How I Learned to Stop Worrying and Love The Central Limit Theorem

ChangHoon Hahn, Florian Beutler, Manodeep Sinha, Andreas Berlind

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

changhoon.hahn@lbl.gov

DRAFT --- 877aac8 --- 2017-11-27 --- NOT READY FOR DISTRIBUTION

ABSTRACT

abstract here

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

1. Introduction

- Talk about the use of Bayesian parameter inference and getting the posterior in LSS cosmology
- Explain the two major assumptions that go into evaluating the likelihood
- Emphasize that we are not talking about non-Gaussian contributions to the likelihood
- Emphasize the scope of this paper is to address whether one of the assumptions matters for galaxy clustering analyses.

2. Gaussian Likelihood Assumption

- Depending on Hogg's paper maybe a simple illustration of how the likelihood asumption

3. Mock Catalogs

Mock catalogs play a key role in standard cosmological anslyses of LSS studies. They're extensively used for testing analysis pipelines ((mock challenege papers) Beutler et al. 2017), testing the effect of systematics (Hahn et al. 2017), and, most relevantly for this paper, for estimating the covariance matrix used in evaluating the likelihood for parameter inference (cite the bunch of other papers Beutler et al. 2017).

Maybe a little paragraph about the advantages of covariance matrices from mocks versus analytic?

Our primary goal in this paper is to test the Gaussian likelihood assumption in two LSS analyses: the powerspectrum multipole full shape analysis of Beutler et al. (2017) and group multiplicity function analysis of Sinha et al. (2017). Throughout the paper we will make extensive use of the mock catalogs used in these analyses. Below, in this section, we give a brief description of these mocks.

3.1. MultiDark-PATCHY Mock Catalog

In their powerspectrum multipole full shape analysis, Beutler et al. (2017) use the MultiDark-PATCHY mock catalogs from Kitaura et al. (2016).

The dark matter fields for these mocks are generated using approxiate gravity solvers on a mesh.

something about the approximate gravity solver These mock catalogs have been calibrated to high fidelity BigMultiDark simulations (Rodríguez-Torres et al. 2016; **?**)

In total Beutler et al. (2017)

- 2048 mocks for both north and south
- Note that 3 mocks were fixed

3.2. Sinha et al. (2017) Mock Catalog

3.3.

$\mathbf{X}_i = [P_0(k)_i, P_2(k)_i, P_4(k)_i]$
$\mathbf{X}_i = \zeta(N)_i$
$\mathbf{X} = \{\mathbf{X}_i\}$

4. Quantifying the Likelihood non-Gaussianity

In Sellentin et al. (2017) **discuss how sellentin and hartlap's stuff is an attempt to quantify the divergence**

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 $\mathbf{C}$. The following are two of the most commonly used divergences: the Kullback-Leibler (hereafter KL) divergence

$$D_{KL}(p \parallel q) = \int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x \tag{1}$$

and the Rényi-α divergence

$$D_{R-\alpha}(p \parallel q) = \frac{1}{\alpha - 1} \log \int p^{\alpha}(x) q^{1-\alpha}(x)\, \mathrm{d}x. \tag{2}$$

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 distribution 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 estimators (**?**Póczos et al. 2012; Krishnamurthy et al. 2014). These estimators, allows 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: $\hat{D}_{\alpha}(X_{1:n} \parallel Y_{1:m})$.

For instance, the estimator presented in Póczos et al. (2012) allows us to estimate the kernel function of the Rényi-α divergence,

$$D_{\alpha}(p \parallel q) = \int p^{\alpha}(x) q^{1-\alpha}(x)\, \mathrm{d}x. \tag{3}$$

using kth 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 $D_{\alpha}(p \parallel q)$ can be estimated as

$$\hat{D}_{\alpha}(X_{1:n} \parallel Y_{1:m}) = \frac{B_{k,\alpha}}{n} \left(\frac{n-1}{m}\right)^{1-\alpha} \sum_{i=1}^{n} \left(\frac{\rho_k^d(X_i)}{\nu_k^d(X_i)}\right)^{1-\alpha}, \tag{4}$$

where $B_{k,\alpha} = \frac{\Gamma(k)^2}{\Gamma(k-\alpha+1)\Gamma(k+\alpha-1)}$. Póczos et al. (2012) goes to further prove that this estimated kernel function is asymptotically unbias,

$$\lim_{n,m\to\infty} \mathbb{E}[\hat{D}_{\alpha}(X_{1:n} \parallel Y_{1:m})] = D_{\alpha}(p \parallel q). \tag{5}$$

Plugging $\hat{D}_{\alpha}(X_{1:n} \parallel Y_{1:m})$ into Eq. 2, we get an estimator for the Rényi-α divergence. A similar estimator (**?**) can also be derived for the KL divergence (Eq. 1). We note that while the divergence estimators converge to the true divergence with a large enough samples, with a limited number of samples from the distribution, the estimators are noisy.

These divergence estimates have been applied to Support Vector Machines and used extensively in the machine learning and astronomical literature with great success **elaborate a lot more**

- Compile papers that use this divergence, Ntampaka et al. (2015, 2016)

For more details on the non-parametric divergence estimators, we refer readers to Póczos et al. (2012) and Krishnamurthy et al. (2014).

Now we can use the divergence estimators above to quantify the non-Gaussianity of the likelihood. More specifically, we're intersted in the divergence between the distribution $p(x)$ sampled by the mock observables ($\mathbf{X}^{\rm mock}$) and the multivariate Gaussian distribution assumed in standard analyses, which is solely described by the mean and covariance calculated from the mocks ($\mathcal{N}(\overline{\mathbf{X}^{\rm mock}}, \mathbf{C})$. Since $\mathbf{X}^{\rm mock}$ is a sample from $p(x)$, we draw a reference sample $\mathbf{Y}^{\rm ref}$ from $\mathcal{N}(\overline{\mathbf{X}^{\rm mock}}, \mathbf{C}$ to use in the estimators. Similar to the experiments detailed in Póczos et al. (2012), we construct $\mathbf{Y}^{\rm ref}$ with a comparable sample size as $\mathbf{X}^{\rm mock}$: 2000 and $10,000$ for the P_ℓ and ζ analyses respectively.

In Figure 2 we compare the distribution of Rényi-α (left) and KL (right) divergence estimates (orange) $\hat{D}_{R\alpha}$ and $\hat{D}_{KL}$ between the mock data $\mathbf{X}^{\rm mock}$ and a reference sample $\mathbf{Y}^{\rm ref}$ for the $P_\ell(k)$ (top) and $\zeta(N)$ (bottom) analyses. As a reference point for the comparison, we include in Figure 2 (blue) the distribution of Rényi-α and KL divergence estimates if $\mathbf{X}^{\rm mock}$ were actually sampled from $\mathcal{N}(\overline{\mathbf{X}^{\rm mock}}, \mathbf{C}$ and $\mathbf{Y}^{\rm ref}$. All distributions were constructed using 100 divergence estimates.

The discrepancy between the blue and orange distributions illustrates the non-Gaussianity of $p(x)$ sampled by $\mathbf{X}^{\rm mock}$.

- Describe the discrepancy.

5. A More Accurate Likelihood

- Gaussian Mixture Model
- expectation maximization algorithm **?**
- Bayesian Information Criteria
- Figure illustrating both methods on highest N GMF bin

5.1. Independent Component Analysis

Curse of dimensionality! 2048 mocks in Beutler not enough to directly estimate the 37-dimensional space, so we use independent component analysis Hartlap et al. (2009)

- equation
- Similar figure to Hartlap et al. (2009) that tests the independence?
- Kernel Density Estimation
-

Figure 2

6. Impact on Parameter Inference

6.1. MCMC

- details of each of the MCMC runs

6.2. Importance Sampling

- equations explaining importance sampling framework

Figure 3

Figure 4

7. Discussion

- Will it matter for future surveys?
- Likelihood free inference (cite justin's paper)

8. Summary

Acknowledgements

It's a pleasure to thank Simone Ferraro, David W. Hogg, Emmaneul Schaan, Roman Scoccimarro Zachary Slepian

REFERENCES

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

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

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

Kitaura, F.-S., Rodríguez-Torres, S., Chuang, C.-H., et al. 2016, Monthly Notices of the Royal Astronomical Society, 456, 4156

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

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

—. 2016, The Astrophysical Journal, 831, 135

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

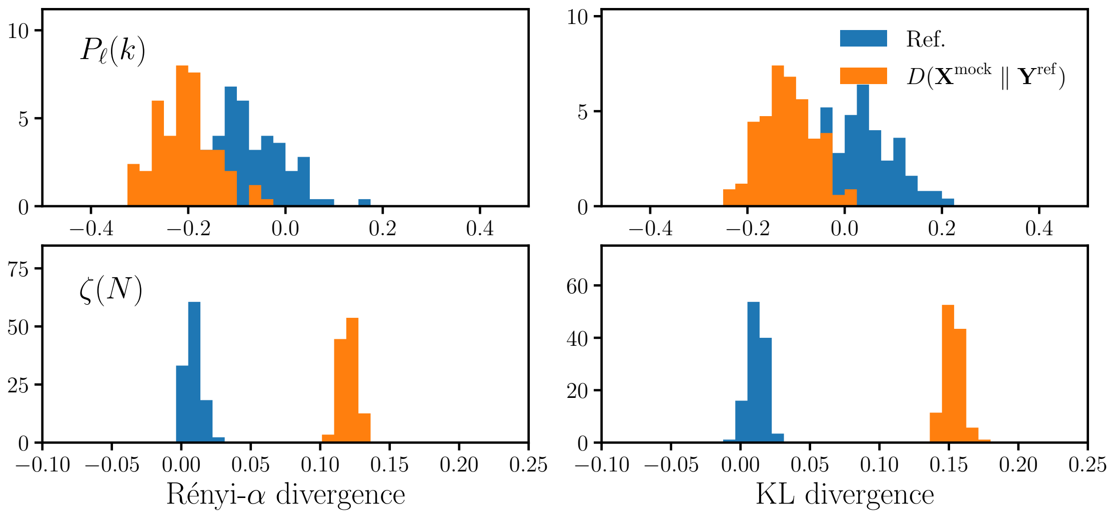

Fig. 1.— Rényi-α and KL divergence estimates, ($D_{R\alpha}$ and D_{KL}), between the mock data $\mathbf{X}^{\rm mock}$ and a reference sample $\mathbf{Y}^{\rm ref}$ for the $P_\ell(k)$ (left) and $\zeta(N)$ (right) analyses.

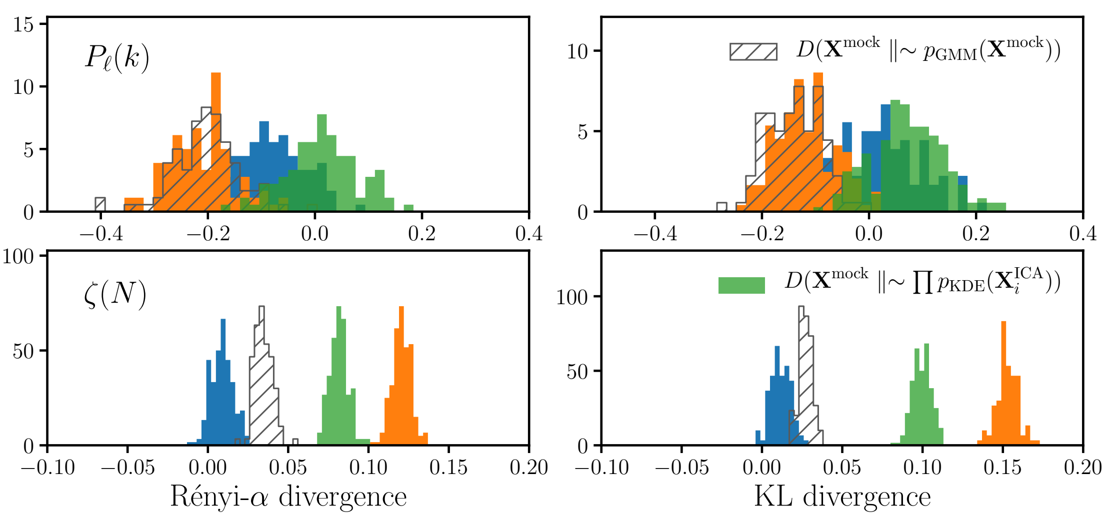

Fig. 2.—

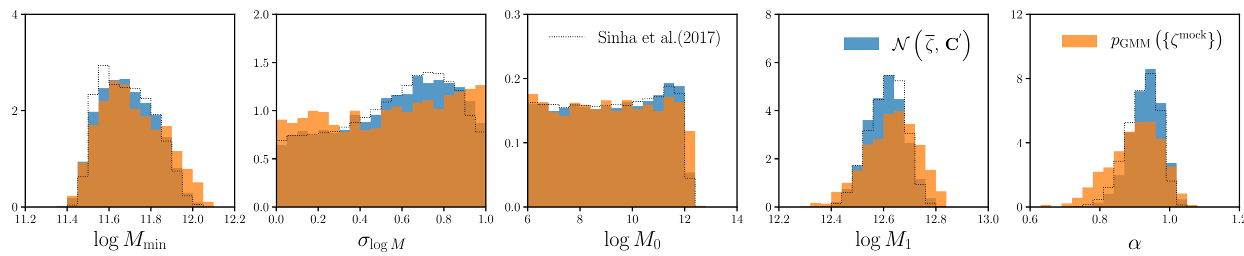
Sinha et al.(2017)
$\mathcal{N}\left(\bar{\zeta}, \mathbf{C}'\right)$
$p_{\rm GMM}\left(\{\zeta^{\rm mock}\}\right)$
$\log M_{\rm min}$
$\sigma_{\log M}$
$\log M_0$
$\log M_1$
α

Fig. 3.—

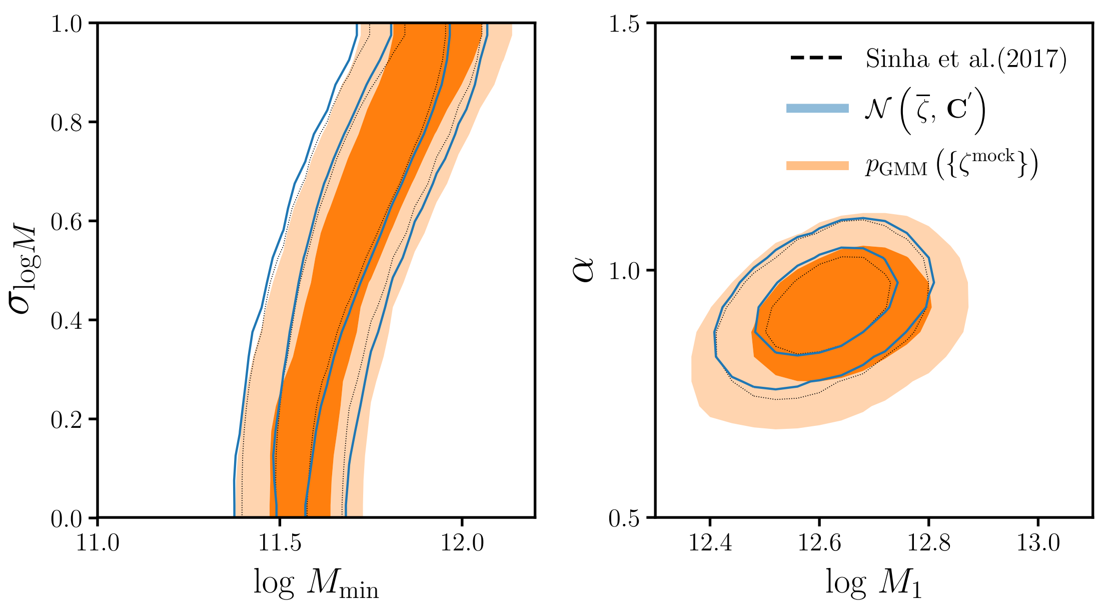
Sinha et al.(2017)
$\mathcal{N}\left(\bar{\zeta}, \mathbf{C}'\right)$
$p_{\rm GMM}\left(\{\zeta^{\rm mock}\}\right)$
$\sigma_{\log M}$
$\log M_{\rm min}$
α
$\log M_1$

Fig. 4.—

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

Sellentin, E., Jaffe, A. H., & Heavens, A. F. 2017, arXiv:1709.03452 [astro-ph, stat], arXiv:1709.03452 [astro-ph, stat]

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

This preprint was prepared with the AAS LATEX macros v5.2.