Science-driven 3D data compression

David Alonso[1]
[1] *Oxford Astrophysics, Department of Physics, Keble Road, Oxford, OX1 3RH, UK*

Photometric redshift surveys map the distribution of matter in the universe through the positions and shapes of galaxies with poorly resolved measurements of their radial positions. While a tomographic analysis can be used to recover some of the large-scale radial modes present in the data, this approach suffers from a number of practical shortcomings, and the criteria to decide on a particular binning scheme are commonly blind to the ultimate science goals. We present a method designed to separate and compress the data into a small number of uncorrelated radial modes, circumventing the main problems of standard tomographic analyses. The method is based on the Karhunen-Loèvetransform, and is connected to other 3D data compression bases advocated in the literature, such as the Fourier-Bessel or Fourier-Laguerre bases.

I. INTRODUCTION

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

II. METHOD

A common method to draw cosmological constraints from photometric redshift surveys is to divide the galaxy sample into bins in photo-z space and use the information encoded in all the relevant auto- and cross-correlations between different bins, making use of various calibration methods in order to estimate the true redshift disitribution of each bin. Several criteria can be followed in order to select these redshift bins, such as minimising the correlation between non-neighbouring bins or preserving a roughly constant number density on all bins. Here we present a formalism to determine an optimal scale-dependent binning scheme based on maximizing the amount of cosmological information.

A. Tomographic galaxy clustering

Let us start by assuming that we have split the galaxy sample into N_s subsamples. As mentioned above, we will think of each of these subamples as some kind of redshift binning (e.g. binning galaxies in terms of their maximum-likelihood redshift), but the formalism applies to any set of subsamples. Let $f^\alpha(\hat{\mathbf{n}})$ be the a field on the sphere at the angular position $\hat{\mathbf{n}}$ and defined in terms of the properties of the sources in the α-th sample (e.g. the cosmic shear field γ^α or the galaxy overdensity δ^α), and let $\phi^\alpha(z)$ be the redshift distribution of these sources. Finally, let $a^\alpha_{\ell m}$ be the spherical harmonic coefficients of f^α[1]. The power spectrum for our set of subsamples is defined as the two-point correlator of a^α:

$$\left\langle \mathbf{a}_{\ell m}\, \mathbf{a}^\dagger_{\ell' m'} \right\rangle \equiv \delta_{\ell\ell'}\delta_{mm'}\mathsf{C}_\ell, \qquad (1)$$

where we have packaged a^α as a vector for each (ℓ, m): $\mathbf{a}_{\ell m} \equiv (a^1_{\ell m}, ..., a^{N_s}_{\ell m})$. In general, the observed field will receive contributions from the true cosmological signal ($\mathbf{s}$) and measurement noise ($\mathbf{n}$). Assuming both components are uncorrelated, the same split applies in the power spectrum:

$$\mathsf{C} = \mathsf{S} + \mathsf{N}, \quad \mathsf{S} \equiv \left\langle \mathbf{s}\,\mathbf{s}^\dagger \right\rangle, \quad \mathsf{N} \equiv \left\langle \mathbf{n}\,\mathbf{n}^\dagger \right\rangle. \qquad (2)$$

Once the choice of subsamples α is chosen, the standard analysis method would proceed by performing a likelihood evaluation of the two-point statistics of these subsamples. While this procedure is relatively simple, it suffers from a number of drawbacks, an incomplete list of which is:

1. It is not clear what the optimal strategy should be to define the sub-samples. One could make sure to exploit all of the information present in the data by using a large number of very narrow redshift bins, and let the likelihood evaluation pick up the information encoded in them.

2. $C^{\alpha\beta}_\ell$ is a $N_s \times N_s \times N_\ell$ data vector. Thus increasing N_s will increase the computational time required for each likelihood evaluation like N_s^2 and number of elements of the covariance matrix of $C^{\alpha\beta}_\ell$ like

[1] Spin-2 fields, such as the cosmic shear, will be decomposed in spin-2 spherical harmonics, however the discussion below holds for fields of arbitrary spin.

N_s^4, with the corresponding increase in complexity needed to estimate this covariance. Although this can be partially alleviated by considering only correlations between neighbouring redshift shells, the amount of information lost by ditching all correlations beyond a given neighbouring index is not clear in a general setting.

3. Estimating the redshift distribution for a large number of subsamples can be inaccurate, depending on the method used to do so, on the quality of the photometric redshift posterior information and on the statistics of the available spectroscopic sample.

B. The Karhunen-Loeve basis

The idea of the method considered here is to form a small number of linear combinations of the data distributed in narrow redshift bins designed to contain the maximum amount of cosmological information. This can be expressed as a generic Karhunen-Loeve eigenvalue problem (e.g. see [1] for details), and the procedure to determine the coefficients of these linear combinations is relatively simple:

1. We start by assuming that the field $\mathbf{a}$ has been measured in a number of narrow redshift bins, and by defining the inverse-variance weighted field $\tilde{\mathbf{a}}_{\ell m} \equiv \mathsf{N}_\ell^{-1}\,\mathbf{a}_{\ell m}$.

2. Let us consider a (possibly ℓ-dependent) set of linear combinations of the weighted field measured on narrow redshift bins:

$$\mathbf{b}_{\ell m} = \mathsf{E}_\ell \cdot \tilde{\mathbf{a}}_{\ell m} \equiv \mathsf{E}_\ell \circ \mathbf{a}, \qquad (3)$$

where E_ℓ is a yet-unspecified matrix and we have defined the non-standard dot product: $\mathbf{v}_\ell^\dagger \circ \mathbf{w}_\ell \equiv \mathbf{v}_\ell^\dagger \cdot \mathsf{N}_\ell^{-1} \cdot \mathbf{w}_\ell$. The power spectrum for this new observable would then simply be given by:

$$\mathsf{D}_\ell \equiv \left\langle \mathbf{b}_{\ell m}\,\mathbf{b}_{\ell m}^\dagger \right\rangle = \mathsf{E}_\ell^\dagger \circ \mathbf{C}_\ell \circ \mathsf{E}_\ell. \qquad (4)$$

3. Let us now find the set of orthonormal vectors $\mathbf{v}_\ell^p$) (with $\mathbf{v}^{p\dagger} \circ \mathbf{v}^q = \delta^{pq}$) that diagonalize the cross-shell power spectrum C_ℓ[2]. We then identify $(\mathsf{E}_\ell)_\alpha^p$ with the elements of $\mathbf{v}^p(\ell)$. Note that, after this transformation and without any further optimization, some of the practicalities of the original problem the original problem are already simplified, since we can now focus on the diagonal elements of the new power spectrum and its covariance.

4. Let's assume that we are interested in measuring a set of cosmological parameters $\Theta \equiv \{\theta_1, ...\}$. The information regarding this set of parameters encoded in a given data vector $\mathbf{x}$ can be quantified in terms of its Fisher matrix (the Hessian of the log-likelihood with respect to Θ), which assuming $\langle \mathbf{x} \rangle = 0$ reads

$$F_{ij} \equiv \langle \partial_i \partial_j \mathcal{L} \rangle = \frac{1}{2} \mathrm{Tr}\left(\partial_i \mathsf{X} \mathsf{X}^{-1} \partial_j \mathsf{X} \mathsf{X}^{-1}\right), \qquad (5)$$

where $\mathsf{X} \equiv \langle \mathbf{x}\mathbf{x}^T \rangle$ is the covariance matrix of the data. Since the power spectrum of b defined above is diagonal, this expression gets simplified further, and the Fisher matrix can be decomposed into the independent contributions of each mode: $F_i j = \sum_p F_{ij}^p$, where

$$F_{ij}^p \equiv \sum_\ell \frac{2\ell+1}{2} \left(\partial_i \log D_\ell^p\right) \left(\partial_j \log D_\ell^p\right). \qquad (6)$$

We can thus rank the eigenvectors $(\mathsf{E}_\ell)_\alpha^p$ in terms of their information content (in a Fisher-matrix sense).

5. We then truncate the number of modes to analyze to the first M eigenmodes thus defined, which contain the bulk of the information needed to constrain Θ.

This strategy therefore allows one to reliably and significantly reduce the dimensionality of the data vector from $N_s^2 \times N_\ell$ to $M \times N_\ell$ while minimising the loss of information. Note that, although the method is based on an initial thin-slicing of the galaxy distribution, the fact that the final datased comprises only a small set of samples means that the method is not penalized in terms of photometric redshift uncertainties. Once the set of K-L eigenmodes E_ℓ, are found for a fiducial cosmological model, they can be directly applied as weights to all the objects in the survey to generate the b_ℓ^p modes. Furthermore, using E_ℓ for the fiducial cosmology as model-agnostic weights and inserting them in Eq. 4, the theoretical prediction for the power spectrum of each mode D_ℓ^p can be computed in a model-independent way.

The same methods used to calibrate photo-z uncertainties in the standard tomographic analysis hold in this case with slight modifications (e.g. weighed and ℓ-dependent stacking of photo-z pdfs, or cross-correlations of the weighed maps with a spectroscopic survey in the case of clustering redshifts).

[2] Note that this can always be found also in the case of non-standard dot products. Let P be the matrix defining the product, and $\mathsf{P} = \mathsf{L}\mathsf{L}^T$ its Cholesky decomposition. The problem of diagonalizing a symmetric matrix A with respect to this product is equivalent to diagonalizing $\mathsf{A}' \equiv \mathsf{L}^T \mathsf{A} \mathsf{L}$ with the standard dot product and multiplying the resulting eigenvectors by $\left(\mathsf{L}^T\right)^{-1}$.

III. PERFORMANCE AND PARTICULAR EXAMPLES

Here we explore the performance of this method in a number of specific science cases.

A. Special case: the harmonic-bessel basis

Let us consider a simplified case where f is the overdensity field of a non-evolving galaxy population for which, furthermore, and for which we neglect the effect of redshift-space distortions. Let us further assume that we have perfect redshift information, such that we can split the sample into thin radial slices of equal width $\delta\chi$, which we label by their comoving radius χ. The noise in the measurement of f is given purely by shot noise, and since (as per our initial assumptions) the number density of sources does not change with r, the noise power spectrum is diagonal and scales like

$$N_\ell(r,r') \propto \frac{\delta_{r,r'}}{r^2}. \tag{7}$$

Thus, the dot product is just given by:

$$\mathbf{b}^\dagger \circ \mathbf{c} \propto \int dr\, r^2\, b(r)^*\, c(r). \tag{8}$$

In this case, the cross-shell power spectrum is given by,

$$C_\ell^{rr'} = \frac{2}{\pi}\int_0^\infty dk\, k^2\, P_k\, j_\ell(kr) j_\ell(kr'), \tag{9}$$

and it is trivial to show that the K-L eigenmodes are simply given by the spherical Bessel functions: $(\mathsf{E}_\ell)^k_r \propto \sqrt{2/\pi} j_\ell(kr)$:

$$D_\ell^{kk'} \propto \frac{2}{\pi}\int dr\, r^2 \int dr'\, r'^2 j_\ell(kr) j_\ell(k'r') C_\ell^{rr'} \tag{10}$$

$$= \int dq\, q^2 P_q \left[\frac{2}{\pi}\int dr\, r^2\, j_\ell(qr) j_\ell(kr)\right]\left[\frac{2}{\pi}\int dr'\, r'^2\, j_\ell(qr') j_\ell(k'r')\right] \tag{11}$$

$$= \int dq\, q^2 P_q \frac{\delta(k-q)}{q^2}\frac{\delta(k'-q)}{q^2} = P_k \frac{\delta(k-k')}{k^2} = \frac{P_k}{k^2\Delta k}\delta_{k,k'} \tag{12}$$

This choice of basis defines the so-called harmonic-Bessel (or Fourier-Bessel) decomposition, and has been postulated as a possible data-compression method for the analysis of photometric redshift data (CITES here). In any realistic scenario (e.g. in the presence of redshift uncertainties, RSDs or for the analysis of weak lensing data), this basis is, however, non-optimal in terms of uncorrelatedness, as opposed to the K-L basis described above.

B. Weak lensing - K-L basis for dark energy

To quantify the performance of the K-L modes for weak lensing we study the case of an LSST-like survey. The survey specifications and the characteristics of the galaxy sample are described in detail in CITE. In summary, we assume a sample with ~ 29 objects per arcmin^2 with the redshift distribution shown in the left panel of Figure 1. We also approximate the photo-z distribution as Gaussian with a scatter $\sigma_z = 0.05\,(1+z)$.

The signal part of the cross-power spectrum between the cosmic shear measurements made in two different redshift shells is given by:

$$S_\ell^{\alpha\beta} = \frac{2}{\pi}\int_0^\infty dk\, k^2\, \Delta_\ell^\alpha(k)\Delta_\ell^\beta(k), \tag{13}$$

where the transfer functions Δ_ℓ^α take the form:

$$\Delta_\ell^{\gamma,\alpha}(k) \equiv \frac{3H_0^2\Omega_M}{2k^2}\sqrt{\frac{(\ell+2)!}{(\ell-2)!}}\int d\chi\, W^\alpha(\chi)\frac{j_\ell(k\chi)}{\chi\, a(\chi)}\sqrt{P(k,z(\chi))}, \quad W^\alpha(\chi) \equiv \int_{z(\chi)}^\infty dz\, \phi^\alpha(z)\frac{\chi(z')-\chi}{\chi(z')\chi}. \tag{14}$$

Here $\phi^\alpha(z)$ is the redshift distribution of sources in the α-th bin. The noise power spectrum is white and simply given by the intrinsic ellipticity scatter weighed by the number density of sources in each redshift bin $\bar{n}^\alpha$:

$$N_\ell^{\alpha\beta} = \delta_{\alpha\beta}\frac{\sigma_\gamma^2}{\bar{n}^\alpha}, \tag{15}$$

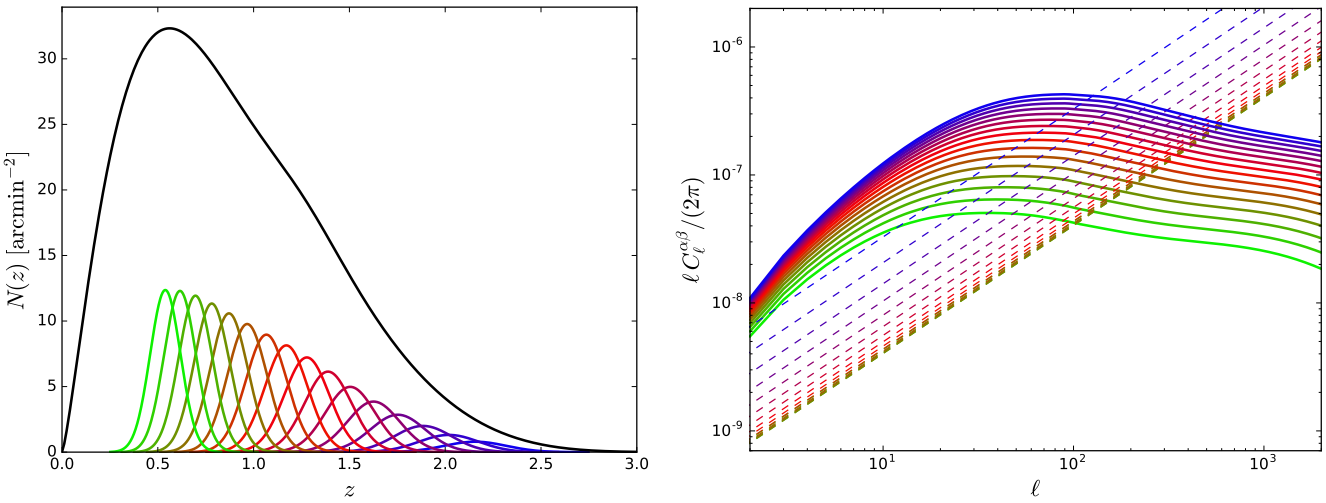

FIG. 1.

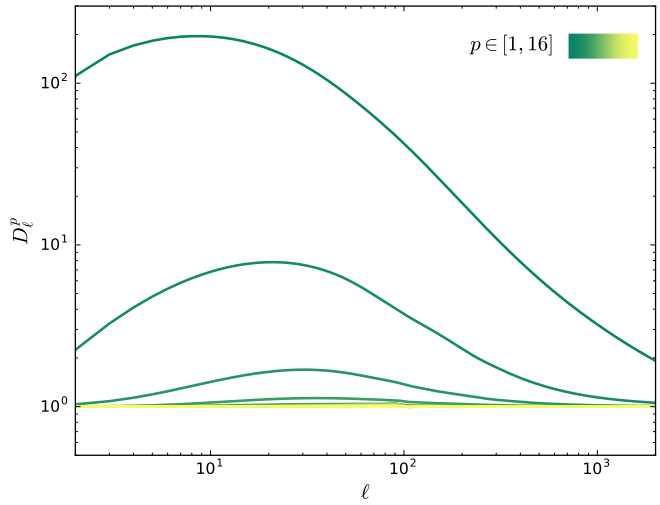

FIG. 2.

with $\bar{n}^\alpha$ in units of srad^{-1}. We use $\sigma_\gamma = 0.28$.

As our initial set of narrow redshift bins, we select top-hat bins in photo-z space for $z_{\rm ph} > 0.5$ with a width given by the value of σ_z at the center of the bin. The resulting set of 16 bins is shown in the left panel of Figure 1. The large overlap between bins implies that a choice of thinner slices is unlikely to unveil significantly more information, and we have verified that the results shown below do not change after doubling the number of bins. The lensing auto-power spectra (both signal and noise) for these bins are shown in the right panel of Figure 3. The elements of $C_\ell^{\alpha\beta}$ were estimated using a modified version of the code presented in CITE.

We compute the K-L modes for this setup and rank them according to their information content on the dark energy equation of state w. The power spectra of the resulting set of modes are shown in the left panel of Figure 2. Comparing against the right panel of Fig. 1 we can see that the K-L decomposition effectively separates the signal-dominated and noise-dominated modes, with all modes $p > 3$ dominated by noise (note that the noise power spectrum gets mapped into 1 under the K-L transform). The fractional contribution of each mode to the total constraint on w (i.e. its contribution to the corresponding Fisher matrix element) is shown in the left panel of Figure 3. Most of the information ($\sim 95\%$) is contained within a single mode, and the first two modes are able to recover more than 99% of the total. The eigenvectors corresponding to the first and second modes for different values of ℓ are shown in the right panel of the same figure. Firstly, we observe that the eigenvectors preserve roughly the same shape for all ℓ, and converge to the same shape at large ℓ. The first eigenvector upweights the parts of the redshift range with the highest signal-to-noise, penalising the low-z regime due to its poor lensing signal and the high-z bins due to their high shot noise. The second eigenmode then recovers part of this information by marginally upweighting these regions.

C. Galaxy clustering - measuring $f_{\rm NL}$

$$\Delta_\ell^{\delta,\alpha}(k) \equiv \int dz\, \phi^\alpha(z)\, [b^\alpha(z) j_\ell(k\,\chi(z)) - f(z) j_\ell''(k\,\chi(z))]\, \sqrt{P(k,z)} \tag{16}$$

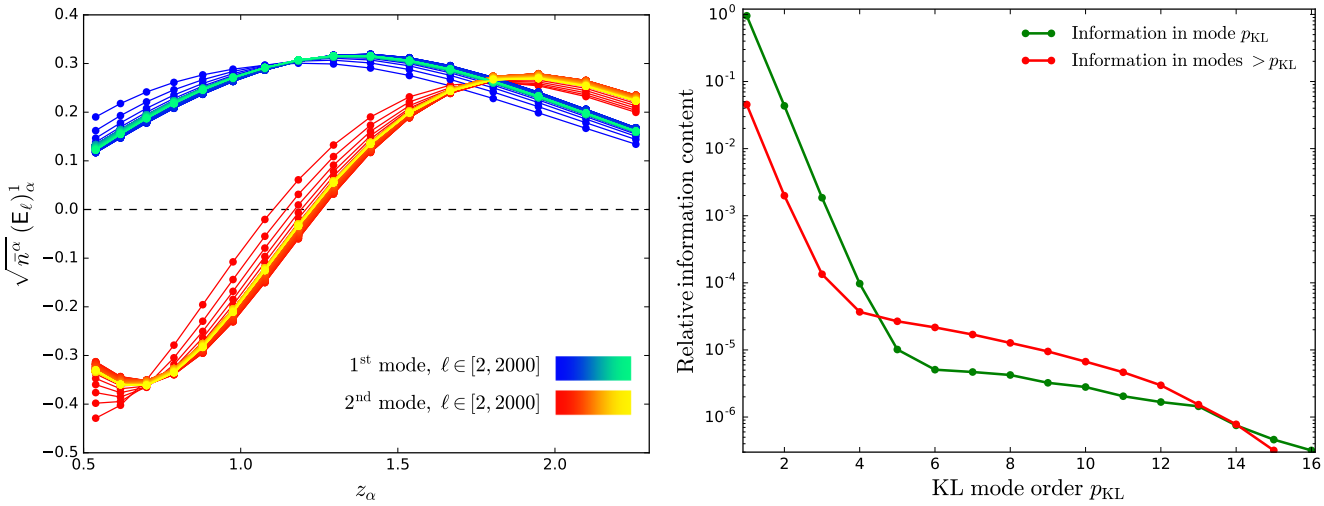

FIG. 3.

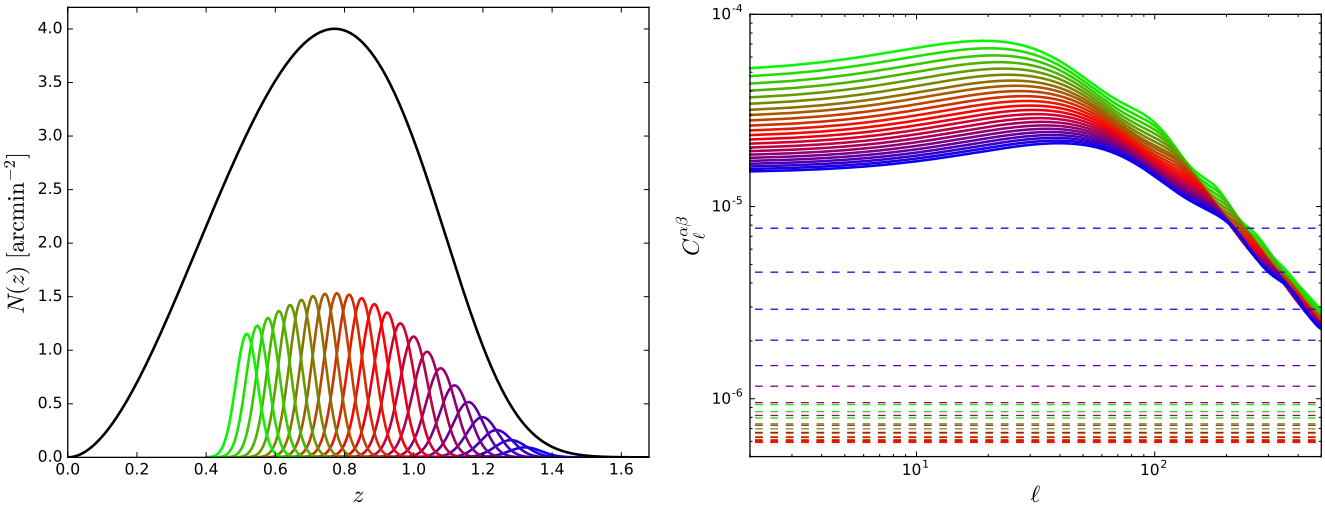

FIG. 4.

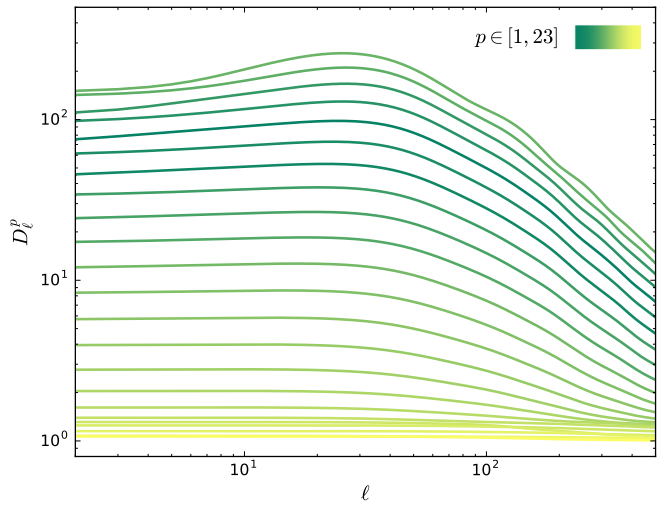

FIG. 5.

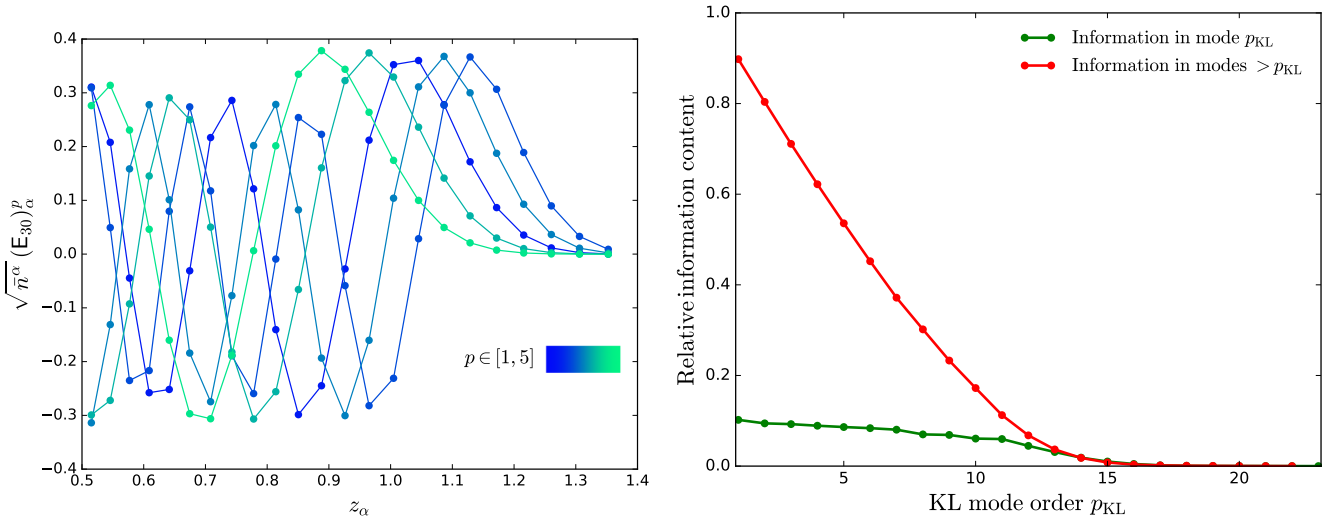

FIG. 6.

Here $b^\alpha(z)$ is the galaxy bias and $f(z) \equiv d\log\delta/d\log a$ is the growth rate of structure (we have kept the contribution from redshift-space distortions at linear order but ignored the effect of magnification bias).

IV. DISCUSSION

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

ACKNOWLEDGEMENTS

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

[1] M. Tegmark, A. N. Taylor, and A. F. Heavens, Astrophys. J. **480**, 22 (1997), astro-ph/9603021.

Appendix A: Pseudo-C_ℓ estimation of the K-L modes

One of the standard methods to estimate the angular power spectrum of any two quantities in the cut sky is the so-called pseudo-C_ℓ estimator. This section adapts this method to the modes resulting from the K-L decomposition described before.

The standard pseudo-C_ℓ method is based on computing the spherical harmonic coefficients of the mask field:

$$\tilde{a}^\alpha_{\ell m} = \int d\hat{\mathbf{n}}\, a^\alpha(\hat{\mathbf{n}})\, w^\alpha(\hat{\mathbf{n}}), \tag{A1}$$

where w^α is the weights map characterizing the mask of the field a^α. One then estimates the power spectrum of

this object by averaging over m for each ℓ:

$$\tilde{C}_\ell^{\alpha\beta} \equiv \frac{\sum_m \tilde{a}^\alpha_{\ell m}\tilde{a}^{\beta *}_{\ell m}}{2\ell+1}. \tag{A2}$$

This object is then related to the true underlying power spectrum through a mode-coupling matrix $M^{\alpha\beta}_{\ell\ell'}$ such that

$$\tilde{C}_\ell^{\alpha\beta} = \sum_{\ell'} M^{\alpha\beta}_{\ell\ell'} C^{\alpha\beta}_{\ell'}, \quad M^{\alpha\beta}_{\ell\ell'} \equiv \sum_{\ell''} \frac{(2\ell'+1)(2\ell''+1)}{4\pi} W^{\alpha\beta}_{\ell''} \begin{pmatrix} \ell & \ell' & \ell'' \\ 0 & 0 & 0 \end{pmatrix}^2 \tag{A3}$$

where the coupling matrix M depends solely on the power spectrum of the masks $W_\ell^{\alpha\beta} \equiv (2\ell+1)^{-1}\sum_m w^\alpha_{\ell m} w^{\beta *}_{\ell m}$.

The extension of this estimator to the power spectrum of the K-L modes is straightforward: we project the masked harmonic coefficients $\tilde{a}^\alpha$ over the K-L eigenvectors E (i.e. $\tilde{\mathbf{b}}_{\ell m} \equiv \mathsf{E}_\ell \circ \tilde{\mathbf{a}}_{\ell m}$) and compute their power spectra by averaging over m. The resulting estimator takes the form $\tilde{D}^p_\ell = \sum_{\ell'} M^{pp'}_{\ell\ell'} D^{p'}_{\ell'}$, where the new mode-coupling matrix is given by:

$$M^{pp'}_{\ell\ell'} \equiv M^{\alpha\beta}_{\ell\ell'} \left[(\mathsf{E}_\ell)^p_\alpha (\mathsf{N}^{-1})_{\alpha\alpha'} (\mathsf{E}_{\ell'})^{p'}_{\alpha'}\right] \left[(\mathsf{E}_\ell)^p_\beta (\mathsf{N}^{-1})_{\beta\beta'} (\mathsf{E}_{\ell'})^{p'}_{\beta'}\right] = M_{\ell\ell'} \tag{A4}$$

where the second equality holds only if all the maps a^α_ℓ share the same mask w.[3]

[3] Note that, for full-sky coverage $M_{\ell\ell'} = \delta_{\ell\ell'}$ and using the orthonormality of E we get $M^{pp'}_{\ell\ell'} = \delta_{\ell\ell'}\delta_{pp'}$.