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 --- 0d551fe --- 2017-11-23 --- 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.

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

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**

One way of quantifying the non-Gaussianity of the likelihood is to calculate the divergence between the distribution $p(x)$, which generated the observable, and $q(x)$ a multivariate Gaussian described by the average of the mocks and the covariance matrix $\mathbf{C}$.

One of the most commonly used divergence is the Kullback-Leibler (hereafter KL) divergence:

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

Another is the Rényi-α divergence between $p(x)$ and $q(x)$:

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

We note that in the limit as α approaches 1, the Rényi-α divergence is KL divergence. To evaluate the Rényi-α divergence, we need to evaluate the following kernel function

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

Of course, in our case, we do not know $p(x)$. 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 by using nonparametric estimators (**?**Póczos et al. 2012; Krishnamurthy et al. 2014). The estimator presented in Póczos et al. (2012), which is based on kth nearest neighbor density estimators, allows us to estimate the kernel function Eq. 3 directly from samples $X_{1:n} = \{X_1, ...X_n\}$ and $Y_{1:m} = \{Y_1, ...Y_m\}$ drawn from p and q respectively: $\hat{D}_\alpha(X_{1:n} \parallel Y_{1:m})$.

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) proves that this estimator is asymptotically unbias, *i.e.*

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

Plugging $\hat{D}_\alpha(X_{1:n} \parallel Y_{1:m})$ into Eq. 2, we can estimate the Rényi-α divergence. A similar estimator (**?**) can also be used to estimate the KL divergence (Eq. 1). 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).

In Figure 2 we compare the Rényi-α (top) and KL (bottom) divergence estimates $D_{R\alpha}(\mathbf{X}^{\rm mock} \parallel \mathbf{Y}^{\rm ref})$ and $D_{KL}(\mathbf{X}^{\rm mock} \parallel \mathbf{Y}^{\rm ref})$ 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. $\mathbf{Y}^{\rm ref}$ is drawn from a multivariate Gaussian distribution described by the covariance matrix $\mathbf{C}$ — $\mathbf{Y}^{\rm ref} \sim \mathcal{N}(\mathbf{C})$.

- Describe the discrepancy.

- Figure that compare $D_{R\alpha}(\mathbf{X}^{\rm mock} \parallel \mathbf{Y}^{\rm ref})$

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

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.

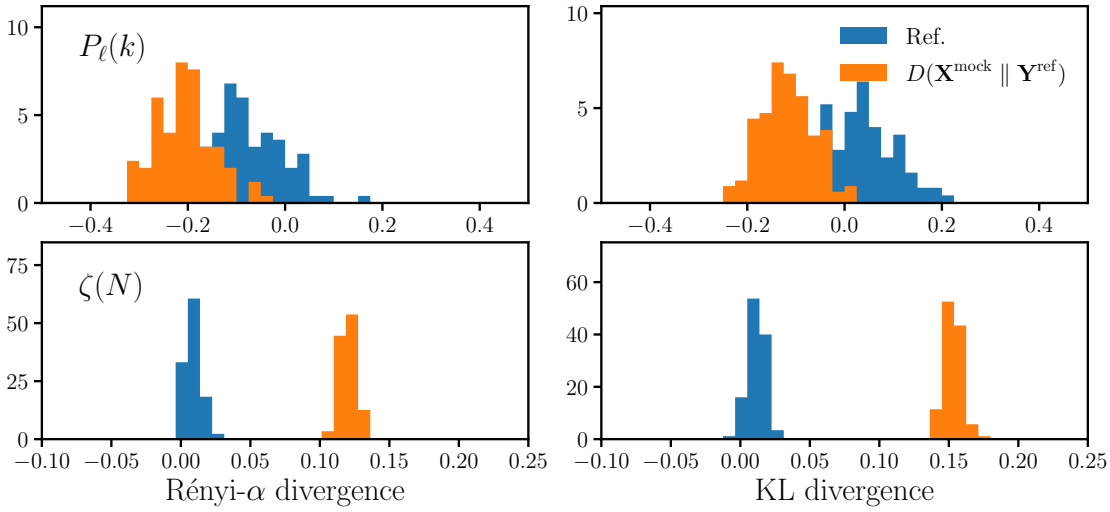

Fig. 1.— Rényi-α and KL divergence estimates, ($D_{R\alpha}$ and D_{KL}), between the mock data $\mathbf{X}^{\rm mock}$ and a reference sample $\mathbf{Y}^{\rm ref}$ for the $P_\ell(k)$ (left) and $\zeta(N)$ (right) analyses.

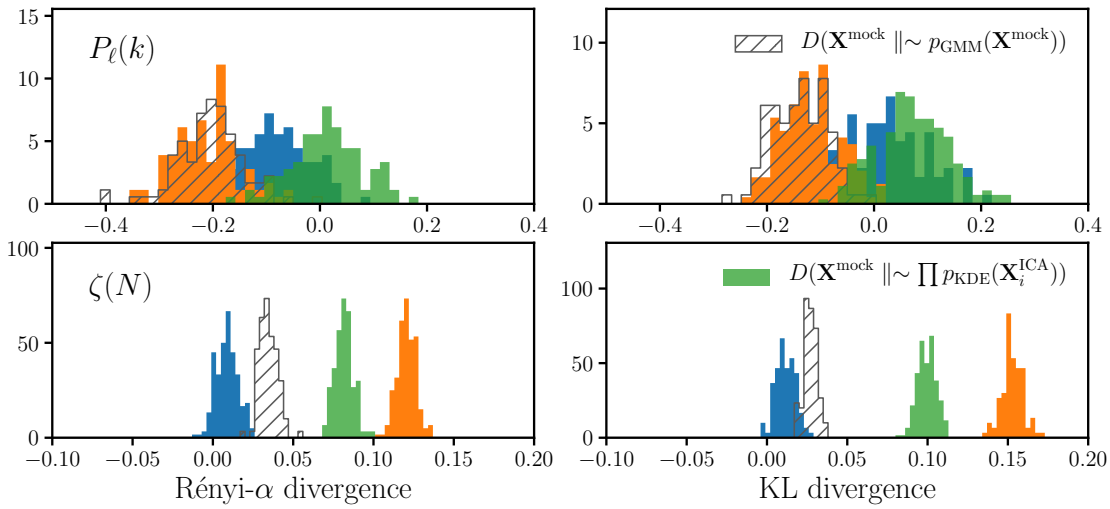

Fig. 2.—

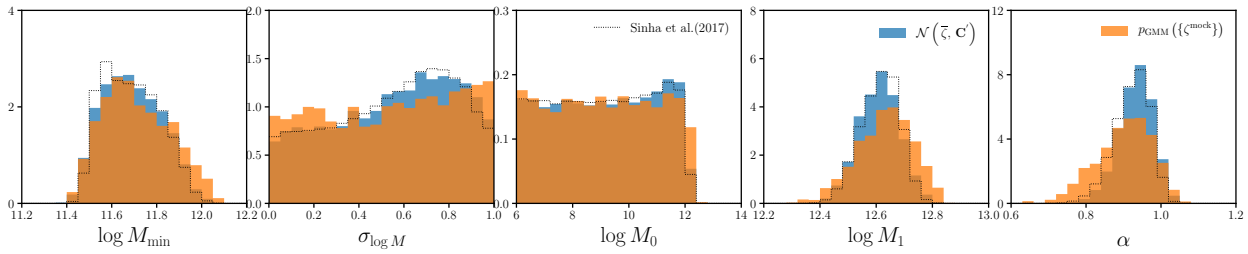
Sinha et al.(2017)
$\mathcal{N}\left(\overline{\zeta}, \mathbf{C}'\right)$
$p_{\rm GMM}\left(\{\zeta^{\rm mock}\}\right)$
$\log M_{\rm min}$
$\sigma_{\log M}$
$\log M_0$
$\log M_1$
α

Fig. 3.—

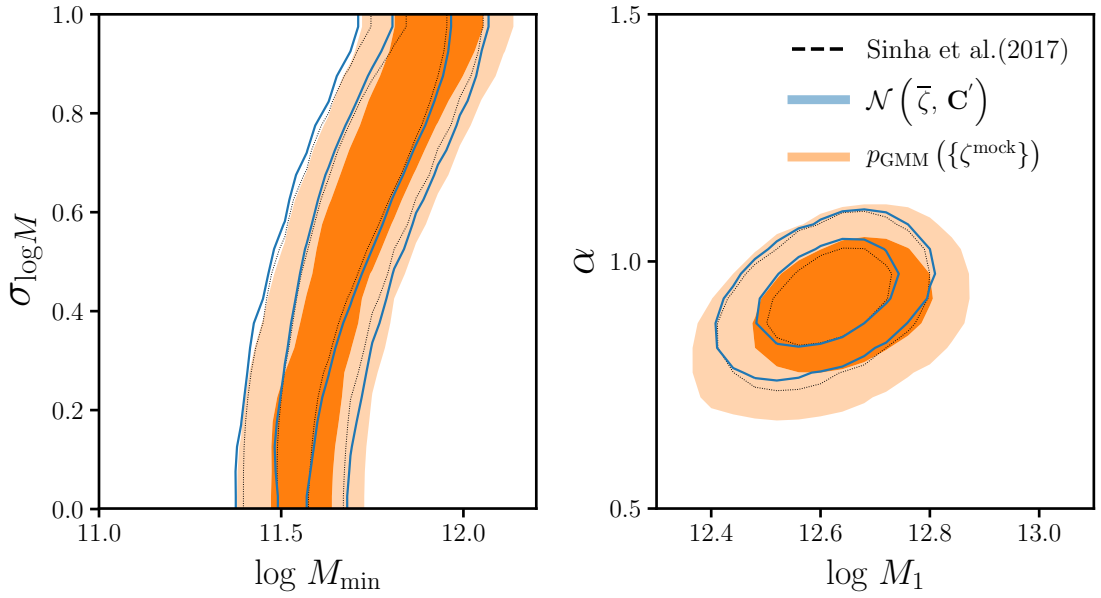
Sinha et al.(2017)
$\mathcal{N}\left(\overline{\zeta}, \mathbf{C}'\right)$
$p_{\rm GMM}\left(\{\zeta^{\rm mock}\}\right)$
$\sigma_{\log M}$
$\log M_{\rm min}$
α
$\log M_1$

Fig. 4.—