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 some of the problems of standard tomographic analyses. The method is based on the Karhunen-Loève transform, and is connected to other 3D data compression bases advocated in the literature, such as the Fourier-Bessel decomposition. We apply this method to both weak lensing and galaxy clustering. In the case of galaxy clustering, we show that the resulting optimal basis is closely associated with the Fourier-Bessel basis, and that for certain observables, such as the effects of magnification bias or primordial non-Gaussianity, the bulk of the signal can be compressed into a small number of modes. In the case of weak lensing we show that the method is able to compress the vast majority of the signal-to-noise information into a single mode, and that optimal cosmological constraints can be obtained considering only three uncorrelated KL eigenmodes, considerably simplifying the analysis with respect to a traditional tomographic approach.

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. The Karhunen-Loeve transform

The idea behind the Karhunen-Loève transform, as developed within the field of cosmological data analysis in e.g. [1, 2], is to compress a given data vector into a small set of modes containing most of the useful information on a particular parameter (or set of parameters). Let $\mathbf{x}$ be a data vector of dimension N_s, and let θ be a particular parameter we want to measure. Under the assumption that $\mathbf{x}$ is Gaussianly distributed with mean 0 and covariance C, a set of linear combinations $y_p \equiv \mathbf{e}_p^\dagger \mathbf{x}$ can be found such that the y_p are white and uncorrelated ($\langle y_p y_q^* \rangle = \delta_{pq}$), and such that the first $m < N_s$ combinations contain most of the information about θ. This is done by solving the generalized eigenvalue problem [2]:

$$\partial_\theta \mathsf{C}\, \mathbf{e}_p = \lambda_p\, \mathsf{C}\, \mathbf{e}_p, \tag{1}$$

where $\partial_\theta \equiv \partial/\partial_\theta$.

Although the Karhunen-Loève transform can be used to compress the information on any particular parameter, it has been most commonly used to separate signal-dominated and noise-dominated modes by optimizing for the amplitude of the signal, as we explore below. Before moving on, however, it is worth noting that a generalized eigenvalue problem such as Eq. 1 can always be recast as a standard eigenvalue problem of the form $\mathsf{A}\, \tilde{\mathbf{e}}_p = \lambda_p\, \tilde{\mathbf{e}}_p$, where

$$\mathsf{A} \equiv \mathsf{C}^{-1/2}\, (\partial_\theta \mathsf{C})\, \mathsf{C}^{-1/2}, \quad \tilde{\mathbf{e}}_p \equiv \mathsf{C}^{1/2} \mathbf{e}_p, \tag{2}$$

and we have made use of the fact that C is positive-definite (and therefore $\mathsf{C}^{1/2}$ is well-defined and invertible).

1. The K-L transform for the signal-to-noise

Let us decompose the data vector $\mathbf{x}$ into uncorrelated signal and noise components $\mathbf{x} = \mathbf{s} + \mathbf{n}$ where, in this context, the signal is the part of the data containing any information of cosmological interest, and the noise is any contaminant preventing us from accessing it. In this particular case, the data covariance matrix can be split into their independent contributions $\mathsf{X} = \mathsf{S} + \mathsf{N}$.

The K-L transform has traditionally been used to design an eigenbasis that maximizes the overall signal-to-noise ratio (e.g [1, 3]). This can be done by defining a fictitious parameter α multiplying the signal part of the data with fiducial value $\alpha = 1$ (i.e. $\mathbf{x} = \alpha\mathbf{s} + \mathbf{n}$). In this case, after some trivial manipulations, the eigenvalue equation (Eq. 1) takes the form:

$$(\mathsf{S} + \mathsf{N})\mathbf{e}_p = \lambda_p \mathsf{N} \mathbf{e}_p, \tag{3}$$

where we have redefined $2/(2-\lambda_p) \to \lambda_p$. This can be cast into a standard eigenvalue equation using the Cholesky decomposition of the noise covariance matrix $\mathsf{N} = \mathsf{L}\mathsf{L}^\dagger$:

$$\left[\mathsf{L}^{-1}\mathsf{C}\,(\mathsf{L}^{-1})^\dagger\right]\,\tilde{\mathbf{e}}_p = \lambda_p \tilde{\mathbf{e}}_p, \tag{4}$$

where $\tilde{\mathbf{e}}_p \equiv \mathsf{L}^\dagger \mathbf{e}_p$.

At this point it is worth noting that the generalized eigenvalue problem in Eq. 3 can be understood as the problem diagonalizing C under a non-standard dot product $\circ$ given by the inverse noise covariance matrix (i.e. $\mathbf{a}\circ\mathbf{b} \equiv \mathbf{a}^\dagger\mathsf{N}^{-1}\mathbf{b}$). Under this dot product, an eigenbasis $\mathsf{F} \equiv (\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_{N_s})$ can be found such that F is orthonormal $\mathsf{F}\circ\mathsf{F} = \mathsf{I}$, and the covariance of the transformed data vector $\mathbf{y} \equiv \mathsf{F}\circ\mathbf{x}$ is diagonal:

$$\langle \mathbf{y}\,\mathbf{y}^\dagger\rangle = \mathsf{F}^\dagger\mathsf{N}^{-1}\mathsf{C}\mathsf{N}^{-1}\mathsf{F} = \Lambda \equiv \mathrm{diag}(\lambda_1, ..., \lambda_{N_s}). \tag{5}$$

Using the orthonormality of F (with respect to the non-standard dot product), this can be cast into the same form as Eq. 4, where $\mathbf{f}_p = \mathsf{L}\tilde{\mathbf{e}}_p = \mathsf{N}\mathbf{e}_p$.

Finally, note that, because both S and N are positive-definite matrices, their eigenvalues will also be positive. Since the eigenvalues of N under the K-L transform are, by construction, 1, the elements of Λ above will all be greater than 1, and converging to 1 for the noise-dominated modes.

2. *The K-L transform with correlated contaminants*

Let us now consider a more general case in which we further split the noise into two parts $\mathbf{n} \to \mathbf{n} + \mathbf{m}$, where $\mathbf{m}$ is a contaminant with a non-zero correlation with the signal. The covariance matrix of the data is then given by:

$$\langle \mathbf{x}\,\mathbf{x}^\dagger\rangle = \alpha^2\mathsf{S} + 2\alpha\mathsf{M}_s + \mathsf{M} + \mathsf{N}, \tag{6}$$

where $\mathsf{M}_s \equiv (\langle \mathbf{m}\,\mathbf{s}^\dagger\rangle + \langle \mathbf{s}\,\mathbf{m}^\dagger\rangle)/2$, $\mathsf{M} \equiv \langle \mathbf{m}\,\mathbf{m}^\dagger\rangle$ and we have kept the fictitious parameter α defined in the previous section. Eq. 1 then reads:

$$(\mathsf{S} + \mathsf{M}_s)\,\mathbf{e}_p = \frac{\lambda_p}{2}\mathsf{C}\,\mathbf{e}_p. \tag{7}$$

Unfortunately, in this case the manipulation that lead us to Eq. 3 in the previous section cannot be performed. If we were to do so, the matrix remaining on the right hand side of this equation would not be positive-definite, and the corresponding generalized eigenvalue problem would be ill-defined. This is not a problem, since the solutions to Eq. 7 still separate the modes with the highest signal. The separation of the noise-dominated modes becomes less obvious, however, since the resulting eigenvalues cannot be simply compared with 1, corresponding to noise-dominated modes in the previous section.

The eigenvector solutions to the generalized eigenvalue problem in Eq. 7 can be collected as columns of a matrix E that simultaneously satisfy the equations:

$$\mathsf{E}^\dagger\,(\mathsf{S} + \mathsf{M}_s)\,\mathsf{E} = \Lambda, \quad \mathsf{E}^\dagger\mathsf{C}\mathsf{E} = \mathsf{I}, \tag{8}$$

where I is the identity and $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_{N_s})$. Since the second equation implies $\mathsf{C}\mathsf{E} \equiv (\mathsf{E}^\dagger)^{-1}$, the original vector $\mathbf{x}$ can be recovered from the coefficients $\mathbf{y} \equiv (y_1, ..., y_{N_s})$ as $\mathbf{x} = \mathsf{C}\,\mathsf{E}\,\mathbf{y}$. More interestingly, one can identify the principal eigenvectors of the Eq. 7 (e.g. those with associated eigenvalues λ_p above a given threshold $\lambda_{\rm thr}$) and project out the remaining modes, which are presumably more contaminated by $\mathbf{m}$. This procedure defines a filter $\mathsf{W} \equiv \mathsf{C}\,\mathsf{E}\,\mathsf{P}\,\mathsf{E}^\dagger$, where P is a projection matrix with 1s in the diagonal elements corresponding to the principal eigenmodes and zeros everywhere else. The filtered data vector is therefore $\tilde{\mathbf{x}} = \mathsf{W}\mathbf{x}$.

B. Application to tomographic datasets

The standard 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 CITES, making use of various calibration methods in order to estimate the true redshift distribution 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. Other approaches ([4–6]) involve projecting the main observable (e.g. galaxy overdensity or shear) onto the Fourier-Bessel eigenbasis. None of these schemes are manifestly optimal from the point of view of S/N, final cosmological constraints or contaminant deprojection, however. This section presents an alternative slicing scheme addressing these shortcomings, based on the K-L transform.

1. *Tomographic analyses*

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 $a^\alpha(\hat{\mathbf{n}})$ 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 a^α[1]. The power spectrum for our set of subsamples is

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

defined as the two-point correlator of $a^\alpha_{\ell m}$:

$$\left\langle \mathbf{a}_{\ell m}\,\mathbf{a}^\dagger_{\ell' m'}\right\rangle \equiv \delta_{\ell\ell'}\delta_{mm'}\mathsf{C}_\ell, \tag{9}$$

where we have packaged $a^\alpha_{\ell m}$ as a vector for each (ℓ, m): $\mathbf{a}_{\ell m} \equiv (a^1_{\ell m}, ..., a^{N_s}_{\ell m})$. Usually the observed field can be decomposed into uncorrelated signal and noise component $\mathbf{a} = \mathbf{s}+\mathbf{n}$, with a similar decomposition in the power spectrum, $\mathsf{C}_\ell = \mathsf{S}_\ell + \mathsf{N}_\ell$.

Once the choice of subsamples α is made, 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. The brute-force solution to make sure one exploits all of the information present in the data would be to use 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 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 neglecting all correlations beyond a given neighbouring index is not clear a priori.

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.

2. *Optimal radial eigenbasis*

Following the description in Section II A 1, it is straightforward to derive an optimal set of radial, uncorrelated eigenmodes.

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 $\tilde{\mathbf{a}}_{\ell m} \equiv \mathsf{N}^{-1}_\ell\,\mathbf{a}_{\ell m}$.

2. Let us consider a set of linear combinations of the weighted field measured on narrow redshift bins:

$$\mathbf{b}_{\ell m} = \mathsf{F}^\dagger_\ell \cdot \tilde{\mathbf{a}}_{\ell m} \equiv \mathsf{F}_\ell \circ \mathbf{a}, \tag{10}$$

where F_ℓ is a yet-unspecified matrix and, as in Section II A 1, we have let N^{-1}_ℓ define the non-standard dot product $\mathbf{v}_\ell \circ \mathbf{w}_\ell \equiv \mathbf{v}^\dagger_\ell \cdot \mathsf{N}^{-1}_\ell \cdot \mathbf{w}_\ell$. The power spectrum for this new observable would then simply be given by:

$$\mathsf{D}_\ell \equiv \left\langle \mathbf{b}_{\ell m}\,\mathbf{b}^\dagger_{\ell m}\right\rangle = \mathsf{F}^\dagger_\ell \circ \mathsf{C}_\ell \circ \mathsf{F}_\ell. \tag{11}$$

3. Requiring that the new modes be uncorrelated, we can identify Eq. 11 with the generalized eigenvalue equation 5, which defines the K-L eigenbasis F_ℓ by additionally requiring that it be orthonormal ($\mathsf{F}_\ell \circ \mathsf{F}_\ell = \mathsf{I}$). 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. The data can be further compressed by assuming 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 expectation value of the Hessian of the log-likelihood with respect to Θ), which assuming $\langle \mathbf{x}\rangle = 0$ reads

$$\mathcal{F}_{ij} \equiv \langle \partial_i \partial_j \mathcal{L}\rangle = \frac{1}{2}\mathrm{Tr}\left(\partial_i \mathsf{X}\,\mathsf{X}^{-1}\partial_j \mathsf{X}\,\mathsf{X}^{-1}\right), \tag{12}$$

where $\mathsf{X} \equiv \langle \mathbf{x}\,\mathbf{x}^\dagger\rangle$ is the covariance matrix of the data. Since the power spectrum of $\mathbf{b}_{\ell m}$ defined above is diagonal, this expression gets simplified further, and the Fisher matrix can be decomposed into the independent contributions of each mode: $\mathcal{F}_{ij} = \sum_p \mathcal{F}^p_{ij}$, where

$$\mathcal{F}^p_{ij} \equiv \sum_\ell \frac{2\ell+1}{2}\left(\partial_i \log D^p_\ell\right)\left(\partial_j \log D^p_\ell\right). \tag{13}$$

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

5. The final set of uncorrelated modes can then be truncated to the first M defined by this procedure, which will contain the bulk of the information needed to constrain Θ.

Besides the elegance of this method in defining a natural set of radial basis functions for the particular dataset under study, analogous to the Fourier-Bessel basis in a translationally-invariant system (see Section III A), its

for fields of arbitrary spin.

merits are better evaluated in terms of data compression. This strategy 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. This can lead, for instance, to a substantial reduction of the computational costs of likelihood sampling and covariance estimation.

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 K-L eigenmodes F_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^p modes. These modes are be characterized by their own window function:

$$\tilde{\phi}^p_\ell(z) = \sum_\alpha \frac{(\mathsf{F}_\ell)^p_\alpha\, \phi^\alpha(z)}{N_\ell^{\alpha\alpha}}, \tag{14}$$

where we have assumed a diagonal noise power spectrum for simplicity. The same methods used to calibrate photo-z uncertainties in the standard tomographic analysis can be applied on b^p to calibrate $\tilde{\phi}^p$ with minor modifications (e.g. weighed and ℓ-dependent stacking of photo-z pdfs, or cross-correlations of the b^p maps with a spectroscopic survey in the case of clustering redshifts). Furthermore, using F_ℓ for the fiducial cosmology as model-agnostic weights and inserting them in Eq. 11, the theoretical prediction for the power spectrum of each mode D_ℓ^p can be computed in a model-independent way.

III. PERFORMANCE AND PARTICULAR EXAMPLES

This Section explores the performance of the K-L decomposition in a number of specific science cases.

A. Special case: the harmonic-Bessel basis

Let us consider a simplified case where the field a is the overdensity field of a non-evolving galaxy population 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 δr, which we label by their comoving radius r. The noise in the measurement of a 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 $N_\ell(r,r') \propto \delta_{r,r'}\, r^{-2}$. Thus, the dot product is just given by:

$$\mathbf{b}^\dagger \circ \mathbf{c} \propto \int dr\, r^2\, b(r)^*\, c(r). \tag{15}$$

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

$$S_\ell^{rr'} = \frac{2}{\pi}\int_0^\infty dk\, k^2\, P_k\, j_\ell(kr) j_\ell(kr'), \tag{16}$$

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

$$\begin{aligned} D_\ell^{kk'} &\equiv \sum_{r,r'} (F_\ell)^k_r\, (F_\ell)^{k'}_{r'}\, S_\ell^{rr'} \\ &\propto \frac{2}{\pi}\int dr\, r^2 \int dr'\, r'^2 j_\ell(kr) j_\ell(k'r') S_\ell^{rr'} \\ &= \int dq\, q^2 P_q \left[\frac{2}{\pi}\int dr\, r^2\, j_\ell(qr) j_\ell(kr)\right]\left[\frac{2}{\pi}\int dr'\, r'^2\, j_\ell(qr') j_\ell(k'r')\right] \\ &= \int dq\, q^2 P_q \frac{\delta(k-q)}{q^2}\frac{\delta(k'-q)}{q^2} = P_k \frac{\delta(k-k')}{k^2} = \frac{P_k}{k^2 \Delta k}\delta_{k,k'} \end{aligned} \tag{17}$$

This choice of basis defines the so-called harmonic-Bessel (or Fourier-Bessel) decomposition, and has been used as a data-compression method for the analysis of photometric redshift datasets (e.g. [6]). In any realistic scenario – e.g. in the presence of redshift uncertainties, redshift-space distortions or in the analysis of weak lensing data – this basis is non-optimal, since different k-modes will be correlated, as opposed to the K-L basis described in the previous section.

B. Galaxy clustering - Bessel-like eigenfunctions

The assumptions used in the previous section are an ideal limit of the data collected by a photometric survey. In a more realistic (although still idealized) scenario, the information about the radial position of a given source is encoded in its posterior photo-z distribution $p(z|\alpha)$, where α is a continuous variable determining the properties of the photo-z (e.g. the mean of the posterior). The cross-power spectrum of two samples with photo-z

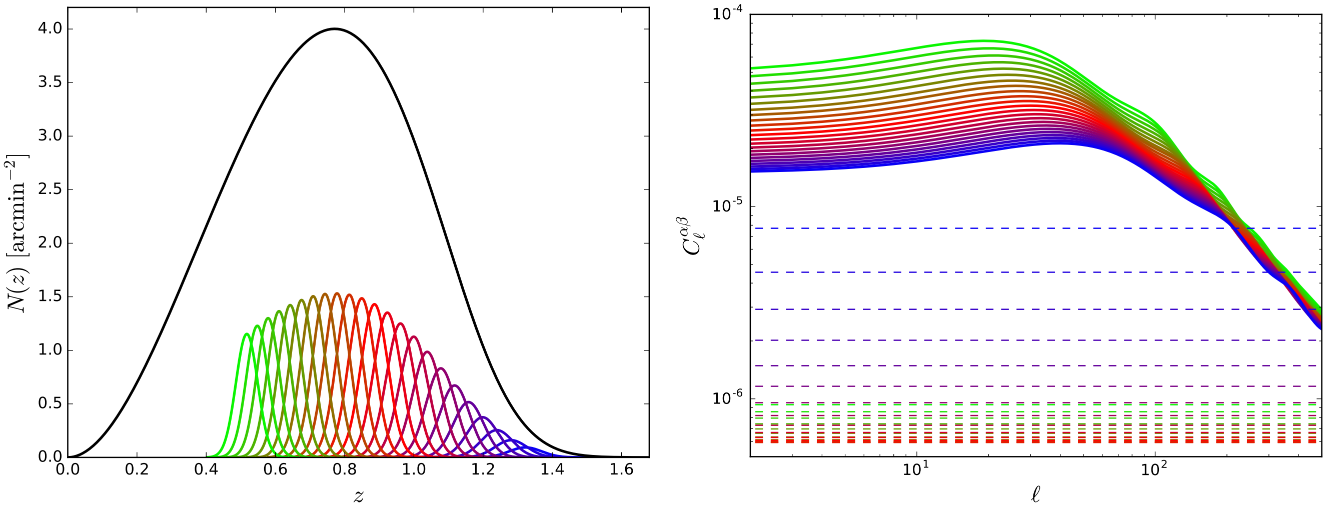

FIG. 1. *Left*: redshift distribution and bins considered for the K-L analysis of a strawman large-scale-structure survey targeting a sample of red galaxies. *Right*: clustering auto-power spectra of the redshift bins shown in the left panel. The signal and noise power spectra are shown as thick solid and thin dashed lines respectively.

properties α and β is given by

$$C_\ell^{\alpha\beta} = S_\ell^{\alpha\beta} + N_\ell^{\alpha\beta}, \tag{18}$$

$$S_\ell^{\alpha\beta} = \frac{2}{\pi}\int_0^\infty dk\, k^2\, \Delta_\ell^\alpha(k)\, \Delta_\ell^\beta(k), \tag{19}$$

$$N_\ell^{\alpha\beta} = \frac{\delta(\alpha-\beta)}{n_t\, p(\alpha)}, \tag{20}$$

where n_t is the total angular number density of sources,

$$\Delta_\ell^\alpha(k) \equiv \int dz\, p(z|\alpha)\, \Psi_\ell(k,z)\, \sqrt{P(k,z)},$$
$$\Psi_\ell(k,z) = b^\alpha(z) j_\ell(k\,\chi(z)) - f(z) j_\ell''(k\,\chi(z)). \tag{21}$$

Here $b^\alpha(z)$ is the linear galaxy bias, $f(z) = d\log\delta/d\log a$ is the growth rate of structure, $P(k,z)$ is the matter power spectrum at redshift z, $p(\alpha)$ is the probability that a source has photo-z properties α, and $p(z|\alpha)$ is the conditional redshift distribution of these sources (we have labelled this quantity $\phi^\alpha(z)$ in previous sections). Note that, for simplicity, we have kept the contribution of redshift-space distortions at linear order and neglected the effect of magnification (this will be studied in Section REF).

For a continuous variable α, the generalized eigenvalue problem in Eq. 3 becomes a homogeneous Fredholm integral equation of the second kind:

$$\int d\beta\, C_\ell^{\alpha\beta} e_\ell^p(\beta) = \lambda_p \int d\beta\, N_\ell^{\alpha\beta} e_\ell^p(\beta) \Rightarrow \tag{22}$$

$$\Rightarrow \int d\beta\, n_t\, p(\alpha) S_\ell^{\alpha\beta}\, e_\ell^p(\beta) = (\lambda_p - 1) e_\ell^p(\alpha). \tag{23}$$

In the limit of perfect photo-z ($p(z|\alpha) = \delta(z-\alpha)$), and in the absence of redshift-space distortions, the solution to this equation are the spherical Bessel functions, as proven in the previous section. For general kernels, however, no analytical solution to the homogeneous Fredholm equation can usually be found, and the standard procedure to solve it is through discretization, which is equivalent to taking finite bins in α. We will use this method here to find the K-L eigenmodes that maximize the signal content for galaxy clustering.

To do so, we have considered a specific strawman photometric survey targeting a sample of red galaxies, characterized by their higher bias and better photo-z uncertainties than their blue counterparts (and therefore better suited for clustering analyses). The sample we consider is compatible with what could be observed by LSST, characterized by the redshift distribution shown in the left panel of Fig. 1 (full details can be found in [7]). We assume a photo-z uncertainty of $\sigma_z = 0.02\,(1+z)$ and split the sample into redshift bins in photo-z space with $z_{\rm ph} > 0.5$ and a width given by the photo-z uncertainty at the bin centre. The auto-power spectra for our set of 23 bins are shown in the right panel of Fig. 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.

Using the prescription described in Section II A, we find the K-L eigenmodes and associated power spectra, and rank them according to their contribution to the total signal-to-noise ratio (defined here as the Fisher matrix element of the signal amplitude). The power spectra of the resulting K-L modes are shown in Figure 2. Unlike the case of weak lensing, explored in Section III E, the information encoded in the galaxy overdensity is local in redshift, and thus the correlation between different bins decays rapidly with redshift separation. The signal-to-noise is therefore spread over ~ 15 signal-dominated

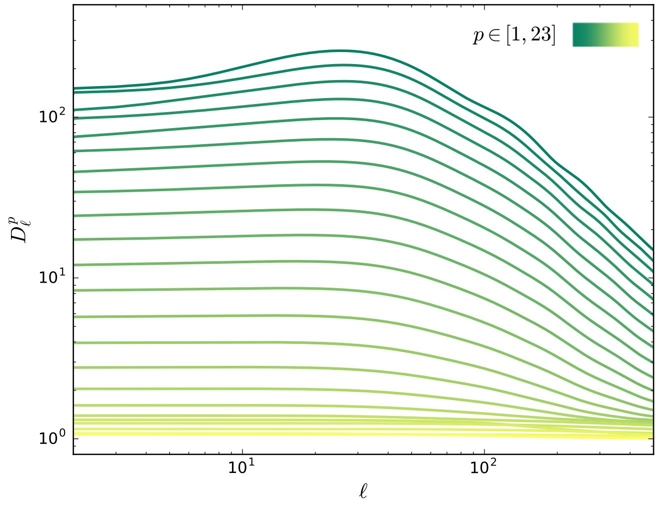

FIG. 2. Power spectra of the K-L eigenmodes for the strawman weak large-scale-structure survey. Unlike in the case of weak lensing, a large number of eigenmodes are signal-dominated. This is due to the overall higher signal-to-noise ratio of galaxy clustering with respect to galaxy shear as well as to the smaller correlations between distant bins.

modes, and the noise-dominated modes can be thought of as the radial scales filtered out by the finite photo-z uncertainty (as we mentioned in Section II A 1, the noise power spectrum gets mapped into 1 under the K-L transform). The relative contribution of each mode to the total signal-to-noise is shown in the top panel of Fig. 3. 90% of the total constraining power can be achieved by considering the first 13 eigenvectors. The form of the first 7 of these eigenvectors for $\ell = 30$ are shown in the right panel of Fig. 3. The eigenmodes are sinusoids with increasing frequencies, in agreement with the expectation that, in the limit of $\sigma_z \to 0$ and no background redshift dependence, the K-L decomposition is achieved by the spherical Bessel functions. A Fourier-Bessel decomposition is therefore probably a near-optimal analysis method, although the K-L decomposition allows a more precise determination of the truly orthogonal radial modes.

C. Galaxy clustering - optimal basis for $f_{\rm NL}$

It is expected that future large-scale photometric surveys will make the search for primordial non-Gaussianity one of their main science cases. This can be achieved by measuring the excess power on large scales caused by a non-zero value of $f_{\rm NL}$[2] generates in the two-point statistics of biased tracers of the matter distribution CITES.

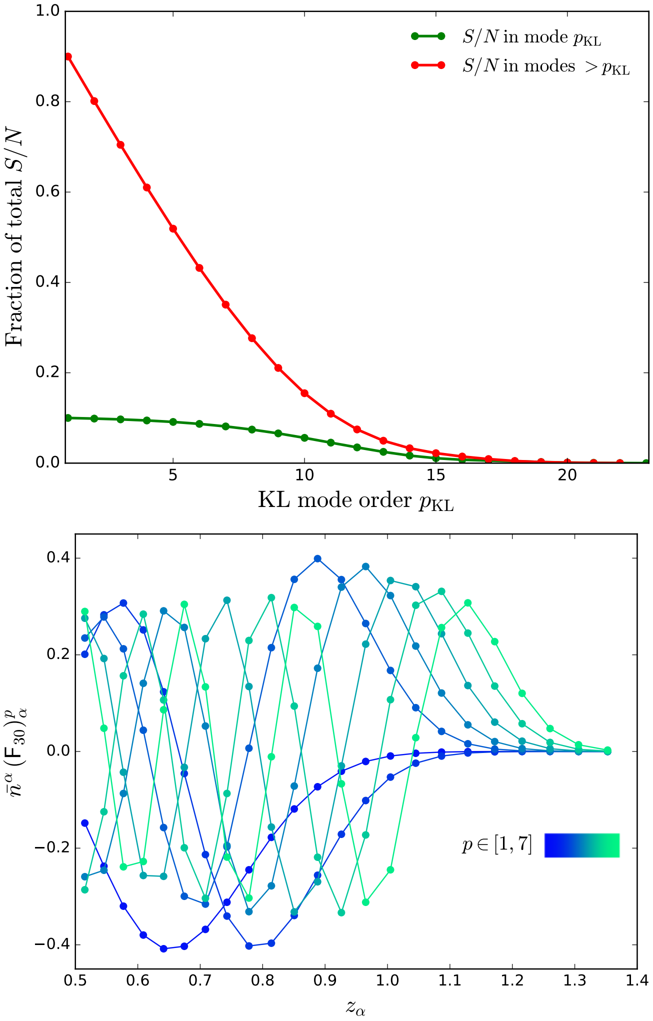

FIG. 3. *Left*: information content of the different K-L eigenmodes for the strawman galaxy clustering survey. The bulk of the information ($> 90\%$) on $f_{\rm NL}$ is encoded in the first 12 modes. *Left*: the first 5 K-L modes for $\ell = 30$. The sinusoidal shape of the modes agrees with the expectation that, in the limit of $\sigma_z \to 0$ and no background redshift dependence, the K-L modes should be given by the spherical Bessel functions.

Since the signal is most relevant on large scales, we can expect the bulk of it to be concentrated in a small number of radial modes, which makes the general K-L decomposition outlined in Section II A an ideal analysis method. Similar approaches have been explored in the literature to devise optimal weights for spectroscopic galaxy surveys CITE.

We again consider the red galaxy sample used in the previous section, but now estimate the K-L basis of eigenmodes that optimize the information content on $f_{\rm NL}$ instead of the overall signal amplitude. I.e. we solve the generalized eigenvalue problem in Eq. 1 where $\theta = f_{\rm NL}$. We compare the performance of this basis with other choices of radial modes as follows: for a given number of modes, we estimate the associated uncertainty on $f_{\rm NL}$,

[2] The reader is referred to CITES for a thorough review of non-Gaussianity and a definition of $f_{\rm NL}$.

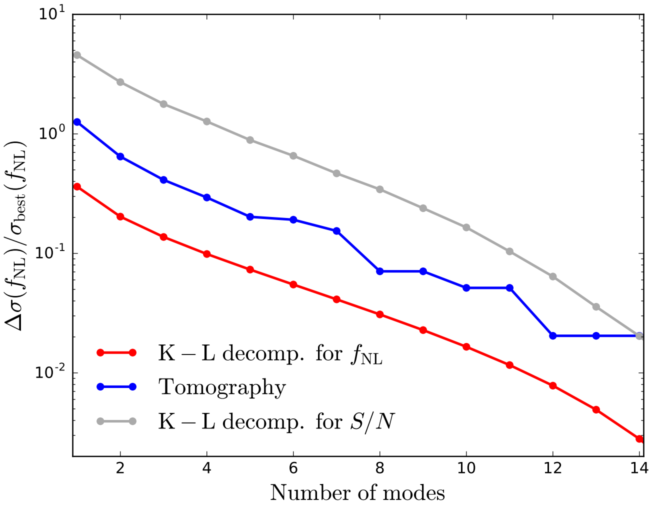

FIG. 4. *Left*: redshift distribution and bins considered for the K-L analysis of a strawman lensing survey

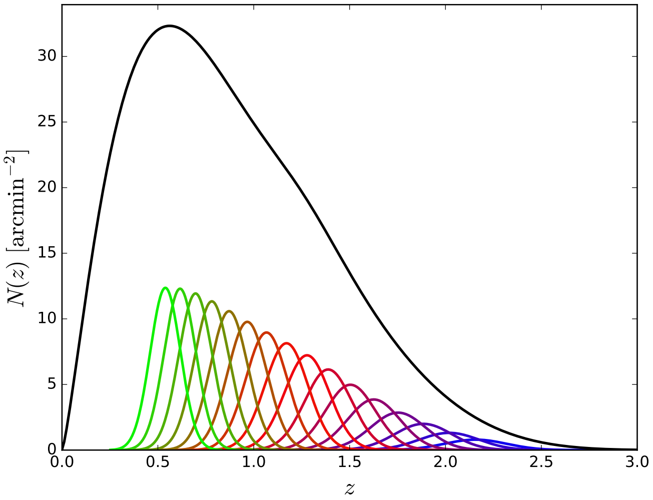

FIG. 5. *Left*: redshift distribution and bins considered for the K-L analysis of a strawman lensing survey

$\sigma(f_{\rm NL})$ by summing the contributions to the corresponding Fisher matrix element of those modes, and compute the excess of $\sigma(f_{\rm NL})$ with respect to the best achievable constraint $\sigma_{\rm best}(f_{\rm NL})$. The results are shown in Fig. 4 for three choices of radial functions:

- The K-L eigenbasis resulting from optimizing the information content on $f_{\rm NL}$ discussed in this section. The results are shown in red.

- The K-L eigenbasis resulting from optimizing the overall signal-to-noise of the galaxy clustering signal, as discussed in the previous section. The results are shown in gray.

- Photo-z tomography: the result of dividing the galaxy sample into a number of top-hat photo-z bins of equal width. The results are shown in blue.

As demonstrated by this Figure, for a fixed number of modes the optimal K-L basis always outperforms any other data compression prescription. In particular, the constraints on $f_{\rm NL}$ are only degraded by $\sim 30\%$ when considering only the first principal eigenmode, and almost 90% of the total constraining power is contained in the first three. Interestingly, a naive tomographic approach achieves the same uncertainty on $f_{\rm NL}$ with a smaller number of modes (redshift bins) than the K-L eigenbasis for the S/N. However, since the tomographic bins are not orthogonal, unlike the K-L modes, for a fixed $\sigma(f_{\rm NL})$ both K-L bases typically outperform the tomographic approach in terms of the size of the associated power spectrum. In any case, this example serves to stress the fact that the optimal radial basis in terms of overal S/N is not necessary optimal in terms of final constraints for cosmological parameters that depend on specific features of the power spectrum.

D. Galaxy clustering - magnification bias

Gravitational lensing of the galaxy positions alters the clustering pattern of galaxies through the so-called magnification bias effect. This appears as an extra term in the galaxy clustering transfer function (Eq. 21):

$$\Delta^{M,\alpha}_\ell(k) = -2\ell(\ell+1)\int d\chi W^{M,\alpha}(\chi)\frac{j_\ell(k\chi)}{k^2a(\chi)}\sqrt{P(k,z(\chi))},$$
$$W^{M,\alpha}(\chi) = \frac{3H_0^2\Omega_M}{2}\int_{z(\chi)}^\infty dz'\,\phi^\alpha(z')\frac{2-5\,s}{2}\,\frac{\chi(z')-\chi}{\chi(z')\chi}, \tag{24}$$

where s is the tilt in the number counts of sources as a function of magnitude limit. This effect, commonly labeled "magnification bias" [8–10], can be used as an alternative measurement of gravitational lensing, through galaxy positions instead of shapes. The contribution of the magnification term is, however, weak in comparison with the density and RSD terms (Eq. 21), and therefore its measurement can be hampered by the cosmic variance contribution of these terms.

One can therefore think of the density and RSD terms as correlated contaminants of the magnification signal, and use the K-L formalism described in Section II A 2 to devise an optimal basis of radial eigenmodes containing the bulk of its signal-to-noise.

To test this approach we consider, as in the previous section, an LSST-like survey. Since lensing magnification is an integrated effect, it is less hampered by poor photo-z uncertainties, and it is most easily measured by cross-correlating high-redshift and low-redshift data CITES. For this reason, in this case we consider a sample of blue galaxies, with inferior photo-z errors but wider redshift support. Full details can be found in [7]. In summary, we consider a sample with ~ 40 objects per arcmin^2 with

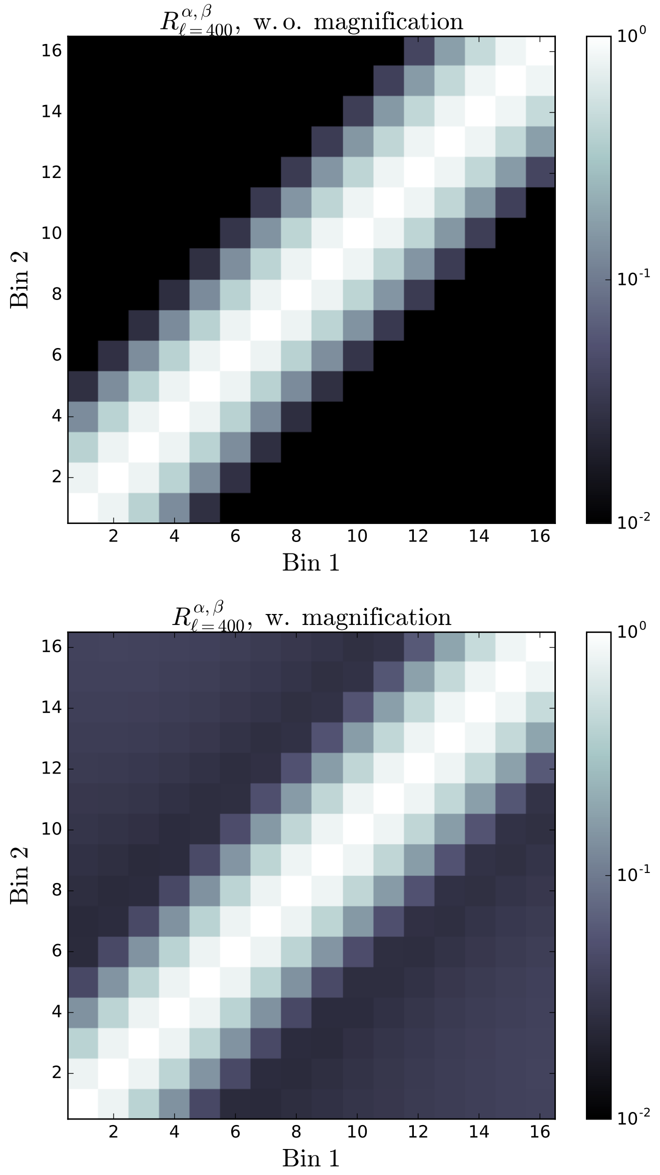

FIG. 6. *Left*: information content of the different K-L eigenmodes for the strawman galaxy clustering survey. The bulk of the information ($> 90\%$) on $f_{\rm NL}$ is encoded in the first 12 modes. *Left*: the first 5 K-L modes for $\ell = 30$. The sinusoidal shape of the modes agrees with the expectation that, in the limit of $\sigma_z \to 0$ and no background redshift dependence, the K-L modes should be given by the spherical Bessel functions.

the redshift distribution shown in Figure 5. We also approximate the photo-z distributions as Gaussians with a scatter $\sigma_z = 0.05(1+z)$, and divide the sample into 16 top-hat bins in photo-z space with $z_{\rm ph} < 0.5$ and widths given by the value of σ_z at the bin center (again, we verified that our conclusions did not change after decreasing the width by a factor 2).

A key property of the magnification bias effect is the fact that, since gravitational lensing is caused by the integrated matter distribution between source and

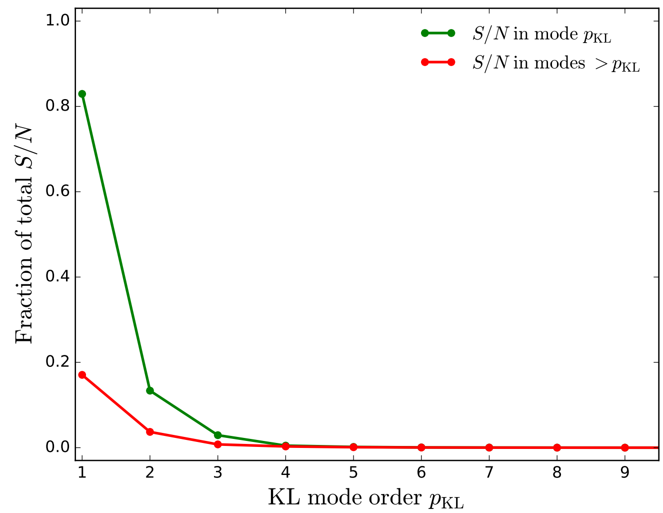

FIG. 7. *Left*: information content of the different K-L eigenmodes for the strawman galaxy clustering survey. The bulk of the information ($> 90\%$) on $f_{\rm NL}$ is encoded in the first 12 modes. *Left*: the first 5 K-L modes for $\ell = 30$. The sinusoidal shape of the modes agrees with the expectation that, in the limit of $\sigma_z \to 0$ and no background redshift dependence, the K-L modes should be given by the spherical Bessel functions.

observer, the magnification signals in widely separated redshift bins can be tightly correlated. This is shown explicitly in Figure 6. The figure shows the correlation coefficients between the 16 redshift bins, defined as $R^{\alpha\beta}_\ell = C^{\alpha\beta}_\ell/\sqrt{C^{\alpha\alpha}_\ell C^{\beta\beta}_\ell}$, at $\ell = 400$, with (right panel) and without (left panel) the magnification bias effect. Although the contribution of lensing magnification to the correlation between neighbouring bins is subdominant, it produces noticeable correlations between distant ones.

This property is particularly interesting in the context of the K-L decomposition: a signal that is tightly correlated accross samples will contribute significantly only to a small set of eigenmodes. To explore this possibility, we follow the prescription outlined in Section II A 2 for correlated contaminants. The contribution of each eigenmode to the total signal-to-noise of the magnification bias (in a Fisher-matrix sense) is shown in Figure 7. As expected, most of the signal ($> 80\%$) is contained in the first eigenvalue, with the practical totality of it concetrated in the first three modes.

We finish this section by noting that this approach is similar to the "nulling" method of [11], and that an analogous treatment could be carried out on the cosmic shear field to separate the lensing and intrinsic alignment contributions [12].

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

The effects of gravitational lensing can be measured directly by studying the correlation it induces on the

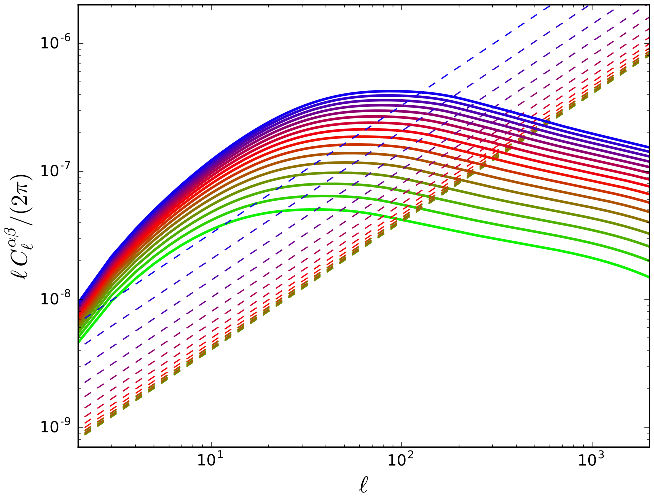

FIG. 8. Shear auto-power spectra of the redshift bins shown in the left panel. The signal and noise power spectra are shown as thick solid and thin dashed lines respectively.

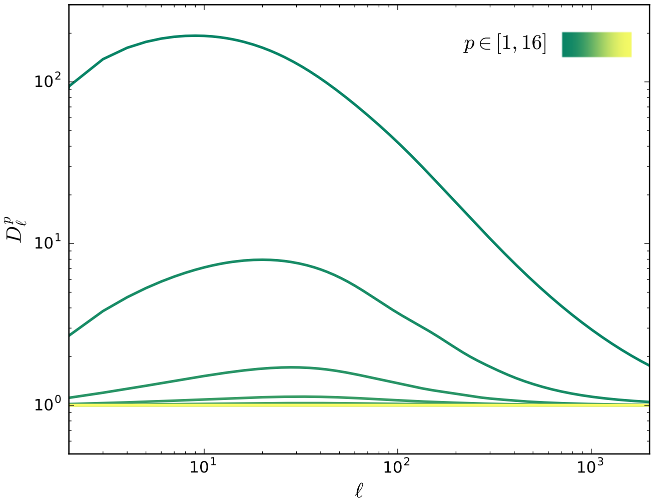

FIG. 9. Power spectra of the K-L eigenmodes for the strawman weak lensing survey. All but the first three modes are noise-dominated, and most of the information is encoded in the first mode.

shapes and orientation of galaxy images. This effect, labeled "cosmic shear" is arguably the most promising observational probe for photometric redshift surveys, and therefore we will discuss the K-L analysis of this signal in particular detail.

As in the case of lensing magnification, and unlike the dominant galaxy clustering terms, the cosmic shear signal is correlated between widely separated redshift bins due to the integrated nature of gravitational lensing. Thus we can expect that a K-L transform should be able to compress most of the signal to noise into a small set of radial eigenmodes. To quantify this mothed we consider the same survey configuration used in Section III D. The signal part of the cross-power spectrum between the cosmic shear measurements made in two different redshift shells is given again by Eq. 19, where now the transfer functions Δ_ℓ^α take the form:

$$\Delta_\ell^\alpha(k) \equiv \sqrt{\frac{(\ell+2)!}{(\ell-2)!}}\int d\chi\, W^\alpha(\chi)\frac{j_\ell(k\chi)}{k^2 a(\chi)}\sqrt{P(k,z(\chi))},$$
$$W^\alpha(\chi) \equiv \frac{3H_0^2\Omega_M}{2}\int_{z(\chi)}^\infty dz\,\phi^\alpha(z')\frac{\chi(z')-\chi}{\chi(z')\chi}. \qquad (25)$$

The noise power spectrum is white and simply given by the intrinsic ellipticity scatter weighed by the angular number density of sources in each redshift bin $\bar{n}^\alpha$:

$$N_\ell^{\alpha\beta} = \delta_{\alpha\beta}\frac{\sigma_\gamma^2}{\bar{n}^\alpha}, \qquad (26)$$

with $\bar{n}^\alpha$ in units of srad^{-1}. We use $\sigma_\gamma = 0.28$. The lensing auto-power spectra (both signal and noise) for these bins are shown in Figure 8.

We compute the K-L modes for this setup and rank them according to their contribution to the total lensing signal (in a Fisher matrix sense). The power spectra of the resulting set of modes are shown in the left panel of Figure 9. Comparing against Fig. 8 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. The fractional contribution of each mode to the total signal-to-noise is shown in the left panel of Figure 10. Most of the signal ($\sim 95\%$) 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 three principal modes for different values of ℓ are shown in the right panel of the same figure. 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 and third eigenmodes then recover part of this information by marginally upweighting these regions.

Since the cosmological lensing signal at a given redshift z is caused by matter distribution at all redshifts $< z$, different redshift bins are tightly correlated at the signal level. This is the main reason for the large degree of data compression achievable in this case, a statement that is in principle largely independent of the properties of the photometric redshifts. Exploiting this property could therefore significantly simplify the analysis of weak lensing datasets.

IV. PRACTICAL EXAMPLE: WEAK LENSING

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.

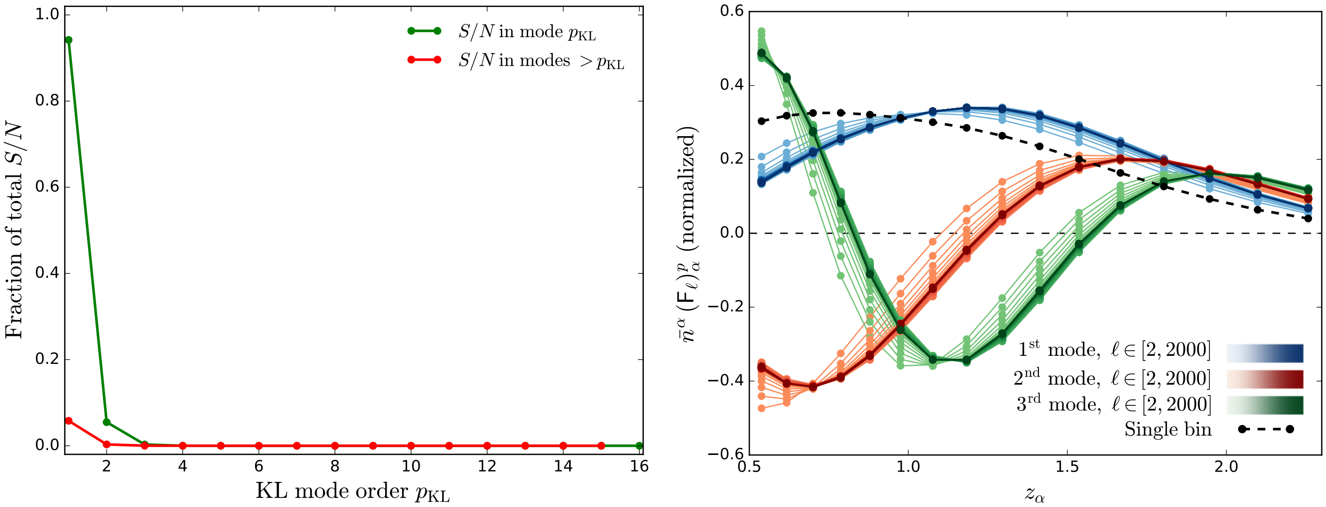

FIG. 10. *Left*: the first (blue-cyan) and second (red-yellow) eigenmodes of the strawman weak-lensing survey for different ℓ. The redshift dependence of the modes stays roughly constant across ℓ and converges to a fixed shape for large ℓ. *Right*: information content of the different eigenmodes. Most of the information $\sim 95\%$ is encoded in the first bin, and more than 99% of it can be recovered considering only the first 2 modes.

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.

V. DISCUSSION

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.

open problems: contaminant deprojection, mode coupling from incomplete sky, IA in single-mode lensing

ACKNOWLEDGEMENTS

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

[1] M. S. Vogeley and A. S. Szalay, Astrophys. J. **465**, 34 (1996), astro-ph/9601185.

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

[3] J. R. Bond, Physical Review Letters **74**, 4369 (1995), astro-ph/9407044.

[4] A. Heavens, MNRAS **343**, 1327 (2003), astro-ph/0304151.

[5] T. D. Kitching, A. F. Heavens, A. N. Taylor, M. L. Brown, K. Meisenheimer, C. Wolf, M. E. Gray, and D. J. Bacon, MNRAS **376**, 771 (2007), astro-ph/0610284.

[6] T. D. Kitching, A. F. Heavens, J. Alsing, T. Erben, C. Heymans, H. Hildebrandt, H. Hoekstra, A. Jaffe, A. Kiessling, Y. Mellier, L. Miller, L. van Waerbeke, J. Benjamin, J. Coupon, L. Fu, M. J. Hudson, M. Kilbinger, K. Kuijken, B. T. P. Rowe, T. Schrabback, E. Semboloni, and M. Velander, MNRAS **442**, 1326 (2014), arXiv:1401.6842.

[7] D. Alonso, P. Bull, P. G. Ferreira, R. Maartens, and M. G. Santos, Astrophys. J. **814**, 145 (2015),

arXiv:1505.07596.

[8] J. E. Gunn, Astrophys. J. **147**, 61 (1967).

[9] T. Matsubara, ApJL **537**, L77 (2000), astro-ph/0004392.

[10] M. Loverde, L. Hui, and E. Gaztañaga, Phys. Rev. D **77**, 023512 (2008), arXiv:0708.0031.

[11] A. F. Heavens and B. Joachimi, MNRAS **415**, 1681 (2011), arXiv:1101.3337.

[12] B. Joachimi and P. Schneider, A&A **488**, 829 (2008), arXiv:0804.2292.

[13] J. D. McEwen and B. Leistedt, ArXiv e-prints (2013), arXiv:1307.1307 [cs.IT].

[14] B. Leistedt, J. D. McEwen, T. D. Kitching, and H. V. Peiris, Phys. Rev. D **92**, 123010 (2015), arXiv:1509.06750.

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:

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

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

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

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

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:

$$M^{pp'}_{\ell\ell'} \equiv M^{\alpha\beta}_{\ell\ell'} \left[(\mathsf{E}_\ell)^p_\alpha (\mathsf{N}^{-1})_{\alpha\alpha'} (\mathsf{E}_{\ell'})^{p'}_{\alpha'}\right] \left[(\mathsf{E}_\ell)^p_\beta (\mathsf{N}^{-1})_{\beta\beta'} (\mathsf{E}_{\ell'})^{p'}_{\beta'}\right] = M_{\ell\ell'} \left[(\mathsf{E}_\ell)^p_\alpha (\mathsf{N}^{-1}_\ell)_{\alpha\beta} (\mathsf{E}_{\ell'})^{p'}_\beta\right]^2 \tag{A4}$$

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

[13, 14]

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