Likelihood Non-Gaussianity in Large Scale Structure Analyses

ChangHoon Hahn, et al.

Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley CA 94720, USA

`changhoon.hahn@lbl.gov`

ABSTRACT

Standard present day large-scale structure (LSS) analyses make a major assumption in their Bayesian parameter inference — the likelihood has a Gaussian functional-form. We investigate the impact of this assumption on two recent LSS analyses: the Beutler et al. (2017) power spectrum multipole (P_ℓ) analysis and the Sinha et al. (2017) group multiplicity function (ζ) analysis. Using non-parametric divergence estimators on mock catalogs originally constructed for covariance matrix estimation we identify significant non-Gaussianity in both the P_ℓ and ζ likelihoods. We then use Gaussian mixture density estimation and Independent Component Analysis to construct likelihood estimates that approximate the true likelihood better than the Gaussian *pseudo*-likelihood. Using these likelihood estimates, we derive the true posterior probability distribution of the B2017 and S2017 parameters. Likelihood non-Gaussianity shifts the $f\sigma_8$ constraint by 0.44σ, but otherwise, does not significantly impact the overall parameter constraints of B2017. For the ζ analysis, using the pseudo-likelihood significantly underestimates the uncertainties and biases the constraints of S2017 halo occupation parameters: $\log M_{\rm min}$, $\log M_1$ and α. The divergence and likelihood estimation methods we present provides a straightforward framework for quantifying the impact of likelihood non-Gaussianity and deriving more accurate parameter constraints.

Subject headings: methods: statistical — galaxies: statistics — methods: data analysis — cosmology: observations — cosmological parameters — large-scale structure of universe

1. Introduction

Bayesian parameter inference provides the standard framework for deriving cosmological parameters from observation in large scale structure (LSS) studies and cosmology in general. Using Bayes' rule,

$$p(\theta \,|\, x) \propto p(x \,|\, \theta) \, p(\theta) \tag{1}$$

the posterior probabilisty distributions of cosmological parameters can be derived from observed measurements such as the galaxy power spectrum. All that's required is the prior distribution of the parameters, $p(\theta)$, and the likelihood, $p(x\,|\,\theta)$ — probability of the data (obsevation) given the theoretical model. Priors are selected in analyses; so parameter inference ultimately boils down to evaluating the likelihood. It consequently goes without say that our analyses can only lead to unbiased constraints if the our likelihood is correct.

In present day LSS analyses, two major assumptions go into evaluating the likelihood. First, the likelihood is assumed to have a Gaussian functional form:

$$p(x\,|\,\theta) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(C(\theta))}}\,\exp\left[\frac{1}{2}(x - m(\theta))^t C(\theta)^{-1}(x - m(\theta))\right] \tag{2}$$

where d is the dimension of the data vector x, $m(\theta)$ is the theoretical model, and C is the covariance matrix. Second, the covariance matrix used in evaluating the Gaussian pseudo-likelihood is assumed to be independent of cosmology. The covariance matrix is evaluated only at a selected fiducial cosmology and used throughout the analysis. In principle, the covariance matrix depends on θ, and the dependence has been shown to have a significant effect on parameter constraints (*e.g.* Eifler et al. 2009; Morrison & Schneider 2013; White & Padmanabhan 2015). In this paper, we focus on the, first, Gaussian pseudo-likelihood assumption. Even if analyses use covariance matrices that account for non-Gaussiance covariance (*e.g.* Scoccimarro et al. 1999; Hu & White 2001; O'Connell et al. 2016), they still use a Gaussian functional form (Eq. 2) for the likelihood. They therefore still employ a Gaussian *pseudo*-likelihood. We will test the assumption and quantify the impact of it on cosmological parameter constraints.

The motivation for the Gaussian pseudo-likelihood ultimately stems from the Central Limit Theorem. Take the power spectrum of the density field for example. On large scales, the density field is approximately a Gaussian random field. Then the power spectrum of a specific Fourier mode would follow a chi-squared distribution, *not* a Gaussian. However, with sufficiently many independent modes contributing, by the Central Limit Theorem the likelihood of the power spectrum would *approach* a Gaussian distribution. In practice, we expect the Gaussian assumption to fail in low signal-to-noise regimes. The assumption is also further invalidated by finite survey volume, shot noise, and systematic effects, which correlate different modes. The breakdown of Gaussianity is clearly illustrated in earlier surveys such as IRAS where limited survey volume and sparse sampling cause the probability distribution function of the galaxy power spectrum to deviate significantly from Gaussian (see Figure 9 in Scoccimarro 2000). Hartlap et al. (2009) and Sellentin & Heavens (2017) similiarly illustrate the breakdown of the Gaussian likelihood assumption for the cosmic shear correlation function likelihood.

Even if the likelihood is Gaussian, Sellentin & Heavens (2016) argue that since an estimate of the covariance matrix is used for the likelihood, the true covariance matrix must

be marginalized for accurate parameter inference. This leads to a likelihood that is no longer Gaussian, but rather a multivariate t-distribution. Fortunately, the Gaussian pseudo-likelihood assumption is not necessary for parameter inference. Outside of LSS, in CMB powerspecturm analyses for instance, the Planck collaboration uses a hybrid likelihood, which only assumes a Gaussian pseudo-likelihood for C_ℓ on small scales (Ade et al. 2014; Aghanim et al. 2016; see also Efstathiou 2004, 2006). On large scales (low-ℓ), the likelihood is instead computed directly in pixel-space and extensively validated. Testing for likelihood non-Gaussianity and non-Gaussian likelihoods in general are not currently part of standard practice in LSS studies. For more precise parameter constraints from LSS, however, analyses must go beyond the Gaussian pseudo-likelihood.

In this paper we investigate the impact of the likelihood Gaussianity assumption on the two recent LSS analyses of Beutler et al. (2017) (hereafter B2017) and Sinha et al. (2017) (hereafter S2017). B2017 analyzes the power spectrum multipoles (P_ℓ; monopole, quadrupole, and hexadecapole) to measure redshift-space distortions along with the Alcock-Paczynski effect and baryon acoustic oscillation scale. Meanwhile S2017 analyzes the group multiplicity function (ζ) in order to constrain parameters of the halo model. Using the B2017 and S2017 analyses, we show in this paper that the assumption of likelihood Gaussianity in LSS is not necessary. We will also show that the mock catalogs used in standard LSS analyses for covariance matrix estimation can be used to quantify the non-Gaussianity. More importantly we will use the mocks to estimate the "true" non-Gaussian likelihood.

We begin in Section 2 by describing the mock catalogs that we use throughout the paper, constructed originally for covariance matrix estimation in B2017 and S2017. Next in Section 3, we present non-parametric divergence estimators and quantify the non-Gaussianity of the P_ℓ and ζ likelihoods using them. Then in Section 4, we introduce two methods for estimating the "true" likelihood using the mock catalogs. We then use the likelihood estimates to quantify the impact of likelihood non-Gaussianity on the posterior parameter constraints of B2017 and S2017in Section 5. We discuss and conclude the paper in Section 6.

2. Mock Catalogs

Mock catalogs play an indispensable role in standard cosmological anslyses of LSS studies. They're used for testing analysis pipelines (Beutler et al. 2017; Grieb et al. 2017; Tinker & et al. in preparation), testing the effect of systematics (Guo et al. 2012; Vargas-Magaña et al. 2014; Hahn et al. 2017a; Pinol et al. 2017; Ross et al. 2017), and, most relevantly for this paper, estimating the covariance matrix (Parkinson et al. 2012; Kazin et al. 2014; Grieb et al. 2017; Alam et al. 2017; Beutler et al. 2017; Sinha et al. 2017). In fact, nearly all current state-of-the-art LSS analyses use covariance matrices estimated from mocks to evaluate the likelihood for parameter inference.

While some argue for analytic estimates of the covariance matrix (e.g. Mohammed et al.

2017) or estimates directly from data by subsampling (e.g. Norberg et al. 2009), covariance matrices from mocks have a number of advantages. Mocks allow us to incorporate detailed systematic errors present in the data and also variance beyond the volume of the data. Even for analytic estimates, a large ensemble of mocks are crucial for validation (*e.g.* Slepian et al. 2017). Moreover, as we show later in this paper, mocks present an additional advantage: they allow us to quantify the non-Gaussianity of the likelihood and more accurately estimate the true likelihood distribution.

In this paper, we focus on two LSS analyses: the power spectrum multipole (P_ℓ) analysis of B2017 and group multiplicity function (ζ) analysis of S2017. Throughout the paper we will make extensive use of the mock catalogs used in these analyses. In this section, we give a brief description of these mocks and how the observables used in the analysis — $P_\ell(k)$ and $\zeta(N)$ — are calculated from them. Afterwards, we will describe how we compute the covariance matrix from the mocks and pre-process the mock observable data.

2.1. MultiDark-PATCHY Mock Catalog

B2017 use the MultiDark-PATCHY mock catalogs from Kitaura et al. (2016) mocks generated using the PATCHY code (Kitaura et al. 2014, 2015). They rely on large-scale density fields generated using augmented Lagrangian Perturbation Theory (ALPT; Kitaura & Heß 2013) on a mesh, which they then populate with galaxies based on a combined non-linear deterministic and stochastic biases. The mocks from the PATCHY code are then calibrated to reproduce the galaxy clustering in the high-fidelity BigMultiDark N-body simulation (Rodríguez-Torres et al. 2016; Klypin et al. 2016). Afterwards, the galaxies are assigned stellar masses using the HADRON code (Zhao et al. 2015). and the SUGAR code (Rodríguez-Torres et al. 2016) is applied to combine different boxes, incorporate selection effects and masking to produce mock light-cone galaxy catalogs. The statistics of the resulting mocks are then compared to observations and the process is iterated to reach desired accuracy. We refer readers to Kitaura et al. (2016) for further details.

In total, Kitaura et al. (2016) generated 12,228 mock light-cone galaxy catalogs for BOSS Data Release 12. In B2017, they use 2045 and 2048 for the northern galactic cap (NGC) and southern galactic cap (SGC) of the LOWZ+CMASS combined sample. B2017 excluded 3 mock realizations due to notable issues. These issues have since been addressed so in our analysis we use all 2048 mocks for both the NGC and SGC of the LOWZ+CMASS combined sample. In B2017, they conduct multiple analyses, some using only the power spectrum monopole and quadrupole and others using monopole, quadrupole, and hexadecapole. They also separately analyze three redshift bins: $0.2 < z < 0.5$, $0.4 < z < 0.6$, and $0.5 < z < 0.75$. In this paper we focus on one of these analyses: the analysis of the power spectrum monopole, quadrupole, and hexadecapole for the $0.2 < z < 0.5$ bin.

2.2.   Sinha et al. (2017) Mocks

The simulations used in the Sinha et al. (2017) analysis are from the Large Suite of Dark Matter Simulations project (LasDamas; McBride et al. 2009), which were designed to model galaxy samples from SDSS DR7. They use initial conditions derived from second order Lagrangian Perturbation Theory using the 2LTPIC code (Scoccimarro 1998; Crocce et al. 2006), which are then evolved using the N-body GADGET-2 code (Springel 2005). Halos are then identified from the dark matter distribution outputs using the ntropy − fofsv code (Gardner et al. 2007), which uses a friend-of-friends algorithm (FoF; Davis et al. 1985) with a linking length of 0.2 times the mean inter-particle separation. S2017 uses two configurations of the LasDamas simulations for the SDSS DR7 samples with absolute magnitude limits $M_r < -19$ and $M_r < -21$. The Consuelo simulation contains 1400^3 dark matter particles with mass of $1.87 \times 10^9 \, h^{-1} M_\odot$ in a cubic volume of $420 \, h^{-1} Mpc$ per side evolved from $z_{\rm init} = 99$. The Carmen simulation contains 1120^3 dark matter particles with mass of $4.938 \times 10^{10} \, h^{-1} M_\odot$ in a cubic volume of $1000 \, h^{-1} Mpc$ per side evolved from $z_{\rm init} = 49$.

The FoF halo catalogs are populated with galaxies using the 'Halo Occupation Distribution' (HOD) framework. The number, positions, and velocities of galaxies are described statistically by an HOD model. S2017 adopts the 'vanilla' HOD model of Zheng & Weinberg (2007), where the mean number of central and satellite galaxies are described by the halo mass and five HOD parameters: $M_{\rm min}, \sigma_{\log M}, M_0, M_1,$ and α. Lastly, once the simulation boxes are populated with galaxies, observational systematic effects are imposed. The peculiar velocities of galaxies are used to impose redshift-space distortions. Galaxies that lie outside the redshift limits or sky footprint of the SDSS sample are removed. For further details regarding the mocks, we refer readers to S2017.

To calculate their covariance matrix, S2017 produced 200 independent mock catalogs from 50 simulations using a single set of HOD model parameters. To take advantage of the methods we later present, we use a large number of mocks. Our methods rely on using the mocks to sample high dimensional distributions, so incorporating more mocks improves their accuracy. We utilize an addition 99 sets of HOD parameters of 200 mocks each. The HOD parameters are sampled from the MCMC chain used to produce the posterior in S2017. In total we use 20, 000 mocks. Among the multiple analyses in S2017, in this paper we focus on the GMF analysis of the SDSS DR7 $M_r < -19$ sample.

2.3.   Mock Observable $\mathbf{X}^{\rm mock}$ and Covariance Matrix $\mathbb{C}$

To get from the mock catalogs described above to the covariance matrices used in B2017 and S2017, the observables are measured for each mock in the *same* way as the observations. We briefly describe how $P_\ell(k)$ and $\zeta(N)$ and their covariance matrices are measured in B2017 and S2017. We then describe how we pre-process the mock observables for

the methods we describe in the next sections.

To measure the power spectrum multipoles of the BOSS DR12 galaxies and the MutliDark-PATCHY mocks (Section 2.1), B2017 uses a fast Fourier transform (FFT) based anisotropic power spectrum estimator based on Bianchi et al. (2015) and Scoccimarro (2015). This estimator estimates the monopole, quadrupole, and hexadecapole ($\ell = 0, 2, 4$) of the power spectrum using FFTs of the overdensity field multipoles for a given survey geometry. For further details on the estimator we refer readers to Section 3 of B2017. The power spectrum is computed in bins of $\Delta k = 0.01\,h\,\mathrm{Mpc}^{-1}$ over the range $k = 0.01-0.15\,h\,\mathrm{Mpc}^{-1}$ for $\ell = 0$, and 2 and $k = 0.01 - 0.10\,h\,\mathrm{Mpc}^{-1}$. for $\ell = 4$. From the $\vec{P}^{(n)} = \left[P_0^{(n)}(k), P_2^{(n)}(k), P_4^{(n)}(k) \right]$ of the MultiDark-PATCHY mocks, B2017 computes the (i, j) element of the covariance matrix of all multipoles as

$$\mathbb{C}_{i,j} = \frac{1}{N_{\mathrm{mock}} - 1} \sum_{n=1}^{N_{\mathrm{mock}}} \left[\vec{P}_i^{(n)} - \overline{P}_i \right] \times \left[\vec{P}_j^{(n)} - \overline{P}_j \right]. \tag{3}$$

$N_{\mathrm{mock}} = 2048$ is the number of mocks and $\overline{P}_i$ is the mean of the mock powerspectra: $\overline{P}_i = \frac{1}{N_{\mathrm{mock}}} \sum_{n=1}^{N_{\mathrm{mock}}} \vec{P}_i^{(n)}$. Since P_0 and P_2 each have 14 bins and P_4 has 9 bins, $\mathbb{C}$ is a 37×37 matrix. In this work, we compute the $P_\ell(k)$ using a similar FFT-based estimator of Hand et al. (2017b) instead of the B2017 estimator. Our choice is purely based on computational convenience. A python implementation of the Hand et al. (2017b) estimator is publicly available in the NBODYKIT package[3] (Hand et al. 2017a). We confirm that the resulting $P_\ell(k)s$ and covariance matrices from the Hand et al. (2017b) and B2017 estimators are consistent with one another.

Next, for the S2017 group multiplicity function analysis, they start with the Berlind et al. (2006) FoF algorithm to identify groups in the SDSS and mock data. S2017 adopts the Berlind et al. (2006) linking lengths in units of mean inter-galaxy separation: $b_\perp = 0.14$ and $b_\parallel = 0.75$. In comoving lengths, the linking lengths for the SDSS DR7 $M_r < -19$ sample correspond to $(r_\perp, r_\parallel) = (0.57, 3.05)h^{-1}\,\mathrm{Mpc}$. Once the groups are identified in the SDSS and mock data, $\zeta(N)$ is derived by calculating the comoving number density of groups in bins of richness N — the number of galaxies in the group. For the $M_r < 19$ sample, S2017 uses eight N bins: $(5 - 6), (7 - 9), (10 - 13), (14 - 19), (20 - 32), (33 - 52), (53 - 84), (85 - 220)$. For further details on the GMF calculation, we refer readers to Section 4.2 of S2017. From the $\zeta^{(n)}(N)s$ of each mock, S2017 computes the (i, j) element of the covariance matrix as

$$\mathbb{C}_{i,j} = \frac{1}{N_{\mathrm{mock}} - 1} \sum_{n=1}^{N_{\mathrm{mock}}} \left[\zeta^{(n)}(N_i) - \bar{\zeta}(N_i) \right] \times \left[\zeta^{(n)}(N_j) - \bar{\zeta}(N_j) \right]. \tag{4}$$

In S2017, they compute the covariance matrix using 200 mocks generated using a single fiducial set of HOD parameters. As we describe in Section 2.2, in this paper we use 20,000

---

[3]http://nbodykit.readthedocs.io/en/latest/index.html

mocks from 100 different set of HOD parameters sampled from the MCMC chain. The GMF covariance matrix we use in this paper is computed with $N_{\rm mock} = 20,000$ mocks.

For the rest of the paper, in order to discuss the two separate analyses of B2017 and S2017 in a consistent manner, we define the matrix $\mathbf{D}^{\rm mock}$ of the mock observables (P_ℓ and ζ) as

$$\mathbf{D}^{\rm mock} = \left\{ \mathbf{D}^{\rm mock}_n \right\} \qquad \text{where } \mathbf{D}^{\rm mock}_n \begin{cases} \vec{P}^{(n)} & \text{for B2017,} \\ \zeta^{(n)} & \text{for S2017.} \end{cases} \tag{5}$$

$\mathbf{D}^{\rm mock}$ has dimensions of 2048×37 and $20,000 \times 8$ for B2017 and S2017 respectively.

For the methods in Sections 4.1 and 4.2, the mock observable data ($\mathbf{D}^{\rm mock}$) need to be pre-processed. This pre-processing involves two steps: mean-subtraction (centering) and whitening. For mean subtraction, the mean of the observable is subtracted from $\mathbf{D}^{\rm mock}$. Then $\mathbf{D}^{\rm mock} - \overline{\mathbf{D}}^{\rm mock}$ is whitened using a linear transformation to remove the Gaussian correlation between the bins of $\mathbf{D}^{\rm mock}$:

$$\mathbf{X}^{\rm mock} = L \left(\mathbf{D}^{\rm mock} - \bar{\mathbf{D}}^{\rm mock} \right). \tag{6}$$

The linear transformation is derived so that covariance matrix of the whitened data, $\mathbf{X}^{\rm mock}$, is the identity matrix $\mathbb{I}$. Such a whitening linear transformation can be derived in infinite ways. One way to derive the linear transformation is through the eigen-decomposition of the covariance matrix (*e.g.* Hartlap et al. 2009; Sellentin & Heavens 2017). We, alternatively, derive the linear transformation $\mathbf{L}$ using Choletsky decomposition of the inverse covariance matrix (Press et al. 1992): $\mathbb{C}^{-1} = \mathbf{L}\,\mathbf{L}^T$. The different methods for whitening do not impact the results of the paper. Now that we have the pre-processed data of the mock observables, we proceed in the next section to quantifying the non-Gaussianity of the P_ℓ and ζ likelihoods.

3. Quantifying the Likelihood non-Gaussianity

The standard approach to parameter inference in LSS studies neglects to account for likelihood non-Gaussianity. However, we are not the first to investigate likelihood non-Gaussianity in LSS analyses. Nearly two decades ago, Scoccimarro (2000) examined the likelihood non-Gaussianity for the power spectrum and reduced bispectrum using mock catalogs of the IRAS redshift catalogs. More recently, Hartlap et al. (2009) and Sellentin & Heavens (2017) examined the non-Gaussianity of the cosmic shear correlation function likelihood using simulations of the Chandra Deep Field South and CFHTLenS, respectively.

While these works present different methods for identifying likelihood non-Gaussianity, they do not present a concrete way of quantifying it. Hartlap et al. (2009), for instance, identifies the non-Gaussianity of the cosmic shear likelihood by looking at the statistical independence/dependence of PCA components of the mock observable. In Sellentin & Heavens

(2017), they use the Mean Integrated Squared Error (MISE) as a distance metric between Gaussian random variables and the whitened mock observable data vector to characterize non-Gaussian correlations between elements of the data vector. These indirect measures of non-Gaussianity are challenging to interpret and extend more generally to LSS studies.

A more direct approach can be taken to quantify the non-Gaussianity of the likelihood. We can calculate the divergence between the distribution of our observable, $p(x)$, and $q(x)$ a multivariate Gaussian described by the average of the mocks and the covariance matrix — *i.e.* the pseudo-likelihood. The following are two of the most commonly used divergences: the Kullback-Leibler (KL) divergence

$$D_{KL}(p \parallel q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx \tag{7}$$

and the Rényi-α divergence

$$D_{R-\alpha}(p \parallel q) = \frac{1}{\alpha - 1} \log \int p^\alpha(x) \, q^{1-\alpha}(x) \, dx. \tag{8}$$

In the limit as α approaches 1, the Rényi-α divergence is equivalent to the KL divergence.

Of course, in our case, we don't know $p(x)$ — *i.e.* the probability distribution function of our observable. If we did, we would simply use that instead of bothering with the covariance matrix or this paper. We can, however, still estimate the divergence using nonparametric divergence estimators (Wang et al. 2009; Poczos et al. 2012; Krishnamurthy et al. 2014). These estimators allow us to estimate the divergence directly from samples $X_{1:n} = \{X_1, ...X_n\}$ and $Y_{1:m} = \{Y_1, ...Y_m\}$ drawn from p and q respectively: $\hat{D}(X_{1:n} \parallel Y_{1:m})$. For instance, the estimator presented in Poczos et al. (2012) allows us to estimate the kernel function of the Rényi-α divergence,

$$D_\alpha(p \parallel q) = \int p^\alpha(x) q^{1-\alpha}(x) \, dx. \tag{9}$$

using k^{th} nearest neighbor density estimators. Let $\rho_k(x)$ denote the Euclidean distance of the kth nearest neighbor of x in the sample $X_{1:n}$ and $\nu_k(x)$ denote the Euclidean distance of the kth nearest neighbor of x in the sample $Y_{1:m}$. Then

$$D_\alpha(p \parallel q) \approx \hat{D}_\alpha(X_{1:n} \parallel Y_{1:m}) = \frac{B_{k,\alpha}}{n} \left(\frac{n-1}{m}\right)^{1-\alpha} \sum_{i=1}^n \left(\frac{\rho_k^d(X_i)}{\nu_k^d(X_i)}\right)^{1-\alpha}, \tag{10}$$

where $B_{k,\alpha} = \frac{\Gamma(k)^2}{\Gamma(k-\alpha+1)\Gamma(k+\alpha-1)}$. This estimator, as Poczos et al. (2012) proves, is asymptotically unbias,

$$\lim_{n,m\to\infty} \mathbb{E}\left[\hat{D}_\alpha(X_{1:n} \parallel Y_{1:m})\right] = D_\alpha(p \parallel q). \tag{11}$$

Plugging $\hat{D}_\alpha(X_{1:n} \parallel Y_{1:m})$ into Eq. 8, we get an estimator for the Rényi-α divergence. Wang et al. (2009) derives a similar estimator for the KL divergence (Eq. 7). These divergence

estimates have been applied to Support Distribution Machines and used in the machine learning and astronomical literature with great success (*e.g.* Póczos et al. 2011; Poczos et al. 2012; Póczos et al. 2012; Xu et al. 2013; Ntampaka et al. 2015, 2016; Ravanbakhsh et al. 2017). For more details on the non-parametric divergence estimators, we refer readers to Poczos et al. (2012) and Krishnamurthy et al. (2014).

With these estimators, we can now explicitly quantify the non-Gaussianity of the likelihood by computing the divergence between the likelihood distribution and the Gaussian pseudo-likelihood distribution, $\mathcal{L}^{\mathrm{pseudo}}$. We draw a reference sample $\mathbf{Y}^{\mathrm{ref}}$ from $\mathcal{L}^{\mathrm{pseudo}}$ then since $\mathbf{X}^{\mathrm{mock}}$ is in principle sampled from $p(x)$, we can use the estimators to compute $D(\, p(x) \parallel \mathcal{L}^{\mathrm{pseudo}}) \approx \hat{D}(\mathbf{X}^{\mathrm{mock}} \parallel \mathcal{L}^{\mathrm{pseudo}})$. Similar to the experiments detailed in Poczos et al. (2012), we construct $\mathbf{Y}^{\mathrm{ref}}$ with a comparable sample size as $\mathbf{X}^{\mathrm{mock}}$: 2000 and 10,000 for the P_ℓ and ζ analyses respectively. Based on the sample size of $\mathbf{X}^{\mathrm{mock}}$ we calculate the divergences using the $k = 10$ nearest neighbors.

In Figure 1, we present the resulting Rényi-α (left) and KL (right) divergences (orange) between the likelihood and the Gaussian pseudo-likelihood for the B2017 P_ℓ (top) and S2017 ζ (bottom) analyses: $\hat{D}_{R\alpha}$ and $\hat{D}_{KL}$. For reference, we also include (in blue) divergence estimates of the pseudo-likelihood onto itself, which we calculate as $\hat{D}(\mathbf{X}^{\mathrm{ref}} \parallel \mathbf{Y}^{\mathrm{ref}})$. $\mathbf{X}^{\mathrm{ref}}$ is a data vector with the same dimension as $\mathbf{X}^{\mathrm{mock}}$ sampled from the pseudo-likelihood. $\hat{D}$s are *estimates* of the true divergence, therefore we resample $\mathbf{Y}^{\mathrm{ref}}$ and compute each $\hat{D}$ estimate 100 times. In Figure 1, we present the resulting distributions of $\hat{D}$, which illustrate the uncertainty of $\hat{D}$. The discrepancy between the $\hat{D}(\, p(x) \parallel \mathcal{L}^{\mathrm{pseudo}})$ distributions and the reference $\hat{D}(\mathbf{X}^{\mathrm{ref}} \parallel \mathbf{Y}^{\mathrm{ref}})$ distributions $(\Delta\hat{D})$ quantify the discrepancy between the pseudo-likelihood and the likelihood. Each panel of Figure 1 shows significant discrepancy between the two distributions — *both the $P_\ell(k)$ and $\zeta(N)$ likelihoods are significantly non-Gaussian.*

The Gaussian pseudo-likelihood assumption for P_ℓ is motivated by the Central Limit Theorem. If enough modes contribute to the power spectrum, then the likelihood approaches a Gaussian. Given the survey volume of BOSS DR12 and the restrictive k range of the B2017 analysis ($0.01 < k < 0.15$ for $\ell = 0$ and 2; $0.01 < k < 0.15$ for $\ell = 4$), one would expect this to be mostly true. Although relatively small, we find significant $\Delta\hat{D}$ and therefore likelihood non-Gaussinaity. In order to better understand the source of this non-Gaussianity, we repeat the divergence comparisons for different k ranges. If we exclude the largest scales and set $k_{\mathrm{min}} = 0.05$, $\Delta\hat{D}$ decreases. Meanwhile, if we exclude the smallest scales and set $k_{\mathrm{max}} = 0.1$ for all multipoles, $\Delta\hat{D}$ increases. This suggests that the largest scales (low k) contribute most to the P_ℓ likelihood non-Gaussianity. Furthermore, when we compare the divergences for just the monopole and quadrupole, $\Delta\hat{D}$ decreases. Among the multipole, the hexadecapole contributes most to the non-Gaussianity of the P_ℓ likelihood. In both the low k regimes and the hexadecapole, the contribution to the non-Gaussianity is likely due to an insufficient number of modes for satisfying the Central Limit Theorem.

For ζ, the discrepancies between the $\hat{D}$ distributions are consistent with the fact that the true ζ likelihood distribution is likely Poisson — not Gaussian — similar to the likelihood of observed cluster counts (Cash 1979; Collaboration et al. 2014; Ade et al. 2016). Although the groups identified with a FoF algorithm do not correspond to clusters, we nevertheless expect the likelihood to be non-Gaussian. We again repeat the divergence comparison for different N ranges to better understand the source of non-Gaussianity. Excluding the lowest N bin does not significantly impact $\Delta\hat{D}$. However, when we exclude the highest N bin, $\Delta\hat{D}$ decreases significantly. We therefore find that the high richness end of ζ contibute most to the non-Gaussianity of the ζ likelihood. The contribution to the non-Gaussianity, similar to the P_ℓ case, comes most from the low signal-to-noise regime.

4. Estimating the Non-Gaussian Likelihood

In the previous section, we estimate the divergence between the P_ℓ and ζ likelihoods and their respective Gaussian pseudo-likelihoods. These divergences identify and quantify the significant non-Gaussianity in the likelihoods of LSS studies. Our ultimate goal, however, is to quantify the impact of likelihood non-Gaussianity on the final cosmological parameter constrains and to develop more accurate methods for parameter inference in LSS. From the divergence estimates alone, it's not obvious how they propagate onto the final parameter constraints. Therefore in this section, we present two methods for more accurately estimating the true non-Gaussian likelihoods of P_ℓ and ζ from their corresponding mocks. These methods provide more accurate estmiates of the likelihood than the Gaussian pseudo-likelihood. Moreover, we will use them later to quantify the impact of likelihood non-Gaussianity on the B2017 and S2017 parameter constraints.

4.1. Gaussian Mixture Likelihood Estimation

When mock catalogs are used for parameter inference in LSS analyses, they essentially serve as data points sampling the likelihood distribution. For the pseudo-likelihood, this distribution is assumed to have a Gaussian functional form, which is why we estimate the covariance matrix from mocks. However, the Gaussian functional form, or any functional form for that matter, is *not* necessary to estimate the likelihood distribution. Instead, the multi-dimensional likelihood distribution can be directly estimated from the set of mock catalogs — for instance using Gaussian mixture density estimation (Press et al. 1992; McLachlan & Peel 2000). Besides its extensive use in machine learning and statistics, in astronomy, Gaussian mixture density estimation has been used for inferring the velocity distribution of stars from the Hipparcos satellite (Bovy et al. 2011), classifying galaxies in the Galaxy And Mass Assembly Survey (Taylor et al. 2015), classifying pulsars (Lee et al. 2012), and much more (see also Hogg et al. 2010; Kuhn & Feigelson 2017).

Fig. 1.— Rényi-α and KL divergence estimates ($\hat{D}_{R\alpha}$ and $\hat{D}_{KL}$; orange) between the likelihood distribution and the Gaussian pseudo-likelihood for the B2017 P_ℓ (top) and S2017 ζ (bottom) analyses. We include in blue, as reference, the divergence estimates of the pseudo-likelihood onto itself. $\hat{D}_{R\alpha}$ and $\hat{D}_{KL}$ are computed using the non-parametric k-NN estimator (Section 3) on the mock data $\mathbf{X}^{\mathrm{mock}}$ and a reference sample $\mathbf{Y}^{\mathrm{ref}}$ drawn from the pseudo-likelihood. We compute $\hat{D}_{R\alpha}$ and $\hat{D}_{KL}$ 100times and plot their distribution in order to illustrate the uncertainty of the $\hat{D}$ estimator. The significant discrepancy between the two divergence distributions in each of the panels, identifies the *significant non-Gaussianity of the $P_\ell(k)$ and $\zeta(N)$ likelihoods.*

Gaussian mixture density estimation is a "semi-parametric" method that uses a weighted sum of k Gaussian component densities, a Gaussian mixture model (hereafter GMM)

$$\hat{p}(x; \boldsymbol{\theta}) = \sum_{i=1}^{k} \pi_i \, \mathcal{N}(x; \boldsymbol{\theta}_i), \tag{12}$$

to estimate the density. The component weights (π_i; also known as mixing weights) and the component parameters $\boldsymbol{\theta}_i$ are free parameters of the mixture model. Given some data set $\mathbf{X}_N = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, these free GMM parameters are, most popularly, estimated through an expectation-maximization algorithm (EM; Dempster et al. 1977; Neal & Hinton 1998). The EM algorithm begins by randomly assigning $\boldsymbol{\theta}_i^0$ to the k Gaussian components. The algorithm then iterates between two steps. In the first step, the algorithm computes for every data point, $\mathbf{x}_n$, a probability of being generated by each component of the model. These probabilities can be thought of as weighted assignments of the points to the components. Next, given the $\mathbf{x}_n$ assignment to the components, $\boldsymbol{\theta}_i^t$ of each component are updated to $\boldsymbol{\theta}_i^{t+1}$ to maximize the likelihood of the assigned points. At this point, π_i can also be updated by summing up the assignment weights and normalizing it by the total number of data points, N. This entire process is repeated until convergence — *i.e.* when the log-likelihood of given the mixture model $\log p(\mathbf{X}_N; \boldsymbol{\theta}^t)$ converges. The EM algorithm is guaranteed to converges to a local maximum of the likelihood (Wu 1983). In practice, instead of arbitrarily assigning the initial condition, $\boldsymbol{\theta}_i^0$ is derived from a k-means clustering algorithm (Lloyd 1982). Without going into details, the k-means algorithm clusters the data $\mathbf{X}_N$ into k clusters, each described by the mean (or centroid) μ_i of the samples in the cluster. The algorithm then iteratively chooses centroids that minimize the average squared distance between points in the same cluster. For our GMMs, we initialize the EM algorithm using the k-means++ algorithm of Arthur & Vassilvitskii (2007).

So far in our description of GMMs, we have kept the number of components k fixed. k, however, is a free parameter and selecting it is a crucial step in Gaussian mixture density estimation. With too many components the model may overfit the data; with too few components, the model may not be flexible enough to approximate the true underlying distribution. In order to address this model selection problem of selecting k, we make use of the Bayesian Information Criterion (BIC; Schwarz 1978). BIC has been widely used for determining the number of components in mixture modeling (*e.g.* Leroux 1992; Roeder & Wasserman 1997; Fraley & Raftery 1998; Steele & Raftery 2010) and for model selection in general in astronomy (*e.g.* Liddle 2007; Broderick et al. 2011; Wilkinson et al. 2015; Vakili & Hahn 2016). According to BIC, models with higher likelihood are preferred; however, to address the concern of overfitting, BIC introduces a *penalty* term for the number of parameters in the model:

$$\mathrm{BIC} = -2 \ln \mathcal{L} + N_{\mathrm{par}} \ln N_{\mathrm{data}}. \tag{13}$$

We select k based on the number of components in the model with the lowest BIC.

With Gaussian mixture density estimation we can directly estimate the likelihood distribution using the mock catalogs. We first fit GMMs with ks < 30 components to the whitened mock data $\mathbf{X}^{\mathrm{mock}}$ using the EM algorithm for each model. For each of the converged GMMs, we calculate the BIC and then select the model with the lowest BIC as the best density estimate of the likelihood distribution: $\hat{p}_{\mathrm{GMM}}(x)$. The selected density estimate can then be used to calculate the likelihood and quantify the impact of likelihood non-Gaussianity on the parameter constraints of B2017 and S2017. But before we do that, we have to confirm whether $\hat{p}_{\mathrm{GMM}}$ is in fact a better estimate of the likelihood over the Gaussian pseudo-likelihood. To do this, we return to the divergence estimates of Section 3.

To estimate the divergence between our Gaussian mixture density estimate, $\hat{p}_{\mathrm{GMM}}$, and the likelihood distribution, we take the same approach as our $\hat{D}(\mathbf{X}^{\mathrm{mock}}|\mathbf{Y}^{\mathrm{ref}})$ calculation in Section 3. Instead of $\mathbf{Y}^{\mathrm{ref}}$ drawn from the pseudo-likelihood, we draw samples from $\hat{p}_{\mathrm{GMM}}(x)$ with the same dimensions. Then we calculate k-NN Rényi-α and KL divergence estimates between this sample and $\mathbf{X}^{\mathrm{mock}}$. As we did in Figure 1, we repeat this process 100times, resampling $\hat{p}_{\mathrm{GMM}}$ each time, in order get a distribution of divergence estimates that reflects the scatter in the estimator. In Figure 2, we present the resulting distribution of divergences between $\hat{p}_{\mathrm{GMM}}$ and the likelihood distribution in purple for the $P_\ell(k)$ (top) and $\zeta(N)$ (bottom) analyses. For comparison, we include the $\hat{D}$ distributions from Figure 1.

For the $\zeta(N)$ analysis of S2017, our Gaussian mixture density estimate significantly improves the divergence discrepancy compared to the pseudo-likelihood. In other words, *our Gaussian mixture density estimate is a significant better estimate of the ζ likelihood distribution than the pseudo-likelihood.* On the other hand, our Gaussian mixture density estimate for the $P_\ell(k)$ analysis of B2017 does not significantly improve the divergence discrepancy. This difference in the performance of Gaussian mixture density estimation is not surprising. One would expect a direct density estimation to be more effective for the S2017 case, where we estimate an 8-dimensional distribution with $N_{\mathrm{mock}} = 20,000$ samples, compared to the B2017 case where we estimate a 37-dimensional distribution with only $N_{\mathrm{mock}} = 2048$ samples. Given the unconvincing accuracy of the Gaussian mixture density estimate of the P_ℓ likelihood, in the next section we present an alterative method for estimating the non-Gaussian likelihood.

4.2. Independent Component Analysis

Gaussian mixture density estimation fails to accurately estimate the 37-dimensional P_ℓ likelihood distribution of B2017. Rather than estimating the likelihood distribution directly, if we can transform the observable $\mathbf{x}$ (*e.g.* P_ℓ) into statistically independent components $\mathbf{x}^{\mathrm{IC}}$ the problem becomes considerably simpler. Since $\mathbf{x}^{\mathrm{IC}}$ is statistically independent, the
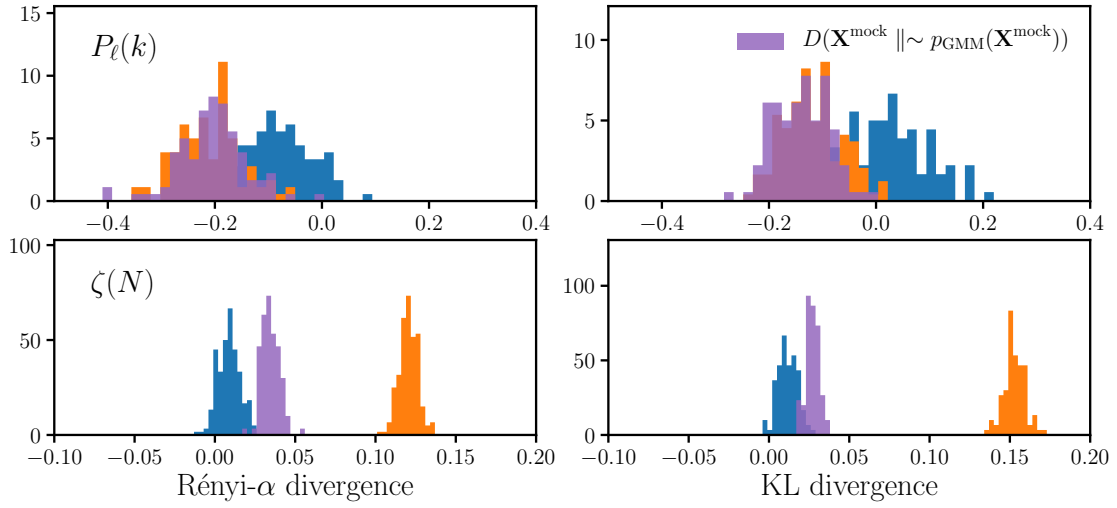
Fig. 2.— Rényi-α and KL divergence estimates ($\hat{D}_{R\alpha}$ and $\hat{D}_{KL}$; purple) between the likelihood distribution and the Section 4.1 GMM likelihood estimate for the B2017 P_ℓ (top) and S2017 ζ (bottom) analyses. We include the divergence estimates from Figure 1 for comparison. The Gaussian mixture likelihood does not significantly improve the discrepancy in divergence for the P_ℓ analysis. This is due to the high-dimensionality (37 dimensions) of the P_ℓ likelihood. For the ζ analysis, *our Gaussian mixture likelihood estimate is a significantly better estimate of the likelihood than the pseudo-likelihood.*

likelihood distribution becomes

$$p(x) = \prod_{n=1}^{N_{\text{bin}}} p_{x_n^{\text{IC}}}(x) \tag{14}$$

where N_{bin} is the number of bins in the observable and the number of independent components. For the B2017 case, this reduces the problem of estimating a 37 dimensional distribution with 2048 samples to a problem of estimating 37 one dimensional distributions with 2048 samples each. The challenge, however, is in finding the transformation.

Efforts in the past have attempted to tackle this sort of high-dimensional problem (*e.g.* Scoccimarro 2000; Eisenstein & Zaldarriaga 2001; Gaztañaga & Scoccimarro 2005; Norberg et al. 2009; Sinha et al. 2017). They typically use singular value decomposition or principal component analysis (PCA; Press et al. 1992). For a Gaussian likelihood, the PCA components of it are statistically independent. However, when the likelihood is *not* Gaussian, the PCA components are uncorrelated but *not necessarily statistically independent* (Hartlap et al. 2009). Since the P_ℓ and ζ likelihoods are non-Gaussian, we cannot use PCA. Instead, we follow Hartlap et al. (2009) and use Independent Component Analysis (ICA Hérault & Ans 1984; Comon 1994; Hyvärinen & Oja 2000; Hyvarinen 2001).

In order to find the transformation of $\mathbf{x}$ to $\mathbf{x}^{\text{IC}}$ we first assume that $\mathbf{x}$ is generated by some linear transformation $\mathbf{x} = \mathbf{M}\,\mathbf{x}^{\text{IC}}$. Then the goal of ICA is to invert this problem, $\mathbf{y} = \mathbf{W}\,\mathbf{x}$, and find $\mathbf{W}$ and $\mathbf{y}$ that best estimate $\mathbf{x}^{\text{IC}} \approx \mathbf{y}$. The basic premise of ICA is simple, *maximizing non-Gaussianity maximizes the statistical independence*. Consider a single component of $\mathbf{y}$:

$$\mathbf{y}_n = \mathbf{w}_n^t\,\mathbf{x} = \mathbf{w}_n^t\,\mathbf{M}\,\mathbf{x}^{\text{IC}} \tag{15}$$

where $\mathbf{w}_n^t$ is the nth row of $\mathbf{W}$. Since $\mathbf{y}_n$ is a linear combination of the independent components $\mathbf{x}^{\text{IC}}$, from the Central Limit Theorem $\mathbf{y}_n$ is necessarily more Gaussian than any of the components *unless* $\mathbf{y}_n$ is equal to one of the $\mathbf{x}^{\text{IC}}$ components. In other words, we can achieve $\mathbf{x}^{\text{IC}} \approx \mathbf{y}$ by finding $\mathbf{W}$ that maximizes the non-Gaussianity of $\mathbf{y}$. For a more rigorous justification of ICA we refer readers to Hyvarinen (2001). In practice, non-Gaussianity is commonly measured using differential entropy — "negentropy". For $\mathbf{y}_n$ with density function p_{y_n} the entropy is defined as

$$H_{y_n} = -\int p_{y_n}(y)\,\log p_{y_n}(y)\,\mathrm{d}y. \tag{16}$$

Since the Gaussian distribution has the largest entropy among all distributions with a given variance, the negentropy can be defined as,

$$J_{y_n} = H_{y_n^{\text{Gauss}}} - H_{y_n}. \tag{17}$$

Finding the statistically independent components is now a matter of finding the $\mathbf{W}$ that maximizes $\sum_n J_{\mathbf{y_n}}$ — the negentropy of $\mathbf{y}$. In this paper, we make use of the `FastICA` fixed-point iteration algorithm (Hyvarinen 1999). The algorithm starts with randomly selected

$\mathbf{w}_n s$, then it uses approximations of negentropy from Hyvärinen (1998) and Newton's method to iteratively solve for $\mathbf{W}$ that maximizes negentropy. For details on the `FastICA` algorithm, we refer readers to Hyvarinen (1999).

Performing ICA on the whitened observable data $\mathbf{X}^{\mathrm{mock}}$, we derive the matrix $\mathbf{W}$ that transforms $\mathbf{X}^{\mathrm{mock}}$ into N_{bin} approximately independent components:

$$\mathbf{X}^{\mathrm{ICA}} = \mathbf{W}\,\mathbf{X}^{\mathrm{mock}} = \{\mathbf{X}_1^{\mathrm{ICA}}, ..., \mathbf{X}_{N_{\mathrm{bin}}}^{\mathrm{ICA}}\}. \tag{18}$$

From these statistically independent components and Eq. 14, we can estimate the likelihood distribution. $p_{x_n^{\mathrm{IC}}}(x)$, from Eq. 14, is the 1-dimensional distribution function of the nth ICA component. This distribution is sampled by $\mathbf{X}_n^{\mathrm{ICA}}$, the transformed mock data. That means $\mathbf{X}_n^{\mathrm{ICA}}$ can be used to estimate $p_{x_n^{\mathrm{ICA}}}$ using a method like kernel density estimation (KDE; Hastie et al. 2009; Feigelson & Babu 2012). With KDE, the density estimate, $\hat{p}_{x_n^{\mathrm{ICA}}}$, is constructed by smoothing the empirical distribution of the ICA component x_n^{ICA} using a smooth kernel:

$$\hat{p}_{x_n^{\mathrm{ICA}}}(x) = \frac{1}{N_{\mathrm{mock}}b} \sum_{j=1}^{N_{\mathrm{mock}}} K\left(\frac{x - \mathrm{X}_n^{(j),\mathrm{ICA}}}{b}\right). \tag{19}$$

b is the bandwidth and K is the kernel function. Following the choices of Hartlap et al. (2009), we use a Gaussian distribution for K and the "rule of thumb" bandwidth (also known as Scott's rule; Scott 1992; Davison 2008) for b. Combining the $\hat{p}_{x_n^{\mathrm{ICA}}}$ estimates for all $n = 1, ..., N_{\mathrm{bin}}$ into Eq. 14, we can estimate the likelihood distribution $p(x) \approx \prod_n \hat{p}_{x_n^{\mathrm{ICA}}}(x)$

We again check whether the likelihood estimate from ICA is actually a better estimate of the true likelihood distribution than the Gaussian pseudo-likelihood. Following the same procedure as we did for the Gaussian mixture likelihood in Section 4.1, we calculate the divergence between our ICA likelihood, $\prod \hat{p}_{x_n^{\mathrm{ICA}}}(x)$, and the likelihood distribution, $p(x)$. We draw a sample from $\prod \hat{p}_{x_n^{\mathrm{ICA}}}$ with the same dimensions as $\mathbf{Y}^{\mathrm{ref}}$ (Section 3), apply the mixing matrix (undoing the ICA transformation), and then calculate the k-NN Rényi-α and KL divergence estimates between the sample and $\mathbf{X}^{\mathrm{mock}}$. We repeat these steps 100times to get the distribution of estimates that reflects the scatter in the estimator. In Figure 3, we present the resulting distribution of $\hat{D}\left(\mathbf{X}^{\mathrm{mock}} \|\sim \prod \hat{p}_{x_n^{\mathrm{ICA}}}\right)$ in green for the $P_\ell(k)$ (top) and $\zeta(N)$ (bottom) analyses. For comparison, we include the distributions from Figure 1.

For *both* B2017 and S2017, our ICA likelihood significantly improves the divergence discrepancy compared to the pseudo-likelihood. For S2017, however, the ICA likelihood proves to be less accurate than the Gaussian mixture likelihood in Section 4.1. More importantly, for B2017where the Gaussian mixture likelihood did not improve upon the pseudo-likelihood, the ICA method provides a significantly more accurate likelihood estimate. This demonstrates that the ICA method is an effective alterative to the more direct Gaussian mixture method. Furthermore, the effectiveness of the ICA method in estimating higher dimensional likelihoods with fewer samples (mocks) is particularly appealing for LSS, since analyses continue
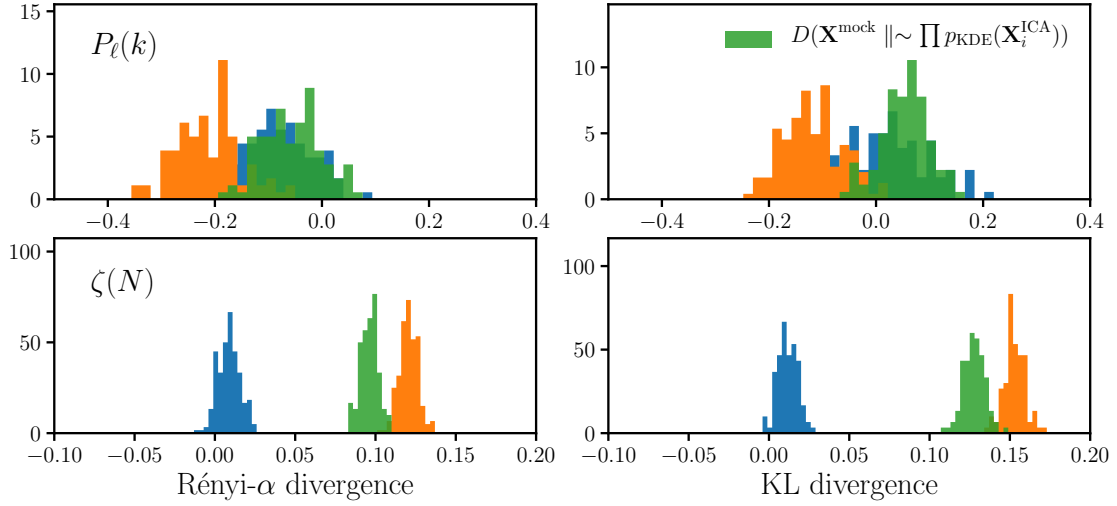
Fig. 3.— Rényi-α and KL divergence estimates ($\hat{D}_{R\alpha}$ and $\hat{D}_{KL}$; green) between the likelihood distribution and the Section 4.2 ICA likelihood estimate for the B2017 P_ℓ (top) and S2017 ζ (bottom) analyses. We include the divergence estimates from Figure 1 for comparison. The ICA likelihood significantly improves the divergence discrepancy for both the P_ℓ and ζ analyses. For ζ, the improvement of the ICA likelihood over the pseudo-likelihood is more modest than our GMM estimate from Section 4.1. However, for P_ℓ where the GMM method struggled, our ICA likelihood provides a significantly better estimate of the true P_ℓ likelihood than the pseudo-likelihood.

to increase the size of their observable data vector. As the divergence estimation framework we present makes it easy to test the accuracy of different methods, a hard choice is not necessary and multiple methods can easily be tested to construct the best estimate of the likelihood distribution for each specific analysis. Based on the performances of the GMM and ICA methods, in the following sections we use the ICA likelihood for the B2017 analysis and the GMM likelihood for the S2017 analysis.

5.   Impact on Parameter Inference

To derive the posterior distribution of their model parameters, both B2017 and S2017 use the standard Monte Carlo Markov Chain (MCMC) approach with the Gaussian pseudo-likelihood. The B2017 analysis includes 11 parameters,

$$\left\{ f\sigma_8, \ \alpha_\parallel, \ \alpha_\perp, \ b_1^{\mathrm{NGC}}\sigma_8, \ b_1^{\mathrm{SGC}}\sigma_8, \ b_2^{\mathrm{NGC}}\sigma_8, \ b_2^{\mathrm{SGC}}\sigma_8, \ \sigma_v^{\mathrm{NGC}}, \ \sigma_v^{\mathrm{SGC}}, \ N^{\mathrm{NGC}}, \ \mathrm{and} \ N^{\mathrm{SGC}} \right\},$$

while the S2017 analysis includes 5 parameters,

$$\left\{ \log M_{\mathrm{min}}, \ \sigma_{\log M}, \ \log M_0, \ \log M_1, \ \mathrm{and} \ \alpha \right\}.$$

Using the improved likelihood estimates of Sections 4.1 and 4.2, we can now better estimate the true posterior distributions of these parameters and quantify the impact of likelihood non-Gaussianity on parameter constraints. The most straightforward approach to do this would be to use the likelihood estimates to compute MCMC samples from scratch. While this is relatively tractable for the B2017 analysis, for S2017 this is *significantly* more involved. Rather than a perturbation theory based model from B2017, the S2017 model is a forward model, same as their mocks (Section 2.2). Rerunning the MCMC samples would involve evaluating their computationally costly forward model $> 10^5$ times.

Without having to re-run the MCMC chains, we instead use importance sampling to derive the new posteriors from the original chains (see Wasserman 2004, for details on importance sampling). The *target* distribution we want is the new posterior. To sample this distribution, we sample the original posterior as the *proposal* distribution with importance weights, which in our case is the ratio of our likelihood estimates over the pseudo-likelihood. If we let $\mathrm{P}(\mathbf{x}|\boldsymbol{\theta})$ be the original pseudo-likelihood and $\mathrm{P}'(\mathbf{x}|\boldsymbol{\theta})$ be our "new" likelihood, then the new marginal likelihood can be calculated through importance sampling:

$$\mathrm{P}'(\mathbf{x}|\theta_1) = \int \mathrm{P}'(\mathbf{x}|\boldsymbol{\theta})\, \mathrm{d}\theta_2...\mathrm{d}\theta_m = \int \frac{\mathrm{P}'(\mathbf{x}|\boldsymbol{\theta})}{\mathrm{P}(\mathbf{x}|\boldsymbol{\theta})}\, \mathrm{P}(\mathbf{x}|\boldsymbol{\theta})\, \mathrm{d}\theta_2...\mathrm{d}\theta_m. \tag{20}$$

Then through Monte Carlo integration,

$$\mathrm{P}'(\mathbf{x}|\theta_1) \approx \sum_{\boldsymbol{\theta}^{(i)} \in S} \frac{\mathrm{P}'(\mathbf{x}|\boldsymbol{\theta}^{(i)})}{\mathrm{P}(\mathbf{x}|\boldsymbol{\theta}^{(i)})}. \tag{21}$$

where S is the sample drawn from $\mathrm{P}(\mathbf{x}|\boldsymbol{\theta})$. In our case S is just the original MCMC chain. The only calculation required is the importance weights in Eq. 21, $\mathrm{P}'(\mathbf{x}|\boldsymbol{\theta}^{(i)})/\mathrm{P}(\mathbf{x}|\boldsymbol{\theta}^{(i)})$ for each sample $\boldsymbol{\theta}^{(i)}$ of the original MCMC chain. For B2017, $\mathrm{P}(\mathbf{x}|\boldsymbol{\theta}^{(i)})$ is the ICA likelihood; for S2017, $\mathrm{P}(\mathbf{x}|\boldsymbol{\theta}^{(i)})$ is the GMM likelihood.

We present in Figure 4 the resulting posterior distributions using the non-Gaussian ICA likelihood for the $\left\{ f\sigma_8, \alpha_\parallel, \alpha_\perp, b_1^{\mathrm{NGC}}\sigma_8, b_1^{\mathrm{SGC}}\sigma_8, b_2^{\mathrm{NGC}}\sigma_8, b_2^{\mathrm{SGC}}\sigma_8 \right\}$ parameters in the B2017 P_ℓ analysis (orange). We include the original B2017 posteriors for comparison in blue. On the bottom of each panel, we also include box plots marking the confidence intervals of the updated and original posteriors. The boxes and "whiskers" repesent the 68% and 95% confidence intervals, respectively. The median and 68% confidence intervals of the posteriors are also listed in Table 1. $f\sigma_8$ and $b_2^{\mathrm{SGC}}\sigma_8$ are the main parameters with noticeable change in their posteriors. The posterior of $b_2^{\mathrm{SGC}}\sigma_8$ broadens from $0.476^{+1.262}_{-1.175}$ to $0.422^{+1.517}_{-1.377}$ once likelihood non-Gaussianity is incorporated. More importantly, the $f\sigma_8$ posterior shifts from $0.478^{+0.053}_{-0.049}$ to $0.456^{+0.059}_{-0.049}$, which corresponds to a shift of 0.44σ. The other parameter constraints, however, remain largely unaffected by likelihood non-Gaussianity.

Focusing on the main cosmological parameters $f\sigma_8$, $\alpha_\parallel$, and $\alpha_\perp$, we present their joint posterior distributions in Figure 5. The contours mark the 68% and 95% confidence intervals of the posteriors. The shift in the $f\sigma_8$ distribution is reflected in the $(f\sigma_8, \alpha_\parallel)$ and $(\alpha_\perp, f\sigma_8)$ contours (left and middle panels respectively). The $(\alpha_\parallel, \alpha_\perp)$ distribution (right), however, show nearly no change from the non-Gaussian likelihood. Despite its impact on $f\sigma_8$ and $b_2^{\mathrm{SGC}}\sigma_8$, overall, likelihood non-Gaussianity does *not* significantly impact the parameter constraints of the P_ℓ analysis. $b_2^{\mathrm{SGC}}\sigma_8$ is a poorly constrained nuisance parameter and although using the pseudo-likelihood biases $f\sigma_8$, the impact relative to its uncertainty is small — less than 0.5σ. The fact that the P_ℓ analysis is largely unaffected by likelihood non-Gaussianity is consistent with the divergences in Figure 1, which found relatively small discrepancies between the likelihood and pseudo-likelihood. It also illustrates the remarkable effectiveness of the Central Limit Theorem.

Next in Figure 6, we present the posterior distributions calculated using the non-Gaussian GMM likelihood for the HOD parameters in the S2017 ζ analysis (orange). We include the posteriors calculated using the pseudo-likelihood for comparison in blue. The box plots on the bottom of each plot mark the 68% and 95% confidence intervals of the posteriors. In the dotted lines, we plot the original S2017 posteriors, which differ slightly from the blue distribution. This discrepancy is caused by the difference in the covariance matrix we use in the pseudo-likelihood (see Section 2.3). The difference, however, is negligible.

Besides the poorly constrained parameters $\sigma_{\log M}$ and $\log M_0$, likelihood non-Gaussianity significantly impacts the posterior distributions of the HOD parameters. Each of the parameter constraints for $\log M_{\min}$, $\log M_1$, and α, are significantly broadened and shifted from the pseudo-likelihood constraints (see Table 1 for details). The $\log M_1$ constraint, for instance,
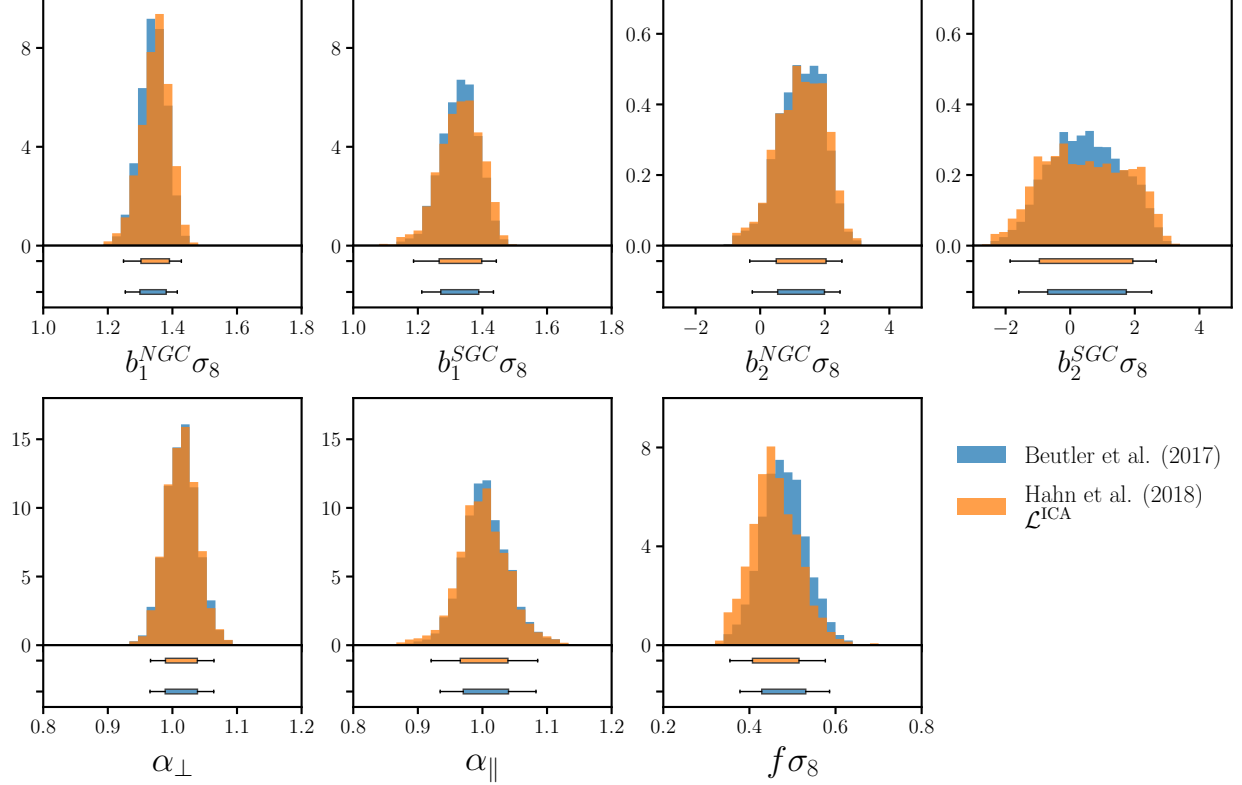
Fig. 4.— The posterior distribution for $\left\{ f\sigma_8,\ \alpha_\parallel,\ \alpha_\perp,\ b_1^{\mathrm{NGC}}\sigma_8,\ b_1^{\mathrm{SGC}}\sigma_8,\ b_2^{\mathrm{NGC}}\sigma_8,\ b_2^{\mathrm{SGC}}\sigma_8, \right\}$ in the B2017 P_ℓ analysis using the non-Gaussian ICA likelihood (orange). We include in blue the original B2017 posteriors for comparison. On the bottom of each panel we include box plots that mark the 68% and 95% confidence intervals of the posterior. The discrepancies between the posteriors are most evident for the parameters $f\sigma_8$ and and $b_2^{\mathrm{SGC}}\sigma_8$. The $f\sigma_8$ constraint shifts by 0.44σ. Hence, using the pseudo-likelihood in the P_ℓ analysis biases the posteriors of these parameters. Overall, however, likelihood non-Gaussianity does *not* have a siginificant impact on the parameter constraints of the P_ℓ analysis.

is shifted by 0.43σ and its 68% confidence interval is expanded by 42%. Similarly, the α constraint is shifted by 0.51σ and its 68% confidence interval is expanded by 66%. The impact of likelihood non-Gaussianity is further emphasized in the joint posterior distributions in Figure 7. The $\log M_{\min}$ versus $\sigma_{\log M}$ and $\log M_{\min}$ versus α contours are both shifted and broadened compared to the $\mathcal{L}^{\mathrm{pseudo}}$ posterior. Figures 6 and 7 reveal that *using the Gaussian pseudo-likelihood significantly underestimates the uncertainty and biases the HOD parameter constraints of the S2017 ζ analysis.*

The contrast between the pseudo-likelihood posteriors and our posteriors in Figures 6 and 7 reflect the divergences in Figure 1, which revealed significant discrepancy between the ζ likelihood and the pseudo-likelihood. These divergences and posteriors are consistent with the expectation that the true ζ likelihood distribution is likely Poisson. Although we expect the likelihood to be similar to the observed cluster count likelihood, the complicated connection between FoF groups and the underlying matter overdensity makes writing down the exact ζ likelihood function tremendously difficult. Nonetheless, the GMM likelihood estimation method we present provides an accurate estimate of the non-Gaussian likelihood.

The posteriors of the S2017 ζ analysis highlight the importance of accounting for likelihood non-Gaussianity in parameter inference of LSS studies. One of the main results of the S2017 HOD analysis is that the ΛCDM + HOD model can successful fit either $\zeta(N)$ or the project two-point correlation function $w_p(r_p)$, separately but struggles to jointly fit both (see Figure 10). This tension suggests that the 'vanilla' HOD model is insufficiently flexible in describing the galaxy-halo connection. Likelihood non-Gaussianity impacts this result. Once the non-Gaussianity is included in the analysis, the posteriors are broadened and shifted towards relaxing the tensions. We examine the effect of likelihood non-Gaussianity for HOD parameter constraints in more detail in Hahn et al. (in prep.).

Even for the P_ℓ analysis, the impact of likelihood non-Gaussianity on the parameter constraints cannot be easily dismissed as we demand increasingly more precise constraints from future experiments. Using the pseudo-likelihood biases the $f\sigma_8$ constraints by $\sim 0.5\%$. Meanwhile, the Dark Energy Spectroscopic Instrument (DESI; Levi et al. 2013), for instance, seeks to constrain $f\sigma_8$ to within a percent[4]. The future, however, is encouraging in this regard. The next surveys will expand the cosmic volumes probed by galaxies and therefore increase the number of modes on all scales. Even as they seek to extend the k range of analyses, thanks to the Central Limit Theorm, we consequently expect likelihood non-Gaussianity to have a smaller effect. For higher order statistics, however, likelihood non-Gaussianity will still have a significant effect. After all, for a given survey, Scoccimarro (2000) found that the reduced bispectrum likelihood is significantly more non-Gaussian than the power spectrum likelihood. Fortunately the methods we present in this paper can easily be extended to other

---

[4]DESI Final Design Report: http://desi.lbl.gov/wp-content/uploads/2014/04/fdr-science-biblatex.pdf

observables and analyses.

6. Summary and Discussion

Current standard LSS analyses makes a major assumption in their parameter inference — the likelihood has a Gaussian functional-form. Although this assumption is motivated by the Central Limit Theorem, in detail the assumption cannot be true. In this paper, we investigate the impact of this assumption on two recent LSS analyses: the B2017 power spectrum multipole ($\ell = 0, 2$, and 4) analysis and the S2017 group multiplicity function analysis. We combine mock catalogs, originally constructed for covariance matrix estimation in these analyses, with non-parametric divergence estimators, used in Machine Learning, to measure the divergences between the P_ℓ and ζ likelihoods and the Gaussian pseudo-likelihoods from B2017 and S2017. For both the P_ℓ and ζ likelihoods, the divergences reveal significant likelihood non-Gaussianity. For the P_ℓ likelihood, large scales (low k) and the hexadecapole contribute most to the relatively small non-Gaussianity. Meanwhile, for the ζ likelihood, the high richness end of ζ contribute most to the non-Gaussianity.

From the same mock catalogs of B2017 and S2017, we use non-parametric density estimation to construct estimates of the likelihood distribution that better approximates the true non-Gaussian P_ℓ and ζ likelihoods. For the ζ likelihood, the direct method of estimating the likelihood distribution with the mocks and Gaussian mixture density estimation results in an accurate estimate. For the B2017 P_ℓ analysis, which has fewer mocks and a higher dimensional likelihood, we use Independent Component Analysis to transform the likelihood distribution into statistically independent components. By estimating the one-dimensional distribution of these independent components, we derive an estimate of the high-dimensional likelihood distribution. The divergence between our two likelihood estimates and the P_ℓ and ζ likelihoods demonstrate that we derive *more accurate* estimates of the true likelihoods than the pseudo-likelihoods.

Finally, using the P_ℓ and ζ likelihood estimates and importance sampling, we derive posterior parameter constraint for the B2017 and S2017 analyses. Since our likelihood estimates are closer to the true likelihood, they allow us to derive more accurate posterior probability distributions than with Gaussian pseudo-likelihoods. By comparing our posteriors to the parameter constraints from B2017 and S2017, we find that likelihood non-Gaussianity does not significantly impact the P_ℓ analysis of B2017. Among the non-nuisance parameters, only $f\sigma_8$ is impacted by 0.44σ. Meanwhile for the S2017 ζ analysis, likelihood non-Gaussianity significant impacts the posterior distributions of the HOD parameter. Using the pseudo-likelihood significantly underestimates the width of the $\log M_{\min}$, $\log M_1$, and α posteriors. It also significantly biases their constraints. Accounting for likelihood non-Gaussianity, significantly eases the tension between the ζ and $w_p(r_p)$ constraints found in S2017. Our comparisons of the posteriors highlight the importance of incorporating likelihood non-Gaussianity in

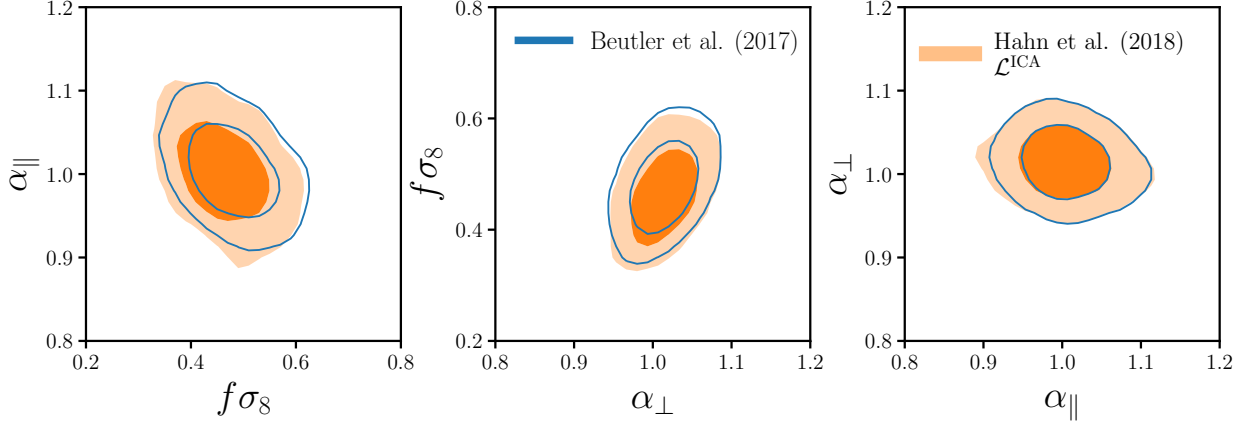Fig. 5.— Joint posterior distributions of $f\sigma_8$, $\alpha_\parallel$, and $\alpha_\perp$ in the B2017 P_ℓ analysis, compute using the non-Gaussian ICA likelihood (orange). We include, in blue, the original B2017 posteriors for comparison. The contours in the left and middle panels reflect the shift in $f\sigma_8$ caused by likelihood non-Gaussianity. Otherwise, the contours illustrate that likelihood non-Gaussianity has little impact on the cosmological parameters for the P_ℓ analysis.

Table 1: Impact of likelihood non-Gaussianity on the posterior parameter constraints of B2017 and S2017.

	B2017 P_ℓ analysis			
	$b_1^{\rm NGC}\sigma_8$	$b_1^{\rm SGC}\sigma_8$	$b_2^{\rm NGC}\sigma_8$	$b_2^{\rm SGC}\sigma_8$
	$f\sigma_8$	$\alpha_\parallel$	$\alpha_\perp$	
B2017	$1.341^{+0.040}_{-0.042}$	$1.333^{+0.056}_{-0.062}$	$1.293^{+0.697}_{-0.752}$	$0.476^{+1.262}_{-1.175}$
	$0.478^{+0.053}_{-0.049}$	$1.003^{+0.038}_{-0.032}$	$1.014^{+0.025}_{-0.025}$	
non-Gaussian $\mathcal{L}_{\rm ICA}$	$1.351^{+0.040}_{-0.049}$	$1.335^{+0.063}_{-0.069}$	$1.295^{+0.746}_{-0.798}$	$0.422^{+1.517}_{-1.377}$
	$0.456^{+0.059}_{-0.049}$	$1.001^{+0.039}_{-0.035}$	$1.014^{+0.024}_{-0.025}$	

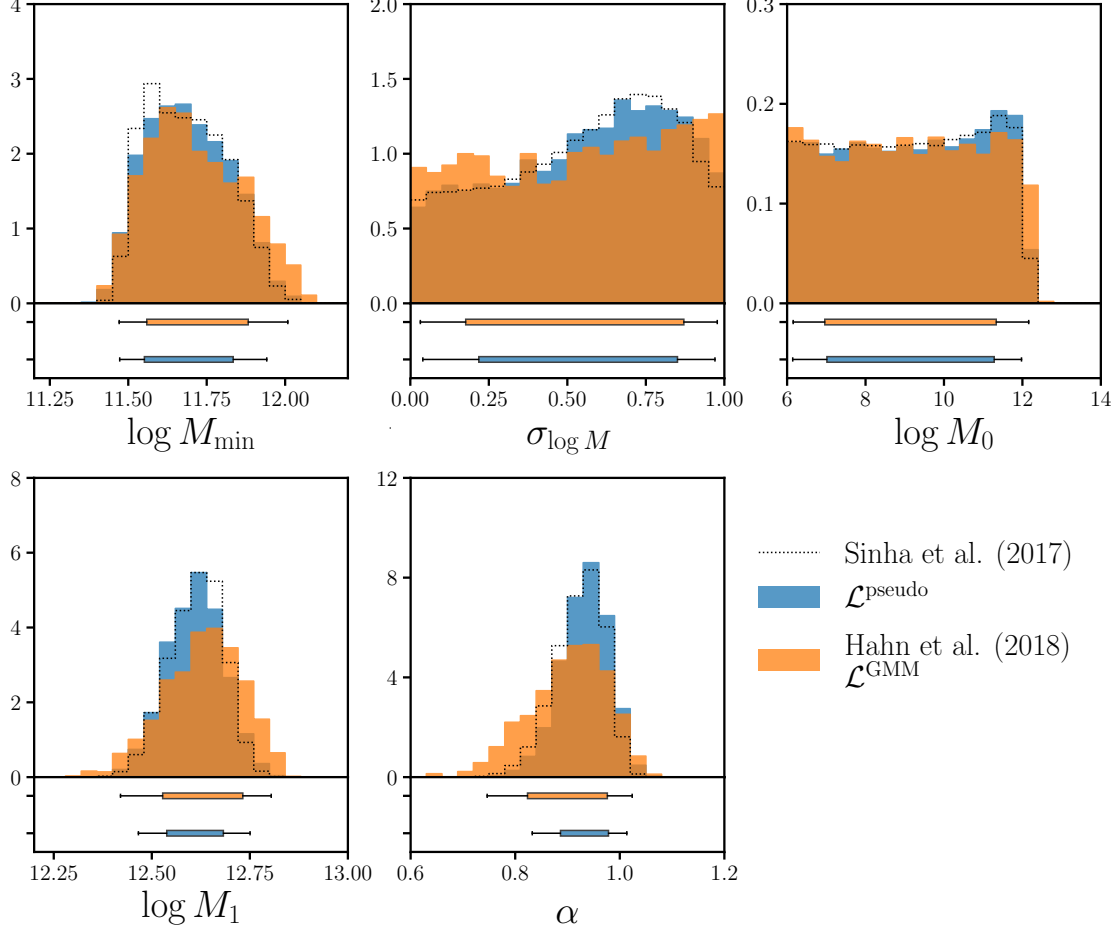|	S2017 ζ analysis				
	$\log M_{\rm min}$	$\sigma_{\log M}$	$\log M_0$	$\log M_1$	α
S2017	$11.68^{+0.148}_{-0.128}$	$0.585^{+0.255}_{-0.367}$	$9.154^{+2.074}_{-2.162}$	$12.62^{+0.064}_{-0.077}$	$0.928^{+0.042}_{-0.054}$
Gaussian $\mathcal{L}_{\rm pseudo}$	$11.68^{+0.152}_{-0.131}$	$0.586^{+0.264}_{-0.369}$	$9.195^{+2.086}_{-2.180}$	$12.61^{+0.070}_{-0.074}$	$0.936^{+0.043}_{-0.049}$
non-Gaussian $\mathcal{L}_{\rm GMM}$	$11.69^{+0.188}_{-0.135}$	$0.554^{+0.317}_{-0.378}$	$9.159^{+2.174}_{-2.198}$	$12.64^{+0.095}_{-0.109}$	$0.909^{+0.067}_{-0.086}$

Fig. 6.— The posterior distribution for HOD parameters log $M_{\rm min}$, $\sigma_{\log M}$, log M_0, log M_1, and α in the S2017 ζ analysis using the non-Gaussian GMM likelihood (orange). We include in blue the posteriors calculated from the pseudo-likelihood for comparison. We also include the original S2017 posterior (dotted; see text for details). On the bottom of each panel we include box plots that mark the 68% and 95% confidence intervals of the posterior. Besides the poorly constrained parameters $\sigma_{\log M}$ and log M_0, the posteriors of log $M_{\rm min}$, log M_1, and α, are significantly broader and shifted compared to the pseudo-likelihood constraints. Likelihood non-Gaussianity significantly impacts the parameter constraints of the *zeta* analysis. Therefore using the pseudo-likelihood underestimates the uncertainty and biases the HOD parameter constraints.
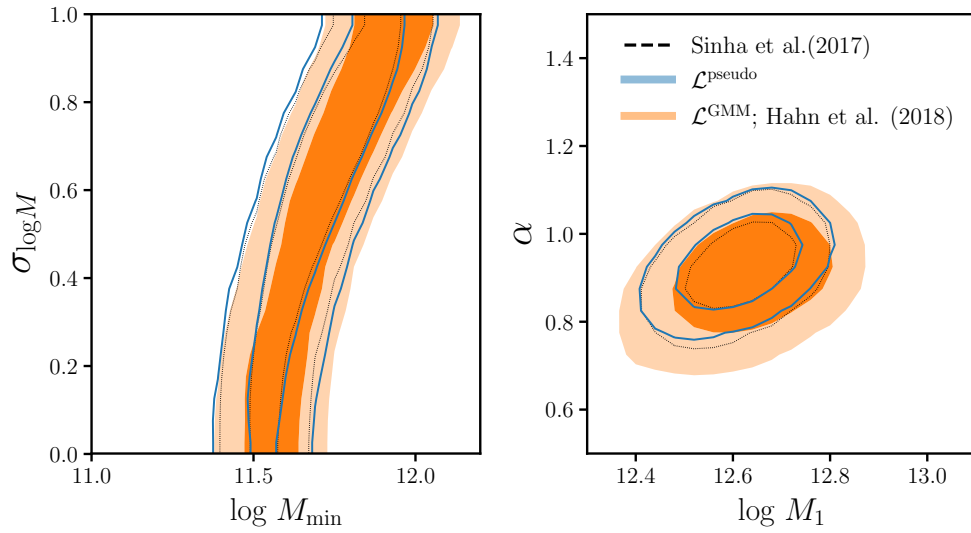
Fig. 7.— Joint posterior distributions of select HOD parameters in the S2017 ζ analysis, compute using the non-Gaussian GMM likelihood (orange). We include, in blue, the posteriors computed using the pseudo-likelihood; we also include the original S2017 posterior (dotted; see text for details). The contours confirm that that *due to likelihood non-Gaussianity, posteriors from the pseudo-likelihood underestimate the uncertainties and significantly biases the parameter constraints of the S2017 analysis.*

parameter inference of LSS studies.

Based on our results, P_ℓ analyses using future surveys (*e.g.* DESI) will not be significantly impacted by likelihood non-Gaussianity. While these analyses seek to extend the k ranges probed, future surveys will also expand the cosmic volumes probed by galaxies and increase the number of modes on all scales. We therefore expect Central Limit Theorem to reduce the impact of likelihood non-Gaussianity. Likewise, we expect future surveys to also reduce the impact of likelihood non-Gaussianity on ζ analysis with the same multiplicity range, since larger cosmic volumes will probe more high multiplicity groups. For a wider multiplicity range, however, likelihood non-Gaussianity may still be a significant effect. For higher order statistics such as the galaxy bispectrum or three-point function, even for future surveys, likelihood non-Gaussianity will likely be a significant effect to consider for parameter inference. Regardless of our expectation, for ever more precise parameter inference the Gaussian likelihood assumption must be extensively tested. The divergence and likelihood estimations we introduce in this paper provide a straightforward framework for testing and quantifying the impact of likelihood non-Gaussianity on the final parameter constraints.

Our likelihood estimation methods also allow us to go beyond the pseudo-likelihood and derive more accurate estimates of the likelihood. With a similar motivation at addressing likelihoods that are non-Gaussian or difficult to write down, methods for likelihood-free inference such as Approximate Bayesian Computation (ABC; Hahn et al. 2017b; Kacprzak et al. 2017; Alsing et al. 2018) have recently been introduced to LSS studies. Although as a likelihood-free inference method, ABC has the advantage of relaxing *any* assumption on the likelihood, even with smart sampling methods like Population Monte Carlo, it requires an expensive generative forward model to be computed far more times than the number of mocks required for covariance matrix estimation. Our methods (especially the ICA method) do not require any more mocks than those already constructed for accurate covariance matrix estimation. Therefore, the methods for likelihood estimation we present in this paper provide both accurate and practical methods for Bayesian parameter inference in LSS.

REFERENCES

Ade, P. a. R., Aghanim, N., Armitage-Caplan, C., et al. 2014, Astronomy & Astrophysics, 571, A15

Ade, P. a. R., Aghanim, N., Arnaud, M., et al. 2016, Astronomy & Astrophysics, 594, A24

Aghanim, N., Arnaud, M., Ashdown, M., et al. 2016, Astronomy & Astrophysics, 594, A11

Alam, S., Ata, M., Bailey, S., et al. 2017, Monthly Notices of the Royal Astronomical Society, 470, 2617

Alsing, J., Wandelt, B., & Feeney, S. 2018, arXiv:1801.01497 [astro-ph], arXiv:1801.01497 [astro-ph]

Arthur, D., & Vassilvitskii, S. 2007, in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07 (Philadelphia, PA, USA: Society for Industrial and Applied Mathematics), 1027

Berlind, A. A., Frieman, J., Weinberg, D. H., et al. 2006, The Astrophysical Journal Supplement Series, 167, 1

Beutler, F., Seo, H.-J., Saito, S., et al. 2017, Monthly Notices of the Royal Astronomical Society, 466, 2242

Bianchi, D., Gil-Marín, H., Ruggeri, R., & Percival, W. J. 2015, Monthly Notices of the Royal Astronomical Society, 453, L11

Bovy, J., Hogg, D. W., & Roweis, S. T. 2011, The Annals of Applied Statistics, 5, 1657

Broderick, A. E., Fish, V. L., Doeleman, S. S., & Loeb, A. 2011, The Astrophysical Journal, 735, 110

Cash, W. 1979, The Astrophysical Journal, 228, 939

Collaboration, P., Ade, P. A. R., Aghanim, N., et al. 2014, Astronomy & Astrophysics, 571, A20

Comon, P. 1994, Signal Processing, 36, 287

Crocce, M., Pueblas, S., & Scoccimarro, R. 2006, Monthly Notices of the Royal Astronomical Society, 373, 369

Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, The Astrophysical Journal, 292, 371

Davison, A. C. 2008, Statistical Models (Cambridge Series in Statistical and Probabilistic Mathematics) (Cambridge University Press)

Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977, Journal of the Royal Statistical Society. Series B (Methodological), 39, 1

Efstathiou, G. 2004, Monthly Notices of the Royal Astronomical Society, 349, 603

—. 2006, Monthly Notices of the Royal Astronomical Society, 370, 343

Eifler, T., Schneider, P., & Hartlap, J. 2009, Astronomy and Astrophysics, 502, 721

Eisenstein, D. J., & Zaldarriaga, M. 2001, The Astrophysical Journal, 546, 2

Feigelson, E. D., & Babu, G. J. 2012, Modern Statistical Methods for Astronomy

Fraley, C., & Raftery, A. E. 1998, The Computer Journal, 41, 578

Gardner, J. P., Connolly, A., & McBride, C. 2007, in Astronomical Data Analysis Software and Systems XVI, Vol. 376, 69

Gaztañaga, E., & Scoccimarro, R. 2005, Monthly Notices of the Royal Astronomical Society, 361, 824

Grieb, J. N., Sánchez, A. G., Salazar-Albornoz, S., et al. 2017, Monthly Notices of the Royal Astronomical Society, 467, 2085

Guo, H., Zehavi, I., & Zheng, Z. 2012, The Astrophysical Journal, 756, 127

Hahn, C., Scoccimarro, R., Blanton, M. R., Tinker, J. L., & Rodríguez-Torres, S. A. 2017a, Monthly Notices of the Royal Astronomical Society, 467, 1940

Hahn, C., Vakili, M., Walsh, K., et al. 2017b, Monthly Notices of the Royal Astronomical Society, 469, 2791

Hand, N., Feng, Y., Beutler, F., et al. 2017a

Hand, N., Li, Y., Slepian, Z., & Seljak, U. 2017b, Journal of Cosmology and Astro-Particle Physics, 07, 002

Hartlap, J., Schrabback, T., Simon, P., & Schneider, P. 2009, Astronomy and Astrophysics, 504, 689

Hastie, T., Tibshirani, R., & Friedman, J. 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) (Springer)

Hérault, J., & Ans, B. 1984, Comptes Rendus de l'Académie des Sciences Paris, Série III, Life Sciences, 299, 525

Hogg, D. W., Bovy, J., & Lang, D. 2010, ArXiv e-prints, 1008, arXiv:1008.4686

Hu, W., & White, M. 2001, The Astrophysical Journal, 554, 67

Hyvärinen, A. 1998, in Advances in Neural Information Processing Systems 10, ed. M. I. Jordan, M. J. Kearns, & S. A. Solla (MIT Press), 273

Hyvarinen, A. 1999, IEEE Transactions on Neural Networks, 10, 626

—. 2001, Independent Component Analysis (New York: J. Wiley)

Hyvärinen, A., & Oja, E. 2000, Neural Networks: The Official Journal of the International Neural Network Society, 13, 411

Kacprzak, T., Herbel, J., Amara, A., & Réfrégier, A. 2017, arXiv:1707.07498 [astro-ph], arXiv:1707.07498 [astro-ph]

Kazin, E. A., Koda, J., Blake, C., et al. 2014, Monthly Notices of the Royal Astronomical Society, 441, 3524

Kitaura, F.-S., Gil-Marín, H., Scóccola, C. G., et al. 2015, Monthly Notices of the Royal Astronomical Society, 450, 1836

Kitaura, F.-S., & Heß, S. 2013, Monthly Notices of the Royal Astronomical Society, 435, L78

Kitaura, F.-S., Yepes, G., & Prada, F. 2014, Monthly Notices of the Royal Astronomical Society, 439, L21

Kitaura, F.-S., Rodríguez-Torres, S., Chuang, C.-H., et al. 2016, Monthly Notices of the

Royal Astronomical Society, 456, 4156

Klypin, A., Yepes, G., Gottlöber, S., Prada, F., & Heß, S. 2016, Monthly Notices of the Royal Astronomical Society, 457, 4340

Krishnamurthy, A., Kandasamy, K., Poczos, B., & Wasserman, L. 2014, arXiv:1402.2966 [math, stat], arXiv:1402.2966 [math, stat]

Kuhn, M. A., & Feigelson, E. D. 2017, arXiv:1711.11101 [astro-ph, stat], arXiv:1711.11101 [astro-ph, stat]

Lee, K. J., Guillemot, L., Yue, Y. L., Kramer, M., & Champion, D. J. 2012, Monthly Notices of the Royal Astronomical Society, 424, 2832

Leroux, B. G. 1992, The Annals of Statistics, 20, 1350

Levi, M., Bebek, C., Beers, T., et al. 2013, arXiv:1308.0847 [astro-ph], arXiv:1308.0847 [astro-ph]

Liddle, A. R. 2007, Monthly Notices of the Royal Astronomical Society, 377, L74

Lloyd, S. 1982, IEEE Transactions on Information Theory, 28, 129

McBride, C., Berlind, A., Scoccimarro, R., et al. 2009, in Bulletin of the American Astronomical Society, Vol. 213, 425.06

McLachlan, G., & Peel, D. 2000, Finite Mixture Models (Wiley-Interscience)

Mohammed, I., Seljak, U., & Vlah, Z. 2017, Monthly Notices of the Royal Astronomical Society, 466, 780

Morrison, C. B., & Schneider, M. D. 2013, Journal of Cosmology and Astro-Particle Physics, 11, 009

Neal, R. M., & Hinton, G. E. 1998, in Learning in Graphical Models, NATO ASI Series (Springer, Dordrecht), 355

Norberg, P., Baugh, C. M., Gaztañaga, E., & Croton, D. J. 2009, Monthly Notices of the Royal Astronomical Society, 396, 19

Ntampaka, M., Trac, H., Sutherland, D. J., et al. 2015, The Astrophysical Journal, 803, 50

—. 2016, The Astrophysical Journal, 831, 135

O'Connell, R., Eisenstein, D., Vargas, M., Ho, S., & Padmanabhan, N. 2016, Monthly Notices of the Royal Astronomical Society, 462, 2681

Parkinson, D., Riemer-Sørensen, S., Blake, C., et al. 2012, Physical Review D, 86, 103518

Pinol, L., Cahn, R. N., Hand, N., Seljak, U., & White, M. 2017, Journal of Cosmology and Astroparticle Physics, 2017, 008

Póczos, B., Szabó, Z., & Schneider, J. 2011, in 2011 19th European Signal Processing Conference, 1718

Poczos, B., Xiong, L., & Schneider, J. 2012, arXiv:1202.3758 [cs, stat], arXiv:1202.3758 [cs, stat]

Póczos, B., Xiong, L., Sutherland, D. J., & Schneider, J. 2012, in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2989

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, Numerical Recipes in C (2Nd Ed.): The Art of Scientific Computing (New York, NY, USA: Cambridge University Press)

Ravanbakhsh, S., Lanusse, F., Mandelbaum, R., Schneider, J., & Poczos, B. 2017, in Thirty-First AAAI Conference on Artificial Intelligence

Rodríguez-Torres, S. A., Chuang, C.-H., Prada, F., et al. 2016, Monthly Notices of the Royal Astronomical Society, 460, 1173

Roeder, K., & Wasserman, L. 1997, Journal of the American Statistical Association, 92, 894

Ross, A. J., Beutler, F., Chuang, C.-H., et al. 2017, Monthly Notices of the Royal Astronomical Society, 464, 1168

Schwarz, G. 1978, The Annals of Statistics, 6, 461

Scoccimarro, R. 1998, Monthly Notices of the Royal Astronomical Society, 299, 1097

—. 2000, The Astrophysical Journal, 544, 597

—. 2015, Physical Review D, 92, arXiv:1506.02729

Scoccimarro, R., Couchman, H. M. P., & Frieman, J. A. 1999, The Astrophysical Journal, 517, 531

Scott, D. W. 1992, Multivariate Density Estimation: Theory, Practice, and Visualization (Wiley)

Sellentin, E., & Heavens, A. F. 2016, Monthly Notices of the Royal Astronomical Society, 456, L132

—. 2017, arXiv:1707.04488 [astro-ph], arXiv:1707.04488 [astro-ph]

Sinha, M., Berlind, A. A., McBride, C. K., et al. 2017, arXiv:1708.04892 [astro-ph], arXiv:1708.04892 [astro-ph]

Slepian, Z., Eisenstein, D. J., Brownstein, J. R., et al. 2017, Monthly Notices of the Royal Astronomical Society, 469, 1738

Springel, V. 2005, Monthly Notices of the Royal Astronomical Society, 364, 1105

Steele, R. J., & Raftery, A. E. 2010

Taylor, E. N., Hopkins, A. M., Baldry, I. K., et al. 2015, Monthly Notices of the Royal Astronomical Society, 446, 2144

Tinker, J. L., & et al. in preparation

Vakili, M., & Hahn, C. H. 2016, arXiv:1610.01991 [astro-ph], arXiv:1610.01991 [astro-ph]

Vargas-Magaña, M., Ho, S., Xu, X., et al. 2014, Monthly Notices of the Royal Astronomical Society, 445, 2

Wang, Q., Sanjeev, K., & Sergio, V. 2009, IEEE TRANSACTIONS ON INFORMATION THEORY, 55, 2392

Wasserman, L. 2004, All of Statistics: A Concise Course in Statistical Inference (Springer Texts in Statistics) (Springer)

White, M., & Padmanabhan, N. 2015, Journal of Cosmology and Astro-Particle Physics,

12, 058

Wilkinson, D. M., Maraston, C., Thomas, D., et al. 2015, Monthly Notices of the Royal Astronomical Society, 449, 328

Wu, C. F. J. 1983, The Annals of Statistics, 11, 95

Xu, X., Ho, S., Trac, H., et al. 2013, The Astrophysical Journal, 772, 147

Zhao, C., Kitaura, F.-S., Chuang, C.-H., et al. 2015, Monthly Notices of the Royal Astronomical Society, 451, 4266

Zheng, Z., & Weinberg, D. H. 2007, The Astrophysical Journal, 659, 1