Draft version November 5, 2015
Preprint typeset using LaTeX style emulateapj v. 12/16/11

ESTIMATING INTRINSIC SHAPE DISTRIBUTION OF GALAXIES USING GAUSSIAN MIXTURE MODELS

Sowmya Kamath
Kavli Institute for Particle Astrophysics and Cosmology & Physics Department, Stanford University, Stanford, CA 94305, USA
Draft version November 5, 2015

ABSTRACT

Knowledge of intrinsic shape distribution of galaxies is vital for cosmic shear measurements. In this paper we attempt to measure the intrinsic shapes from a simulated catalog of observed shapes of galaxies, that have a constant applied shear. The two component intrinsic shape is modeled as a Gaussian Mixture Model (GMM). We derive the joint probability distribution of the GMM parameters and reduced shear by marginalizing over the intrinsic shape. We run an inference with the joint posterior probability on the catalog data using Markov Chain Monte Carlo(MCMC) method. We present the results obtained by this simplistic model.

1. INTRODUCTION

Cosmic shear is the distortion of galaxies due to weak gravitational lensing by the large-scale structure in the Universe (Figure 1). Measuring these tiny distortions can help us understand the properties and evolution of structure on large scales. Typically galaxy images are stretched by only a few per cent, for example an intrinsically circular galaxy image would become an ellipse with major-to-minor axis ratio of about 1.06 Voigt et al. (2012). However, galaxies are not circular in shape in absence of lensing; they have an intrinsic ellipticity which is about two orders of magnitude larger than the gravitational shear.

To measure the cosmic shear by correlating the observed ellipticities of galaxies or by hierarchical probabilistic models, it is essential that we accurately know the intrinsic shape distribution of galaxies. The shear measurements are subject to possible systematic errors including incomplete correction for seeing and optical distortions, selection effects, and noise-rectification biases Bernstein & Jarvis (2002), and their cosmological interpretation relies on accurate knowledge of the redshift distribution of the source galaxies. Another possible systematic error is intrinsic (i.e. not lensing-induced) correlations among the ellipticities of neighboring source galaxies Hirata & Seljak (2004), which could arise if the galaxy ellipticities are affected by large scale tidal fields. For simplicity we ignore all these factors here and assume that our observed shape data is free from all bias and known to a certain precision and that all the galaxies are randomly oriented.

The paper is organized as follow: In section 2, we introduce the notation and describe the nature of the problem applied in the model used in section 4. In section 3, we describe the relevant dataset which is used in Section section 5 to obtain GMM parameters of intrinsic shape and the cosmic shear value. The results obtained are presented in section 6. The results are briefly discussed in section 7. We present our conclusions in section 8.

2. DESCRIPTION OF THE PROBLEM

We start by decribing the notations used. The reduced shear g, describes the shape distortion of the observed galaxies which is described by a 2-component quantity $[g_1, g_2]$, written as a complex number g for convenience.

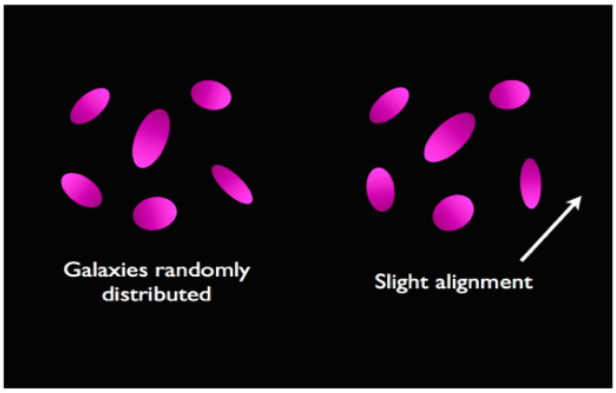

Figure 1. Alignment of galaxies due to cosmic shear

However, galaxies are not intrinsically spherical, and their shapes can be quantified on the basis of their intrinsic ellipticiity e.

$$g = g_1 + ig_2 = |g|e^{i2\phi} \quad (1)$$

$$e = e_1 + ie_2 = |e|e^{i2\beta} \quad (2)$$

The reason for the factor 2 in the phase is the fact that an ellipse transforms into itself after a rotation by 180^o

In the case of an axis-symmetric galaxy with semi-major axis a and b the semi-minor axis, the ellipticity of the profile is defined as

$$e = \frac{ab}{a+b} \quad (3)$$

The observed ellipticity is a combination of intrinsic ellipticity and shear given by eqn 4.

$$e^{lens} = \frac{e+g}{1+g^*e} \quad (4)$$

The goal of the project is given a distribution of observed galaxy shapes, to model the intrinsic shape distribution. As a simplistic case, we begin with the assumption that intrinsic galaxy ellipticities are isotropically distributed and the mean of their sheared ellipticities is related to the reduced shear by $< e^{lens} = g >$.

3. DATA

A catalog of observed shapes of galaxies measured in the GRavitational lEnsing Accuracy Testing 3 (GREAT3) challenge is taken as the input data (source: private communication Josh Meyers,Stanford University).The GREAT3 challenge is the third in a series of image analysis challenges, for testing and facilitating the development of methods for analyzing astronomical images to measure weak gravitational lensing Mandelbaum et al. (2014). The catalog contains measured shapes of 1000 galaxies with uniform applied shear. The data is visualized in Figure 2. Panel (a) shows the distribution of magnitude ($|e|$) and phase(2β) of observed shapes.Note the large spike at $|e| = 1$ is an artifact of the fitting algorithms, since $|e| > 1$ is not physical. So, we apply a cutoff at $|e| = 0.9$, and the shape measurements in panel 3 is what we will refer hereafter to as *data* for this paper.

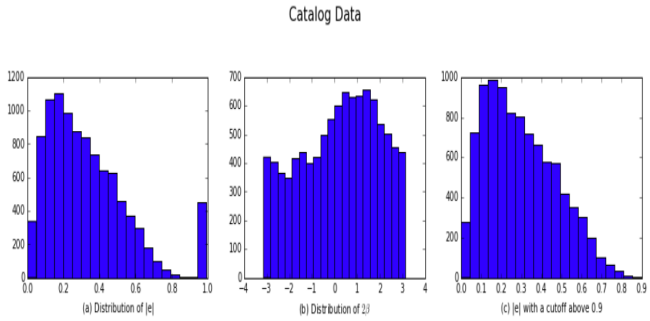

Figure 2. Catalog Data: Observed shape magnitude(left) and phase(center). Magnitudes after applying cutoff at —e—=0.9(right)

4. MODEL

As discussed in the previous section, galaxies are randomly oriented , hence the intrinsic shape should not have any preferred alignment. As a simplistic case, we assume that the intrinsic shape can be modeled as a sum of N_g Gaussians $\mu_i^g, \sigma_i^g, w_i^g$, centered at 0 i.e, $\mu_i^g = 0$. Each of these K galaxies are then sheared by a certain value g. The measurement error in the observed shapes e_k^{obs} is assumed to have a normal distribution with standard deviation σ_k. The Probabilistic Graphic Model of the problem is given by Figure 3.

5. IMPLEMENTATION

Since galaxies have no preferred alignment, intrinsic shape distributions e_1 and e_2 are chosen to be identical with mean $\mu^g = 0$.

Posterior Probability of the GMM parameters and reduced shear,g, from Baye's theorem is, given by

$$Pr(\{\mu_i^g, \sigma_i^g, w_i^g\}, g|e^{obs}) \propto Pr(e^{obs}|\{\mu_i^g, \sigma_i^g, w_i^g, \}, g) \times Pr(\{\mu_i^g, \sigma_i^g, w_i^g, \}) Pr(g) \quad (5)$$

The sampling probability for the k^{th} datapoint, $Pr(e_k^{obs}|e_k^{lens})$ is given by eqn (6).

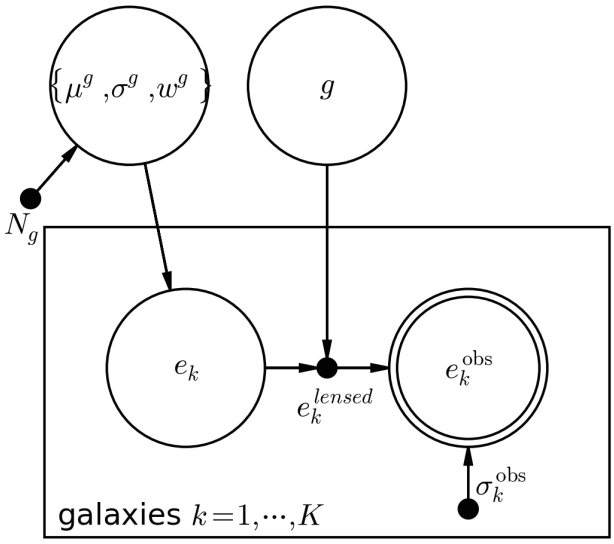

Figure 3. Probabilistic graphical model

$$Pr(e_k^{obs}|e_k^{lens}, \sigma_k) = \frac{1}{2\pi\sigma_k^2} e^{\frac{(e_{1,k}^{obs} e_{1,k}^{lens})^2}{2\sigma_k^2} + \frac{(e_{2,k}^{obs} e_{2,k}^{lens})^2}{2\sigma_k^2}} \quad (6)$$

for each of the k=1...K galaxies, where we take e_1 and e_2 to be independent of each other. For the entire dataset,

$$Pr(e^{obs}|e^{lens}, \sigma) = \prod_k \frac{1}{2\pi\sigma_k^2} e^{\frac{(e_{1,k}^{obs} e_{1,k}^{lens})^2}{2\sigma_k^2} + \frac{(e_{2,k}^{obs} e_{2,k}^{lens})^2}{2\sigma_k^2}} \quad (7)$$

The relation between e_k and e_k^{lens} is deterministic, hence

$$Pr(e_k^{lens}|e_k, g) = \delta(e_k^{lens} - \frac{e_k + g}{1 + g^* e_k}) \quad (8)$$

In this form, the likelihood probability is not analytic. In order to be able to compute the likelihood function analytically, we make an approximation, $e_k^{obs} = e_k + g$ for $g << 1$. The probability then evaluates to eqn(9).

$$Pr(e_k^{lens}|e_k, g) = \delta(e_k^{lens} - (e_k + g)) \quad (9)$$

Each intrinsic shape component e_1, e_2 is a sum of N_g Gaussians

$$Pr(e_k|\{w_i, \sigma_i\}) = \sum_i^{N_g} w_i^g \mathcal{N}(0, \sigma_i^g) \times \sum_i^{N_g} w_i^g \mathcal{N}(0, \sigma_i^g) \quad (10)$$

Since we are not interested in the intermediate values of e_k and e_k^{lens}, we marginalize over them.

$$
\begin{aligned}
Pr(e^{obs}|\{\sigma_i^g, w_i^g\}, g) &= \prod_k \int\int Pr(e_k^{obs}|e_k^{lens}) \times Pr(e_k^{lens}|e_k, g) \times Pr(e_k|\{w_i^g, \sigma_i^g\})\, de_k\, de_k^{lens} \\
&= \prod_k \int\int Pr(e_k^{obs}|e_k^{lens}) \times \delta(e_k^{lens}(e_k + g)) \times Pr(e_k|\{w_i^g, \sigma_i^g\})\, de_k\, de_k^{lens} \\
&= \prod_k \frac{1}{2\pi\sigma_k^2} e^{\frac{(e_{1,k}^{obs} e_{1,k}^{lens})^2}{2\sigma_k^2} + \frac{(e_{2,k}^{obs} e_{2,k}^{lens})^2}{2\sigma_k^2}} \times \sum_i^{N_g} w_i^g \mathcal{N}(0, \sigma_i^g) \times \sum_i^{N_g} w_i^g \mathcal{N}(0, \sigma_i^g) \\
&= \prod_k \sum_i^{N_g} \int\int (\mathcal{N}((e_{1,k}^{obs} - (e_{1,k} + g)), \sigma_k) \times w_i^g \mathcal{N}(0, \sigma_{1,i}^g)) \times (\mathcal{N}((e_{2,k}^{obs} - (e_{2,k} + g)), \sigma_k) \times w_i^g \mathcal{N}(0, \sigma_i^g)) de_{1,k} de_2
\end{aligned}
\tag{11}
$$

The joint likelihood function is calculated as,

$$
Pr(e^{obs}|\{w_i, \sigma_i\}, g) = \prod_k \sum_i^{N_g} \frac{w_i^g w_i^g \exp\left(\frac{-(g_1 - e_{1,k}^{obs})^2 - (g_2 - e_{2,k}^{obs})^2}{2(\sigma_k^2 + \sigma_i^g)}\right)}{2\pi(\sigma_k^2 + \sigma_i^g)} \tag{12}
$$

Substituting in Equation 5 gives the posterior probability.

GMM parameters and g are assumed to have uniform priors ($Pr(\{\mu_i^g, \sigma_i^g, w_i^g, \})Pr(g)$). For simplicity, the measurement error σ_k is assumed to be constant.

6. RESULTS

An inference was run with the joint posterior probability calculated in section 4 on the Data using MCMC with 5000 steps. The step sizes were set at follows $g : 1e-3; \sigma^g : 1e-3; w^g : 1e-4$. The parameter values after a preset burn-in period, were used to model the intrinsic shape and the applied shear.

The intrinsic ellipticities were modeled as GMM σ_i, w_i with $N_g = 5$ and $N_g = 3$. The initial GMM values for the MCMC we obtained by applying scikit-learn GMM Pedregosa et al. (2011) . The fit values are plotted in Figure 4 . As shown above, When $|g| < 1 :< eobs >= g$ The mean of Data $e1$ and e_2 was calculated and set as the initial point for MCMC; $g = [0.03661, 0.024804]$

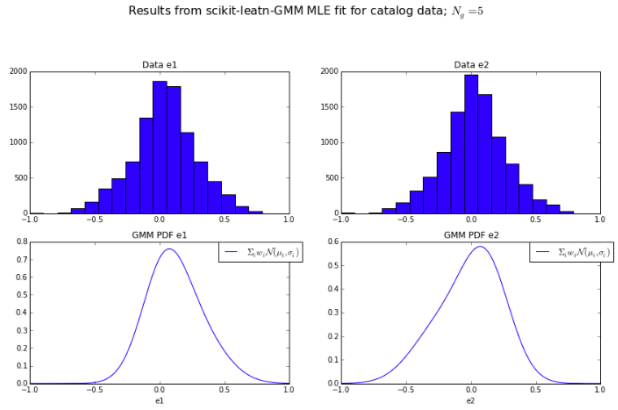

Figure 4. Plot of samples of the MCMC sampler for $N_g = 5$

Figure 5 gives the samples obtained from the MCMC chain for GMM $\{\sigma_i, w_i\}$ and $g(g_1, g_2)$. The yellow line is the denotes the burn-in point, values left of which, were not included the parameter estimation Table 1

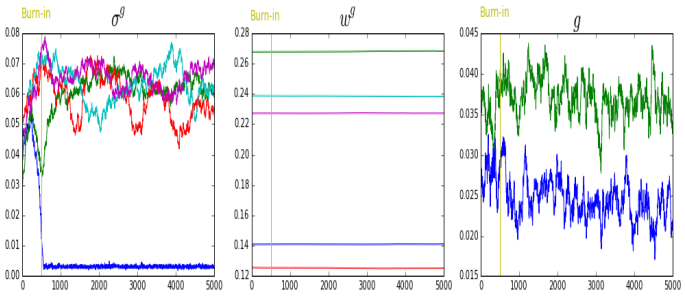

Figure 5. Plot of samples of the MCMC sampler for $N_g = 5$

Table 1
Parameter values; N_g=5

Parameter	i=1	i=2	i=3	i=4	i=5
σ^g	0.0701	0.0776	0.0697	0.0085	0.0723
	±0.0028	±0.0077	±0.0021	±0.0005	±0.0068
w^g	0.2501	0.1079	0.2826	0.2270	0.1322
	±0.0051	±0.0085	±0.0085	±0.0015	±0.0061
g	0.03759	0.02472			
	±0.0025	±0.025			

Using the results from Table 1 we create intrinsic shape distribution with $N_g = 5$ GMM Figure 6. The top-left histogram gives the PDF of the intrinsic shape, while top-right is a histogram of samples drawn from this PDF. When a shear g is applied to this distribution and measured with an error σ_k, we obtain the observed ellipticities $e1$ and e_2 distribution plots in Figure 6 lower panels. The catalog Data e_1 and e_2 are over-plotted for comparison.

Results from $N_g = 3$ GMM model have also been presented here for comparison.Figure 7 gives the samples from the MCMC chain for GMM $\{\sigma_i, w_i\}$ and $g(g_1, g_2)$.

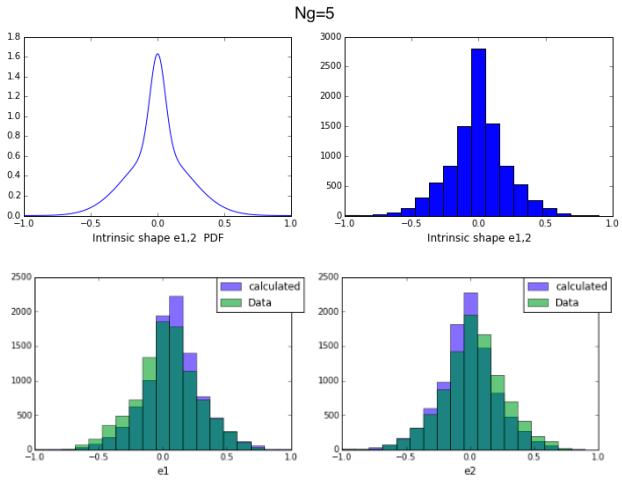

Figure 6. Results for $N_g = 5$. Top panel gives the PDF of intrinsic shape distribution(left) and the histogram of samples drawn from it. Bottom panel gives the observed shape after applying shear g, compared to the catalog Data

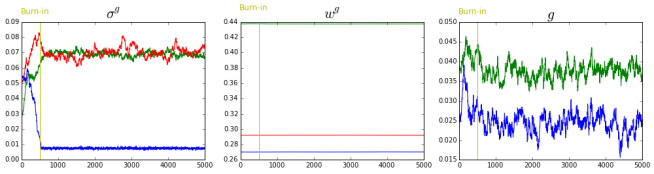

Figure 7. Plot of samples of the MCMC sampler for $N_g = 5$

Table 2
Parameter values; N_g=3

Parameter	i=1	i=2	i=3
σ^g	0.0834	0.0336	0.0841
	± 0.0026	± 0.0012	± 0.0023
w^g	0.2702	0.4375	0.2922
	$\pm 7.87e-05$	$\pm 1.27e-04$	$\pm 9.09e-05$
g	0.03810	0.0243	
	± 0.0024	± 0.0243	

The parameters obtained are shown in Table 2. The intrinsic shape PDF, sampled values and sheared observed shapes are plotted in Figure 8

7. DISCUSSION

The model studied here is a very simplistic one and does not allow different profiles for the intrinsic shape distribution. However the assumptions made are reasonable and the results look promising. The MCMC chains appear to converge partially. The intrinsic shape distributions are narrower than the observed data components Figure 4. The PDF calculated for $N_g = 3$ is slightly broader than for $N_g = 3$.

8. CONCLUSIONS

We have estimated the intrinsic galaxy shape distributions from a catalog of uniformly sheared observed galaxy shapes, while simultaneously measuring the applied shear. We have fit the intrinsic shapes for two Gaussian Mixture Models, $N_g = 5, 3$ and compared the result

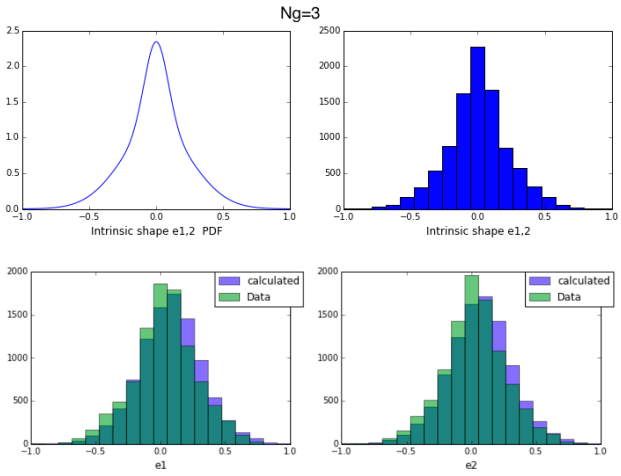

Figure 8. Results for $N_g = 5$. Top panel gives the PDF of intrinsic shape distribution(left) and the histogram of samples drawn from it. Bottom panel gives the observed shape after applying shear g, compared to the catalog Data

We can draw the following conclusions from the analysis:

- The intrinsic shape can be successfully modeled as a sum of Gaussians.
- Both GMM models give similar results for shear g and are able to reproduce the catalog data.
- $N_g = 5$ models the galaxies to have smaller ellipticities than the catalog data.
- GMM with $N_g = 3$ gives lower error on GMM parameter estimates.

The comparisons between the two GMM models are qualitative. A better quantitative statistic for the fit can be tested. The method can also be tested with respect to other data sets. The case of uniform shear is a simplistic assumption. However for more realistic situations, better techniques to handle the variability of the shear must be applied.

I would like to express my gratitude to Josh Meyers and Phill Marshall for their guidance and feedback

REFERENCES

Bernstein, G. M., & Jarvis, M. 2002, AJ, 123, 583
Hirata, C. M., & Seljak, U. 2004, Phys. Rev. D, 70, 063526
Mandelbaum, R., Rowe, B., Bosch, J., et al. 2014, ApJS, 212, 5
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825
Voigt, L. M., Bridle, S. L., Amara, A., et al. 2012, MNRAS, 421, 1385