Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **473,** 4306–4317 (2018) doi:10.1093/mnras/stx2644
Advance Access publication 2017 October 10

Science-driven 3D data compression

David Alonso★
Oxford Astrophysics, Department of Physics, Keble Road, Oxford OX1 3RH, UK

Accepted 2017 October 6. Received 2017 September 25; in original form 2017 August 8

ABSTRACT

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 coordinates. 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 (KL), 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 ratio 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.

Key words: methods: data analysis – large-scale structure of Universe.

1 INTRODUCTION

Astronomical data are inherently three-dimensional (3D): the main observable is the intensity of the electromagnetic emission in the sky as a function of wavelength and line-of-sight direction, determined by two angles. An idealized cosmological analysis would therefore use, as a data vector, the full cube $I(\lambda, \theta, \phi)$ probed over sufficiently well-resolved angular and frequency scales (de Putter et al. 2014). However, the operational costs of obtaining such a data set imply that we can realistically only access a compressed version of it, where the compression method comes in different flavours:

(i) We can decrease the measurement noise by integrating the sky intensity over large frequency bands. This approach has been used, for instance, in cosmic microwave background observations (Bennett et al. 2013; Planck Collaboration I 2016; Abazajian et al. 2016).

(ii) Angular resolution can also be sacrificed for wider sky and frequency coverage, as has been proposed for future intensity mapping experiments (Battye et al. 2004; Bandura et al. 2014; Santos et al. 2015).

(iii) The size of the data set can also be reduced by collecting only the flux associated with the brightest extragalactic objects. An accurate measurement of their spectra then allows a determination of their redshift, producing the well-known spectroscopic redshift surveys (Laureijs et al. 2011; Levi et al. 2013; Alam et al. 2017).

(iv) Measuring individual spectra is a costly operation, however, and can usually only be done for a small subsample of all available sources. This problem can be mitigated by inferring the source's redshift from their emission in a small number of wider frequency bands, in what is known as a photometric redshift survey (Chambers et al. 2016; Dark Energy Survey Collaboration et al. 2016; Aihara et al. 2017). The redshifts thus determined are far less precise than their spectroscopic counterparts, and usually only an imperfect estimation of the redshift probability distribution for each galaxy is accessible.

Even after this first compression stage, the size of the data set makes a direct analysis of it as a data vector a computationally intractable problem. Typically, this should not be an issue in terms of information loss, since large portions of the data are usually dominated by measurement noise, contaminated by sources of systematic uncertainty (both observational and theoretical) or contain only redundant information. An efficient data compression method will therefore identify these sections of data space and

★ E-mail: David.Alonso@physics.ox.ac.uk

eliminate them, or collect them into summary statistics, while minimizing the loss of meaningful cosmological information. An example of this is the standard analysis of cosmological data sets in terms of their two-point statistics (Tegmark 1997). However, even in this case, the resulting data vector can be large enough to present an important computational challenge in terms of likelihood evaluation and covariance estimation. The complexity of latter problem, in particular, scales with the square of the data vector size, and can become an important drain of computational resources (Dodelson & Schneider 2013; Heavens et al. 2017).

In this work, we will concern ourselves with the topic of 3D data compression: the problem of identifying the uncorrelated angular and radial modes of the data that optimally contain the maximum amount of information. This problem has been previously addressed in the literature (Heavens 2003; McEwen & Leistedt 2013; Khalid et al. 2014; Leistedt et al. 2015), and a number of approaches have been proposed depending on the definition of uncorrelatedness used, and on the type of information one wishes to preserve. Here, we will present a method to derive a set of uncorrelated radial eigenmodes that are manifestly optimal in terms of information compression for any quantity, such as individual cosmological parameters or the amplitude of the cosmological signal over any set of known contaminants. The method is based on the well-understood Karhunen–Loève transform (KL) (Vogeley & Szalay 1996; Tegmark et al. 1997), and is similar in spirit to the derivation of optimal weighting schemes for the analysis of spectroscopic surveys (Feldman et al. 1994; Mueller et al. 2017). Although we will focus here on the case of photometric redshift surveys, the method can be applied to any set of cosmological data sets.

This paper is structured as follows: Section 2 describes the Karhunen–Loève (KL transform and its applicability in the context of 3D data compression. Section 3 shows the performance of the method in a number of science cases, such as the derivation of optimal radial bases for galaxy-clustering (Section 3.2) and weak-lensing (Section 3.5) observations, and the use of the KL eigenmodes to measure the effects of primordial non-Gaussianity (Section 3.3) and lensing magnifications (Section 3.4) with a small number of modes. Finally, Section 4 summarizes our findings and discusses the advantages and shortcomings of the method.

2 METHOD

2.1 The KL transform

The idea behind the Karhunen–Loève transformKL, as developed within the field of cosmological data analysis in, for example, Vogeley & Szalay (1996) and Tegmark et al. (1997) 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 $\boldsymbol{x}$ be a data vector of dimension $N_{\rm s}$, and let θ be a particular parameter we want to measure. Under the assumption that $\boldsymbol{x}$ is Gaussianly distributed with mean 0 and covariance $\mathbf{C}$, a set of linear combinations $y_p \equiv \boldsymbol{e}_p^\dagger \boldsymbol{x}$ 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 $m < N_{\rm s}$ combinations contain most of the information about θ. This is done by solving the generalized eigenvalue problem (Tegmark et al. 1997):

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

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

Although the Karhunen–Loève transformKL 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 equation (1) can always be recast as a standard eigenvalue problem of the form $\mathbf{A}\,\tilde{\boldsymbol{e}}_p = \lambda_p\,\tilde{\boldsymbol{e}}_p$, where

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

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

2.1.1 The KL transform for the signal-to-noise ratio

Let us decompose the data vector $\boldsymbol{x}$ into uncorrelated signal and noise components $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n}$ 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.[1] In this particular case, the data covariance matrix can be split into their independent contributions $\mathbf{C} = \mathbf{S} + \mathbf{N}$.

The KL transform has traditionally been used to design an eigenbasis that maximizes the overall signal-to-noise ratio (S/N; e.g Bond 1995; Vogeley & Szalay 1996). This can be done by defining a fictitious parameter ρ multiplying the signal part of the data with fiducial value $\rho = 1$ (i.e. $\boldsymbol{x} = \rho\boldsymbol{s} + \boldsymbol{n}$). In this case, after some trivial manipulations, the eigenvalue equation (equation 1) takes the form

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

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

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

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

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

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

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

Finally, note that, because both $\mathbf{S}$ and $\mathbf{N}$ are positive-definite matrices, their eigenvalues will also be positive. Since $\mathbf{N}$ becomes the identity under the KL transform, the elements of Λ above will all be greater than 1, and converging to 1 for the noise-dominated modes.

2.1.2 The KL transform with correlated contaminants

Let us now consider a more general case in which we further split the noise into two parts $\boldsymbol{n} \rightarrow \boldsymbol{n} + \boldsymbol{m}$, where $\boldsymbol{m}$ is a contaminant with

[1] $\boldsymbol{n}$ could include, for instance, the contribution of foregrounds in intensity mapping experiments, which motivates the use of the KL transform as a foreground cleaning method (Shaw et al. 2014).
[2] Effectively, $\mathbf{L}$ acts as the square root of $\mathbf{N}$.

a non-zero correlation with the signal. The covariance matrix of the data is then given by

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

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

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

Unfortunately, in this case, the manipulation that lead us to equation (3) 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 equation (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 equation (7) can be collected as columns of a matrix $\mathbf{E}$ that simultaneously satisfies the equations:

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

where $\mathbf{I}$ is the identity and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{N_s})$. Since the second equation implies $\mathbf{C}\mathbf{E} \equiv (\mathbf{E}^\dagger)^{-1}$, the original vector $\boldsymbol{x}$ can be recovered from the coefficients $\boldsymbol{y} \equiv (y_1, \ldots, y_{N_s})$ as $\boldsymbol{x} = \mathbf{C}\,\mathbf{E}\,\boldsymbol{y}$. More interestingly, one can identify the principal eigenvectors of the equation (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 $\boldsymbol{m}$. This procedure defines a filter $\mathbf{W} \equiv \mathbf{C}\,\mathbf{E}\,\mathbf{P}\,\mathbf{E}^\dagger$, where $\mathbf{P}$ 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 $\tilde{\boldsymbol{x}} = \mathbf{W}\boldsymbol{x}$.

2.2 Application to tomographic data sets

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 (Asorey et al. 2012; Becker et al. 2016; Hildebrandt et al. 2017), 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 minimizing the correlation between non-neighbouring bins or preserving a roughly constant number density on all bins. Other approaches (Heavens 2003; Kitching et al. 2007, 2014) involve projecting the main observable (e.g. galaxy overdensity or shear) on to 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 KL transform. The resulting radial basis is optimal from the point of view of data compression, allowing us to concentrate on a few modes containing most of the information.

2.2.1 *Tomographic analyses*

Let us start by assuming that we have split the galaxy sample into $N_{\rm s}$ subsamples. As mentioned above, we will think of each of these subsamples 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 $a^\alpha(\hat{\boldsymbol{n}})$ be the field on the sphere at the angular position $\hat{\boldsymbol{n}}$ and defined in terms of the properties of the sources in the αth sample (e.g. the cosmic shear field γ^α or the galaxy overdensity δ^α), and let $p(z|\alpha)$ be the redshift distribution of these sources. Finally, let $a^\alpha_{\ell m}$ be the spherical harmonic coefficients of a^α.[3] The power spectrum for our set of subsamples is defined as the two-point correlator of $a^\alpha_{\ell m}$:

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

where we have packaged $a^\alpha_{\ell m}$ as a vector for each (ℓ, m), $\boldsymbol{a}_{\ell m} \equiv (a^1_{\ell m}, \ldots, a^{N_{\rm s}}_{\ell m})$, and $\mathbf{C}_\ell$ is an $N_{\rm s} \times N_{\rm s}$ matrix. Usually, the observed field can be decomposed into uncorrelated signal and noise components $\boldsymbol{a} = \boldsymbol{s} + \boldsymbol{n}$, with a similar decomposition in the power spectrum, $\mathbf{C}_\ell = \mathbf{S}_\ell + \mathbf{N}_\ell$.

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

(i) 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.

(ii) $C_\ell^{\alpha\beta}$ is a $N_{\rm s} \times N_{\rm s} \times N_\ell$ data vector. Thus, increasing $N_{\rm s}$ will increase the computational time required for each likelihood evaluation like $N_{\rm s}^2$ and number of elements of the covariance matrix of $C_\ell^{\alpha\beta}$ like $N_{\rm s}^4$, 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 order is not clear a priori.

(iii) 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.

2.2.2 *Optimal radial eigenbasis*

Following the description in Section 2.1.1, it is straightforward to derive an optimal set of radial, uncorrelated eigenmodes.

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

(ii) Let us consider a set of linear combinations of the weighted field measured on narrow redshift bins:

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

where $\mathbf{F}_\ell$ is a yet-unspecified matrix and, as in Section 2.1.1, we have let $\mathbf{N}_\ell^{-1}$ define the non-standard dot product $\boldsymbol{v}_\ell \circ \boldsymbol{w}_\ell \equiv \boldsymbol{v}_\ell^\dagger \cdot \mathbf{N}_\ell^{-1} \cdot \boldsymbol{w}_\ell$. The power spectrum for this new observable would then simply be given by

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

[3] 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.

(iii) Requiring that the new modes be uncorrelated, we can identify equation (11) with the generalized eigenvalue equation (5), which defines the KL eigenbasis $\mathbf{F}_\ell$ by additionally requiring that it be orthonormal ($\mathbf{F}_\ell \circ \mathbf{F}_\ell = \mathbf{I}$). Note that, after this transformation and without any further optimization, some of the practicalities of the original problem are already simplified, so we can now focus on the diagonal elements of the new power spectrum and its covariance.

(iv) The data can be further compressed by assuming that we are interested in measuring a set of cosmological parameters $\Theta \equiv \{\theta_1, \dots\}$. The information regarding this set of parameters encoded in a given data vector $\boldsymbol{x}$ 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 $\langle \boldsymbol{x} \rangle = 0$ reads

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

where $\mathbf{C} \equiv \langle \boldsymbol{x}\, \boldsymbol{x}^\dagger \rangle$ is the covariance matrix of the data. We can thus rank the eigenvectors $(\mathbf{F}_\ell)^p_\alpha$ in terms of their information content (in a Fisher-matrix sense). In the simplest scenario, one may be interested in maximizing the overall S/N, in which case each mode contributes independently to the Fisher matrix element of the signal amplitude.

(v) 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 data set under study, analogous to the Fourier–Bessel basis in a translationally-invariant system (see Section 3.1), its 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 $N_\mathrm{s}^2 \times N_\ell$ to $M \times N_\ell$ while minimizing 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 dataset comprises only a small set of samples means that the method is not penalized in terms of photometric redshift uncertainties. Once the KL eigenmodes $\mathbf{F}_\ell$ 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 to be characterized by their own window function:

$$\tilde{\phi}^p_\ell(z) = \sum_\alpha \frac{(\mathbf{F}_\ell)^p_\alpha\, p(z|\alpha)}{N^{\alpha\alpha}_\ell}, \tag{13}$$

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 $\mathbf{F}_\ell$ for the fiducial cosmology as model-agnostic weights and inserting them in equation (11), the theoretical prediction for the power spectrum of each mode D^p_ℓ can be computed in a model-independent way.

Finally, the method outlined in this section is based on the KL decomposition that maximizes the amplitude of the signal under study. This is the main application advocated in this paper, since it is plausible that the set of modes containing the bulk of the cosmological signal will also drive the constraints on any comprehensive set of cosmological parameters (we explore this in more detail in Section 3.5). However, we must note that for individual parameters, the optimal degree of data compression is achieved by solving the general KL eigenvalue problem (1), which can lead to substantial improvements with respect to the S/N-optimal basis. We explore one particular example of this in Section 3.3.

3 PERFORMANCE AND PARTICULAR EXAMPLES

This Section explores the performance of the KL decomposition in a number of specific science cases.

3.1 Special case: the harmonic-Bessel basis

Let us start by considering a simplified case where the field a is the overdensity field of a non-evolving galaxy population for which we neglect the effects 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 a is given purely by shot noise, and since (as per our initial assumptions) the number density of sources does not change with χ, the noise power spectrum is diagonal and scales like $N_\ell(\chi, \chi') \propto \delta_{\chi,\chi'}\, \chi^{-2}$. Thus, the dot product is just given by

$$\boldsymbol{b}^\dagger \circ \boldsymbol{c} \propto \int \mathrm{d}\chi\, \chi^2\, b(\chi)^*\, c(\chi). \tag{14}$$

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

$$S^{\chi\chi'}_\ell = \frac{2}{\pi} \int_0^\infty \mathrm{d}k\, k^2\, P_k\, j_\ell(k\chi) j_\ell(k\chi'), \tag{15}$$

and it is trivial to show that the KL eigenmodes are simply given by the spherical Bessel functions: $(\mathbf{F}_\ell)^k_\chi \propto \sqrt{2/\pi} j_\ell(k\chi)$:

$$\begin{aligned} D^{kk'}_\ell &\equiv \sum_{\chi,\chi'} (F_\ell)^k_\chi (F_\ell)^{k'}_{\chi'} S^{\chi\chi'}_\ell \\ &\propto \frac{2}{\pi} \int \mathrm{d}\chi\, \chi^2 \int \mathrm{d}\chi'\, \chi'^2 j_\ell(k\chi) j_\ell(k'\chi') S^{\chi\chi'}_\ell \\ &= \int \mathrm{d}q\, q^2 P_q\, \mathcal{D}(q,k)\, \mathcal{D}(q,k') \\ &= \int \mathrm{d}q\, q^2 P_q \frac{\delta(k-q)}{q^2} \frac{\delta(k'-q)}{q^2} = \frac{P_k}{k^2 \Delta k} \delta_{k,k'}, \end{aligned} \tag{16}$$

$$\mathcal{D}(q,k) \equiv \frac{2}{\pi} \int \mathrm{d}\chi\, \chi^2\, j_\ell(q\chi) j_\ell(k\chi) = \frac{\delta(k-q)}{q^2}. \tag{17}$$

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 data sets (e.g. Kitching et al. 2014). 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 (among other things different k-modes will be correlated), as opposed to the KL basis described in the previous section.

3.2 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

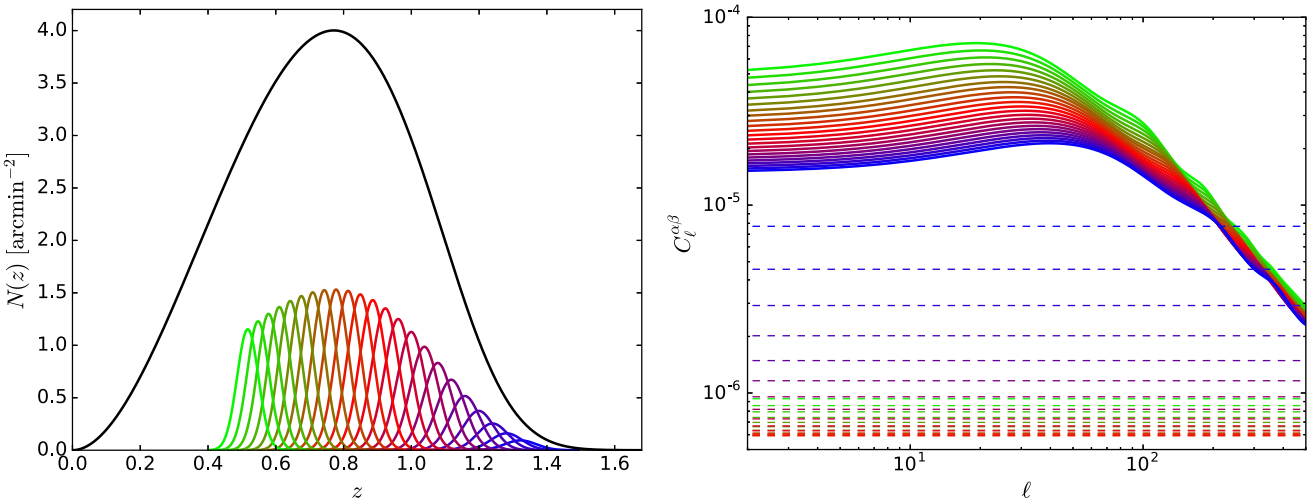

Figure 1. Left-hand panel: redshift distribution and bins considered for the KL analysis of a strawman large-scale-structure survey targeting a sample of red galaxies. Right-hand panel: clustering auto-power spectra of the redshift bins shown in the left-hand panel. The signal and noise power spectra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

of two samples with photo-z properties α and β is given by (Huterer, Knox & Nichol 2001; Tegmark et al. 2002)

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

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

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

where n_t is the total angular number density of sources and

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

Here, $b^\alpha(z)$ is the linear galaxy bias, $f(z) = \mathrm{d}\log\delta/\mathrm{d}\log a$ 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. 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 3.4).

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

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

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

Note that this expression corresponds to the KL basis that optimizes the S/N. The general KL problem in equation (1) would correspond to changing $C_\ell^{\alpha\beta}$ above to $\partial_\theta C_\ell^{\alpha\beta}$, and $N_\ell^{\alpha\beta}$ to $C_\ell^{\alpha\beta}$. In the limit of perfect photo-zs ($p(z|\alpha) = \delta(z-\alpha)$), and in the absence of redshift-space distortions, the solutions 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 KL 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 (making them better suited for clustering analyses). The sample we consider is compatible with what could be observed by the Large Synoptic Survey Telescope (LSST Science Collaboration et al. 2009), characterized by the redshift distribution shown in the left-hand panel of Fig. 1 (full details can be found in Alonso et al. 2015). 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 $z_\mathrm{ph} > 0.5$ 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-hand 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. All power spectra were computed using a modified version of the code presented in Di Dio et al. (2013).

Using the prescription described in Section 2.1.1, we find the KL eigenmodes and associated power spectra, and rank them according to their contribution to the total S/N (defined here as the Fisher matrix element of the signal amplitude). The power spectra of the resulting KL modes are shown in Fig. 2. Unlike the case of weak lensing, explored in Section 3.5, the information encoded in the galaxy overdensity is local in redshift, and thus the correlation between different bins decays rapidly with redshift separation. The S/N is therefore spread over $\sim$15 signal-dominated modes, and the noise-dominated modes can be thought of as the radial scales filtered out by the finite photo-z uncertainty (as mentioned in Section 2.1.1, the noise power spectrum gets mapped into one under the KL transform). The relative contribution of each mode to the total S/N is shown in the top panel of Fig. 3. A total of 90 per cent of the total constraining power can be achieved by considering the first 13 eigenvectors. The forms of the first seven of these eigenvectors for $\ell = 30$ are shown in the bottom 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 KL decomposition is achieved by the spherical Bessel functions. A Fourier–Bessel decomposition is therefore probably a near-optimal analysis method for galaxy clustering,

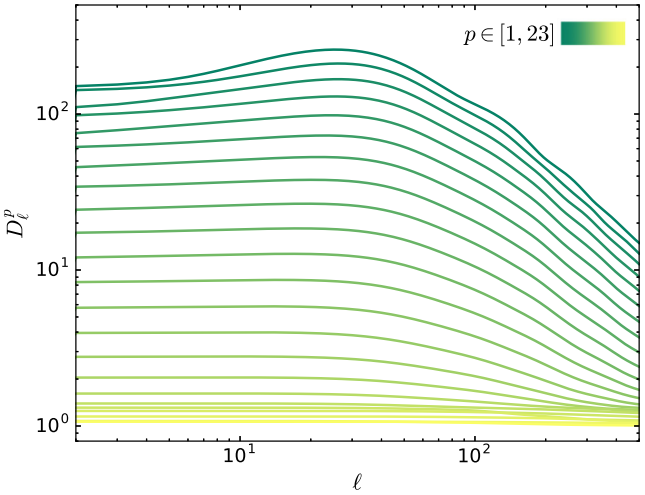

Figure 2. Power spectra of the KL eigenmodes for the strawman large-scale-structure survey. Unlike in the case of weak lensing (see Section 3.5), a large number of eigenmodes are signal-dominated. This is due to the overall higher S/N of galaxy clustering with respect to galaxy shear as well as to the smaller correlations between distant bins.

although the KL decomposition allows a more precise determination of the truly orthogonal radial modes.

3.3 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 $f_{\rm NL}$[4] generates in the two-point statistics of biased tracers of the matter distribution (Matarrese et al. 2000; Dalal et al. 2008). 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 KL decomposition outlined in Section 2.1 an ideal analysis method. Similar approaches have been explored in the literature to devise optimal weights for spectroscopic galaxy surveys (Mueller et al. 2017).

We again consider the red galaxy sample used in the previous section, but now estimate the KL basis of eigenmodes that optimize the information content on $f_{\rm NL}$ instead of the overall signal amplitude, that is, we solve the generalized eigenvalue problem in equation (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 $f_{\rm NL}$, $\sigma(f_{\rm NL})$, 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 the following three choices of radial functions:

(i) The KL eigenbasis resulting from optimizing the information content on $f_{\rm NL}$ discussed in this section (in red).
(ii) The KL eigenbasis resulting from optimizing the overall S/N of the galaxy-clustering signal, as discussed in the previous section (in grey).
(iii) Photo-z tomography: the result of dividing the galaxy sample into a number of top-hat photo-z bins of equal width Δz (in blue).

[4] The reader is referred to Bartolo et al. (2004) for a thorough review of non-Gaussianity and a definition of $f_{\rm NL}$.

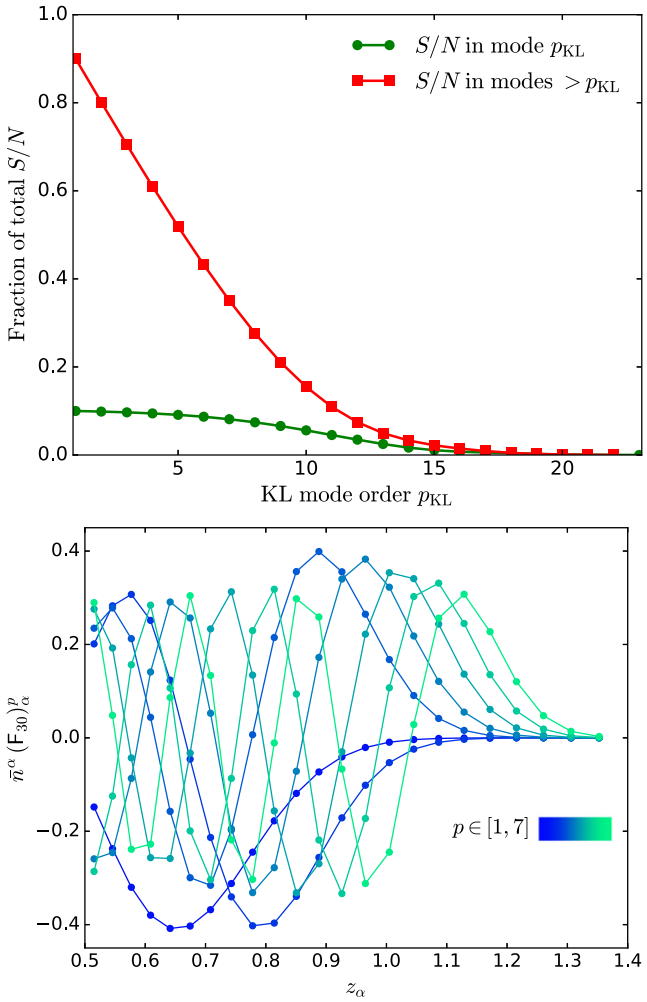

Figure 3. Top panel: fraction of the total S/N of the different KL eigenmodes for the strawman galaxy-clustering survey. The bulk of the S/N (>90 per cent) is encoded in the first 13 modes. Bottom panel: the first seven KL modes for $\ell = 30$. The sinusoidal shape of the modes agrees with the expectation that, in the limit of $\sigma_z \to 0$ and no background redshift dependence, the KL modes should be given by the spherical Bessel functions.

As demonstrated by this figure, for a fixed number of modes, the optimal KL basis always outperforms any other data compression prescription. In particular, the constraints on $f_{\rm NL}$ are only degraded by ∼30 per cent when considering only the first principal eigenmode, and almost 90 per cent 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 KL eigenbasis for the S/N. However, since the tomographic bins are not orthogonal, unlike the KL modes, for a fixed $\sigma(f_{\rm NL})$, both KL 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 overall S/N is not necessary optimal in terms of final constraints for cosmological parameters that depend on specific features of the power spectrum.

3.4 Galaxy clustering – magnification bias

Gravitational lensing of the observed galaxy positions alters their clustering pattern. This appears as an extra term in the galaxy-

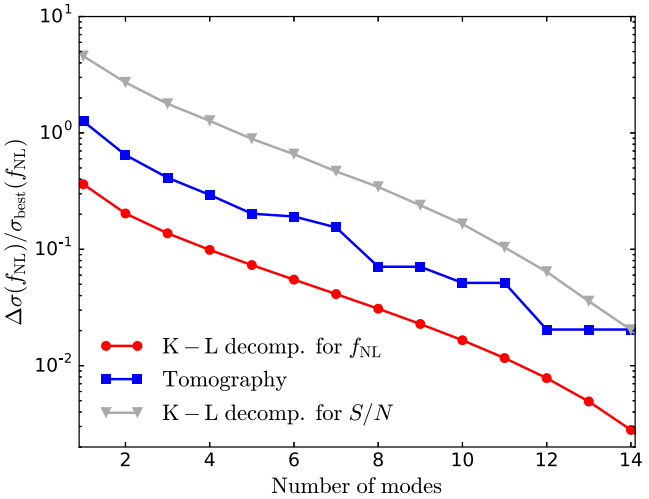

Figure 4. Excess uncertainty on $f_{\rm NL}$ with respect to the best achievable error on this parameter as a function of the number of modes included in the analysis for three different radial decomposition schemes: optimal KL modes for $f_{\rm NL}$ (red), optimal KL modes for the overall S/N of the clustering signal (grey) and tomographic slicing into the corresponding number of bins of equal width (blue). The curves converge to zero when including all 23 modes.

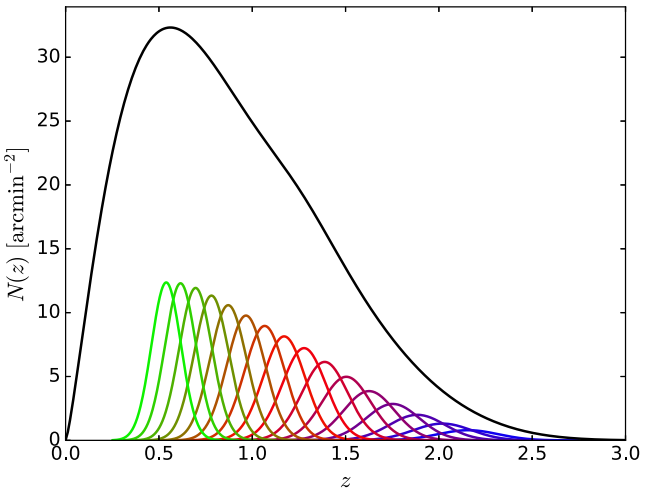

Figure 5. Redshift distribution and bins considered for the KL analysis of a strawman lensing survey (Section 3.5) and for the extraction of the magnification bias signal from galaxy clustering (Section 3.4).

clustering transfer function (equation 21):

$$\Delta_\ell^{M,\alpha}(k) = -2\ell(\ell+1)\int {\rm d}\chi\, W^{M,\alpha}(\chi)\frac{j_\ell(k\chi)}{k^2 a(\chi)}\sqrt{P(k,z(\chi))},$$
$$W^{M,\alpha}(\chi) = \frac{3H_0^2\Omega_{\rm M}}{2}\int_{z(\chi)}^{\infty}{\rm d}z'\, p(z'|a)\frac{2-5\,s}{2}\,\frac{\chi(z')-\chi}{\chi(z')\chi}, \qquad (24)$$

where s is the tilt in the number counts of sources as a function of magnitude limit. This effect, commonly labelled 'magnification bias' (Gunn 1967; Matsubara 2000; Loverde et al. 2008) 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 (equation 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 KL formalism described in Section 2.1.2 to devise an optimal basis of radial eigenmodes containing the bulk of its S/N.

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 (Scranton et al. 2005; Hildebrandt et al. 2009). 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 Alonso et al. (2015). In summary, we consider a sample with ∼40 objects per arcmin^2 with the redshift distribution shown in Fig. 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 centre (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 observer, the magnification signals in widely separated redshift bins can be tightly correlated. This is shown explicitly in Fig. 6. The figure shows the correlation coefficients between the 16 redshift bins, defined as $R_\ell^{\alpha\beta} = C_\ell^{\alpha\beta}/\sqrt{C_\ell^{\alpha\alpha}C_\ell^{\beta\beta}}$, at $\ell = 400$, with (bottom panel) and without (top 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 KL decomposition: A signal that is tightly correlated across samples will contribute significantly only to a small set of eigenmodes. To explore this possibility, we follow the prescription outlined in Section 2.1.2 for correlated contaminants. The contribution of each eigenmode to the total S/N of the magnification bias (in a Fisher-matrix sense) is shown in Fig. 7. As expected, most of the signal (>80 per cent) is contained in the first eigenvalue, with the practical totality of it concentrated in the first three modes.

We finish this section by noting that this approach is similar to the 'nulling' method of Heavens & Joachimi (2011), and that an analogous treatment could be carried out on the cosmic shear field to separate the lensing and intrinsic alignment contributions (Joachimi & Schneider 2008).

3.5 Weak lensing

The effects of gravitational lensing can be measured directly by studying the correlation it induces on the shapes and orientation of galaxy images. This effect, labelled 'cosmic shear', is arguably the most promising observational probe for photometric redshift surveys, and therefore we will discuss the KL 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 KL transform should be able to compress most of the S/N into a small set of radial eigenmodes. To quantify this, we consider the same survey configuration used in Section 3.4. The signal part of the cross-power spectrum between the cosmic shear measurements made in two different redshift shells is given again by equation (18), where now the

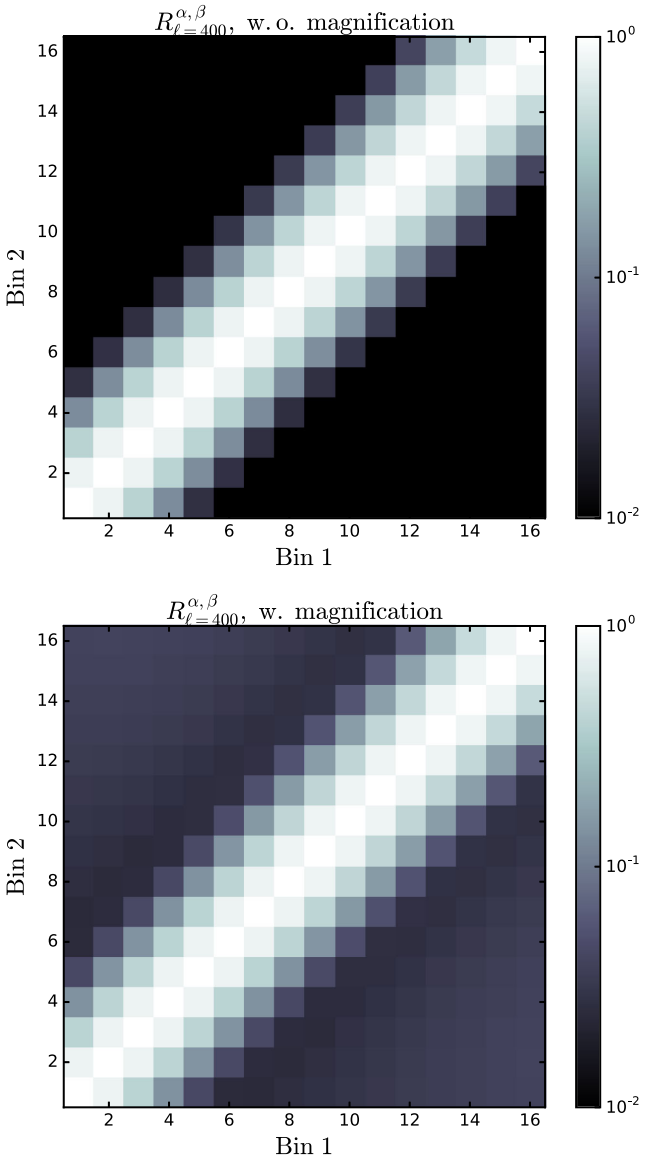

Figure 6. Correlation coefficient between the galaxy overdensities measured in the 16 redshift bins shown in Fig. 5. The top panel shows the contributions of the true matter overdensity and redshift-space distortions alone. In this case, the correlations between neighbouring bins are mostly caused by the overlap in redshift between them, and decays quickly with bin separation. The bottom panel then adds the contribution from lensing magnification, which generates a significant correlation between distant bins.

transfer functions Δ^α_ℓ take the form

$$\Delta^\alpha_\ell(k) \equiv \sqrt{\frac{(\ell+2)!}{(\ell-2)!}} \int \mathrm{d}\chi\, W^\alpha(\chi) \frac{j_\ell(k\chi)}{k^2 a(\chi)} \sqrt{P(k, z(\chi))},$$

$$W^\alpha(\chi) \equiv \frac{3H_0^2\Omega_\mathrm{M}}{2} \int_{z(\chi)}^{\infty} \mathrm{d}z\, p(z'|\alpha) \frac{\chi(z')-\chi}{\chi(z')\chi}. \quad (25)$$

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 $\bar{n}^\alpha$:

$$N^{\alpha\beta}_\ell = \delta_{\alpha\beta} \frac{\sigma^2_\gamma}{\bar{n}^\alpha}, \quad (26)$$

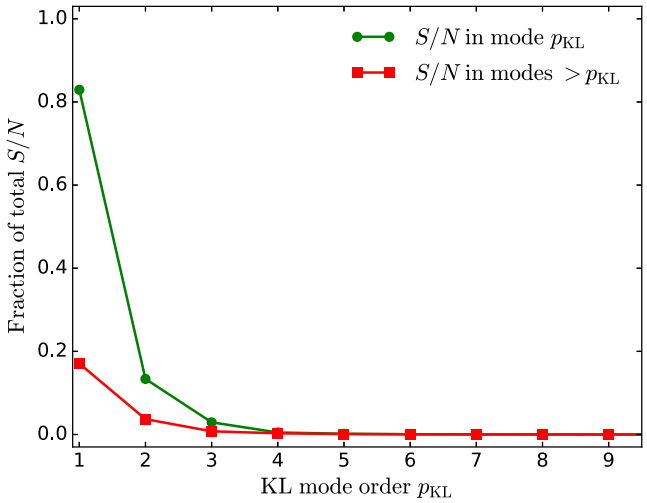

Figure 7. Fraction of the total S/N of the magnification bias effect encoded in each KL eigenmode (green) as well as the cumulative information contained by all higher order modes (red). The principal eigenmode contains ∼80 per cent of the signal, and the first three modes are enough to capture it completely in practice.

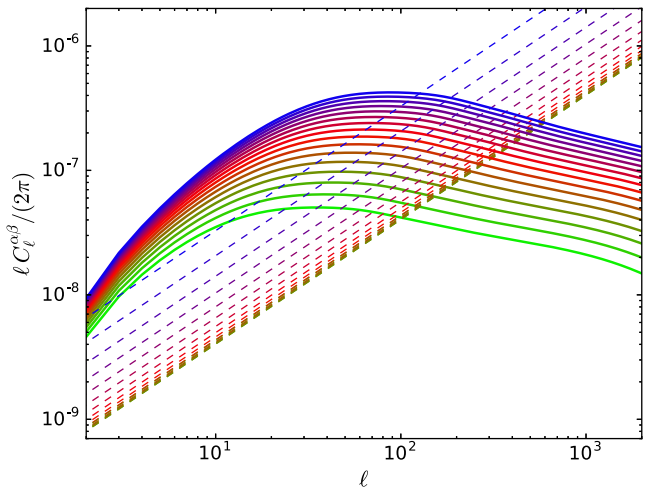

Figure 8. Shear auto-power spectra for the redshift bins shown in Fig. 5. The signal and noise power spectra are shown as thick solid and thin dashed lines, respectively.

with $\bar{n}^\alpha$ in units of $\mathrm{s\,rad}^{-1}$ and $\sigma_\gamma = 0.28$ (LSST Science Collaboration et al. 2009). The lensing auto-power spectra (both signal and noise) for these bins are shown in Fig. 8.

We compute the KL 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 Fig. 9. Comparing against Fig. 8, we can see that the KL 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 S/N is shown in the top panel of Fig. 10. Most of the signal (∼95 per cent) is contained within a single mode, and the first two modes are able to recover more than 99 per cent of the total. The eigenvectors corresponding to the first three principal modes for different values of ℓ are shown in the right-hand 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

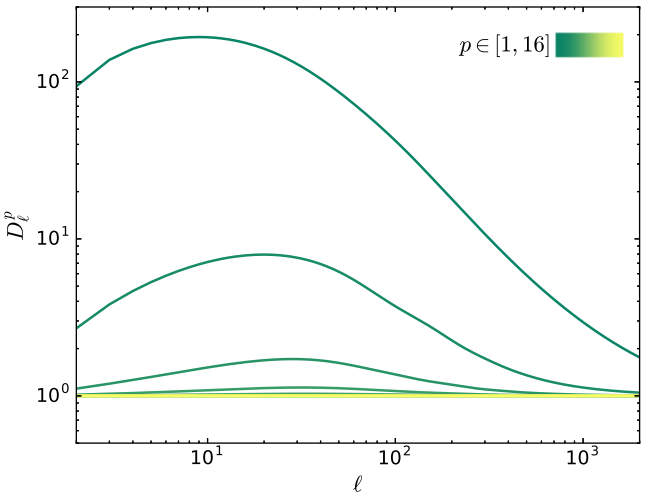

Figure 9. Power spectra of the KL eigenmodes for the strawman weak-lensing survey. All but the first three modes are noise-dominated, and most of the information is encoded in the first mode.

of the redshift range with the highest S/N, penalizing 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. The dashed black line in the same figure shows the weighting scheme associated with the measurement of the lensing signal integrated over a single bin encompassing the redshift range covered by the 16 bins in Fig. 8. These weights are similar to the principal KL eigenmode, and thus the KL decomposition determines, broadly speaking, that the bulk of the S/N is mostly concentrated in the redshift-integrated signal, and extra information regarding the growth of structure can be picked up by up- or down-weighting the contributions in different sections of the redshift range.

As we have discussed in the previous sections, the principal KL eigenmodes that optimize the recovery of the cosmological signal are not necessarily optimal in terms of encoding cosmological information, although it is plausible to expect so in general. In order to study this further, we have performed a Fisher-matrix forecast of the final constraints on cosmological parameters achievable by collecting the information encoded in the first M principal eigenmodes, and compared them with the best possible constraints coming from the use of the full set of 16 redshift bins (or, equivalently, all of the KL eigenmodes). We do so following the approach described in section 3 of Alonso et al. (2015) and using, as observables, the corresponding set of KL modes $b_{\ell m}^p$. For these forecasts, we considered a set of nine parameters: the relative density of cold dark matter ω_c, the relative contribution of baryons ω_b, the normalized local expansion rate h, the amplitude A_s and spectral index n_s of primordial scalar perturbations, the sum of neutrino masses $\Sigma\, m_\nu$, the equation of state of dark energy w and two parameters, $\log_{10} M_c$ and η_b, parametrizing the contribution of baryonic effects in the matter power spectrum as described in Schneider & Teyssier (2015).

Fig. 11 shows the results of this analysis in terms of 1σ contours in the Σm_ν–w plane marginalized over all other parameters. The results are shown for the set of 1, 2 and 3 principal KL eigenmodes in red, blue and yellow, respectively, while the best achievable constraints using all of the modes are shown as a solid black ellipse. We have removed the axis labels to focus the reader's attention on the relative improvement of the constraints with the number of modes. Even though the first eigenmode contains the vast majority

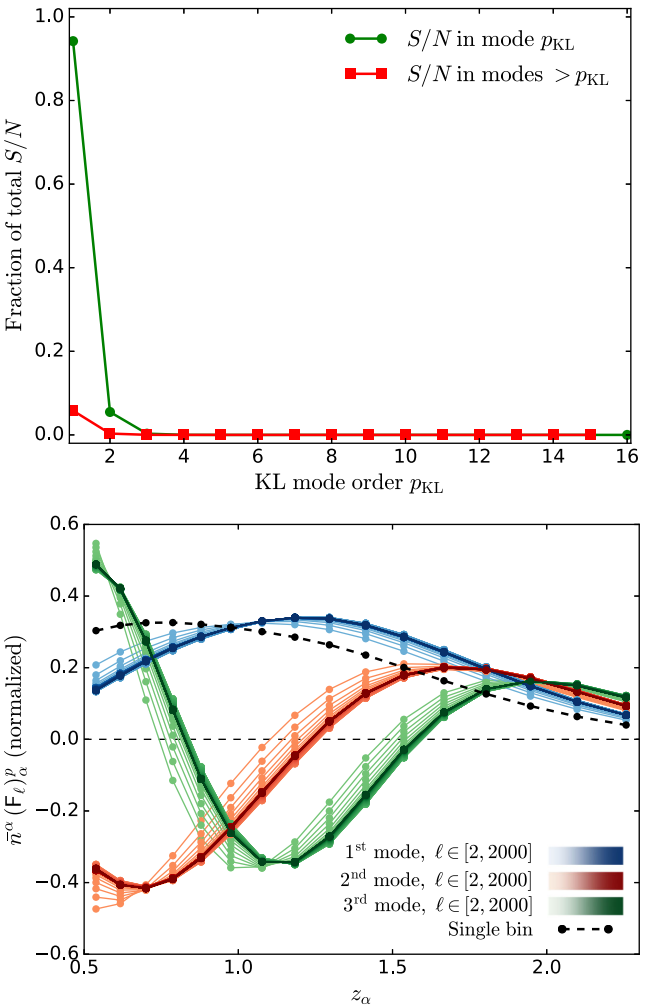

Figure 10. Top panel: fraction of the total weak-lensing S/N borne by each KL eigenmode (green) and the cumulative fraction contained in all higher order modes (red). The first mode contains ∼95 per cent of the signal, and the first three modes are enough to recover most of the information content. Bottom panel: the first (blue), second (red) and third (green) KL eigenmodes of the strawman weak-lensing survey for different ℓ. In all cases, the darkness of the line colour increases with ℓ. The redshift dependence of the modes stays roughly constant across ℓ and converges to a fixed shape for large ℓ. The black dashed line corresponds to the weighting scheme associated with single-bin tomography (i.e. computing the signal integrated over the whole redshift range), which is similar to the weighting associated with the first principal eigenmode.

of the lensing signal, as evidenced by the top panel of Fig. 10, the extra information contained in the second and third eigenmodes is necessary in order to break the degeneracies between cosmological parameters. Three modes are however sufficient to recover the full constraining power with negligible loss of information.

To finalize, we would like to emphasize the fact that, as shown in the bottom panel of Fig. 10, the three principal eigenmodes preserve roughly the same shape as a function of multipole order, converging to the same curve for large ℓ. An ℓ-independent basis of radial functions would be advantageous from the point of view of data analysis since, for instance, the radial window functions associated with each mode (see equation 13) would only have to be calibrated once (independently of ℓ). It is therefore interesting to explore the cosmological constraints achievable by using the radial functions associated with the KL eigenmodes at high ℓ *for all* ℓ,

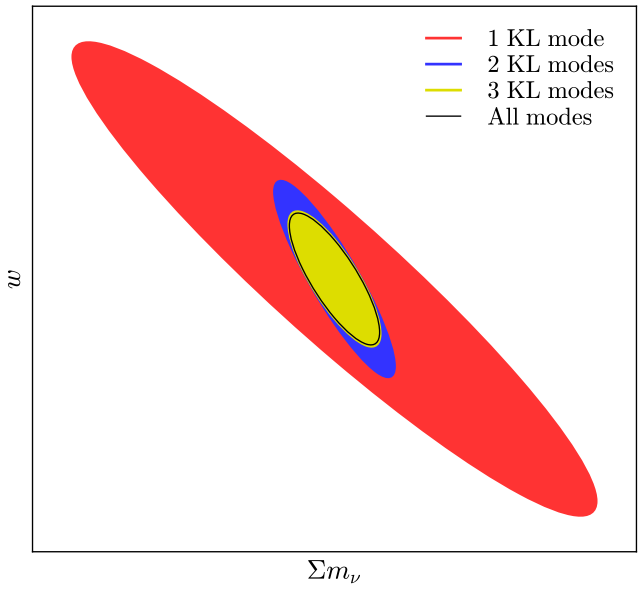

Figure 11. 1σ constraints on the dark energy equation of state w and the sum of neutrino masses Σm_ν achievable by analysing the first (red), first two (blue) and first three (yellow) KL radial eigenmodes, compared with the best achievable constraints (solid black line). These constraints are marginalized over seven other cosmological and nuisance parameters. Although the vast majority of the signal is encoded in the first mode, the next two modes are necessary in order to break the degeneracies between different parameters and recover optimal constraints.

even though, for a fixed multipole order, the corresponding set of modes will not be exactly orthogonal. We have verified that, doing so for the first three KL eigenmodes, the final constraints on either w or Σm_ν degrade by less than 0.5 per cent. This is a reasonable result, given the larger statistical weight of the small-scale (large-ℓ) fluctuations.

4 DISCUSSION

Next-generation cosmological observations will gather their constraining power from a variety of observables, and will therefore have to deal with enormous data vectors. This will present a computational challenge, both from the point of view of likelihood evaluation and in the estimation of the covariance matrix. An efficient data compression scheme would be able to not only alleviate these problems but also to separate the most significant and less contaminated modes in the data.

In this paper, we have studied the problem of 3D data compression in the context of photometric redshift surveys, and presented a method, based on the KL transform, to derive a basis of orthogonal radial functions that optimally separate the data into uncorrelated modes, where optimality can be defined in terms of overall S/N, information content on a particular cosmological parameter or separability between clean signal and contaminants. This basis is a general and natural extension of the well-known harmonic-Bessel (or Fourier–Bessel) decomposition of spherically symmetric and translationally invariant systems, adapted to the particular properties of the data set under study. Even though the definition of this basis requires prior knowledge of some of these properties, including uncertain ones such as a model for the photo-z distributions, once the radial eigenfunctions are selected, the analysis of the resulting data eigenmodes can proceed as usual, including any calibration of these properties. Thus, a sufficiently well-educated model of the survey parameters should preserve the near-optimality of the associated eigenbasis, while not hampering the robustness of the analysis.

We have shown that, for the study of galaxy clustering in an idealized spectroscopic survey, the optimal set of eigenmodes corresponds to the standard harmonic-Bessel basis, and that this would not be the case in the presence of redshift uncertainties, RSDs or in the analysis of weak-lensing observables. For the study of galaxy clustering in a photometric redshift survey, we have shown that the KL basis that maximizes the recovery of the cosmological signal is Bessel-like, although more optimal compression schemes can be derived to optimize the measurement of individual cosmological parameters. In particular, in the case of $f_{\rm NL}$, we have shown that the bulk of the constraining power is concentrated in approximately three radial modes. We have also extended the method to maximize the recovery of a particular signal in the presence of correlated contaminants, and shown that it could be used to simplify the measurement of the effect of lensing magnification as a subdominant contribution to the statistics of the galaxy distribution.

In the case of cosmic shear measurements, we have shown that due to the integrated nature of the gravitational lensing effect, the bulk of the signal (∼95 per cent) is concentrated in a signal radial mode, qualitatively equivalent to the measurement of the weak-lensing effect over the full redshift range of the survey. The next subdominant modes are however needed in order to break degeneracies between different parameters, and we have shown that three modes are enough to recover the best achievable cosmological constraints.

Further work is needed in order to alleviate some of the practical shortcomings of the method: The KL decomposition is arguably less connected to real-space, directly observable quantities (although not less so than standard Fourier-space methods). Some of the usual methods for systematics calibration thus need to be adapted for a KL-based analysis, and this is particularly relevant for photo-z systematics. In the case of weak lensing, however, we have shown that the shape of the radial eigenfunctions converges to the same curve on large multipole orders, and that the use of ℓ-independent eigenfunctions would have a negligible impact on the final cosmological constraints. In this case, photo-z calibration methods could be used in exactly the same manner as in the standard tomographic analysis.

In this work, we have also considered only the case of optimizing data compression for one single parameter (either the signal amplitude or the particular example of $f_{\rm NL}$). Most cosmological analyses will however target a host of cosmological and nuisance parameters, and it is therefore worth considering compression schemes that can also be applied in this case. A direct application of the method outlined in Section 2 would imply optimizing the diagonal of the inverse Fisher matrix that, as noted in Tegmark et al. (1997), is a computationally cumbersome problem. However, under the assumption that the set of parameters of interest is comprehensive enough, it is reasonable to expect that the KL eigenbasis that associated with the overall signal amplitude should be close to optimal in capturing most of the information (although this may not be true for individual parameters as shown in Section 3.3). Other approaches have been advocated in the literature, such as optimizing for each parameter individually and finding the principal directions of the collective set of associated eigenvectors (Tegmark et al. 1997), or identifying the orthogonal directions of the multi-parameter space around the maximum and optimizing the eigenbasis for the minimum-variance orthogonal parameter only (Taylor et al. 2001).

It is also worth emphasizing that, as is the case for the standard harmonic decomposition of fields defined on the sphere, the KL

radial eigenmodes are no longer uncorrelated in the presence of an incomplete sky coverage, and a standard pseudo-C_ℓ analysis reveals non-zero coupling between different multipole orders (ℓ, ℓ') as well as different KL indices (p, p') (see Appendix A and Kitching et al. 2014). The impact of these correlations on the performance of the KL decomposition should be studied in more detail, and well-understood contaminant-deprojection methods, implemented in standard power spectrum methods (Elsner et al. 2017), should be adapted for this analysis.

Finally, although we have explored the applicability of this method to independent galaxy-clustering and weak-lensing measurements, current and upcoming photometric redshift surveys will draw cosmological constraints from a joint analysis of both observables (Krause et al. 2017). The direct application of this method to the joint data vector would, in general, produce eigenmodes that mix both signals. Alternatively, a joint analysis of the KL modes of each observable, taken individually, could be performed, and the merits and drawbacks of each approach should be studied in detail.

ACKNOWLEDGEMENTS

The author would like to thank Pedro Ferreira, Alan Heavens, Boris Leistedt, Jason McEwen, Anže Slosar and David Spergel for useful comments and discussions, and the Center for Computational Astrophysics, part of the Simons Foundation, for their hospitality. He also acknowledges support from the Science and Technology Facilities Council and the Leverhume and Beecroft Trusts.

REFERENCES

Abazajian K. N. et al., 2016, preprint (arXiv:1610.02743)
Aihara H. et al., 2017, preprint (arXiv:1702.08449)
Alam S. et al., 2017, MNRAS, 470, 2617
Alonso D., Bull P., Ferreira P. G., Maartens R., Santos M. G., 2015, ApJ, 814, 145
Asorey J., Crocce M., Gaztañaga E., Lewis A., 2012, MNRAS, 427, 1891
Bandura K. et al., 2014, Proc. SPIE Conf. Ser. Vol. 9145, Ground-based and Airborne Telescopes V. SPIE, Bellingham, p. 914522
Bartolo N., Komatsu E., Matarrese S., Riotto A., 2004, Phys. Rep., 402, 103
Battye R. A., Davies R. D., Weller J., 2004, MNRAS, 355, 1339
Becker M. R. et al., 2016, Phys. Rev. D, 94, 022002
Bennett C. L. et al., 2013, ApJS, 208, 20
Bond J. R., 1995, Phys. Rev. Lett., 74, 4369
Chambers K. C. et al., 2016, preprint (arXiv:1612.05560)
Dalal N., Doré O., Huterer D., Shirokov A., 2008, Phys. Rev. D, 77, 123514
Dark Energy Survey Collaboration et al., 2016, MNRAS, 460, 1270
de Putter R., Holder G. P., Chang T.-C., Dore O., 2014, preprint (arXiv:1403.3727)
Di Dio E., Montanari F., Lesgourgues J., Durrer R., 2013, J. Cosmol. Astropart. Phys., 11, 044
Dodelson S., Schneider M. D., 2013, Phys. Rev. D, 88, 063537
Elsner F., Leistedt B., Peiris H. V., 2017, MNRAS, 465, 1847
Feldman H. A., Kaiser N., Peacock J. A., 1994, ApJ, 426, 23
Gunn J. E., 1967, ApJ, 147, 61
Heavens A., 2003, MNRAS, 343, 1327
Heavens A. F., Joachimi B., 2011, MNRAS, 415, 1681
Heavens A., Sellentin E., de Mijolla D., Vianello A., 2017, MNRAS, 472, 4244
Hildebrandt H., van Waerbeke L., Erben T., 2009, A&A, 507, 683
Hildebrandt H. et al., 2017, MNRAS, 465, 1454
Hivon E., Górski K. M., Netterfield C. B., Crill B. P., Prunet S., Hansen F., 2002, ApJ, 567, 2
Huterer D., Knox L., Nichol R. C., 2001, ApJ, 555, 547
Joachimi B., Schneider P., 2008, A&A, 488, 829
Khalid Z., Kennedy R. A., McEwen J. D., 2014, preprint (arXiv:1403.5553)
Kitching T. D., Heavens A. F., Taylor A. N., Brown M. L., Meisenheimer K., Wolf C., Gray M. E., Bacon D. J., 2007, MNRAS, 376, 771
Kitching T. D. et al., 2014, MNRAS, 442, 1326
Krause E. et al., 2017, preprint (arXiv:1706.09359)
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Leistedt B., McEwen J. D., Kitching T. D., Peiris H. V., 2015, Phys. Rev. D, 92, 123010
Levi M. et al., 2013, preprint (arXiv:1308.0847)
Loverde M., Hui L., Gaztañaga E., 2008, Phys. Rev. D, 77, 023512
LSST Science Collaboration et al., 2009, preprint (arXiv:0912.0201)
Matarrese S., Verde L., Jimenez R., 2000, ApJ, 541, 10
Matsubara T., 2000, 537, L77
McEwen J. D., Leistedt B., 2013, preprint (arXiv:1307.1307)
Mueller E.-M., Percival W. J., Ruggeri R., 2017, preprint (arXiv:1702.05088)
Planck Collaboration I, 2016, A&A, 594, A1
Santos M., Bull P., Alonso D. et al., 2015, Cosmology from a SKA HI Intensity Mapping Survey. p. 19
Schneider A., Teyssier R., 2015, J. Cosmol. Astropart. Phys., 12, 049
Scranton R. et al., 2005, ApJ, 633, 589
Shaw J. R., Sigurdson K., Pen U.-L., Stebbins A., Sitwell M., 2014, ApJ, 781, 57
Taylor A. N., Ballinger W. E., Heavens A. F., Tadros H., 2001, MNRAS, 327, 689
Tegmark M., 1997, Phys. Rev. D, 55, 5895
Tegmark M., Taylor A. N., Heavens A. F., 1997, ApJ, 480, 22
Tegmark M. et al., 2002, ApJ, 571, 191
Vogeley M. S., Szalay A. S., 1996, ApJ, 465, 34

APPENDIX A: PSEUDO-C_ℓ ESTIMATION OF THE KL 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 (Hivon et al. 2002). This method can be directly applied to the two-point statistics of the KL eigenmodes, and reveals the correlation between radial modes generated by an incomplete sky coverage (Kitching et al. 2014).

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

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

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:

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

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

$$\left\langle \hat{C}^{\alpha\beta}_\ell \right\rangle = \sum_{\ell'} M^{\alpha\beta}_{\ell\ell'} C^{\alpha\beta}_{\ell'},$$

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

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 KL modes is straightforward: We project the masked harmonic coefficients $\hat{a}^\alpha$ over the KL eigenvectors $\mathbf{F}_\ell$ (i.e. $\hat{\boldsymbol{b}}_{\ell m} \equiv \mathbf{E}_\ell \circ \hat{\boldsymbol{a}}_{\ell m}$) and compute their power spectra by averaging over m. The resulting estimator takes the form $\hat{D}^p_\ell = \sum_{\ell'} M^{pp'}_{\ell\ell'} D^{p'}_{\ell'}$, where the new

mode-coupling matrix is given by

$$M^{pp'}_{\ell\ell'} \equiv \sum_{\alpha\beta} Q^{pp'}_{\ell\ell',\alpha}\, M^{\alpha\beta}_{\ell\ell'}\, Q^{pp'}_{\ell\ell',\beta} = M_{\ell\ell'} \left(\sum_{\alpha} Q^{pp'}_{\ell\ell',\alpha} \right)^2 , \tag{A4}$$

where

$$Q^{pp'}_{\ell\ell',\alpha} = \sum_{\beta} (\mathbf{F}_\ell)^p_\alpha (\mathbf{N}_\ell^{-1})_{\alpha\beta} (\mathbf{F}_{\ell'})^{p'}_\beta , \tag{A5}$$

and the second equality in equation (A4) holds only if all the maps a^α_ℓ share the same mask w. Note that, for full-sky coverage $M_{\ell\ell'} = \delta_{\ell\ell'}$, and using the orthonormality of $\mathbf{F}$, we recover $M^{pp'}_{\ell\ell'} = \delta_{\ell\ell'}\delta_{pp'}$.

This paper has been typeset from a TEX/LATEX file prepared by the author.