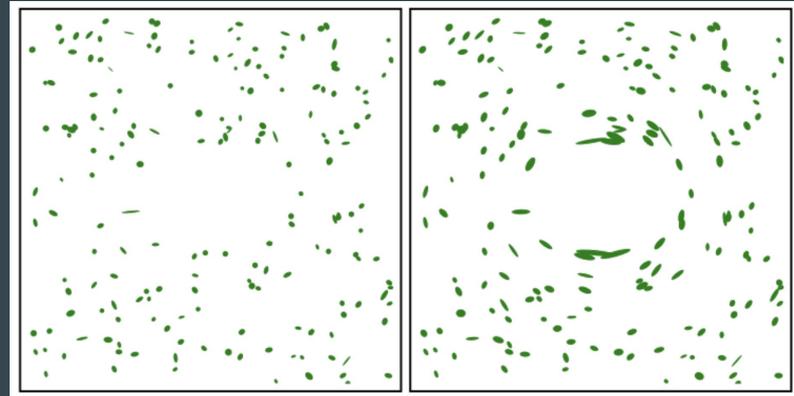
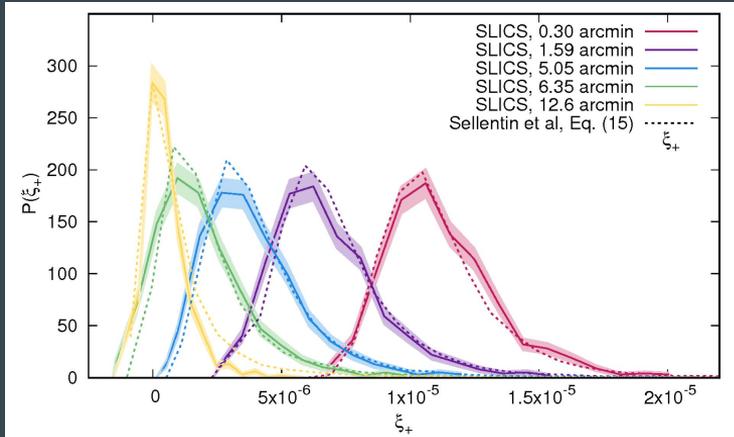


Figure credit: HETDEX

Statistical correlations in the shapes of millions of galaxies



Non-Gaussianity in weak lensing data



$$\mu = C_\ell$$

$$p = C_\ell \left(1 - \frac{2}{\nu}\right) \quad \text{for } \nu > 2$$

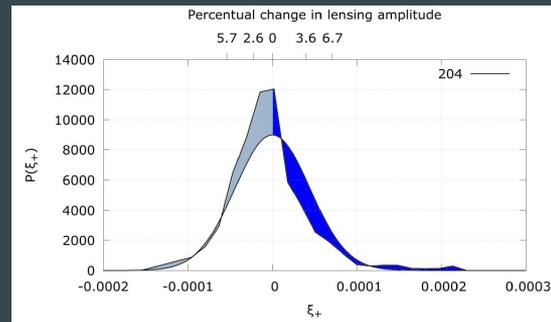
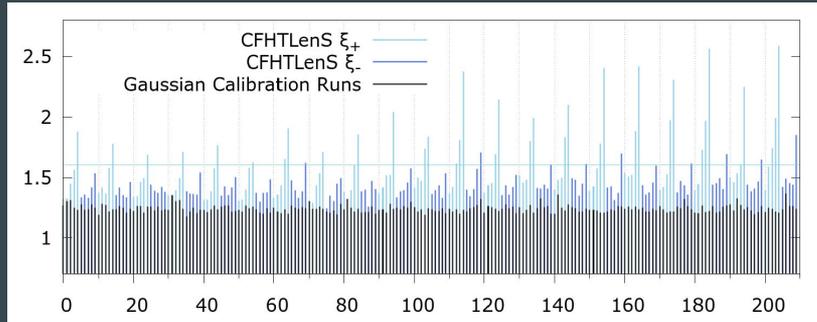
$$v = \frac{2C_\ell^2}{\nu}$$

Weak lensing amplitude

$$\propto$$

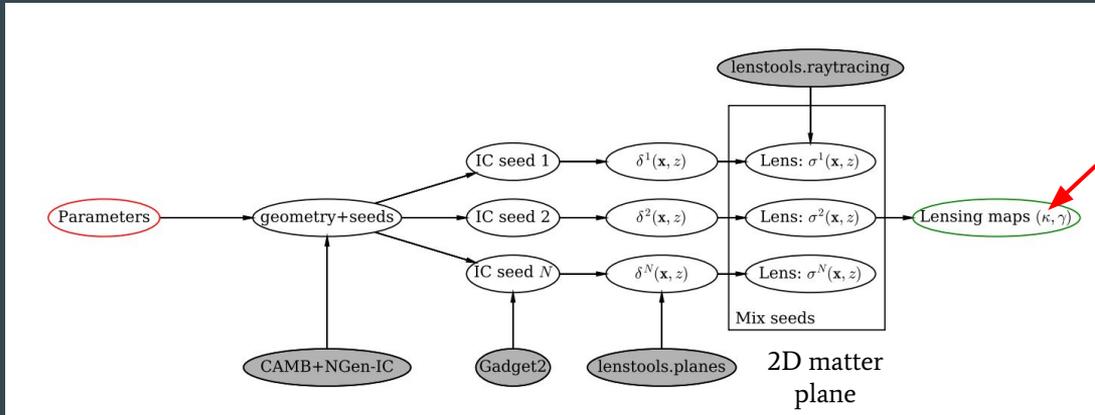
$$S_8 = \sigma_8 \sqrt{\Omega_m}$$

Shear correlation function non-Gaussianity (Sellentin+ 2018, 1 & 2)



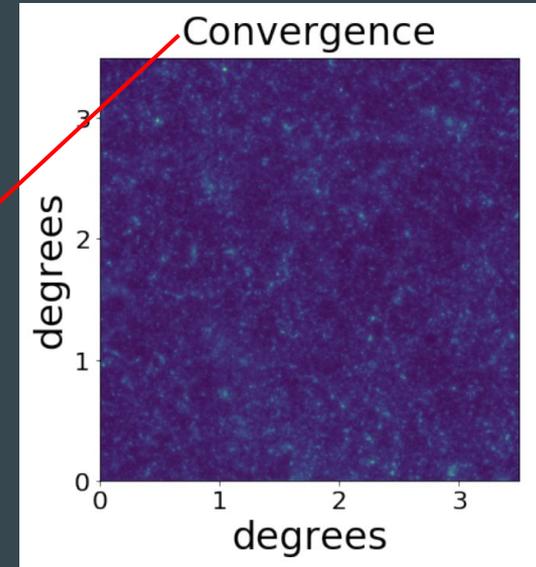
Simulated weak lensing data

Simulated 75,000 mock convergence maps using LensTools (Petri 2016)



Petri (2016)

3D matter
distribution
(N -body sim)

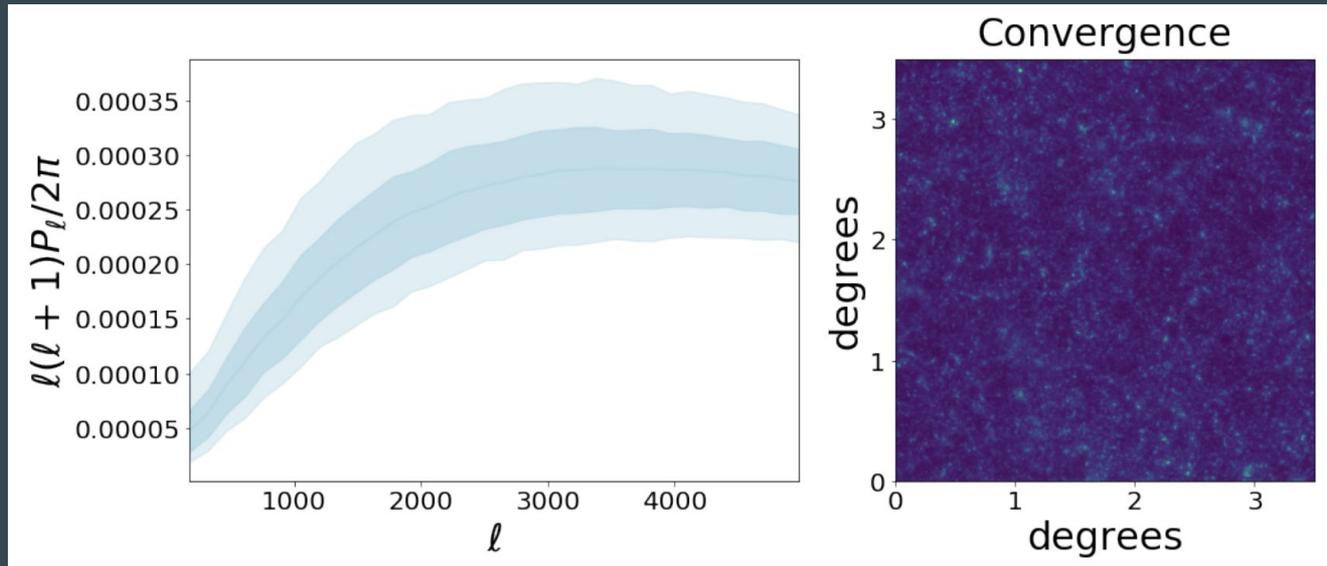


Diaz Rivero & Dvorkin
(2020)

Simulated weak lensing data

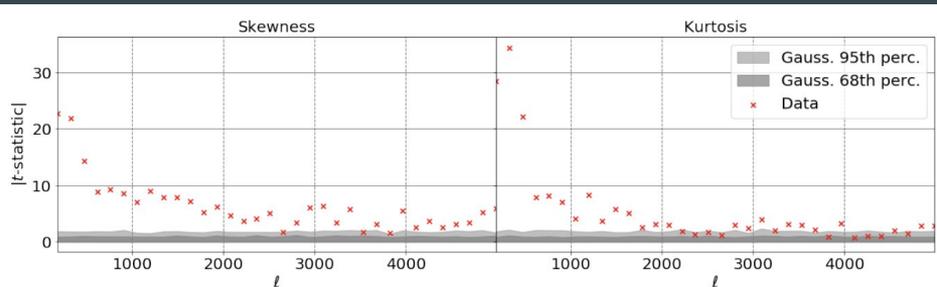
Simulated 75,000 mock convergence maps using LensTools (Petri 2016) and calculated the weak lensing convergence power spectrum:

$$P_\ell = \frac{1}{2\ell + 1} \sum_{m=-\ell}^{\ell} |a_{\ell,m}|^2$$

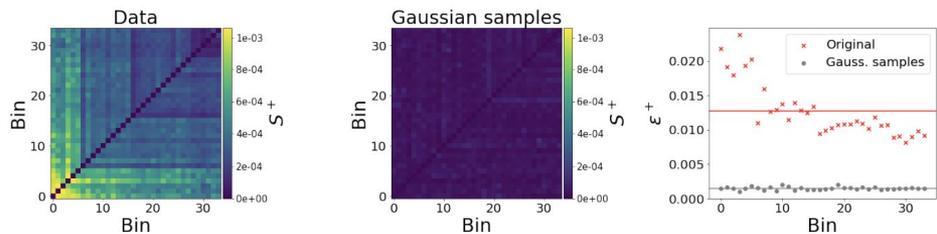


NG in the weak lensing convergence power spectrum

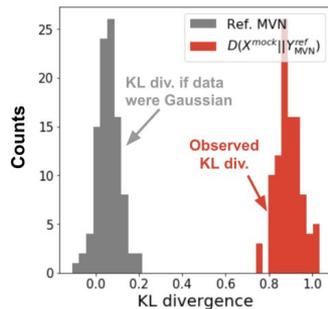
Test 1: t-stat of skewness and kurtosis



Test 2: transcovariance matrix

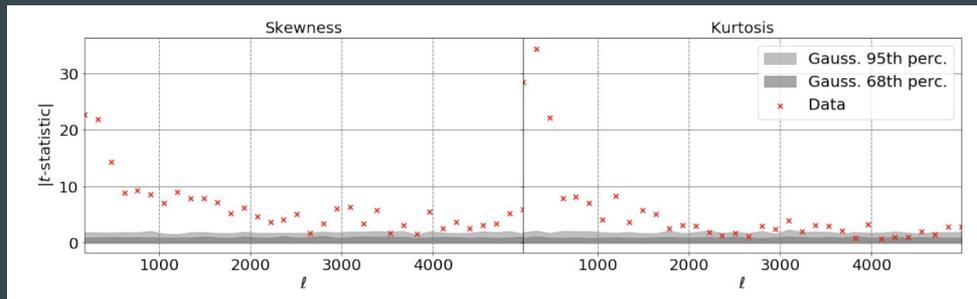


Test 3: KL divergence

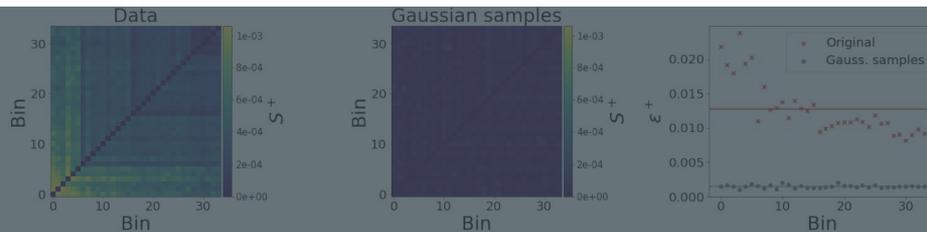


NG in the weak lensing convergence power spectrum

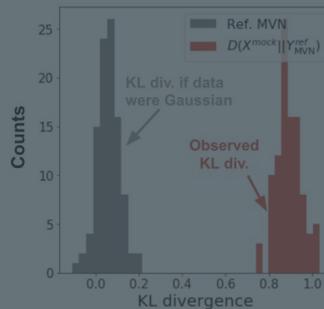
Test 1: t-stat of skewness and kurtosis



Test 2: transcovariance matrix

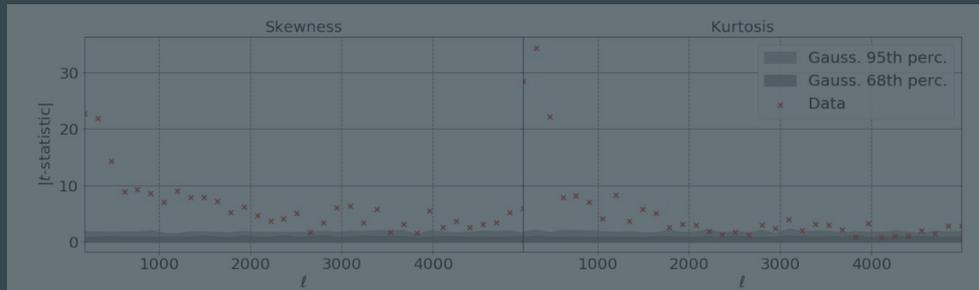


Test 3: KL divergence

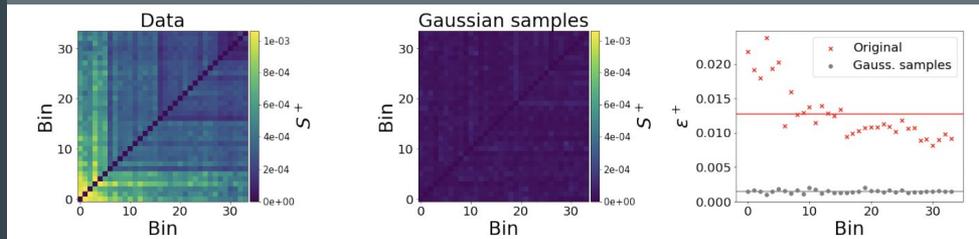


NG in the weak lensing convergence power spectrum

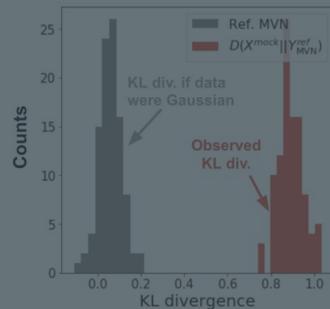
Test 1: t-stat of skewness and kurtosis



Test 2: transcovariance matrix



Test 3: KL divergence

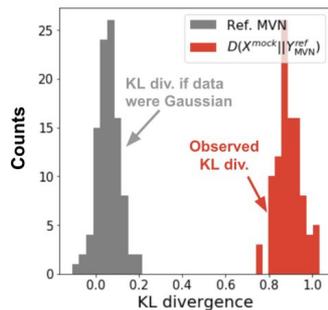
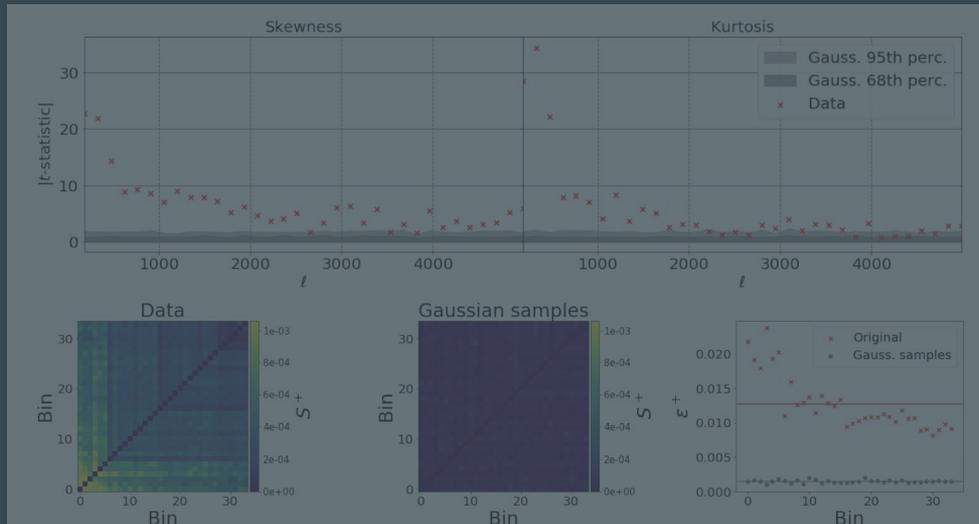


NG in the weak lensing convergence power spectrum

Test 1: t-stat of skewness and kurtosis

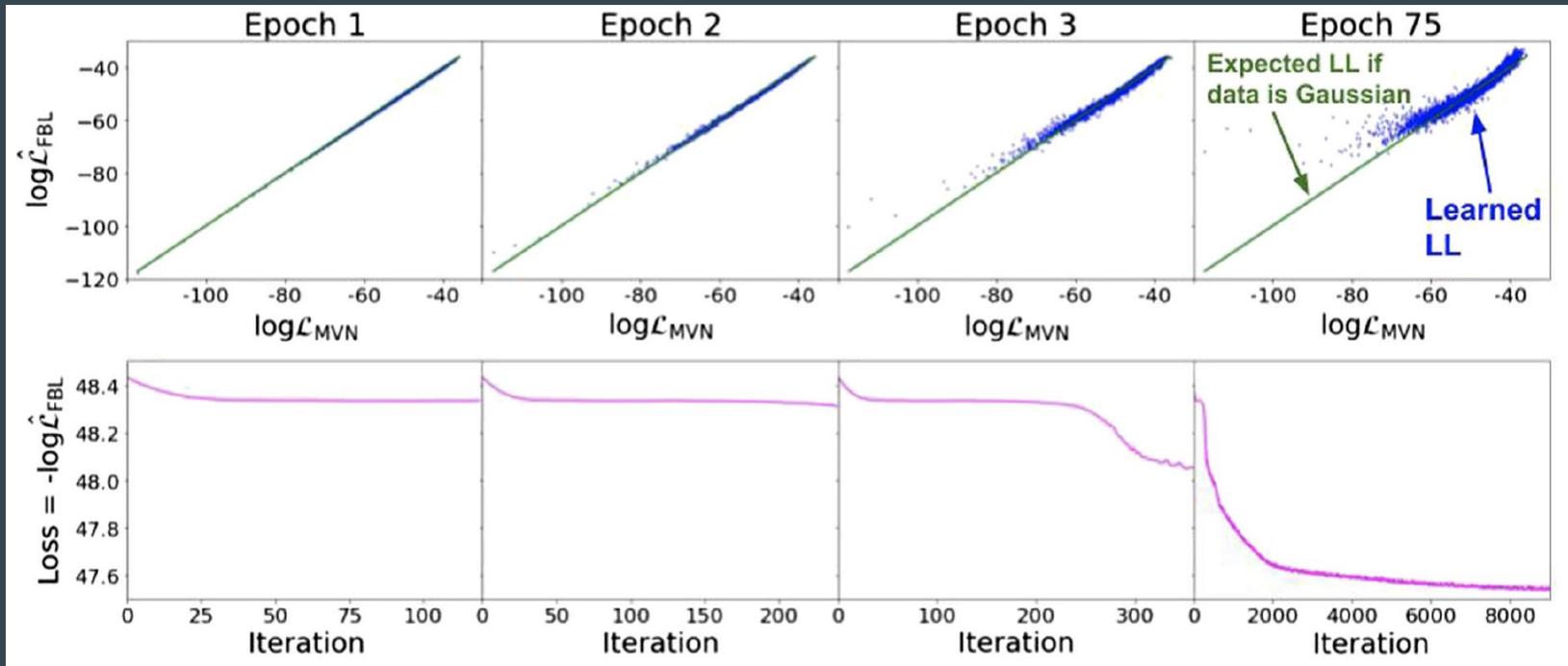
Test 2: transcovariance matrix

Test 3: KL divergence

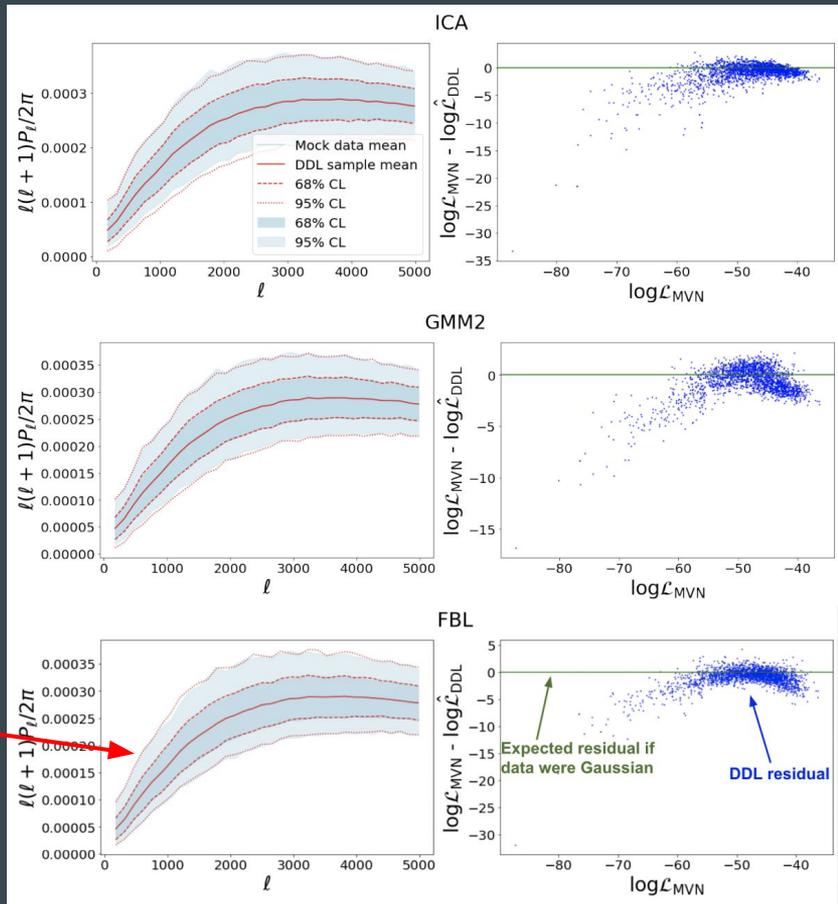


FBL for the convergence power spectrum

Training the model: whitened data, Adam opt., ELU activation function, 80/10/10 split



FBL for the convergence power spectrum

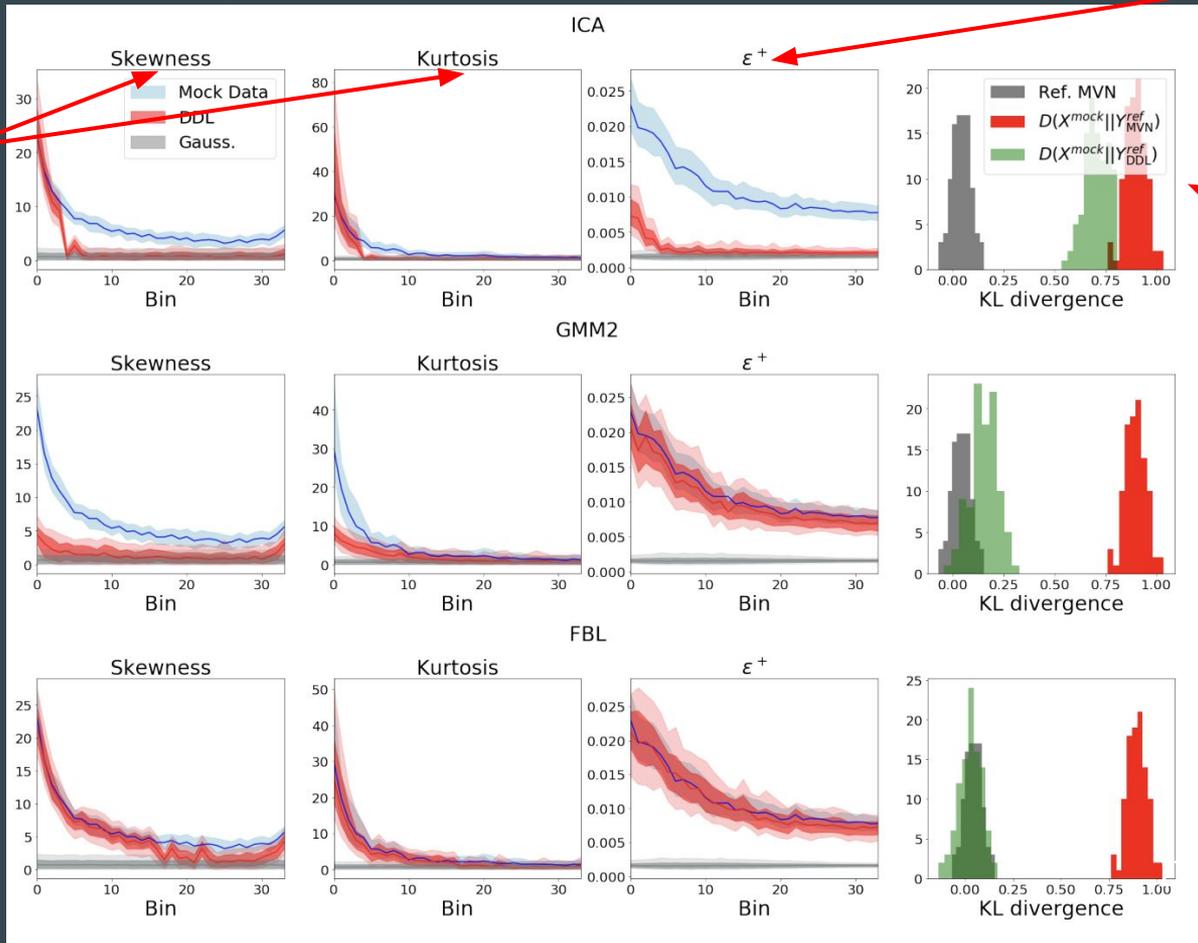


Samples drawn from the DDL



FBL for the convergence power spectrum

Test 1: t-stat of skewness and kurtosis

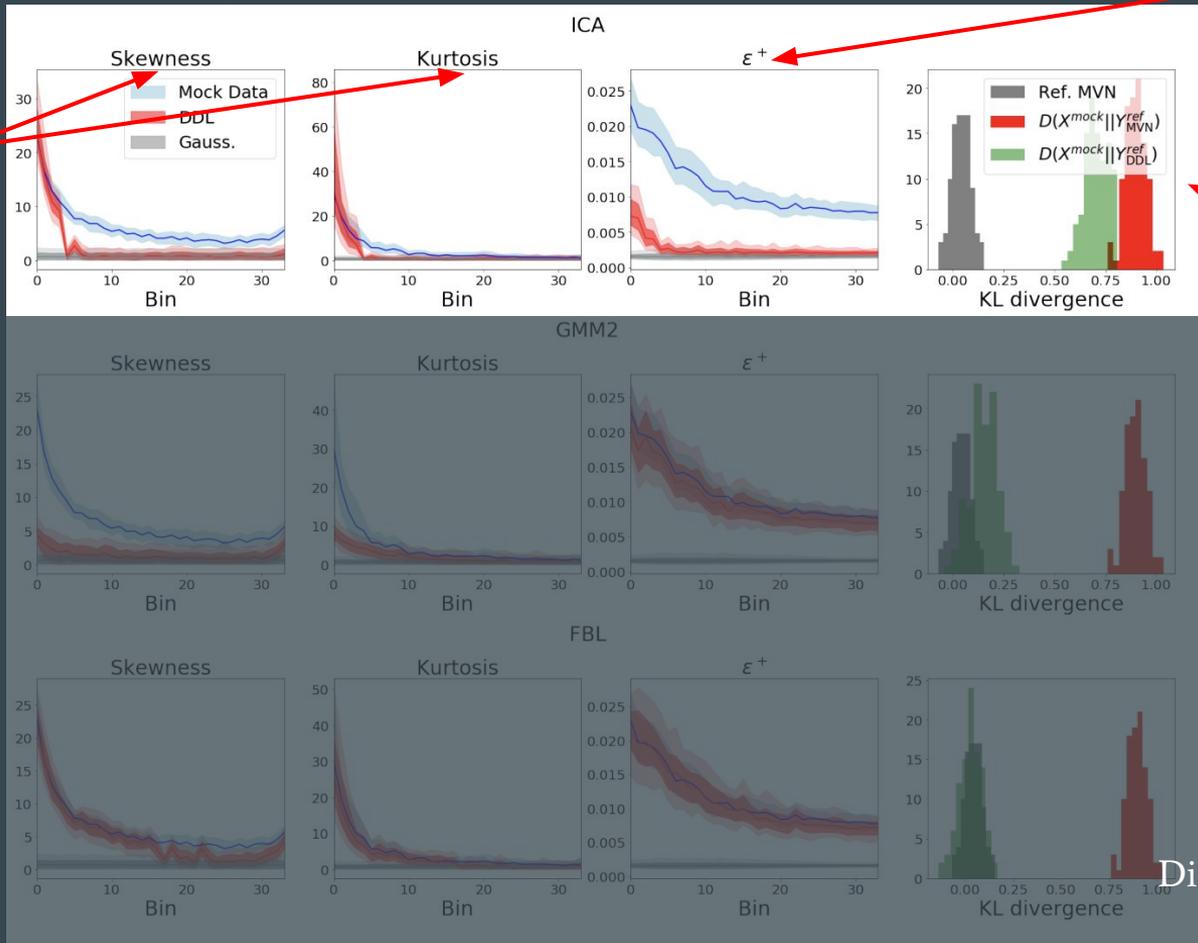


Test 2: transcovariance matrix

Test 3: KL divergence

FBL for the convergence power spectrum

Test 1: t-stat of skewness and kurtosis



Test 2: transcovariance matrix

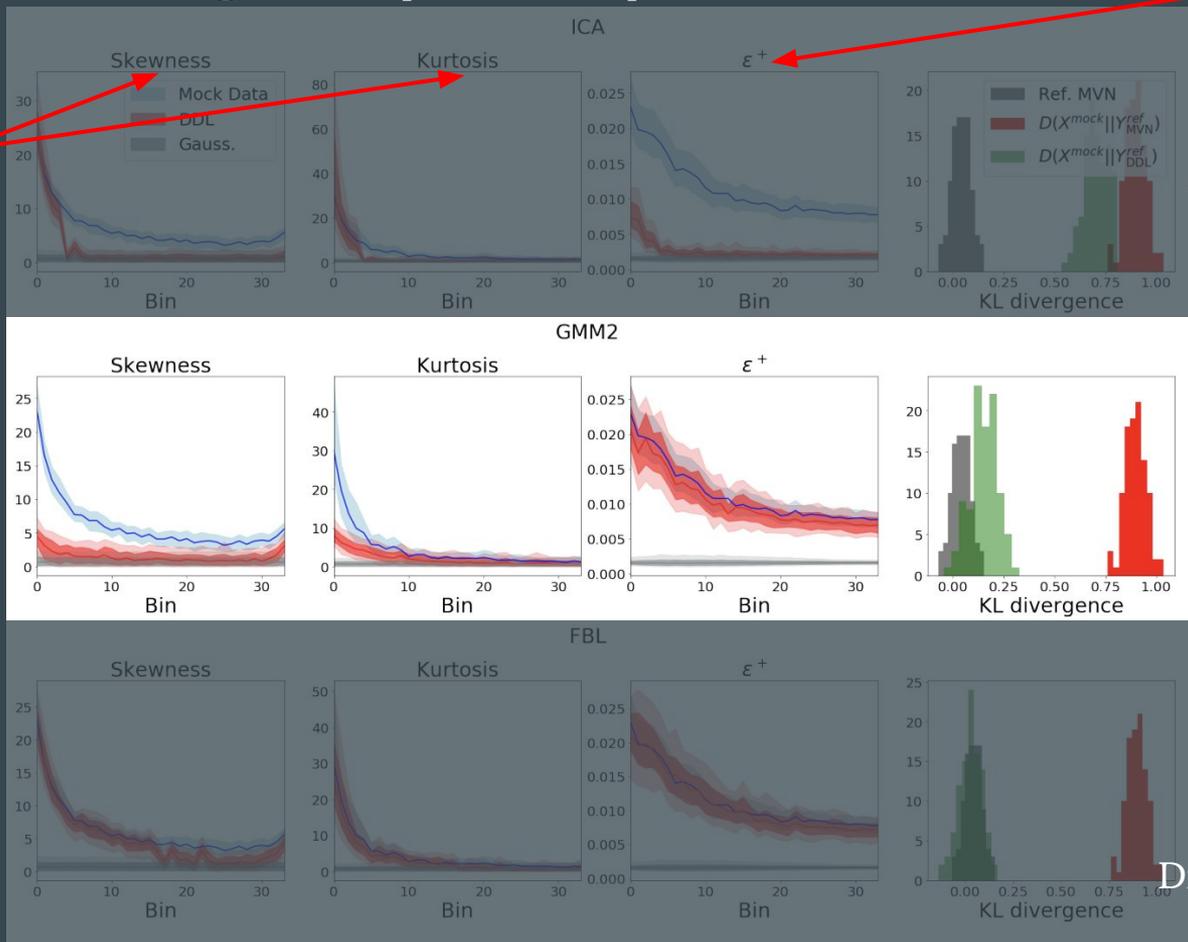
Test 3: KL divergence

FBL for the convergence power spectrum

Test 1: t-stat of skewness and kurtosis

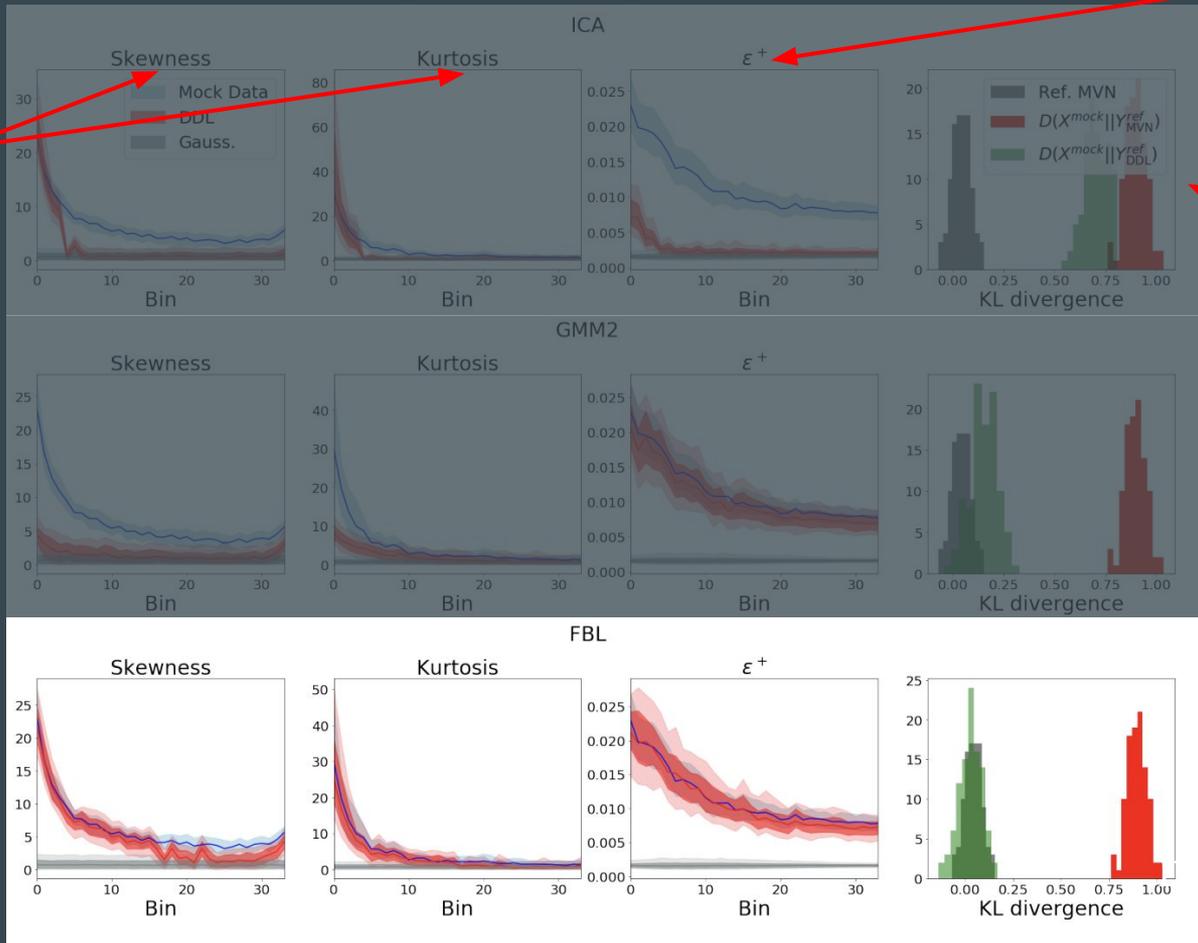
Test 2: transcovariance matrix

Test 3: KL divergence



FBL for the convergence power spectrum

Test 1: t-stat of skewness and kurtosis



Test 2: transcovariance matrix

Test 3: KL divergence

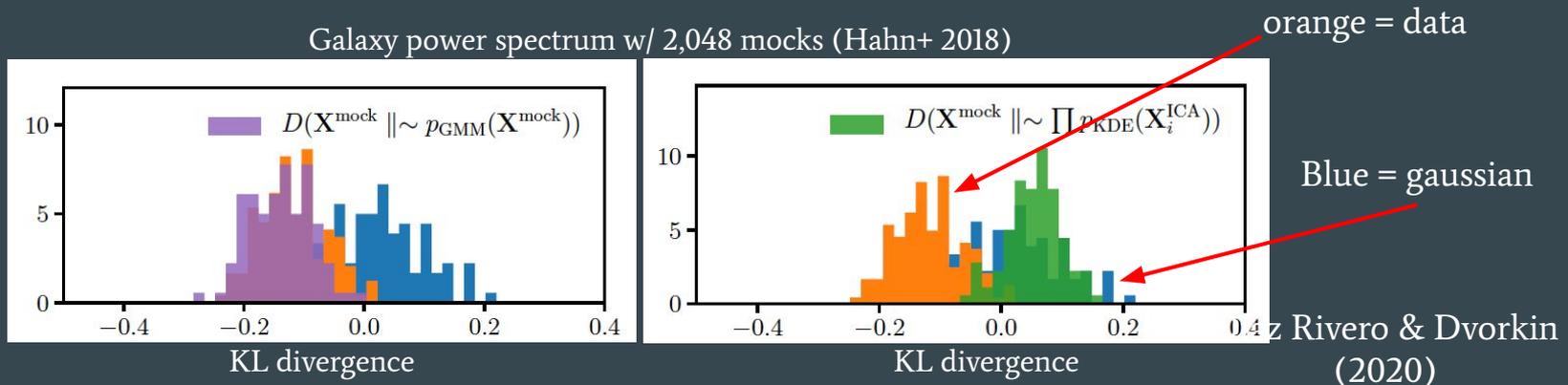
Implications

For our mock weak lensing data, GMM and ICA fail at capturing different NG, while the FBL does much better.

Data volume is not the only thing that determines the success/failure of a DDL: some **understanding of the NG** present in the data is crucial to select the right model.

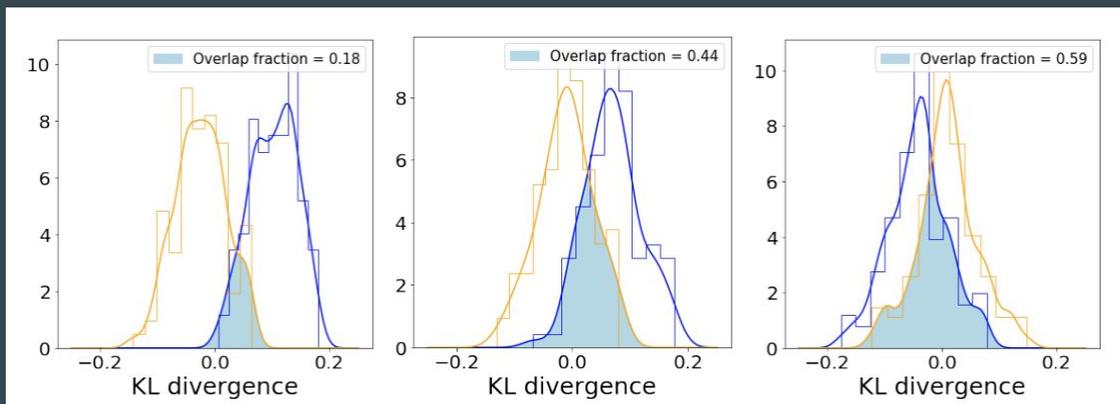
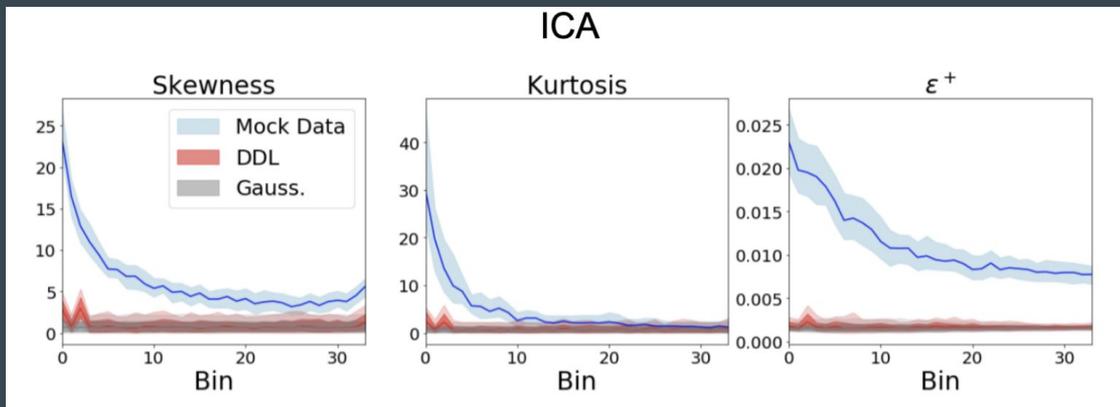
- ◆ E.g. ICA inadequate for NGs across bins

FBL flexibility can preclude them from a trial-and-error procedure that other DDLs can require.



Implications

But **data volume** obviously matters too! **With 2,048 mocks...**



Implications

WL in particular is interesting because:

- Seems to have some **significant non-Gaussianities**, even on scales where cosmic variance doesn't dominate (see also Sellentin+ 2016, 2018 1 & 2).
- Some WL works (inadvertently) **Gaussianize the data** (e.g. combining bins) before inferring parameters, potentially destroying useful information, and conclude NG doesn't shift parameters (Lin+ 2019, Taylor+ 2019, Alsing+ 2019).

the PDF). We apply the appropriate linear transformation to modify the covariance matrices computed in Section IV to account for the final binning choice. As an unfortunate byproduct of this need to “Gaussianize” the likelihood, the power of the ACT PDF to simultaneously constrain σ_8 and P_0 is substantially weakened, simply because the shape of the PDF is not as well constrained when combining so many smaller bins into a single larger bin. A clear goal for future PDF analyses is to implement a more sophisticated, non-Gaussian likelihood function, allowing the full use of the constraining power in the PDF.

ACT Thermal SZ
one-point PDF
(Hill+, 2015)

- **Shortcomings of ICA** in addressing pairwise non-Gaussian correlations in WL data: works have used ICA dimensionality reduction before inferring parameters from weak lensing data and concluded NG don't impact parameter constraints considerably (Gupta+, 2018).

Questions?