Baryonic Effects in Cosmic Shear Tomography: PCA Parameterization and the Importance of Extreme Baryonic Models

Irshad Mohammed1,2 and Nickolay Y. Gnedin1,2,3

1 Theoretical Astrophysics Group, Fermi National Accelerator Laboratory, Batavia, IL 60510, USA; mohammed@fnal.gov
2 Kavli Institute for Cosmological Physics, The University of Chicago, Chicago, IL 60637, USA; gnedin@fnal.gov
3 Department of Astronomy & Astrophysics, The University of Chicago, Chicago, IL 60637, USA

Received 2017 July 6; revised 2018 June 27; accepted 2018 June 29; published 2018 August 22

Abstract

Baryonic effects are among the most severe systematics to the tomographic analysis of weak lensing data, which is the principal probe in many future generations of cosmological surveys like LSST, Euclid, etc. Modeling or parameterizing these effects is essential in order to extract valuable constraints on cosmological parameters. In a recent paper, Eifler et al. suggested a reduction technique for baryonic effects by conducting a principal component analysis and removing the largest baryonic eigenmodes from the data. In this paper, we conducted the investigation further and addressed two critical aspects. First, we performed the analysis by separating the simulations into training and test sets, computing a minimal set of principle components from the training set and examining the fits on the test set. We found that using only four parameters, corresponding to the four largest eigenmodes of the training set, the test sets can be fitted thoroughly with an rms $\sim 0.0011$. Second, we explored the significance of outliers, the most exotic/extreme baryonic scenarios, in this method. We found that excluding the outliers from the training set results in a relatively bad fit and degraded the rms by nearly a factor of 3. Therefore, for a direct employment of this method to the tomographic analysis of the weak lensing data, the principle components should be derived from a training set that comprises adequately exotic but reasonable models such that the reality is included inside the parameter domain sampled by the training set. The baryonic effects can be parameterized as the coefficients of these principle components and should be marginalized over the cosmological parameter space.

Key words: cosmology: theory – gravitational lensing: weak – large-scale structure of universe

1. Introduction

There are diverse pieces of evidence from existing observations, such as the cosmic microwave background (CMB; Spergel et al. 2003; Planck Collaboration et al. 2014), baryonic acoustic oscillations (BAOs; Percival et al. 2010; Anderson et al. 2012; Sánchez et al. 2013), gravitational lensing (Kaiser & Squires 1993), etc., indicating that the major component of the matter content of the universe, over 80%, is the so-called dark matter, which interacts only gravitationally. The mostly collisionless nature of dark matter has empowered our computational capabilities to accurately simulate the universe on linear and quasi-linear scales using dark-matter-only (DMO) simulations. On the other hand, doing so has introduced a major systematic uncertainty in several observational probes due to the different dynamics of baryonic matter, a small but non-negligible contribution to the total matter budget, at sufficiently small scales. These small scales contain a significant amount of information on the cosmological parameters; to extract that information one, therefore, has to correctly model these systematics. Without such modeling, constraints on cosmological parameters and interpretation of observational data may be biased (Zentner et al. 2013).

Bending of light due to intervening matter (a lens) is referred to as gravitational lensing (for a thorough review see Mellier 1999; Bartelmann & Schneider 2001; Schneider 2006). If the gravitational potential of the lens is large and the geometry of the observer-lens-source is favorable, multiple images of the source can be observed, the phenomenon also known as strong gravitational lensing (SL). On the other hand, if the potential of the lens is not strong enough and/or the geometry is not perfect, only the shape of the sources get distorted. This phenomenon is known as weak gravitational lensing (WL). The distortions in the shape of the background galaxies are referred to as shear. The shear signal is small compared to the noise and the intrinsic shape/ellipticity of the source, nearly 1%. Therefore, it can only be measured statistically, by averaging out the noise and the intrinsic ellipticity of a sample of sources together, assuming there is no preferred ellipticity or orientation of the background galaxies. The most commonly used statistic is the two-point correlation function, or its Fourier transform, the power spectrum. The tomographic analysis of the shear power spectra in different redshift bins is among the most promising tool to constrain cosmological parameters, including the equation of state of the dark energy (Weinberg et al. 2013).

As lensing is sensitive to the total matter content of the lens and does not differentiate between dark matter and baryons, this probe is fairly unbiased to any baryonic effects observationally. However, theoretical modeling of the shear power spectra relies on modeling the distribution of all matter in the universe, which is strongly affected by the effects of baryonic physics at small scales. There have been many attempts to model these baryonic effects using hydrodynamical simulations (Jing et al. 2006; Rudd et al. 2008; Guillet et al. 2010; Sembolini et al. 2011, 2013; van Daalen et al. 2011) and theoretical modeling (White 2004; Zhan & Knox 2004; Mohammed & Seljak 2014; Mohammed et al. 2014).

In a recent paper, Eifler et al. (2015) proposed a mitigation technique for baryonic effects by performing a principal component analysis (PCA) and removing the largest baryonic eigenmodes from the data. This technique is equivalent to de-baryonizing the shear power spectra data, and then fitting it
with DMO models to constrain the cosmological parameter. In this paper, we took the analysis further and discussed two important aspects. First, we divided the simulations into training and test sets, performing the PCA on training sets and analyzing the fits on both training and test sets. This shows the universality of the results. Second, we discussed the importance of outliers (or the most exotic/extreme baryonic scenarios) in this analysis.

This paper is organized as follows: in Section 2, we review the theoretical ingredients of the tomographic analysis of the shear power spectrum. In Section 3, we briefly discuss different hydrodynamical simulations used in this work. In Section 4, we briefly discuss the concept of dimensionality reduction and present the main results of this exercise. In Section 4.2, we describe the full analysis pipeline, and in Section 4.3, we emphasize the importance of the outliers. Finally, we discuss our results in Section 5.

2. Theoretical Review—Modeling Shear Power Spectrum

The distortion of the source shape due to WL can be described by two quantities: shear $\gamma$ and convergence $\kappa$. The convergence $\kappa$ is the local isotropic part of the deformation matrix and can be expressed as

$$\kappa(\theta) = \frac{1}{2} \nabla \cdot \alpha(\theta), \quad (1)$$

where $\theta$ is the angular coordinate on the sky and $\alpha$ is the deflection angle. If we know the redshift of the source galaxies, extra information can be gained by separating the sources in various redshift bins. This method is referred to as lensing tomography and is very useful to gain extra constraints on cosmology from the evolution of the weak lensing power spectra (Hu 1999; Takada & Jain 2004, 2009). In a cosmological setting, the convergence field can be represented as the weighted projection of the mass distribution integrated along the line of sight in the $i$th redshift bin,

$$\kappa_i(\theta) = \int_{z_i}^{z_{i+1}} g_i(\chi) \delta(\chi \theta, \chi) d\chi, \quad (2)$$

where $\delta$ is the total three-dimensional matter overdensity, $\chi$ is the comoving distance, and $\chi z_i$ is the comoving distance to the horizon. The lensing weight $g_i$ in the $i$th redshift bin with $z_i < z < z_{i+1}$ is given by

$$g_i(z) = \begin{cases} \frac{z_{i+1} - z_i}{\chi_{z_i} - \chi_{z_{i+1}}} n_i(z') \frac{\chi_{z_i} - \chi_{z_i}}{\chi_{z_i}'} \frac{d\chi'}{z_i'}, & \text{for } z \leq z_{i+1}, \\ 0, & \text{for } z > z_{i+1} \end{cases} \quad (3)$$

with $\chi_{z_i}$ being the comoving distance at redshift $z_i$, $g_0 = (3/2) \Omega_M H_0^2$, and

$$n_i = \int_{z_i}^{z_{i+1}} n_s(z') dz'. \quad (4)$$

Here, $n_i(z)$ is the distribution of sources in redshift. We assume a source distribution along the line of sight of the form

$$n_i(z) = 4z^2 \exp\left(-\frac{z}{z_0}\right) n_0, \quad (5)$$

with $n_0 = 1.18 \times 10^9$ sources per unit steradian and where $z_0$ is fixed such that the mean redshift of the source distribution is $z_m = 3z_0$ and the corresponding projected source density $n_\perp$ resembles the experiment (for a more detailed review, see Takada & Jain 2009)

$$\int_0^\infty n_s(z) dz = n_\perp. \quad (6)$$

Table 1 shows the weak lensing survey parameters for a stage III and stage IV survey in this setting. Finally, the shear power spectrum between redshift bins $i$ and $j$ can be computed as

$$C_{ij}(\ell) = \int_0^{z_{i+1}} \int_0^{z_{j+1}} \frac{n_i(\chi_1) n_j(\chi_2)}{\chi_1^2} P(\ell, \chi_2) d\chi_2, \quad (7)$$

where $P$ is the 3D matter power spectrum. Larger $\ell$ corresponds to the smaller scale and the large contribution of $C_{\ell}$ at higher $\ell$ comes from nonlinear clustering.

3. Baryonic Effects in Simulations

In this section, we describe various hydrodynamical simulations used to study different baryonic corrections on the tomographic analysis of the weak lensing shear power spectra for a DES- and LSST-like survey. We worked with the same sets of simulations examined by Eifler et al. (2015). We thank Tim Eifler for kindly providing us with the shear power spectra corresponding to various DMO and hydrodynamical simulations for DES- and LSST-like surveys. All spectra are computed for five tomographic bins; there are a total of 5(5+1)/2 = 15 spectra available (5 auto-spectra and 10 cross-spectra). Each of these spectra are computed for 12 multipole bins ($\ell$) equally spaced in logarithm. Therefore, the length of each of the simulation vectors is $15 \times 12 = 180$.

The full simulation set is composed of three subsets:

**OWLS simulations:** This suite of simulations contributes a total of nine different baryonic scenarios, which differ for their cooling, supernovae, and AGN feedback. For a detailed prescription of these simulations and the OWLS project see Schaye et al. (2010), van Daalen et al. (2011), Zentner et al. (2013), and Semboloni et al. (2013).

**ART08 simulations:** This subset of simulations performed with the Adaptive Refinement Tree (ART) code (Kravtsov 1999; Kravtsov et al. 2002; Rudd et al. 2008) contributes a total of two different baryonic scenarios, one of which is treated in the non-radiative regime and does not allow any star or galaxy formation, whereas the other simulations allow these processes. For a detailed description of these simulations see Rudd et al. (2008). These simulations were called “Rudd simulations” in Eifler et al. (2015).

**ART14 simulations:** Three more baryonic scenarios modeled with the ART code are contributed by this subset; the first set contains adiabatic hydrodynamic processes; the second set radiative cooling, but not heating, with primordial abundance

| Survey | Stage | $z_0$ | $z_m$ | $n_\perp$ |
|--------|-------|-------|------|----------|
| DES    | III   | 0.23  | 0.7  | 51       |
| LSST   | IV    | 0.4   | 1.2  | 100      |

- $n_0 = 1.18 \times 10^9$ sources per unit steradian
- $z_0$ is fixed such that the mean redshift of the source distribution is $z_m = 3z_0$
- The corresponding projected source density $n_\perp$ resembles the experiment (for a more detailed review, see Takada & Jain 2009)
of hydrogen and helium; whereas the third set contains extreme radiative cooling with a cooling function similar to solar-metallicity gas. These simulations were called “Gnedin simulations” in Eifler et al. (2015).

Figure 1 shows the deviation of the shear power spectra of each of these baryonic scenarios from the dark-matter-only case for the auto-spectra in the first redshift bin for a stage III (DES-like) survey. For a stage IV (LSST-like) survey, we have shear spectra for all but ART08 simulations (they do not have enough redshift coverage to compute predictions for a stage IV experiment). So there are a total of 12 scenarios available for an LSST-like survey and 14 for a DES-like survey.

4. Results

4.1. Dimensionality Reduction

The advancement in our observational and computational abilities has led us into the era of big data sets, which are observed and/or simulated every day. These data sets not only have an enormous sample size but also have high dimension, i.e., a large number of variables measured/computed in each observation/simulation. These variables are often correlated, in a linear or a nonlinear way, which gives us the ability to look for fewer derived variables that can be used to represent the full original data set with a lower dimension. This method is referred to as Dimensionality Reduction. In this paper, we use the term variables interchangeably with features or attributes.

Mathematically, a given data set with \( p \) variables, \( x = (x_1, x_2, x_3, \ldots, x_p) \) can be reduced to a new data set with \( k \) variables, \( s = (s_1, s_2, s_3, \ldots, s_k) \), where \( k < p \), with the ability to reconstruct the original data set using some criteria. The techniques are broadly classified into two categories: linear and nonlinear, depending on whether the mathematical form of the mapping from \( x \) to \( s \) (or vice versa) is a linear or nonlinear function, respectively. One very popular linear dimensional reduction technique is the standard PCA. For a more general overview of both linear and nonlinear technique see Fodor (2002).

PCA is among the most commonly used linear dimensionality reduction techniques. It attempts to reduce the dimension of the data by finding an alternate basis for the original variables with the largest variance preserving the covariance matrix of the variables. The basis is in the form of orthogonal linear combinations, also referred to as principal components (PCs), normalized to the square root of the corresponding eigenvalue.

4.2. Analysis

We start by defining the contrast of baryons \( (\delta C_{ij}(\ell)) \); also interchangeably referred to as boost for the weak lensing auto and cross shear power spectra in different redshift bins, which represent the relative deviation from the DMO case

\[
\delta C_{ij}(\ell) = \frac{C_{ij}^{\text{BAR}}(\ell)}{C_{ij}^{\text{DMO}}(\ell)} - 1.
\]

Second, we split the simulation sets of each case, DES and LSST, into training and test sets. The training set includes the subset of all baryonic scenarios used to compute the minimal set of PCs. For each case, we use a subset of eight baryonic scenarios as the training set, and the remaining four simulations for the LSST case (or six for DES case) as the test set. The top and bottom rows of Figure 2 show the baryon contrast for both training (left panel) and test (right panel) sets for DES and LSST cases, respectively.

Next, we compute the 180 \( \times \) 180 covariance matrix of the training sample and determine the ordered eigenvalues (a total of 180; however, only the first 8 are non-vanishing as there are only 8 samples in the training set). The eigenvalue fraction is defined as the fraction of the total covariance matrix trace not accounted for by first \( k \) eigenvalues:

\[
f_k = 1 - \frac{\sum_{i=1}^{k} E_i}{\text{Tr}(\delta C_{ij})},
\]

where \( E_i \) is the \( i \)th eigenvalue. The eigenvalue fraction quantifies the fractional residue left over after one removes the first \( k \) eigenmodes. The left panels of Figure 3 show the...
eigenvalue fraction for DES-like (top row) and LSST-like (bottom row) surveys. The first eigenvalue contains nearly 80% of the total fraction of the variance; therefore, after removing the first eigenmode, the eigenvalue fraction is $\sim 0.2$. Similarly, if we remove the second eigenvalue, the eigenvalue fraction drops to $\sim 0.002$. For both cases, we chose $k = 4$ as the new reduced dimension, i.e., we remove the first four eigenmodes, which gives the eigenvalue fraction of $\sim 0.0001$. We will discuss this particular choice below. The right panels of Figure 3 show the four PCs extracted from the respective training sets normalized to the square root of the respective eigenvalues.

Using only the first four eigenmodes, we fit both the training and the test sets with a linear combination of the PCs:

$$\delta C_{ij}^{\text{Fits}}(\ell) = \sum_{i=1}^{4} a_i A_i,$$

where $A_i$ is the four PCs (or the four largest eigenvectors) as shown in Figure 3 and $a_i$ is the coefficients. Figure 4 shows the best-fit coefficients for the DES and LSST cases, and Figure 5 shows the corresponding fits.

Finally, the full model of the shear power spectra, including the parametric form of the baryonic effects, can be presented as

$$C_{ij}^{\text{RAR}}(\ell) = C_{ij}^{\text{DMO}}(\ell) (1 + \delta C_{ij}^{\text{Fits}}(\ell)).$$

This model contains a total of four free parameters, in addition to the cosmological parameters.

The particular choice of the four eigenvalues that we adopt above is dictated by the precision requirements for an LSST-like survey. In Figure 6, we show the distributions of fitting errors (differences between dashed and solid lines in Figure 5) for two, three, and four principal components ($k = 2, 3,$ and $4$ in Equation (9)). Only with four principal components do the errors go safely below 1%. In addition, there is nearly a factor of 2 gain in increasing the number of components from three to four. Using five principal components reduces the error even further by up to another 50%, but since the errors are already well below our target of 1% with just four components, it would be overkill to use five or more components. Therefore, we use four components as a fiducial setup. If the error requirements get stricter in the future, the number of principal components will need to be increased accordingly.
4.3. Importance of Outlier Models

Outliers of the data set refer to the most extreme baryonic scenarios, which, nevertheless, cannot be excluded on purely physical grounds. In our simulation sets, we mark the ART14-CX simulation as such an outlier, which is also evident by its shape (see Figure 1). In our analysis in the previous section, we included this outlier in the training set, and therefore the PCs contain its signatures. In this section, we explore its importance in the case where the outlier is not present in the training set but exists in the test set. We perform this exercise for the LSST case only, i.e., with eight training examples and four test examples.

Figure 7 shows the the eigenvalue fraction, PCs, and best-fit coefficients for this case. In order to make a quantitative comparison with the previous case, we also show in Figure 8 the precision with which approximations with four principal components are able to fit the actual simulated baryonic contrasts. When the outlier model is moved from the training set into the test set, errors in the fitted baryonic contrasts increase about threefold, both in the rms sense and in maximum errors.

5. Summary and Discussion

The Eifler et al. (2015) approach to mitigating baryonic effects in weak lensing is based on (1) building up a training set of simulations that model baryonic effects, (2) computing several (three to four) principal components of baryonic contrast, and (3) using these components as fitting functions with free parameters to marginalize over the baryonic modifications in the observed power spectra. The success or
failure of such an approach obviously depends on the fidelity of the training set.

In this paper, we showed that such a training set has to be sufficiently broad in a sense of including extreme, but still physical meaningful models. It is not just enough to include simulations that attempt to model baryonic contrast as accurately as possible, because in that case the principal components may not be sensitive enough to other “degrees of baryonic freedom,” i.e., variations in the power spectra that are not faithfully captured by the simulations (either due to numerical limitations or missing some physical ingredients). However, if “outlier” models are included, the precision of the Eifler et al. (2015) approach increases multifold.

In the specific example that we consider here, the outlier model (ART14-CX) adopts the maximal realistic value for the gas cooling function (that of the solar-metallicity gas). While this is certainly extreme, such a model cannot be excluded from pure physical reasons: cooling is overestimated in the low-density intergalactic medium (IGM), but the weak lensing signal is dominated by clustered, higher-density structures, where such strong cooling is not an obvious overestimate.

Hence, the ideal training set should include sufficiently exotic, but still not outright unreasonable models, to offer sufficiently justified assurance that the reality is included inside the parameter region sampled by the training set. Building such a training set is an important and immediate goal for the cosmological weak lensing community.

We thank the anonymous referee for constructive comments that significantly improved the original manuscript. This paper has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

ORCID iDs
Irshad Mohammed @ https://orcid.org/0000-0003-0784-5447

References
Anderson, L., Aubourg, E., Bailey, S., et al. 2012, MNRAS, 427, 3435
Bartelmann, M., & Schneider, P. 2001, PhR, 340, 291
Eifler, T., Krause, E., Dodelson, S., et al. 2015, MNRAS, 454, 2451
Fodor, I. K. 2002, A Survey of Dimension Reduction Techniques (Livermore, CA: Lawrence Livermore National Laboratory)
Guillet, T., Teyssier, R., & Colombi, S. 2010, MNRAS, 405, 525
Hu, W. 1999, ApJL, 522, L21
Jing, Y. P., Zhang, P., Lin, W. P., Gao, L., & Springel, V. 2006, ApJL, 640, L119
Kaiser, N., & Squires, G. 1993, ApJ, 404, 441
Kravtsov, A. V. 1999, PhD thesis, New Mexico State Univ.
Kravtsov, A. V., Klypin, A., & Hoffman, Y. 2002, ApJ, 571, 563
Mellier, Y. 1999, ARA&A, 37, 127
Mohammed, I., Martizzi, D., Teyssier, R., & Amara, A. 2014, arXiv:1410.6826
Mohammed, I., & Seljak, U. 2014, MNRAS, 445, 3382
Percival, W. J., Reid, B. A., Eisenstein, D. J., et al. 2010, MNRAS, 401, 2148
Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2014, A&A, 571, A16
Rudd, D. H., Zentner, A. R., & Kravtsov, A. V. 2008, ApJ, 672, 19
Sánchez, A. G., Kazin, E. A., Beutler, F., et al. 2013, MNRAS, 433, 1202
Schaye, J., Dalla Vecchia, C., Booth, C. M., et al. 2010, MNRAS, 402, 1536
Schneider, P. 2006, in Saas-Fee Advanced Course 33, Gravitational Lensing: Strong, Weak and Micro, ed. G. Meylan et al. (Berlin: Springer), 269
Semboloni, E., Hoekstra, H., & Schaye, J. 2013, MNRAS, 434, 148
Semboloni, E., Hoekstra, H., Schaye, J., van Daalen, M. P., & McCarthy, I. G. 2011, MNRAS, 417, 2020
Spergel, D. N., Verde, L., Peiris, H. V., et al. 2003, ApJS, 148, 175
Takada, M., & Jain, B. 2004, MNRAS, 348, 897
Takada, M., & Jain, B. 2009, MNRAS, 395, 2065
van Daalen, M. P., Schaye, J., Booth, C. M., & Dalla Vecchia, C. 2011, MNRAS, 415, 3649
Weinberg, D. H., Mortonson, M. J., Eisenstein, D. J., et al. 2013, PhR, 530, 87
White, M. 2004, ApPh, 22, 211
Zentner, A. R., Semboloni, E., Dodelson, S., et al. 2013, PhRvD, 87, 043509
Zhan, H., & Knox, L. 2004, ApJL, 616, L75