Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **435,** 1857–1873 (2013)
Advance Access publication 2013 September 5

doi:10.1093/mnras/stt1359

Estimating the large-scale angular power spectrum in the presence of systematics: a case study of Sloan Digital Sky Survey quasars

Boris Leistedt,[1]⋆ Hiranya V. Peiris,[1] Daniel J. Mortlock,[2,3] Aurélien Benoit-Lévy[1] and Andrew Pontzen[1,4,5]

[1]*Department of Physics and Astronomy, University College London, London WC1E 6BT, UK*
[2]*Astrophysics Group, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK*
[3]*Department of Mathematics, Imperial College London, London SW7 2AZ, UK*
[4]*Oxford Astrophysics, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK*
[5]*Balliol College, Broad Street, Oxford OX1 3BJ, UK*

Accepted 2013 July 19. Received 2013 July 16; in original form 2013 June 10

ABSTRACT

The angular power spectrum is a powerful statistic for analysing cosmological signals imprinted in the clustering of matter. However, current galaxy and quasar surveys cover limited portions of the sky, and are contaminated by systematics that can mimic cosmological signatures and jeopardize the interpretation of the measured power spectra. We provide a framework for obtaining unbiased estimates of the angular power spectra of large-scale structure surveys at the largest scales using quadratic estimators. The method is tested by analysing the 600 CMASS mock catalogues constructed for the Baryon Oscillation Spectroscopic Survey. We then consider the Richards et al. catalogue of photometric quasars from the sixth Data Release of the Sloan Digital Sky Survey, which is known to include significant stellar contamination and systematic uncertainties. Focusing on the sample of ultraviolet-excess sources, we show that the excess clustering power present on the largest scales can be largely mitigated by making use of improved sky masks and projecting out the modes corresponding to the principal systematics. In particular, we find that the sample of objects with photometric redshift $1.3 < \tilde{z}_{\rm p} < 2.2$ exhibits no evidence of contamination when using our most conservative mask and mode projection. This indicates that any residual systematics is well within the statistical uncertainties. We conclude that, using our approach, this sample can be used for cosmological studies.

Key words: galaxies: photometry – quasars: general – galaxies: statistics – large-scale structure of Universe.

1 INTRODUCTION

The cosmic microwave background (CMB) and the large-scale structure (LSS) of galaxies contain a wealth of physical information that can be used to test models of the origin and evolution of the Universe. Both are well described by correlated Gaussian random fields, and therefore can be characterized by two-point statistics (see, e.g. Bond, Jaffe & Knox 1998, 2000; Tegmark et al. 2002b). In particular, the angular power spectrum is a natural tool for CMB data analysis, and has also proved useful for the study of the clustering properties of galaxy surveys in redshift bins. Such tomographic approaches will be essential for exploiting next generation surveys such as the Dark Energy Survey[1] (DES), which will provide large photometric catalogues where the uncertainties on the redshift estimates complicate a full three-dimensional analysis.

However, data unavoidably contain non-cosmological contributions, for example due to instrumental errors and systematic uncertainties. These contaminants result in additional correlations in the measured power spectra, and can compromise our interpretation of the observables if not correctly treated. In the context of galaxy surveys, observational systematics and calibration errors can result in extra clustering power over a wide range of scales (see e.g. Huterer, Cunha & Fang 2013; Ross et al. 2011, 2012b; Thomas, Abdalla & Lahav 2010, 2011). This proves especially problematic at the largest scales, since the corresponding modes need to be constrained from partial sky data. These modes are nonetheless crucial for testing early universe theories such as cosmological inflation (e.g. Guth 1981; Linde 1982; Albrecht & Steinhardt 1982), the standard paradigm for describing the origin of structure in the universe. Future galaxy surveys will be able to test this paradigm very

⋆ E-mail: boris.leistedt.11@ucl.ac.uk
[1] www.darkenergysurvey.org

© 2013 The Authors
Published by Oxford University Press on behalf of the Royal Astronomical Society

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019

precisely, particularly through the search for signatures of primordial non-Gaussianity (PNG). PNG creates a scale-dependent galaxy bias affecting the two-point clustering properties of LSS tracers at the largest scales (Dalal et al. 2008; Matarrese & Verde 2008). Hence, these scales, which can be strongly affected by systematics, require particularly careful treatment.

Quasars, being highly biased tracers of LSS, are excellent candidates to study the scale and redshift dependence of galaxy bias, for example to constrain PNG. However, since current spectroscopic samples have low number densities and cannot compete with PNG constraints derived from the CMB, one must resort to photometric quasar catalogues. The largest catalogues currently available are extracted from the Sloan Digital Sky Survey (SDSS), and were used to study PNG and the integrated Sachs–Wolfe effect (ISW) (Slosar et al. 2008; Giannantonio et al. 2006, 2008; Xia et al. 2010, 2011). However, these studies demonstrated the high sensitivity of the correlation functions to the sky masks under consideration, indicating the presence of significant levels of contamination by stars and calibration-related systematics. In particular, recent work by Pullen & Hirata (2012), corroborated by Giannantonio et al. (2013), confirmed the high levels of contamination in the Richards et al. (2009) catalogue of SDSS photometric quasars, leading to concerns about the use of this sample for clustering measurements. In this work, we use sample reduction, masking and mode projection to identify a sub-set of objects in this catalogue that can be used for cosmological analyses. We concentrate on the main systematics found by previous studies, and analyse their impact on the clustering measurements through auto and cross-correlation of redshift sub-samples with each other, and with templates of the systematics. Note that cross-correlating with external data can also prove useful in identifying and mitigating the systematics in quasar samples.

When analysing photometric catalogues in redshift bins, the theory power spectrum predictions require precise estimates of the redshift distributions, which are compromised by the large uncertainties of the photometric redshifts. This issue is critical for photometric quasars, since their redshift estimates are significantly more uncertain than for other types of galaxies and include a significant fraction of catastrophic failures. We investigate the use of spectroscopic catalogues for calculating robust and unbiased redshift distribution estimates for the photometrically selected quasar sub-samples.

In addition to data quality and modelling issues, various methodological issues arise when estimating the power spectrum of a galaxy survey for comparison with theory. The pseudo-spectrum (Wandelt, Hivon & Górski 2001; Hivon et al. 2002) and quadratic maximum likelihood (QML; Tegmark 1997; Bond et al. 2000) estimators were developed to measure the power spectrum in the presence of sky cuts. However, numerous technical subtleties and constraints due to pixelization or limited computer resources are implicit in these estimators, and can create significant biases if not handled carefully, as concluded by several studies on the CMB (see e.g. Efstathiou 2004a; Eriksen et al. 2007; Pontzen & Peiris 2010; Copi et al. 2011). This paper aims to clarify these technicalities in the context of galaxy surveys.

This article is organized as follows. In Section 2, we define and illustrate the properties of quadratic power spectrum estimators and demonstrate their validity by applying them to a set of mock catalogues. In Section 3, we turn to the Richards et al. (2009) catalogue of SDSS photometric quasars. We present our data samples, redshift distribution estimates, masks and power spectrum measurements, and discuss the impact of the main systematics on these measurements. Our conclusions are presented in Section 4. Furthermore, technical details on smoothing and masking rules, Karhunen–Loève compression and χ^2 measures are contained in appendices.

2 THEORY AND METHODS

2.1 Background

We consider a real signal $x(\boldsymbol{n})$ on the unit sphere S^2, equivalently described in terms of its spherical harmonic coefficients $\{x_{\ell m}\}$ with $\ell \in \mathbb{N}$ and $m \in \{-\ell, \ldots, \ell\}$. The angular power spectrum $\{\mathcal{C}_\ell\}$ of x is defined as

$$\mathcal{C}_\ell = \sum_{m=-\ell}^{\ell} \frac{|x_{\ell m}|^2}{2\ell+1}, \tag{1}$$

and corresponds to the average power in fluctuations on scales of the order of $180/\ell^\circ$ on the sphere. Assuming that x is a realization of an underlying random field denoted by X, the power spectrum of x can be viewed as a compression technique, and used to perform statistical inference on physical models of X. In particular, this compression is lossless if X is an isotropic Gaussian random field, and the power spectrum is then a sufficient statistic containing all the relevant information in the realization x. Moreover, the 'observed' power spectrum $\mathcal{C}_\ell$ is a realization of a 'theory' power spectrum C_ℓ that fully characterizes the field of interest X. The variance of the former, known as cosmic variance, depends on the number of modes on the sky and is given by

$$\mathrm{Var}(\mathcal{C}_\ell) = \frac{2C_\ell^2}{2\ell+1}. \tag{2}$$

In practice, real data contain a finite amount of information, and the continuous signal x is observed at finite resolution on the sphere. In the context of LSS surveys, galaxy catalogues are usually constructed from raw imaging data and then reduced into pixelized overdensity maps $\boldsymbol{x} = (x_0, \ldots, x_{\mathrm{N_{pix}}-1})$, where $x_i = x(\boldsymbol{n}_i)$ and $\boldsymbol{n}_i$ is the centre of the ith pixel on the sphere. More details on the construction of such maps from the source number counts will be given in Section 2.4. The average correlation between the pixels, namely the pixel–pixel covariance matrix, depends on the theory power spectrum through a Legendre expansion, i.e.

$$\boldsymbol{S} = \langle \boldsymbol{x}\boldsymbol{x}^t \rangle = \sum_\ell C_\ell \mathbf{P}^\ell, \tag{3}$$

where $(\mathbf{P}^\ell)_{ij} = (2\ell+1)/4\pi\, P_\ell(\boldsymbol{n}_i \cdot \boldsymbol{n}_j)$ is a useful matrix notation (Tegmark 1997). A quadratic estimator for the power spectrum of full-sky pixelized data is given by the projection of the data on to the Legendre matrices, i.e.

$$\mathcal{C}_\ell = \boldsymbol{x}^t \mathbf{P}^\ell \boldsymbol{x}, \tag{4}$$

which is the pixel-space equivalent of equation (1) (see discussion and references in Pontzen & Peiris 2010).

In this section, we have used an arbitrary equal-area pixelization scheme, but henceforth we will adopt the HEALPIX conventions (Górski et al. 2005). In defining equation (4), we only considered full sky coverage. This assumption will be relaxed in the next section. We also implicitly assumed that the power spectrum of the pixelized map $\boldsymbol{x}$ was equal to that of the continuous signal x. This approximation is only true at high resolution when the pixel size is small compared with $180/\ell$, and the integrals in the spherical harmonics and Legendre transforms are correctly approximated by matrix multiplications through quadrature, as in equation (4). The bias induced by pixelization as a function of ℓ is critical for low-resolution power spectrum estimation, and needs to be corrected.

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019

This issue is investigated in Appendix A, and the following sections will assume that the relevant corrections have been applied.

2.2 Partial sky coverage and quadratic estimators

Due to contamination or inaccessibility of certain regions of the sky, most cosmological applications involve signals that only cover a portion of the sphere. The power spectrum must then be calculated from a cut-sky map $\tilde{\boldsymbol{x}}$. From a theoretical perspective, the latter can be viewed as the restriction of the full sky map $\boldsymbol{x}$ using a binary mask $\boldsymbol{m} = (m_0, \ldots, m_{\mathrm{N_{pix}}-1})$, such that $m_i = 0$ for masked pixels and $m_i = 1$ elsewhere. Masked/unmasked vectors or matrices are related to each other through an operator $\mathbf{K}$, a diagonal matrix such that $(\mathbf{K})_{ij} = m_i\delta_{ij}$ (with δ_{ij} the Kronecker delta) removing pixels that lie inside the mask.[2] In what follows, the addition of a tilde will represent cut-sky quantities.

In the presence of partial sky coverage, applying equation (4) on the cut-sky map $\tilde{\boldsymbol{x}}$ leads to a cut-sky power spectrum $\{\tilde{\mathcal{C}}_\ell\}$ that considers the zones inside the mask as *data*, i.e. as pixels with $x(\boldsymbol{n}) = 0$. Consequently $\tilde{\mathcal{C}}_\ell$ differs from the quantity of interest $\mathcal{C}_\ell$ and is not a realization of the underlying theory spectrum C_ℓ. Inverting this effect involves deconvolving the effect of the mask from the observed power spectrum $\{\tilde{\mathcal{C}}_\ell\}$, leading to the definition of the 'pseudo-spectrum' (PCL) estimator (Wandelt et al. 2001; Hivon et al. 2002; Efstathiou 2004b; Brown, Castro & Taylor 2005),

$$\hat{C}_\ell^{\mathrm{PCL}} = \sum_{\ell'} (\mathbf{M}^{-1})_{\ell\ell'}\tilde{\mathcal{C}}_{\ell'}, \tag{5}$$

where $\tilde{\mathcal{C}}_{\ell'} = \tilde{\boldsymbol{x}}^t\tilde{\mathbf{P}}^\ell\tilde{\boldsymbol{x}}$ are the cut-sky estimates. The coupling matrix is defined as

$$(\mathbf{M})_{\ell\ell'} = \mathrm{Tr}\,\tilde{\mathbf{P}}^\ell\tilde{\mathbf{P}}^{\ell'}, \tag{6}$$

and is a function of the mask only. Since the variance of each ℓ-mode depends on number of times it is observed, the minimum variance, namely the cosmic variance presented in equation (2), is achieved on the full sky only. Partial sky coverage decreases the number of observed modes, and the variance of the PCL estimates in the absence of noise is approximately

$$\mathrm{Var}(\hat{C}_\ell^{\mathrm{PCL}}) \approx \frac{1}{f_{\mathrm{sky}}}\mathrm{Var}(\mathcal{C}_\ell). \tag{7}$$

Here, $f_{\mathrm{sky}} = \sum_i m_i/\mathrm{N_{pix}}$ is the fraction of the sky covered by the mask, with $f_{\mathrm{sky}} = 1$ corresponding to full sky coverage. Equation (7) is a good approximation for small-scale modes, which remain numerous after masking. The exact expression for the variance in the Gaussian framework is given in equation (11), and must be used for low-ℓ modes since they are sensitive to the shape of the mask.

The PCL approach is simply an inversion of the mask and does not attempt to minimize the loss of information caused by the decrease in the number of observed modes. In fact, it is well known that the PCL estimates are only optimal (i.e. unbiased, minimum variance estimates) for a flat power spectrum (see e.g. Efstathiou 2004b and Pontzen & Peiris 2010). This equivalence will prove useful in the context of galaxy surveys, as we shall see in the next sections. To recall the definition of the optimal estimator, we consider the generic class of quadratic estimators of the form

$$\hat{C}_\ell = \tilde{\boldsymbol{x}}^t\mathbf{E}^\ell\tilde{\boldsymbol{x}}. \tag{8}$$

[2] Hence $\tilde{\boldsymbol{b}} = \mathbf{K}\boldsymbol{b}$ for any data vector b, while for any matrix $\mathbf{B}$ we write $\tilde{\mathbf{B}} = \mathbf{KBK}$ implicitly taking advantage of the property $\mathbf{K}^t = \mathbf{K}$.

In this formalism, the PCL estimator reads

$$\mathbf{E}_{\mathrm{PCL}}^\ell = \sum_{\ell'} (\mathbf{M}^{-1})_{\ell\ell'}^{\mathrm{PCL}}\,\tilde{\mathbf{P}}^{\ell'}, \tag{9}$$

with the coupling matrix $(\mathbf{M})_{\ell\ell'}^{\mathrm{PCL}} = \mathrm{Tr}\,\tilde{\mathbf{P}}^\ell\tilde{\mathbf{P}}^{\ell'}$. In the Gaussian case, the expected value of the generic quadratic estimator is given by

$$\langle\hat{C}_\ell\rangle = \mathrm{Tr}\,\tilde{\mathbf{C}}\mathbf{E}^\ell, \tag{10}$$

and its variance by

$$\mathbf{V}_{\ell\ell'} = \langle\hat{C}_\ell\hat{C}_{\ell'}\rangle - \langle\hat{C}_\ell\rangle\langle\hat{C}_{\ell'}\rangle = 2\mathrm{Tr}\,\tilde{\mathbf{C}}\mathbf{E}^\ell\tilde{\mathbf{C}}\mathbf{E}^{\ell'}. \tag{11}$$

As a result, in the presence of sky cuts the uncertainties on the power spectrum estimates are typically significantly correlated. Uncorrelated error bars can be obtained by diagonalizing the covariance matrix and using the resulting rotation matrix to transform the power spectrum estimates and the theory predictions (Tegmark 1997; Tegmark et al. 2002b).

In the previous equations, $\tilde{\mathbf{C}}$ denotes the cut-sky pixel–pixel covariance matrix, which can be modelled as the superposition of a signal part $\tilde{\mathbf{S}}$ calculated with a theory prior $\{C_\ell\}$ and noise, i.e.

$$\tilde{\mathbf{C}} = \langle\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^t\rangle = \tilde{\mathbf{S}} + \tilde{\mathbf{N}}. \tag{12}$$

The pixel–pixel covariance matrix must also incorporate any additional signal present in the data, such as the systematics, as detailed in Section 2.4.

The minimum variance estimator in the Gaussian framework, first introduced in Tegmark (1997), is the so-called QML estimator. The latter reads

$$\mathbf{E}_{\mathrm{QML}}^\ell = \sum_{\ell'} (\mathbf{M}^{-1})_{\ell\ell'}^{\mathrm{QML}}\,\frac{1}{2}\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{P}}^{\ell'}\tilde{\mathbf{C}}^{-1}, \tag{13}$$

and uses the deconvolution matrix

$$(\mathbf{M})_{\ell\ell'}^{\mathrm{QML}} = \frac{1}{2}\mathrm{Tr}\,\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{P}}^\ell\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{P}}^{\ell'}. \tag{14}$$

In the Gaussian, isotropic case, QML is a lossless estimator that recovers all the relevant information contained in the data. The deconvolution and covariance matrices of the estimates are then equal to the inverse of the Fisher information matrix. While being no longer theoretically optimal for anisotropic theories, Pontzen & Peiris (2010) showed that QML remained superior to PCL.

PCL and QML can be contrasted in terms of the computational complexity and quality of the final power spectrum estimates. They both depend on the mask, but QML additionally requires an accurate model of the pixel–pixel covariance matrix. This model requires priors on the fiducial theory $\{C_\ell\}$ and on the additional correlations present in the data, such as noise and systematics. Although the pseudo-spectrum estimator does not explicitly use such priors, it is equivalent to a maximum likelihood analysis when a flat spectrum is assumed in place of a more motivated choice for the pixel–pixel covariance matrix (Efstathiou 2004b; Pontzen & Peiris 2010). As a result, PCL yields nearly optimal estimates when the power spectrum is close to flat and with no anisotropic contributions. Moreover, the sensitivity to the shape of the spectrum decreases at small scales as the number of observed modes increases. A simple inversion of the mask then maximizes the likelihood function and the variance of PCL reaches its minimum, namely the inverse of the Fisher matrix (Efstathiou 2004b, 2006), i.e.

$$\mathbf{V}_{\ell\ell'}^{\mathrm{PCL}} \approx \frac{2\mathcal{C}_\ell\mathcal{C}_{\ell'}}{2\ell+1}(\mathbf{M}^{-1})_{\ell\ell'}^{\mathrm{PCL}} \approx (\mathbf{M}^{-1})_{\ell\ell'}^{\mathrm{QML}} = \mathbf{V}_{\ell\ell'}^{\mathrm{QML}}. \tag{15}$$

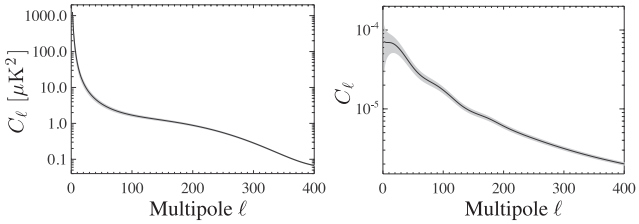
Note that this result is only valid in the regime where the signal dominates in the covariance matrix.

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019



(a) CMB mask (b) LSS mask

Figure 1. Fiducial CMB and galaxy survey masks used to calculate the covariance matrices of Fig. 3. These masks are in Galactic coordinates and approximate the *WMAP* KQ85 mask and the SDSS DR6 sky coverage, respectively.



(a) CMB spectrum (b) CMASS spectrum

Figure 2. Fiducial CMB and CMASS angular power spectra used to calculate the covariance matrices of Fig. 3. The grey bands show the cosmic variance.

To illustrate the contrast between the PCL and QML estimates, we calculated and compared their covariance matrices in three realistic settings of interest, involving typical masks and spectra of CMB and galaxy catalogues. The masks are shown in Fig. 1, the theory power spectra in Fig. 2, and the resulting covariance matrices in Fig. 3. For the CMB, we considered the KQ85 mask (smoothed, galaxy part only) and the best-fitting theory angular power spectrum of the *Wilkinson Microwave Anisotropy Probe* 9 yr data release (*WMAP*, see e.g. Bennett et al. 2012; Hinshaw et al. 2012). As a realistic setting for a galaxy survey, we considered a mask created from the sky coverage of the SDSS sixth Data Release (DR6), enlarged and smoothed for stability of the estimates, shown in Fig. 1 (the variance rapidly becomes unstable for complex mask shapes and requires binning, as detailed in Section 2.3). We used CAMB_SOURCES[3] (Challinor & Lewis 2011) to project the matter power spectrum $P(k)$ (corresponding to the *WMAP* 9 yr cosmology) into an angular power spectrum using the redshift distribution of the CMASS sample, using a fixed astrophysical bias $b_g = 2$. The CMASS sample will be described in further detail in Section 2.5.

The CMB spectrum varies across three orders of magnitude in the range $0 < \ell < 50$ and, as expected, QML performs better than PCL in this range. In particular, the variance of the largest scale modes is up to 20 per cent smaller compared with PCL. For this mask, PCL is a good estimator for $\ell > 50$; its variance is typically within $\sim$10 per cent of that expected for a maximum likelihood estimator, although the degree of sub-optimality will depend on the noise and the geometry of sky cut under consideration (Efstathiou 2004a,b, 2006; Hamimeche & Lewis 2009). The variance of the estimates is much larger for the galaxy survey mask, but since the CMASS spectrum varies by only one order of magnitude in $0 < \ell < 200$, the variance of the PCL and QML estimates in fact only differ by a few per cent for $\ell > 15$. In conclusion, since LSS power spectra are close to flat, PCL yields nearly optimal estimates, when no motivated prior for $\tilde{\mathbf{C}}$ is available. However, the degree of sub-optimality increases when considering masks with complex geometries, and in the presence of non-isotropic contributions (e.g. spatial fluctuations due to calibration errors). In this case, the implicit assumptions in the PCL estimator are poor priors for a maximum likelihood analysis, and a better model of the pixel–pixel covariance matrix must be used in the QML estimator to obtain optimal estimates.

[3] http://camb.info/sources/

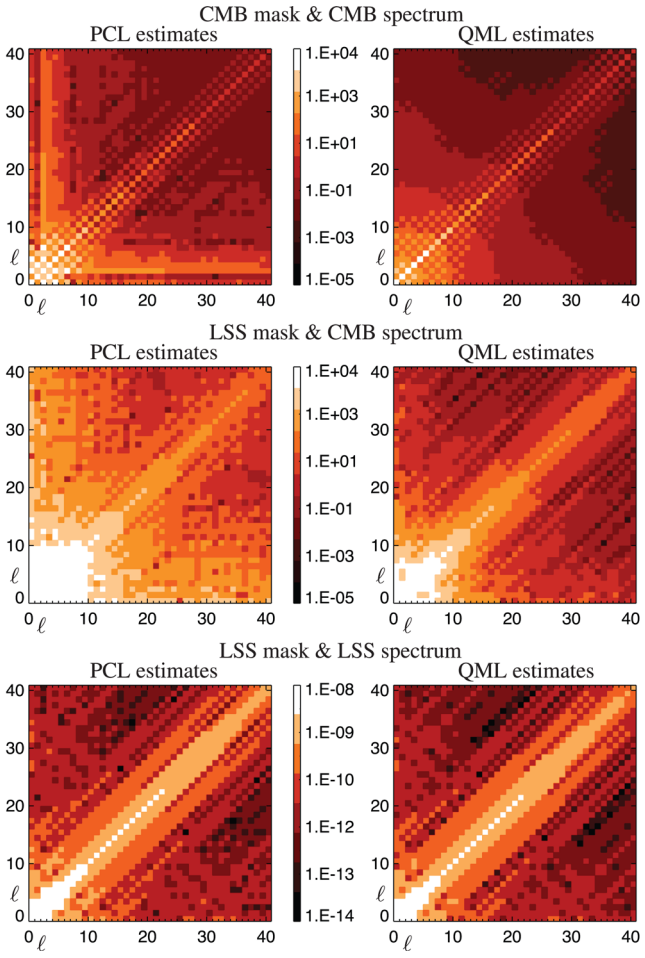


Figure 3. Covariance matrices $|V_{\ell\ell'}|$ of the PCL and QML estimates calculated for the CMB and LSS spectra and masks in Figs 1 and 2 in the absence of noise and systematics. For a CMB spectrum, the QML variances (top and middle right) are smaller and less correlated than PCL variances (top and middle left), and the resulting PCL estimates are significantly sub-optimal compared to QML. For the LSS spectrum, which is considerably flatter, this distinction is much less pronounced (bottom left and right panels), as expected by the reduction of QML to PCL in the case of a flat spectrum.

The complexity and resources involved in the PCL and QML algorithms also differ considerably. PCL benefits from fast, low-memory algorithms, and therefore can be applied to resolutions and multipole ranges which are amply sufficient for galaxy survey analyses. By contrast, QML involves the inversion of large covariance matrices and the execution of non-symmetric matrix multiplications. Although our optimized QML algorithm advantageously balances the work-load across processors and minimizes memory use, typical galaxy surveys such as the SDSS ($f_{\rm sky} \sim 1/6$) can only be analysed at HEALPIX $N_{\rm side} = 64$ on a personal computer. To ensure that the estimates are minimally affected by the resolution, accurate smoothing and masking rules are detailed in Appendix A.

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019

Neglecting these considerations can lead to significant biases if the estimation is performed at low resolution. Since high-ℓ maximum likelihood estimates might be desirable at high resolution in some cases, Gruetjen & Shellard (2012) recently proposed an iterative algorithm (similar to the Newton–Raphson iterative scheme presented in the next section) to converge to the maximum likelihood solution using an augmented PCL basis. This approach will prove useful for obtaining optimal estimates of the damping tail of the CMB spectrum, where a full QML analysis is intractable, and PCL estimates are sub-optimal due to the exponential decay and anisotropic contributions to these modes (Hamimeche & Lewis 2009). However, this improved estimator is unnecessary for the study of galaxy surveys, since PCL is nearly optimal at the smallest scales where shot noise dominates. Alternatively, the complexity of the QML estimator can be reduced by adopting Karhunen–Loève compression (Vogeley & Szalay 1996; Tegmark, Taylor & Heavens 1997; Tegmark et al. 1998; Tegmark, Hamilton & Xu 2002a); for further details see Appendix B. The results of this paper were obtained without this technique since we were able to run the estimator at $N_{\rm side} = 64$ without any other approximation, thus covering the scales which are not dominated by shot noise.

2.3 Likelihood analysis and band-powers

The size and the shape of the mask strongly influence the variance of power spectrum estimates, and thus, the quality of any subsequent analysis. Typical CMB masks only cause a modest increase in the variance (e.g. $\sim$30 per cent for a *WMAP*-style mask with $f_{\rm sky} = 75$ per cent) compared with the full sky case. By contrast, galaxy survey masks typically have $f_{\rm sky} \lesssim 20$ per cent and can increase the variance of the individual multipoles by factors of 5–10.

This problem is usually addressed by estimating the power spectrum in multipole bands, which smoothes the power spectrum, Gaussianises the likelihood function, reduces the sensitivity to the input prior and decreases the variance of the estimates. In particular, this approach proves useful for PCL since its prior can be significantly sub-optimal when using complex masks and in the presence of anisotropic contributions. The PCL and QML estimators in equations (9) and (13) are straightforwardly transformed into band-power estimators by considering binned Legendre matrices $\tilde{\mathbf{P}}^b = \sum_{\ell \in L_b} \tilde{\mathbf{P}}^\ell$, where L_b denotes the range of multipoles included in the b-th bin. The pixel–pixel covariance matrix $\tilde{\mathbf{C}}$ is unchanged and calculated as before, but the coupling matrices in the binned formulation are given by

$$(\mathbf{M})_{bb'} = \sum_{\ell \in L_b} \sum_{\ell' \in L_{b'}} (\mathbf{M})_{\ell\ell'}. \tag{16}$$

The expectation value of each band-power estimate reads

$$\langle \hat{C}_b \rangle = \sum_\ell W_{b\ell} C_\ell, \tag{17}$$

where the window function $W_{b\ell}$ is constructed with the coupling matrix through

$$W_{b\ell} = \sum_{b'} (\mathbf{M})^{-1}_{bb'} (\mathbf{M})_{b'\ell}. \tag{18}$$

Hence, to be confronted with band-power estimates, the theory power spectrum must be transformed into band-powers using these window functions.

Using large band-powers guarantees that the likelihood function is sufficiently close to Gaussian for QML to deliver maximum-likelihood estimates. However, intermediate situations with smaller bins and complex masks might not fulfil this condition, leading to sub-optimal estimates. This issue can be addressed by iteratively converging to the maximum-likelihood solution using a Newton–Raphson scheme (Bond et al. 1998, 2000; Knox, Bond & Jaffe 1998). With the so-called Newton–Raphson maximum likelihood (NRML) estimator, each iteration improves the previous one using

$$\delta C_\ell = \sum_{\ell'} (\mathbf{M}^{-1})^{\rm QML}_{\ell\ell'} \frac{1}{2} {\rm Tr} \left[\left(\boldsymbol{x}\boldsymbol{x}^t - \tilde{\mathbf{C}} \right) \left(\tilde{\mathbf{C}}^{-1} \tilde{\mathbf{P}}^\ell \tilde{\mathbf{C}}^{-1} \right) \right]. \tag{19}$$

This approach is in fact equivalent to calculating the ith estimate by feeding QML with the previous iteration. The formulation of the NRML estimator then simplifies to

$$\hat{C}^{i+1}_\ell = \tilde{\boldsymbol{x}}^t (\mathbf{E}^\ell)^i \tilde{\boldsymbol{x}}, \tag{20}$$

where $(\mathbf{E}^\ell)^i$ explicitly makes use of the covariance matrix calculated from $\hat{C}^i_\ell$, namely $\tilde{\mathbf{C}} = \sum_\ell \tilde{\mathbf{P}}^\ell \hat{C}^i_\ell + \tilde{\mathbf{N}}$. Hence, a critical issue is to construct a reliable covariance matrix at each iteration.

Both PCL and QML estimates can be extended to estimate the angular cross-power spectra (and hence cross band-powers) between two maps denoted by $\tilde{\boldsymbol{x}}_1$ and $\tilde{\boldsymbol{x}}_2$. The PCL estimator reads

$$\hat{C}^{\rm PCL,cross}_\ell = \tilde{\boldsymbol{x}}^t_1 \mathbf{E}^\ell_{\rm PCL} \tilde{\boldsymbol{x}}_2, \tag{21}$$

and implicitly assumes flat autopower spectra and zero cross-power spectrum as priors. On the contrary, since QML derives from a likelihood function, the estimator must be adapted (following e.g. Tegmark & de Oliveira-Costa 2001; Padmanabhan et al. 2005) by considering the input data vector as a concatenation of the two maps,

$$\tilde{\boldsymbol{x}} = \begin{pmatrix} \tilde{\boldsymbol{x}}_1 \\ \tilde{\boldsymbol{x}}_2 \end{pmatrix}, \tag{22}$$

and by using a pixel–pixel covariance matrix which incorporates all the information about the maps and their cross-correlation,

$$\tilde{\mathbf{C}} = \begin{pmatrix} \tilde{\mathbf{C}}_{11} & \tilde{\mathbf{C}}_{12} \\ \tilde{\mathbf{C}}^\dagger_{12} & \tilde{\mathbf{C}}_{22} \end{pmatrix}. \tag{23}$$

This formulation makes use of priors for the three power spectra, and must include models of the additional correlations and noise present in the data. The auto and cross-spectra can be simultaneously estimated from $\tilde{\boldsymbol{x}}$ and $\tilde{\mathbf{C}}$ using the usual formulation of QML, i.e. equations (13) and (14), In particular, the matrices to be used in the estimator with the covariance matrix of equation (23) in order to calculate $\hat{C}^{11}_\ell$, $\hat{C}^{22}_\ell$ and $\hat{C}^{12}_\ell$ read

$$\tilde{\mathbf{P}}^\ell_{11} = \begin{pmatrix} \tilde{\mathbf{P}}^\ell & 0 \\ 0 & 0 \end{pmatrix}, \tilde{\mathbf{P}}^\ell_{22} = \begin{pmatrix} 0 & 0 \\ 0 & \tilde{\mathbf{P}}^\ell \end{pmatrix}, \tilde{\mathbf{P}}^\ell_{12} = \begin{pmatrix} 0 & \tilde{\mathbf{P}}^\ell \\ \tilde{\mathbf{P}}^\ell & 0 \end{pmatrix}. \tag{24}$$

2.4 Galaxy surveys, shot noise and systematics

Galaxy catalogues are usually provided as lists of objects whose positions and properties, such as photometric colours, were measured by an instrument in the context of a sky survey. To relate to the dark matter distribution, a catalogue must be pixelized into a number count map $\tilde{\boldsymbol{G}}$, and then transformed into an overdensity map $\tilde{\boldsymbol{x}}$. Given a pixelization scheme, if G_i denotes the number of objects in the ith pixel (described by its size $\Omega_{\rm i}$ and the position of its centre $\boldsymbol{n}_i$), the overdensities are constructed as

$$\tilde{x}_i = \frac{\tilde{G}_i}{\Omega_{\rm i} \bar{G}} - 1, \tag{25}$$

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019

where $\bar{G} = N_{\rm obj}/\Delta\Omega$ is the average number of objects per steradian. $N_{\rm obj}$ is the total number of objects in the catalogue and $\Delta\Omega$ is the total surface outside the mask. The Poisson sampling of the observed tracers naturally gives rise to shot noise, characterized by a diagonal noise matrix $\tilde{\mathbf{N}}$ (Huterer, Knox & Nichol 2001) such that $(\tilde{\mathbf{N}})_{ij} = \bar{G}^{-1}\delta_{\rm ij}$, where δ is the Kronecker delta. Moreover, the power spectrum estimates of galaxy survey overdensity maps always include a constant bias term due to shot noise,

$$\langle \hat{C}_\ell \rangle = C_\ell + \frac{1}{\bar{G}}. \tag{26}$$

Since the cosmological contribution C_ℓ steadily decreases with ℓ in the linear regime, shot noise usually dominates on small scales for catalogues with small number densities (i.e. for ℓ higher than a limit determined by $\bar{G}$).

So far, we assumed that the map x was the result of cosmological clustering encapsulated by a theory power spectrum $\{C_\ell\}$. In other words, the relevant correlations in the map were due to $\{C_\ell\}$ and the shot noise $1/\bar{G}$. However, observations $\tilde{\boldsymbol{x}}^{\rm obs}$ are often contaminated by various signals introducing spurious correlations and requiring appropriate modelling to avoid sub-optimal estimates. In particular, $\tilde{\boldsymbol{x}}^{\rm obs}$ can always be described as the superposition of the true cosmological signal $\tilde{\boldsymbol{x}}^{\rm true}$ and a contamination part due to systematics. We will assume that templates of the systematics are available, namely $n_{\rm sys}$ maps denoted by $\boldsymbol{c}^k$ with $k = 1, \ldots, n_{\rm sys}$. In this case, one can adopt a model for the contamination signal, and estimate its parameters from the data. The best-fitting contamination model is then subtracted from the measured power spectra. This approach was recently used to correct the angular power spectra and 2-pt correlation functions of the CMASS sample with a best-fitting linear contamination model (Ross et al. 2011; Ho et al. 2012).

A more robust approach to mitigate the influence of the systematics is to incorporate their contribution in the pixel–pixel covariance matrix with large coefficients ξ_k, i.e.

$$\tilde{\mathbf{C}}_{ij} = \tilde{\mathbf{S}}_{ij} + \tilde{\mathbf{N}}_{ij} + \sum_k \xi_k \tilde{c}_i^k \tilde{c}_j^k. \tag{27}$$

This technique, known as mode projection (Slosar, Seljak & Makarov 2004; Ho et al. 2008), assigns a very large variance to the modes corresponding to the systematics in pixel space, such that they do not influence the power spectrum estimates.[4] As a result, the QML estimates are unbiased in the power spectrum of $\tilde{\boldsymbol{x}}^{\rm true}$ as $\xi_i \to \infty$, provided that the contamination signal can be described as a linear combination of the templates. In the Bayesian perspective, mode projection is equivalent to marginalizing over the parameters of a linear model of the contamination. Hence, non-linear contamination or neglected systematics leave residual biases that must be eliminated by other means, for example through masking or modelling of the contamination signal.

2.5 Illustration: recovering the power spectrum of the CMASS sample

The ongoing Baryon Oscillation Spectroscopic Survey (BOSS) is part of SDSS-III (Eisenstein et al. 2011) and aims to measure the spectroscopic redshifts of 1.5 million galaxies, 160 000 quasars and various ancillary targets from SDSS photometry (Gunn et al. 1998, 2006). The CMASS spectroscopic sample includes extended sources selected using colour–magnitude cuts to produce a roughly volume-limited sample in the redshift range $0.4 < z < 0.7$. In the DR9 release of BOSS, CMASS uses data taken up to the end of 2011 July and covers 3344 deg^2 in the Northern and Southern Galactic caps.

Manera et al. (2013) presented a set of 600 mock catalogues for the CMASS DR9 sample, constructed based on a Λ cold dark matter (ΛCDM) cosmology defined by $\{\Omega_{\rm m} = 0.274, \Omega_{\rm b}h^2 = 0.0224, h = 0.70, n_s = 0.95, \sigma_8 = 0.8\}$, and evolved using second-order Lagrangian perturbation theory (Scoccimarro 1997; Crocce, Pueblas & Scoccimarro 2006). These mock catalogues were used to compute accurate covariance matrices for CMASS and constrain cosmological parameters (Sánchez et al. 2012), test deviations from general relativity (Samushia et al. 2013) and measure the scale-dependent halo bias (Ross et al. 2013).

To evaluate the performance of the previously described angular power spectrum estimators, we compared the mean and standard deviation of the power spectrum estimates of the mock catalogues with the theoretical expectations. We only considered the CMASS DR9 mock catalogues in the Northern Galactic cap, covering 2635 deg^2 as shown in Fig. 4. We computed a theory prediction for the angular power spectrum of the underlying dark matter with CAMB_SOURCES (Challinor & Lewis 2011) using a redshift distribution parametrized by an Edgeworth expansion of the redshift histogram of CMASS objects.

We calculated the PCL and QML estimates of the individual mock catalogues at HEALPIX resolutions $N_{\rm side}^{\rm PCL}$=128 and $N_{\rm side}^{\rm QML}$=64. The mean and variance of the estimates compared with theoretical expectations are shown in Fig. 5. Note that the error bars of the power spectrum estimates are correlated; see equation (11). We subtracted the shot noise and also multiplied the theory prediction by a scale-independent bias of 1.9, the value used to construct these mock catalogues (Manera et al. 2013). We were able to recover the theory power spectrum and the covariance of the estimates with good precision using the smoothing, masking and band-limit rules defined in Appendix A. The estimates and the theory predictions from CAMB_SOURCES were insensitive to small changes in the redshift distribution. However, we observed that inconsistent smoothing, masking and band-limit rules led to biases in the recovered power spectra, resulting from a mismatch between the model pixel–pixel covariance matrices and the information content in the data. As expected, due to the flatness of the power spectrum and the simple geometry of the mask, QML performed only marginally better than PCL, which yielded nearly optimal estimates in the absence of



Figure 4. Mask of the CMASS DR9 sample in the Northern Galactic cap, in equatorial coordinates.

[4] This approach does not prevent the extraction of cosmological signals that happen to have the same power spectra as the systematics. In particular, projecting out a systematic is equivalent to ignoring one mode defined in pixel space.

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019

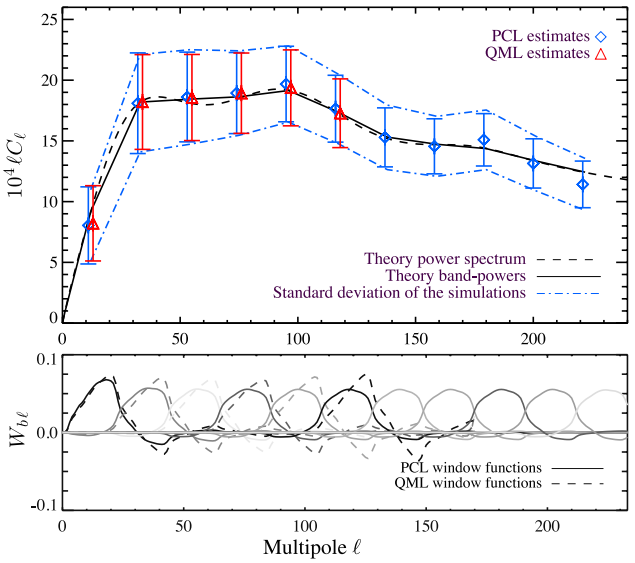


Figure 5. Average PCL and QML estimates of the 600 CMASS mock catalogues (Manera et al. 2013). Shot noise was subtracted from the estimates, and the angular power spectrum was calculated with CAMB_SOURCES (black dashed line) using a scale-independent astrophysical bias $b_g = 1.9$. The theory band-powers (black solid line) were obtained by applying the exact PCL and QML binning window functions of equations (18) and (17) (bottom panel). We were able to recover the theory power spectrum and the covariance of the estimates with good precision using the smoothing, masking and band-limit rules defined in Appendix A.

systematics. The theory spectrum was converted into band-powers using the exact window functions in equations (17) and (18), which are shown in the bottom panel of Fig. 5.

3 APPLICATION TO SDSS PHOTOMETRIC QUASARS

3.1 Data and sub-samples

We considered the Richards et al. (2009) catalogue of photometric quasars, which is based on the DR6 of the SDSS (Adelman-McCarthy et al. 2008). The objects in this catalogue – which we call RQCat as in Pullen & Hirata (2012) – were photometrically selected by a binary Bayesian classifier trained in 4D colour space using several catalogues of spectroscopically confirmed stars and quasars. The final version of the catalogue includes 1172 157 objects with several quality and technical flags, which can be exploited to apply further systematics cuts and obtain cleaner samples. The Bayesian classifier was initially applied to all point sources in the DR6 release without restriction, and explicitly assumed that 95 per cent of the input objects were stars (i.e. constant star and quasar priors $p_{\rm star} = 0.95$ and $p_{\rm quasar} = 0.05$). It was also applied with stronger priors ($p_{\rm star} = 0.98$) to objects with photometric redshifts $\tilde{z}_{\rm p}$ (estimated by the SDSS pipeline, as opposed to the true redshift z) in three redshift ranges $0 < \tilde{z}_{\rm p} \leq 2.2$, $2.2 < \tilde{z}_{\rm p} \leq 3.5$ and $\tilde{z}_{\rm p} > 3.5$ to achieve higher efficiency[5] and completeness. The efficiency is degraded in the two higher redshift ranges since the Bayesian classifier performs worse due to the overlap between the stellar and quasar loci in the colour space (at $z \sim 2.6$). More sophisticated algorithms such as XDQSO (Bovy et al. 2011) were developed to specifically address this issue and identify higher redshift quasars for spectroscopic follow-up in the context of BOSS (Ross et al. 2012a). However, in this work we focused on $z \leq 2.2$ objects, where the performance of the binary Bayesian classifier used by Richards et al. (2009) is satisfactory.

Colour-based selection of quasars is difficult, and RQCat is expected to be significantly contaminated by stars, which are often misclassified as quasars due to similar colours. Hence, the full catalogue cannot be used as a statistical sample for direct power spectrum analysis due to its low efficiency (lower than 80 per cent). In this work, we restricted ourselves to good ultraviolet-excess (UVX) low-redshift objects, defined as $u - g < 1.0$ and $\tilde{z}_{\rm p} < 2.2$ (the corresponding flags are GOOD>0, UVX=1 and LOWZ=1). This sample, denoted by UVX-LOWZ, is the least contaminated by stars and achieves 96.3 per cent ± 1.2 efficiency (Richards et al. 2009).

Previous studies considered the UVX sources in RQCat for cosmological analyses, e.g. studying the environment of quasars (Myers et al. 2006, 2007), detecting the ISW effect (Corasaniti, Giannantonio & Melchiorri 2005; Giannantonio et al. 2006, 2008; Giannantonio et al. 2012) and constraining PNG (Slosar et al. 2008; Xia et al. 2010, 2011). Recent work by Pullen & Hirata (2012), corroborated by Giannantonio et al. (2013), found that the UVX objects were significantly contaminated, as indicated by the cross-spectra of redshift bins which exhibited excess power at the largest scales.

The two main sources of systematics in quasar photometric catalogues are contamination and calibration errors. The origin of contamination lies in the classification stage: selecting quasars based on photometric data is a complex task. Various objects can be misclassified as quasars, and are thus present in the final catalogues. Since the clustering properties of these contaminants differ from those of quasars, they affect the measured power spectra and can jeopardize the interpretation of the data. Calibration errors, on the other hand, are present in the catalogue regardless of the ability of the classifier to separate stars and quasars. In the ideal case, a perfect classifier applied to all point sources detected by a given instrument will lead to a sample with clustering properties purely due to cosmological physics. However, real instruments are not perfectly calibrated, and observing conditions also change with time, introducing spurious correlations due to variations in the number of detected sources on the sky. In addition, calibration errors impact the apparent magnitude estimates, which propagate through the Bayesian classifier (since the latter does not model or account for them), inducing a spatial dependence in its efficiency.

Contamination and calibration issues can be addressed in different ways. First, it is important to reduce the catalogue of interest by selecting the most reliable objects (here UVX sources) and also restricting the analysis to the most reliable areas of the sky. Secondly, corrections can be applied to the power spectrum estimates themselves to minimize the remaining spurious correlations. Alternatively, one can opt for a Bayesian analysis and marginalize over the systematics in the cosmological analysis. In this study, we focused on the sample reduction approach. We separated the UVX-LOWZ sample into four sub-samples by selecting objects with photometric redshifts $\tilde{z}_{\rm p}$ in bins with ranges [0.5, 1.3], [1.3, 1.8], [1.8, 2.2] and [1.3, 2.2]. These samples are called Low-z, Mid-z, High-z and Mid+High-z respectively, as referred to as the RQCat sub-samples. We rejected low-redshift quasars ($\tilde{z}_{\rm p} < 0.5$) because their power spectra were severely contaminated: this can be attributed to strong stellar contamination, and to the fraction of low-z quasars which are extended sources and were therefore

[5] The efficiency, or purity, of a catalogue denotes the fraction of objects which are quasars. It characterizes the ability of the classifier to separate quasars from stars.

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019

not processed by the Bayesian classifier. The corresponding incompleteness is non-trivial and likely to depend on observational effects such as dust absorption and seeing variations.

3.2 Theory predictions

In order to calculate theoretical predictions for the angular power spectra of the four RQCat sub-samples, we used CAMB_SOURCES (Challinor & Lewis 2011), a high-precision code which projects the 3D matter power spectrum $P(k)$ into angular auto and cross-power spectra. Since this study is focused on the impact of the systematics on the observed power spectra, we fixed the cosmological parameters to *Planck* ΛCDM best-fitting values[6] (Planck Collaboration 2013b). We opted for a scale-independent linear bias to relate the observed galaxy clustering to dark matter.

Although the matter power spectrum $P(k)$ only depends on cosmological parameters, computing angular power spectrum predictions requires additional knowledge about the distributions and properties of the samples under consideration. In particular, the auto-angular power spectrum of a sample of interest reads

$$C_\ell = \frac{2}{\pi} \int \mathrm{d}k k^2 P(k) [W_\ell(k)]^2, \tag{28}$$

where the window function $W_\ell(k)$ includes several cosmological effects, as detailed in Challinor & Lewis (2011). In fact, $W_\ell(k)$ requires two quantities in addition to the standard cosmological parameters: the unit-normalized redshift distribution of tracers, denoted by $n(z)$, and the logarithmic slope of the number counts s, which accounts for the effect of magnification due to lensing. A simplified formula for $W_\ell(k)$, which we use for illustration purposes only (see Challinor & Lewis 2011 for the full formalism), reads

$$W_\ell(k) = \int \mathrm{d}z \left[b_\mathrm{g} n(z) + 2(2.5\,s - 1) f(z) \right] D(z) j_\ell(kr), \tag{29}$$

where b_g is the linear galaxy bias, $D(z)$ is the growth factor, j_ℓ is the spherical Bessel function, $r(z)$ is the comoving distance and $f(z)$ is the lensing window function, giving rise to magnification. In this work, we assumed that the logarithmic slope of the number counts, defined as

$$s = \frac{\mathrm{d} \log N(m)}{\mathrm{d}m}, \tag{30}$$

was constant in each redshift range. We estimated its value for the four RQCat sub-samples by calculating the slope of the histogram of number counts in terms of the g-band PSF magnitude at $g = 21$. We found $s = 0.18$, 0.89, 0.89 and 0.87, respectively for these sub-samples, consistent with previous studies in similar redshift ranges (Xia et al. 2009; Pullen & Hirata 2012). The choice of a scale- and redshift-independent linear bias is motivated by previous studies of RQCat and of low-redshift quasars in general (e.g. Myers et al. 2007; Pullen & Hirata 2012; Sherwin et al. 2012; Giannantonio et al. 2013). The redshift evolution of the bias in $0.5 < z < 2.2$ proves to be smaller than the uncertainty on the completeness corrections, and thus marginally affects the predicted angular power spectra.

3.3 Redshift distributions estimates

The quasars in each RQCat sub-sample are characterized by a normalized redshift distribution $n(z)$, with $n(z)\mathrm{d}z$ corresponding to the probability of finding a quasar with redshift between z and $z + \mathrm{d}z$. Consequently, $n(z)$ incorporates the physical distribution (i.e. originating from the quasar luminosity function), the survey characteristics (such as the magnitude limits) and the photometric redshift cuts used to construct the sub-sample under consideration. The simplest estimator for $n(z)$ is a normalized histogram of the photometric redshifts of all objects in each sub-sample of interest. However, in practice, this approach does not yield good estimates due to the large uncertainties in quasar photometric redshift estimates. To illustrate this issue, we used the SDSS-DR7, BOSS and 2SLAQ spectroscopic quasar catalogues (Croom et al. 2009; Schneider et al. 2010; Pâris et al. 2012) to find objects in the four RQCat sub-samples for which reliable spectra, and thus good spectroscopic redshifts, were available. Fig. 6 shows the photometric and spectroscopic redshifts ($\tilde{z}_\mathrm{p}$ and $\tilde{z}_\mathrm{s}$, respectively) of the cross-matched objects. The dispersion of the points $(\tilde{z}_\mathrm{p}, \tilde{z}_\mathrm{s})$ around $\tilde{z}_\mathrm{p} = \tilde{z}_\mathrm{s}$ demonstrates that the photometric redshift estimates of quasars suffer from large uncertainties and catastrophic failures, yielding a large fraction of spectroscopic redshifts outside the photometric windows used to construct the RQCat sub-samples, indicated by the dashed lines. The photometric redshift estimates do not accurately follow the underlying redshift distributions $n(z)$, and cannot be used to compute accurate angular power spectrum predictions.

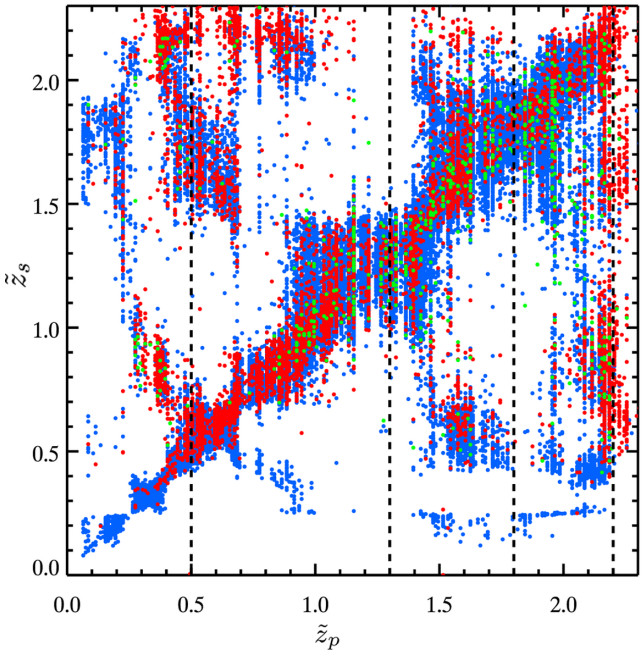


Figure 6. Distributions of the photometric and spectroscopic redshift estimates ($\tilde{z}_\mathrm{p}$ and $\tilde{z}_\mathrm{s}$) of the RQCat UVX-LOWZ sample, cross-matched with the SDSS-DR7, BOSS and 2SLAQ spectroscopic quasar catalogues (blue, red and green dots). The dashed lines indicate the photometric redshift cuts used to assemble the four RQCat sub-samples. The photometric redshift estimates are seen to be unreliable and cannot be used to estimate the redshift distributions of the photometrically selected sub-samples. On the contrary, the cross-matched samples have reliable spectroscopic redshifts, and can be used for this purpose. However, one must apply the relevant completeness corrections in order to account for the change of selection function between the photometric and cross-matched sub-samples (due to e.g. different magnitude limits).

Nevertheless, one can use the redshift distributions of the cross-matched samples, which can be calculated with great accuracy using the spectroscopic redshift estimates, whose uncertainties are negligible compared to the precision required for $n(z)$. However, the

[6] Fixed to $\Omega_\mathrm{c}h^2 = 0.188$, $\Omega_\mathrm{m} = 0.315$, $\Omega_\mathrm{b}h^2 = 0.02\,205$, $H_0 = 67.3\ \mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$, $\ln(10^{10}A_s) = 3.089$, $n_s = 0.9603$ and $\tau = 0.089$.

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019

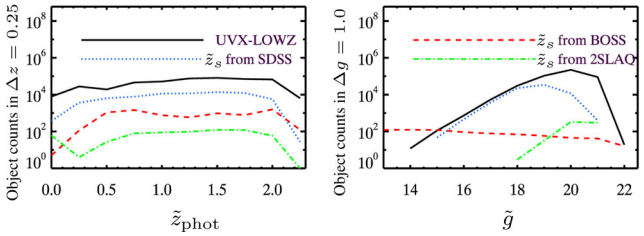


Figure 7. Histograms of the redshifts and apparent magnitudes of UVX-LOWZ objects in RQCat, and of objects with good spectra found in the SDSS-DR7, BOSS and 2SLAQ spectroscopic quasar catalogues. In addition to the higher number of objects, the SDSS-DR7 cross-matched sample has redshift and magnitude distributions close to that of RQCat, indicating similar selection functions. Hence, the redshift distributions of the RQCat sub-samples can be estimated using the cross-matched sample, with only minor completeness corrections to account for the differences in magnitude limits. By contrast, BOSS and 2SLAQ target significantly different redshift and magnitude ranges, and the completeness corrections required to estimate the redshift distributions of RQCat are strongly redshift and magnitude dependent. In addition, the latter are limited by sample variance due to the smaller number of objects.

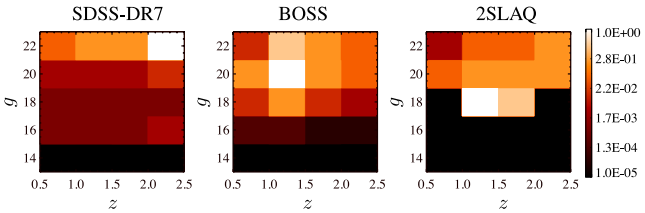


Figure 8. Low-resolution redshift- and magnitude-dependent completeness corrections for estimating the redshift distributions of the RQCat sub-samples through cross-matching with the SDSS-DR7, BOSS and 2SLAQ spectroscopic quasar catalogues. As expected from Fig. 7, the redshift distribution of the cross-matched sample with SDSS-DR7 only requires weak magnitude corrections to relate to that of RQCat. On the contrary, BOSS and 2SLAQ require significant corrections, since they target different redshift and magnitude ranges than RQCat.

redshift, magnitude and spatial distributions of the cross-matched samples may deviate from those of the photometric samples due to differences in their selection functions, caused, e.g. by different magnitude limits, redshift ranges or sky coverage. In order to avoid biases in the redshift distribution estimates, one must include a completeness correction factor, denoted by $f_{\rm c}$, which is a function of redshift z, magnitude g and position on the sky $\boldsymbol{n}$.

In practice, we divide the redshift and magnitude domains into bins denoted by $[z_i^{\rm min}, z_i^{\rm max}]$ (centred at z_i) and $[g_j^{\rm min}, g_j^{\rm max}]$ (centred at g_j), respectively, and the sky into pixels denoted by k. For each RQCat sub-sample, an estimator of $n(z)$ at redshift $z = z_i$ is given by

$$\tilde{n}(z_i) = C \sum_{jk} \frac{N_{\rm obj}({\rm cross}, [z_i^{\rm min}, z_i^{\rm max}], [g_j^{\rm min}, g_j^{\rm max}], k)}{\tilde{f}_c(z_i, g_j, k)}, \qquad (31)$$

where C is a normalization constant (such that $\sum_i \tilde{n}(z_i) = 1$) and $N_{\rm obj}({\rm cross}, [z_i^{\rm min}, z_i^{\rm max}], [g_j^{\rm min}, g_j^{\rm max}], k)$ is the number of objects in the *cross-matched* sample in the ith redshift bin, jth magnitude bin and kth pixel.

With an estimator of $n(z)$ in hand, we now discuss how to construct the completeness correction. A simple estimator for the correction in the (i, j, k)th volume element reads

$$\tilde{f}_c(z_i, g_j, k) = \frac{N_{\rm obj}({\rm cross}, [z_i^{\rm min}, z_i^{\rm max}], [g_j^{\rm min}, g_j^{\rm max}], k)}{N_{\rm obj}({\rm photo}, [z_i^{\rm min}, z_i^{\rm max}], [g_j^{\rm min}, g_j^{\rm max}], k)}, \qquad (32)$$

where $N_{\rm obj}({\rm photo}, [z_i^{\rm min}, z_i^{\rm max}], [g_j^{\rm min}, g_j^{\rm max}], k)$ is the number of objects in the *photometric* sample (namely, one of the RQCat sub-samples), thus constructed using the photometric redshifts.

The resolution of previous multidimensional histograms is limited by sample variance and by the uncertainties on (both photometric and spectroscopic) redshift and magnitude estimates.[7] To estimate the redshift distribution of the RQCat sub-samples, we investigated the use of the SDSS-DR7, BOSS and 2SLAQ spectroscopic quasar catalogues. The redshift and magnitude distributions of the UVX-LOWZ objects in RQCat cross-matched with these catalogues are shown in Fig. 7. We found that the BOSS- and 2SLAQ-based cross-matched samples were small and had selection functions quite different from RQCat. As a result, the estimator required significant completeness corrections, and the multidimensional histograms had to be constructed with large bins to reduce sample variance and biases due to uncertainties on the redshift and magnitude estimates. On the other hand, cross-matching with SDSS-DR7 led to large samples which had selection functions similar to that of the RQCat sub-samples. Fig. 8 shows low-resolution estimates of completeness corrections arising when cross-matching the whole UVX-LOWZ sample (union of the RQCat sub-samples) with SDSS-DR7, BOSS and 2SLAQ, where spatial dependence was neglected for purposes of illustration. Even at such low resolution, the completeness corrections for BOSS and 2SLAQ exhibit strong redshift and magnitude dependences (and are limited by sample variance due to the low number of cross-matched objects). For these reasons, the final redshift distribution estimates of the four RQCat samples were calculated using the SDSS-DR7 catalogue. The right-hand panels of Fig. 9 show number count maps of the RQCat sub-samples and the SDSS-DR7-based cross-matched samples. The numbers of objects in the respective sub-samples are summarized in Table 1.

With the high number of cross-matched objects, we were able to test various assumptions for the completeness corrections, such as the weakness of the redshift dependence, and choose resolutions that yielded the best estimates. Since SDSS-DR7 spans the same redshift range as RQCat, no redshift-dependent corrections were required (Fig. 8 shows that the correction is only weakly redshift dependent, apart from a mild transition at $z = 2$. Neglecting this transition did not significantly impact the final estimates). We used magnitude-dependent completeness corrections at resolution $\Delta g = 1.0$, due to the different magnitude limits in the SDSS photometric and spectroscopic data. Finally, the SDSS-DR7 spectroscopic quasar catalogue was assembled from different data releases of SDSS and is known to be non-uniform on the sky (Schneider et al. 2010). Indeed, Fig. 9 shows that the cross-matched samples contain regions with a greater number of objects, in some cases exploring fainter magnitudes. To address this spatial dependence, we calculated the magnitude-dependent completeness corrections in individual pixels at HEALPIX resolution $N_{\rm side} = 16$. The final estimates $\tilde{n}(z_i)$ are shown in Fig. 9, and were fitted by superpositions of Gaussian distributions for use in CAMB_SOURCES.

Fig. 10 shows the estimates obtained with CAMB_SOURCES (thus using the full formalism from Challinor & Lewis 2011) with our best estimates for s and $n(z)$, compared with the same estimates where magnification and completeness corrections were neglected.

[7] We will neglect the uncertainty in the sky position since it is negligible compared to the pixel size used for the completeness correction.

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019

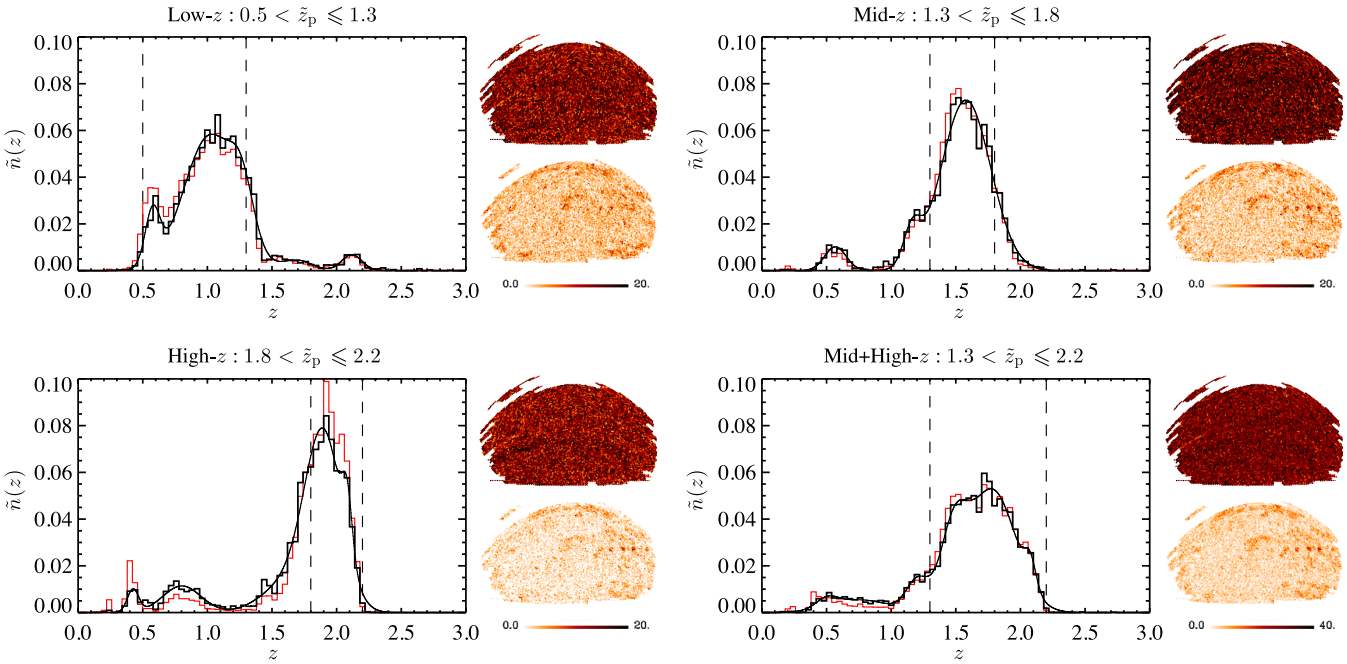


Figure 9. Final estimates of the redshift distributions of the four RQCat photometric sub-samples, calculated by cross-matching with the SDSS-DR7 spectroscopic quasar catalogue (Schneider et al. 2010). The accuracy of the redshift distribution estimates is essential for obtaining robust angular power spectrum estimates. Maps of the photometric and cross-matched samples are shown on the right-hand sub-panels at HEALPIX resolution $N_{\rm side} = 64$. The dashed vertical lines in the main sub-panels indicate the photometric redshift cuts used to construct the RQCat sub-samples. Since the cross-matched samples have different selection functions, their redshift distributions (red thin histograms) must be corrected in order to accurately estimate the redshift distributions of the RQCat samples. The thick black histograms show the final estimates obtained by applying magnitude- and pixel-dependent completeness corrections, and were fitted with a superposition of Gaussian distributions (solid lines) for use in CAMB_SOURCES.

Table 1. Number of objects in the four RQCat redshift sub-samples ($N_{\rm obj}$ photo), and for which good spectra (and thus good spectroscopic redshifts) were found in the SDSS-DR7 quasar catalogue ($N_{\rm obj}$ cross). These cross-matched samples were used to estimate the redshift distributions shown in Fig. 9.

	Low-z	Mid-z	High-z	Mid+High-z
$N_{\rm obj}$ photo	95 185	109 713	92 740	202 453
$N_{\rm obj}$ cross	19 328	17 589	10 654	28 243

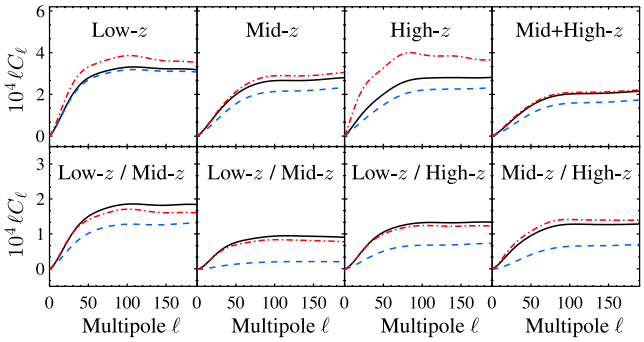


Figure 10. Theory predictions for the four RQCat sub-samples computed with CAMB_SOURCES. The black lines show the final theory predictions, calculated with the best redshift distribution estimates presented in Section 3.3, for a *Planck* cosmology and a fixed galaxy bias $b_{\rm g} = 2.3$. The blue dashed lines show the predictions obtained by neglecting the effect of magnification, and the red dot–dashed lines by neglecting the completeness correction in the estimated redshift distributions (i.e. using the redshift histograms of the cross-matched samples without accounting for the differences in magnitude limits).

The large differences between the resulting angular power spectra demonstrate that these effects must be accounted for, and carefully estimated from the data in order to avoid significant biases in the theory predictions.

3.4 Masks and systematics

We considered five sources of systematics: stellar contamination, dust absorption, seeing, airmass and sky brightness. Following Pullen & Hirata (2012), we constructed the stellar density map from SDSS DR6 point sources with $18.0 < r < 18.5$ and $i < 21.3$. For the extinction, we used the dust maps from Schlegel, Finkbeiner & Davis (1998) with the corrections by Peek & Graves (2010). Templates for seeing, airmass and sky brightness were constructed with the MANGLE software (Hamilton & Tegmark 2004; Swanson et al. 2008) using data retrieved from the FIELDS table in the SDSS CAS server. All maps were binned on to the HEALPIX grid at resolution $N_{\rm side} = 128$, and are shown in Fig. 11.

We designed three sky masks by excluding pixels based on their values in the systematics maps. The thresholds are summarized in Table 2, and the resulting masks are presented in Fig. 12. Following Pullen & Hirata (2012), we also excised rectangular regions with missing data.[8] Our first mask is therefore very similar to those used in previous studies of RQCat and constitutes the reference mask. The two other masks use more aggressive systematics cuts.

[8] In equatorial (J2000) coordinates, the discarded angular rectangles are $(\alpha, \delta) = (122°–139°, -1.5–(-0.5)°)$, $(121°–126°, 0°–4°)$, $(119°–128°, 4°–6°)$, $(111°–119°, 6°–25°)$, $(111.5°–117.5°, 25°–30°)$, $(110°–116°, 32°–35°)$, $(246°–251°, 8.5°, 13.5°)$, $(255°–270°, 20°–40°)$, $(268°–271°, 46°–49°)$, $(232°–240°, 26°–30°)$.

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019

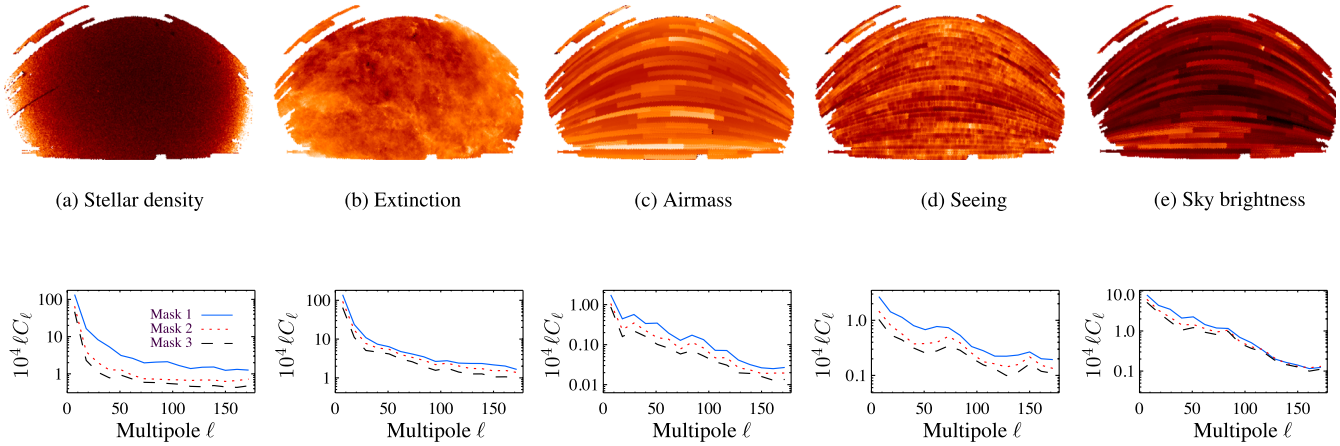


Figure 11. Systematics templates used in this analysis, and the (dimensionless) angular power spectra $\tilde{\mathcal{C}}_\ell$ of their overdensity maps.

Table 2. Systematics thresholds used to restrict the power spectrum analysis of RQCat to the most reliable regions of the sky to minimize contamination from calibration errors. The maps of the systematics are shown in Fig. 11, and the resulting masks in Fig. 12.

Systematic (unit)	Mask 1	Mask 2	Mask 3
Seeing (arcsec)	2.0	1.6	1.55
Reddening (mag)	0.05	0.05	0.045
Stellar density (stars/deg^2)	562	400	350
Airmass (mag)	1.4	1.3	1.25
Sky brightness (nmgy/arcsec^2)	2×10^{-9}	1.8×10^{-9}	1.75×10^{-9}



Figure 12. Masks used for the power spectrum analysis of RQCat, in equatorial coordinates. The retained regions are based on thresholds summarized in Table 2 and the systematics templates of Fig. 11. Additional excised rectangles follow Pullen & Hirata (2012). The three masks respectively have $f_{\rm sky} = 0.148, 0.121and0.101$.

The cut-sky angular power spectra, $\tilde{\mathcal{C}}_\ell$, of the systematics maps for the three masks are shown in Fig. 11. Interestingly, stellar density and dust absorption templates display strong large-scale power ($\ell < 30$), and calibration error templates (seeing, airmass, sky brightness) have notable features at $\ell \sim 70$, 110 and 150. The masked template maps have these features reduced, but not eliminated. Since the data are known to be affected by these systematics, the measured spectra are likely to be contaminated at these multipoles.

3.5 Power spectrum results

We obtained angular band-power estimates with the QML estimator and multipole bins of size $\Delta\ell = 11$, which led to a good balance in terms of multipole resolution and variance of the estimates. We did not use the PCL estimator for the final results because the geometry of the second and third masks, in addition to the presence of systematics, yielded significantly sub-optimal estimates. To illustrate this point, Fig. 13 shows a comparison of the PCL and QML covariance matrices and the band-power estimates of the Mid+High-z subsample for the three masks. Any significant increase of the PCL variance compared to that of QML, especially on diagonal and nearly diagonal elements which contain the most significant contributions, demonstrates the sub-optimality of the PCL prior. For the first mask, the PCL variance of these elements is at most ∼20 per cent greater than the QML variance, indicating that the resulting estimates are nearly optimal. However, for the second and third masks, these elements have a PCL variance up to ∼50 per cent greater than that of QML, and the resulting PCL estimates significantly differ from the optimal QML estimates, as shown in the bottom panel of Fig. 13. This effect is less pronounced for larger multipole bins (e.g. $\Delta\ell = 31$), as the likelihood becomes less sensitive to the priors on the pixel–pixel covariance matrix. However, the resulting loss of resolution prevents the study of localized multipole ranges affected by systematics. For these reasons, we opted for the QML estimator with $\Delta\ell = 11$ in the final analysis. We systematically marginalized over the values of the monopole and the dipole by projecting them out. They are poorly constrained from cut-sky data, and may affect the power spectrum estimates over a wide range of multipoles when deconvoling the cut-sky power spectrum into full sky estimates.[9] We used the values $\bar{G}^{-1} = 1.95 \times 10^{-5}$, 1.55×10^{-5}, 1.85×10^{-5} and 8.15×10^{-6}, respectively for the shot noise of the four RQCat sub-samples, calculated from the average number count per steradian assuming 5 per cent stellar contamination.

The auto and cross-spectra of the four RQCat samples are presented in Figs 14 and 15, and the χ^2 values of the theory prediction are listed in Table 3. We subtracted the shot noise from the autospectra, and used a constant bias, $b_{\rm g} = 2.3$, following previous studies of these data (Giannantonio et al. 2006, 2008; Slosar et al. 2008; Xia et al. 2010; Pullen & Hirata 2012). The theory predictions are summarized in Fig. 10. We also used the exact window functions $W_{b\ell}$ for converting the theory power spectra into band-powers; see equation (17). Fig. 16 shows the cross-correlation power spectra of the quasar samples with the systematics templates, and Table 4 lists the corresponding χ^2 values. Details of the χ^2 computation are contained in Appendix C.

[9] In standard $P(k)$ analyses, this issue is resolved by applying an integral constraint to the power spectrum estimates (see, e.g. Tegmark et al. 2002b).

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019

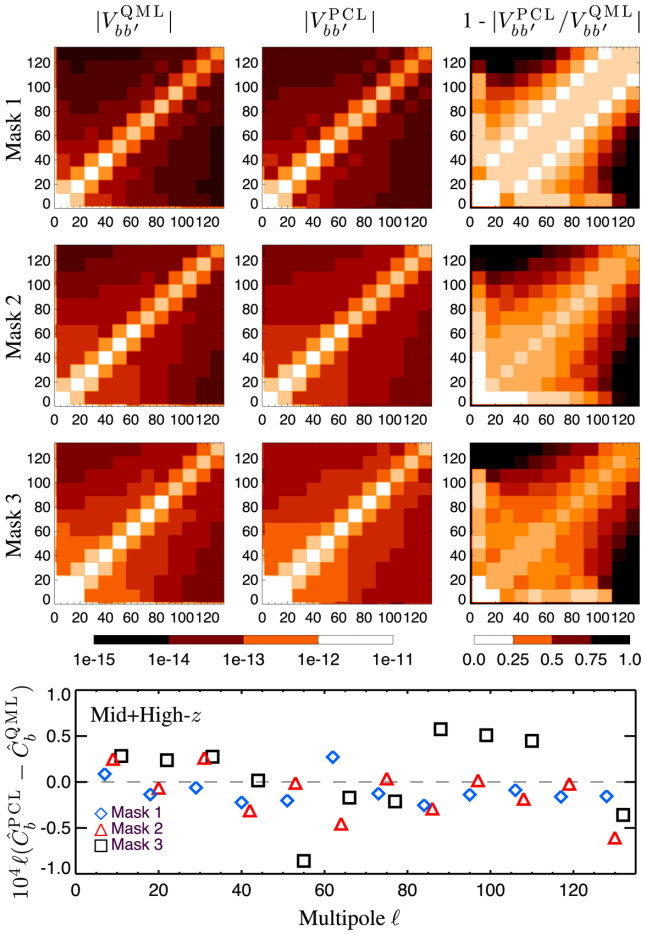


Figure 13. Illustration of the sub-optimality of the PCL estimator in the case of the Mid+High-z sub-sample. The top panels show the covariance matrices of the PCL and QML estimates with $\Delta\ell = 11$ for the three masks of Fig. 12. The sub-optimality of the PCL prior is measured by the fractional increase of variance compared to QML, shown in the right-hand column. The bottom panel shows the resulting effects on the power spectrum estimates, which are more pronounced for the second and third masks due to their complex geometry.

In Figs 14 and 15, the top panels show the final band-power estimates, where the pixel space modes corresponding to the five systematics templates were projected out. The effect of mode projection on the estimates is illustrated in the bottom panels, showing the differences in the QML estimates. Hence, these values can be added to the estimates in the top panels to recover the results without mode projection. The change in the covariance of the estimates due to mode projection is negligible.

3.5.1 Reference mask

Our first mask, which is similar to that used in previous studies of RQCat (Giannantonio et al. 2006, 2008; Slosar et al. 2008; Xia et al. 2010; Pullen & Hirata 2012), is mostly based on extinction, stellar density and seeing cuts, and also excises a few pixels with extreme values of airmass and sky brightness. When using this reference mask, the autospectrum estimates of the four RQCat sub-samples exhibit significant excess power in the first multipole bin. In particular, the cross-correlation of the Low-z sample with the other samples confirm the presence of systematics in common. The cross-spectra of the quasar sub-samples with the systematics templates, shown in Fig. 16, enable us to identify the main sources of contamination responsible for this excess power. In addition to seeing and airmass, which are the main contaminants in the four samples, stellar contamination affects the Low-z sample, and dust extinction and sky brightness contaminate the Mid-z and High-z samples.

The auto- and cross-spectra are marginally improved by projecting out the modes corresponding to the systematics templates, as shown by the small decrease in the χ^2 values, summarized in Tables 3 and 4. In particular, the large-scale power excess persists, confirming the conclusions by Pullen & Hirata (2012) that the contamination must involve non-linear combinations of systematics, or else systematics which have not been accounted for.

3.5.2 Improved masks

Our second mask is based on more restrictive cuts on the systematics, the most important of which are seeing and stellar density cuts. Using this mask not only improves the overall quality of the estimates, as measured by the χ^2, but also eliminates the excess power at low ℓ in all sub-samples except the Low-z one. Interestingly, the cross-spectra of the Low-z sample with the others exhibit no excess power. This indicates that the systematics responsible for the excess in the Low-z sample are successfully mitigated in the Mid-z and High-z samples, and thus in the Mid+High-z sample. The cross-spectra with the systematics templates are significantly decreased, although dust extinction and seeing still affect the Mid-z and High-z samples. Mode-projection further improves the quality of the estimates, but does not eliminate the excess power in the Low-z sample.

The third mask is based even more stringent cuts on the systematics. The cross-spectra with the systematics templates show that using this mask further decreases the influence of extinction, airmass and seeing on the Mid-z and High-z samples. However, it fails to remove the excess power in the Low-z sample, and dust extinction continues to impact the Mid-z sample. Yet, no statistical anomalies are observed in the autospectra of Mid-z, High-z and Mid+High-z samples. Mode-projection further improves the χ^2 values of all auto and cross-spectra. In particular, the χ^2 values for the cross-spectra between the Mid-z and the High-z samples significantly improve, indicating a successful mitigation of the remaining levels of extinction.

In summary, when using the third mask and mode projection, the autospectra of the Mid-z, the High-z and Mid+High-z samples exhibit no evidence of systematics, and are well-fitted by the theory prediction. The cross-spectra of these samples with the Low-z sample are also not anomalous, indicating the absence of systematics in common. Mode-projection eliminates the contributions (to linear order) of the five systematics we have considered, so that any remaining contamination can only involve non-linear combinations or unidentified systematics. Their presence is confirmed in the Low-z sub-sample, but the other sub-samples exhibit no evidence of contamination, indicating that any residual systematics are well within the sample variance.

4 DISCUSSION

We have investigated the problem of estimating angular auto and cross-power spectra on the largest scales in the presence of systematics, and applied this framework to the UVX sources in the RQCat catalogue of SDSS photometric quasars. Previous studies (Xia et al.

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019

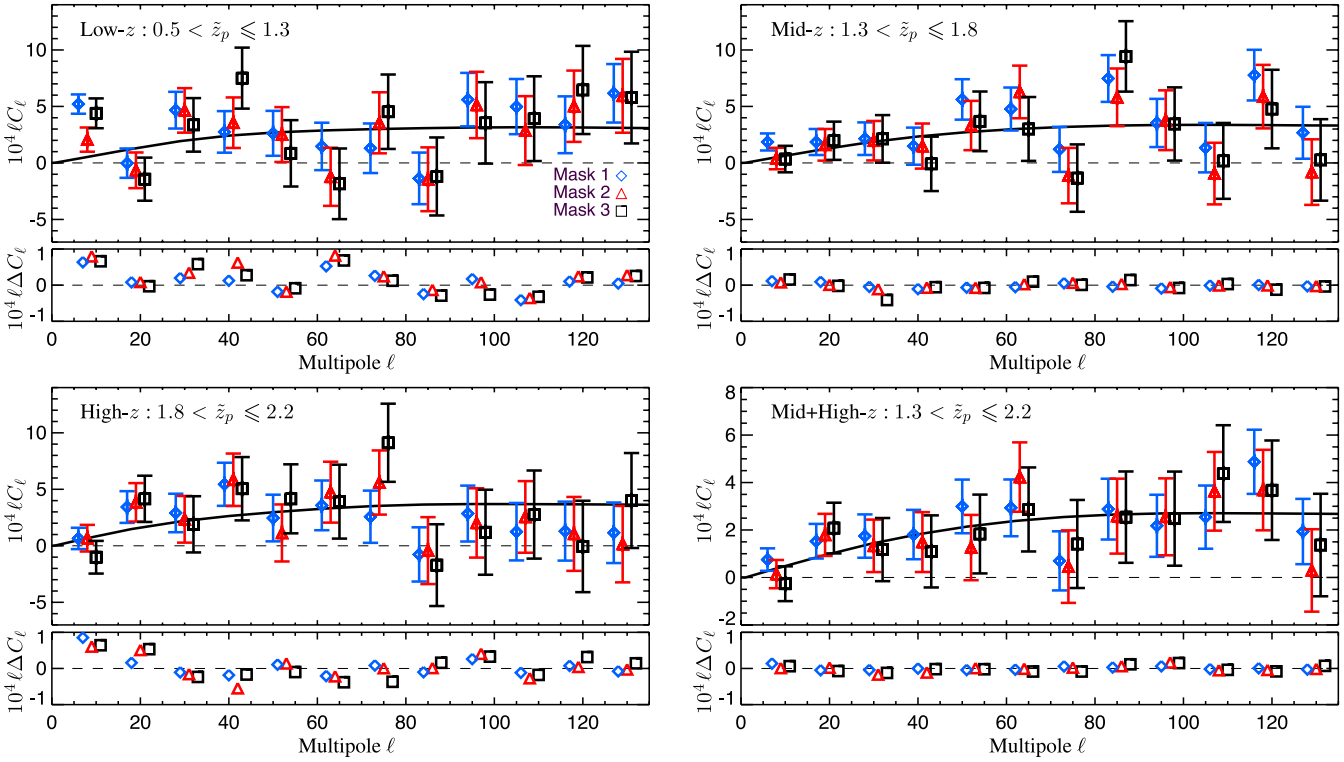


Figure 14. QML estimates of the (dimensionless) autopower spectra of the overdensity maps of the four RQCat samples presented in Fig. 9. The estimates were calculated for the three masks in Fig. 12 (blue diamond, red triangles and black squares) for multipole bins of size $\Delta\ell = 11$, and compared with the theory prediction for a *Planck* cosmology (solid line). The top panel shows the estimates obtaining by mode-projecting the five systematics of Fig. 11. The bottom panel indicates the difference in the estimates when mode projection is not used, i.e. $\hat{C}_b^{\rm no\ mp} - \hat{C}_b^{\rm mp}$, and can be added to the top panel to recover the estimates without mode projection (the change in the covariance of the estimates due to mode projection is negligible).

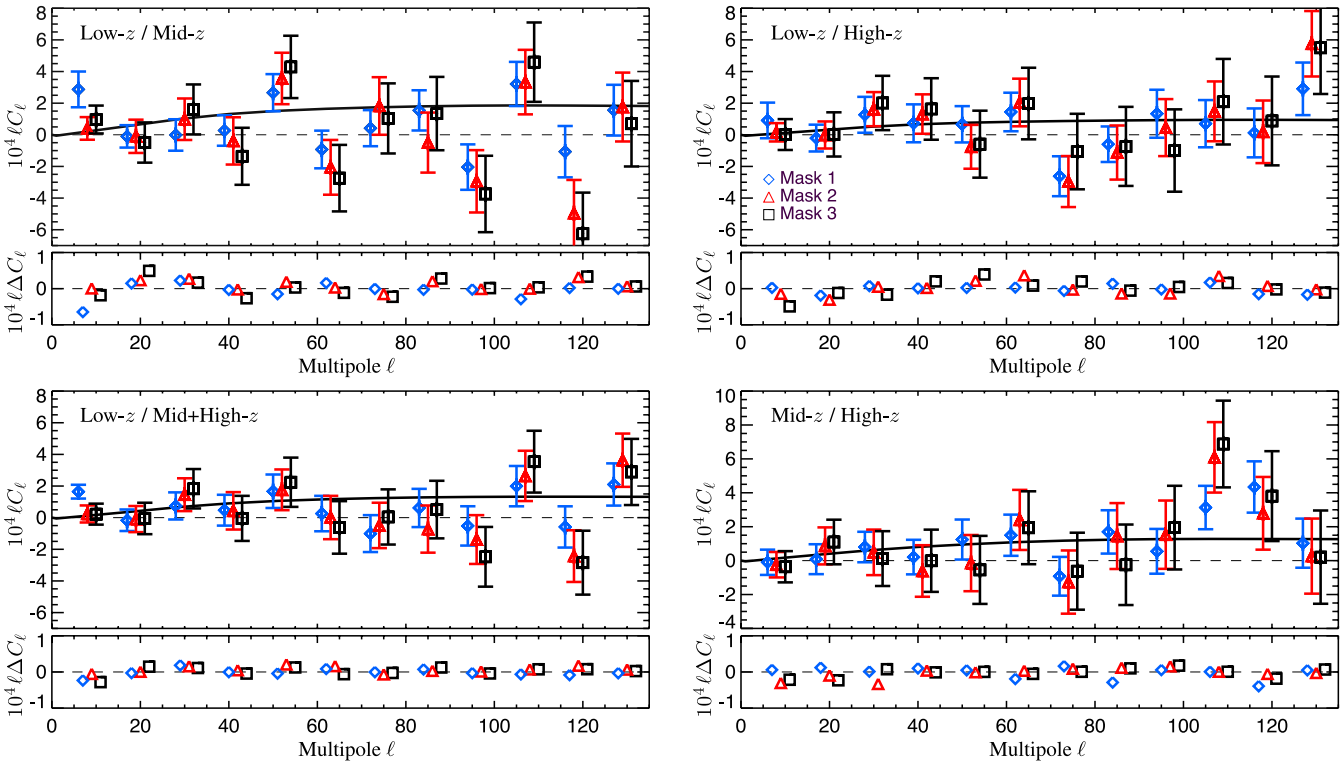


Figure 15. QML estimates of the (dimensionless) cross-power spectra of the RQCat overdensity maps using the same conventions as Fig. 14.

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019

Table 3. The chi square values for the auto and cross-power spectra of the four RQCat samples presented in Figs 14 and 15, with and without mode projection (mp). The number of degrees of freedom is $\nu - p = 13$, and the probability to exceed (PTE) the observed chi squares are shown in parentheses. The results are shown in bold when PTE <1 per cent (corresponding to $\chi^2_{13} = 27.7$).

$\chi^2_{\nu-p}$	Mask 1		Mask 2		Mask 3	
	No mp	mp	No mp	mp	No mp	mp
Low-z	**56.60** (2e-7)	**44.40** (3e-5)	16.43 (0.23)	12.40 (0.49)	25.60 (0.02)	20.24 (0.09)
Mid-z	23.90 (0.03)	22.99 (0.04)	11.20 (0.59)	11.18 (0.60)	8.86 (0.78)	8.72 (0.79)
High-z	16.47 (0.22)	14.46 (0.34)	12.98 (0.45)	11.84 (0.54)	11.05 (0.61)	11.51 (0.57)
Mid+High-z	13.67 (0.40)	13.11 (0.44)	8.25 (0.83)	8.30 (0.82)	4.24 (0.99)	4.54 (0.98)
Low-z / Mid-z	**33.26** (2e-4)	23.76 (5e-3)	**27.72** (1e-4)	15.82 (0.07)	**27.93** (1e-4)	15.35 (0.08)
Low-z / High-z	10.22 (0.68)	11.54 (0.24)	12.13 (0.52)	7.34 (0.60)	6.20 (0.94)	3.52 (0.94)
Low-z / Mid+High-z	26.30 (0.02)	19.97 (0.02)	17.90 (0.16)	9.76 (0.37)	13.94 (0.38)	8.36 (0.50)
Mid-z / High-z	6.18 (0.94)	4.33 (0.89)	11.05 (0.61)	3.92 (0.92)	9.27 (0.75)	2.78 (0.97)

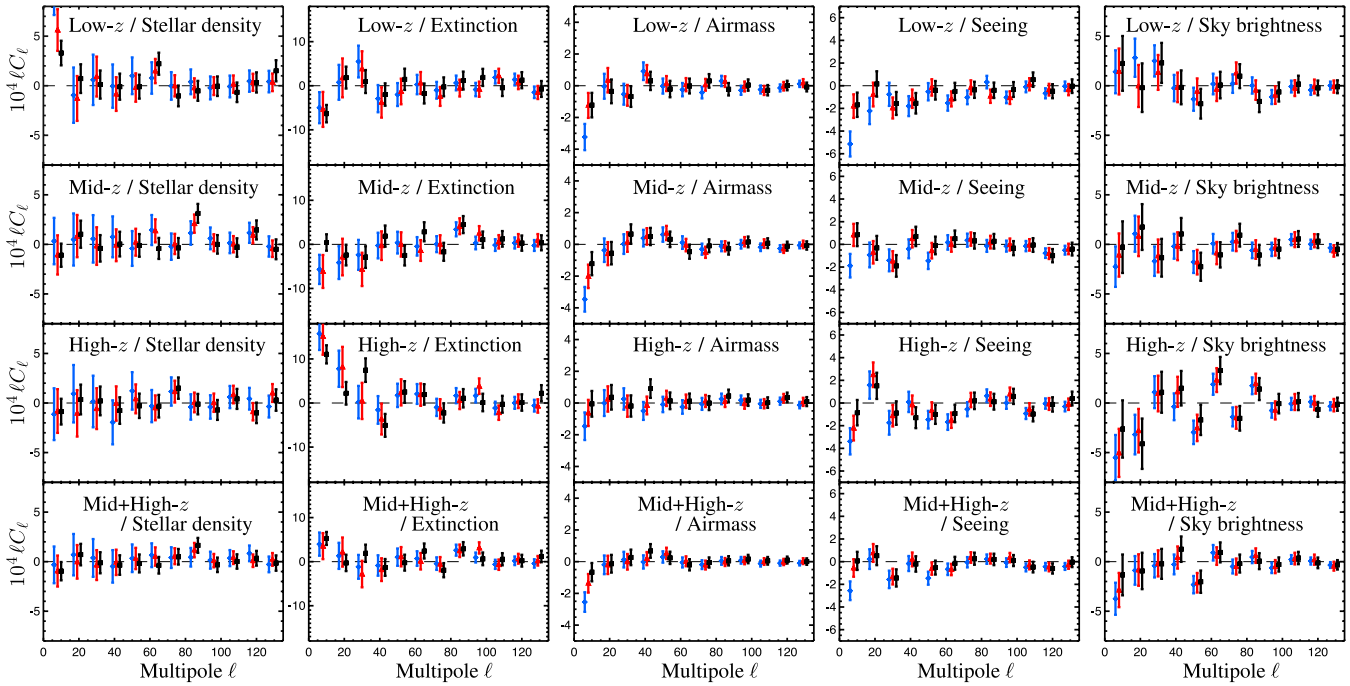


Figure 16. QML estimates for the (dimensionless) cross-power spectra of the four RQCat overdensity maps with the systematics templates using the same conventions as Fig. 14.

Table 4. The chi square values for the cross-power spectra of the four RQCat samples with the systematics templates, presented in Fig. 16, using the same conventions as Table 3.

$\chi^2_{\nu-p}$	Stellar density Masks 1 / 2 / 3	Extinction Masks 1 / 2 / 3	Airmass Masks 1 / 2 / 3	Seeing Masks 1 / 2 / 3	Sky brightness Masks 1 / 2 / 3
Low-z	17.74 / 11.25 / 15.54	10.58 / 9.14 / 14.39	23.97 / 8.24 / 6.79	**50.85** / 25.29 / 12.51	11.31 / 4.16 / 6.99
Mid-z	4.06 / 9.54 / 15.35	12.39 / 17.72 / 13.19	26.97 / 16.14 / 8.21	19.90 / 11.17 / 11.37	8.45 / 6.41 / 7.73
High-z	3.35 / 5.75 / 6.55	**28.33** / 26.91 / **47.14**	5.87 / 3.08 / 6.82	**32.06** / 20.14 / 12.18	**27.98** / 18.64 / 15.31
Mid+High	2.88 / 5.33 / 7.77	11.26 / 14.33 / 23.17	20.26 / 9.11 / 6.15	**34.75** / 14.88 / 9.63	19.80 / 10.84 / 6.65

2011; Pullen & Hirata 2012; Giannantonio et al. 2013) indicated that this catalogue was not suitable for clustering analyses due to the high levels of contamination by the systematics. We examined these conclusions by focusing on $0.5 < \tilde{z}_p \leq 2.2$ objects, divided into four redshift bins, and attempted to remove the influence of the main systematics using improved sky masks and mode projection. We also improved the theoretical predictions by making use of refined redshift distributions estimated by cross-matching objects in RQCat with the SDSS-DR7 spectroscopic catalogue and applying robust completeness corrections.

In agreement with previous studies, we found that $\tilde{z}_p < 1.3$ objects exhibited significant levels of contamination by systematics, in particular dust absorption, airmass and seeing, which could not be eliminated by masking and mode projection. The remaining excess power on large scales points to the presence of unknown or non-linear combinations of the systematics in this redshift bin.

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019

The large-scale excess power observed in the auto and cross-spectra of $1.3 < \tilde{z}_p \leq 2.2$ objects was eliminated by using improved sky masks based on templates of five of the main systematics. The agreement with the theory predictions further improved when projecting out the modes corresponding to these systematics. Within the statistical uncertainties, we found no evidence for remaining contamination in this sample. We conclude that photometric quasar samples can be made suitable for cosmological studies.

We did not attempt to model the contamination signal, but rather constructed a sample (through object selection and masking) which exhibited negligible levels of contamination. This approach relies on the ability to measure these levels of contamination, in our case through auto and cross-spectra of the samples and cross-spectra with systematics templates, which are limited by the variance of the estimates due to the shot noise and sky coverage. Consequently, more contaminated samples (such as the whole RQCat) with higher number densities require a model of the systematics in order to obtain clustering measurements that are not dominated by spurious correlations. In this context, mode projection can be used to marginalize over the parameters of linear contamination models while estimating the power spectrum. However, future surveys may require more sophisticated algorithms, or even Bayesian component separation models similar to COMMANDER-RULER (Eriksen et al. 2006, 2008) used in the context of *Planck* (Planck Collaboration 2013a).

The clustering properties of the large galaxy and quasar catalogues produced by next generation photometric surveys such as DES will put tight constraints on models of galaxy bias and PNG. However, stellar contamination and calibration errors will always be present in galaxy survey data, and may compromise our understanding of the observables if not correctly treated. Thus, as sample variance becomes steadily smaller with increasing catalogue sizes, efficient strategies for mitigating the systematics, such as those presented here, will become critical for the cosmological interpretation of these surveys.

ACKNOWLEDGEMENTS

We thank Stephen Feeney, Filipe Abdalla and Ofer Lahav for useful discussions and comments. BL is supported by the Perren Fund and the IMPACT Fund. HVP is supported by STFC, the Leverhulme Trust and the European Research Council under the European Community's Seventh Framework Programme (FP7/2007- 2013) / ERC grant agreement no 306478-CosmicDawn. ABL is supported by the Leverhulme Trust and STFC. AP is supported by the Oxford Martin School. We acknowledge use of the following public software packages: HEALPIX (Górski et al. 2005); MANGLE (Hamilton & Tegmark 2004; Swanson et al. 2008) and CAMB_SOURCES (Challinor & Lewis 2011). We acknowledge use of the Legacy Archive for Microwave Background Data Analysis (LAMBDA). Support for LAMBDA is provided by the NASA Office of Space Science.

REFERENCES

Adelman-McCarthy J. K. et al., 2008, ApJS, 175, 297
Albrecht A., Steinhardt P. J., 1982, Phys. Rev. Lett., 48, 1220
Aurich R., Lustig S., 2011, MNRAS, 411, 124
Bennett C. L. et al., 2012, preprint (arXiv:1212.5225)
Bond J. R., Jaffe A. H., Knox L., 1998, Phys. Rev. D, 57, 2117
Bond J. R., Jaffe A. H., Knox L., 2000, ApJ, 533, 19
Bovy J. et al., 2011, ApJ, 729, 141
Brown M. L., Castro P. G., Taylor A. N., 2005, MNRAS, 360, 1262
Challinor A., Lewis A., 2011, Phys. Rev. D, 84, 043516
Copi C. J., Huterer D., Schwarz D. J., Starkman G. D., 2011, preprint (arXiv:1103.3505)
Corasaniti P.-S., Giannantonio T., Melchiorri A., 2005, Phys. Rev. D, 71, 123521
Crocce M., Pueblas S., Scoccimarro R., 2006, MNRAS, 373, 369
Croom S. M. et al., 2009, MNRAS, 392, 19
Dalal N., Doré O., Huterer D., Shirokov A., 2008, Phys. Rev. D, 77, 123514
Efstathiou G., 2004a, MNRAS, 348, 885
Efstathiou G., 2004b, MNRAS, 349, 603
Efstathiou G., 2006, MNRAS, 370, 343
Eisenstein D. J. et al., 2011, ApJ, 142, 72
Eriksen H. K. et al., 2006, ApJ, 641, 665
Eriksen H. K. et al., 2007, ApJ, 656, 641
Eriksen H. K., Jewell J. B., Dickinson C., Banday A. J., Górski K. M., Lawrence C. R., 2008, ApJ, 676, 10
Feeney S. M., Peiris H. V., Pontzen A., 2011, Phys. Rev. D, 84, 103002
Giannantonio T. et al., 2006, Phys. Rev. D, 74, 063520
Giannantonio T., Scranton R., Crittenden R. G., Nichol R. C., Boughn S. P., Myers A. D., Richards G. T., 2008, Phys. Rev. D, 77, 123520
Giannantonio T., Crittenden R., Nichol R., Ross A. J., 2012, MNRAS, 426, 2581, available at: http://mnras.oxfordjournals.org/content/426/3/2581.full.pdf+html
Giannantonio T., Ross A. J., Percival W. J., Crittenden R., Bacher D., Kilbinger M., Nichol R., Weller J., 2013, preprint (arXiv:1303.1349)
Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759
Gruetjen H. F., Shellard E. P. S., 2012, preprint (arXiv:1212.6945)
Gunn J. E. et al., 1998, ApJ, 116, 3040
Gunn J. E. et al., 2006, ApJ, 131, 2332
Guth A. H., 1981, Phys. Rev. D, 23, 347
Hamilton A. J. S., Tegmark M., 2004, MNRAS, 349, 115
Hamimeche S., Lewis A., 2009, Phys. Rev. D, 79, 083012
Hinshaw G. et al., 2012, preprint (arXiv:1212.5226)
Hivon E., Górski K. M., Netterfield C. B., Crill B. P., Prunet S., Hansen F., 2002, ApJ, 567, 2
Ho S., Hirata C., Padmanabhan N., Seljak U., Bahcall N., 2008, Phys. Rev. D, 78, 043519
Ho S. et al., 2012, ApJ, 761, 14
Huterer D., Knox L., Nichol R. C., 2001, ApJ, 555, 547
Huterer D., Cunha C. E., Fang W., 2013, MNRAS, 432, 2945
Jaffe A. H., Bond J. R., Ferreira P. G., Knox L. E., 1999 in Maiani L., Melchiorri F., Vittorio N., eds, Proc. AIP Conf. Ser. Vol. 476, Conference on 3K cosmology. Am. Inst. Phys., New York, p. 249
Knox L., Bond J. R., Jaffe A. H., 1998 in Olinto A. V., Frieman J. A., Schramm D. N., eds, Eighteenth Texas Symposium on Relativistic Astrophysics. World Scientific, Singapore, p. 282
Linde A., 1982, Phys. Lett. B, 108, 389
Manera M. et al., 2013, MNRAS, 428, 1036
Matarrese S., Verde L., 2008, ApJ, 677, L77
Myers A. D. et al., 2006, ApJ, 638, 622
Myers A. D., Brunner R. J., Nichol R. C., Richards G. T., Schneider D. P., Bahcall N. A., 2007, ApJ, 658, 85
Padmanabhan N., Hirata C. M., Seljak U., Schlegel D. J., Brinkmann J., Schneider D. P., 2005, Phys. Rev. D, 72, 043525
Pâris I. et al., 2012, A&A, 548, A66
Peek J. E. G., Graves G. J., 2010, ApJ, 719, 415
Planck Collaboration 2013a, preprint (arXiv:1303.5072)
Planck Collaboration 2013b, preprint (arXiv:1303.5076)
Pontzen A., Peiris H. V., 2010, Phys. Rev. D, 81, 103008
Pullen A. R., Hirata C. M., 2012, preprint (arXiv:1212.4500)
Richards G. T. et al., 2009, ApJS, 180, 67
Ross A. J. et al., 2011, MNRAS, 417, 1350
Ross N. P. et al., 2012a, ApJS, 199, 3
Ross A. J. et al., 2012b, MNRAS, 424, 564
Ross A. J. et al., 2013, MNRAS, 428, 1116
Samushia L. et al., 2013, MNRAS, 429, 1514
Sánchez A. G. et al., 2012, MNRAS, 425, 415
Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019

Schneider D. P. et al., 2010, ApJ, 139, 2360
Scoccimarro R., 1997, preprint (astro-ph/9711187)
Sherwin B. D. et al., 2012, Phys. Rev. D, 86, 083006
Slosar A., Seljak U., Makarov A., 2004, Phys. Rev. D, 69, 123003
Slosar A., Hirata C., Seljak U., Ho S., Padmanabhan N., 2008, J. Cosmol. Astropart. Phys., 8, 31
Swanson M. E. C., Tegmark M., Hamilton A. J. S., Hill J. C., 2008, MNRAS, 387, 1391
Tegmark M., 1997, Phys. Rev. D, 55, 5895
Tegmark M., de Oliveira-Costa A., 2001, Phys. Rev. D, 64, 063001
Tegmark M., Taylor A. N., Heavens A. F., 1997, ApJ, 480, 22
Tegmark M., Hamilton A. J. S., Strauss M. A., Vogeley M. S., Szalay A. S., 1998, ApJ, 499, 555
Tegmark M., Hamilton A. J. S., Xu Y., 2002a, MNRAS, 335, 887
Tegmark M. et al., 2002b, ApJ, 571, 191
Thomas S. A., Abdalla F. B., Lahav O., 2010, MNRAS, in press (arXiv:1012.2272)
Thomas S. A., Abdalla F. B., Lahav O., 2011, Phys. Rev. Lett., 106, 241301
Vogeley M. S., Szalay A. S., 1996, ApJ, 465, 34
Wandelt B. D., Hivon E., Górski K. M., 2001, Phys. Rev. D, 64, 083003
Xia J.-Q., Viel M., Baccigalupi C., Matarrese S., 2009, J. Cosmol. Astropart. Phys., 9, 3
Xia J.-Q., Bonaldi A., Baccigalupi C., De Zotti G., Matarrese S., Verde L., Viel M., 2010, J. Cosmol. Astropart. Phys., 8, 13
Xia J.-Q., Baccigalupi C., Matarrese S., Verde L., Viel M., 2011, J. Cosmol. Astropart. Phy.s, 8, 33.

APPENDIX A: PIXELIZATION, BAND-LIMIT AND SMOOTHING ISSUES

The equations defining the PCL and QML estimators do not impose maximum multipoles for the reconstruction, the coupling or the covariance matrices. Such bounds are called band-limits, and are in practice imposed by the finite information content of $\boldsymbol{x}$, which depends on the resolution of the map and any additional operations such as smoothing. In particular, pixelizing the signal x into a (full or cut-sky) map $\boldsymbol{x}$ induces a distortion in the power spectrum estimates, which we parametrize as

$$\mathcal{C}_\ell^{\rm pix} = b_\ell^2 \mathcal{C}_\ell, \tag{A1}$$

where $\mathcal{C}_\ell^{\rm pix}$ and $\mathcal{C}_\ell$ are the power spectra of $\boldsymbol{x}$ and x, respectively. The continuous map x is usually not accessible, but can be approximated using a high-resolution pixelization, which in practice will correspond to the highest resolution at which the data are available (and will only require a small correction of the pixelization-induced bias as detailed below). $\boldsymbol{x}$ will then refer to a smoothed, lower resolution map constructed from x, which will be used to estimate the power spectrum on a specific range of scales. This approach is motivated by the complexity of the PCL and QML pixel–space estimators which depends on the number of pixels in the mask. It is usually desirable to use the lowest resolution for which the power spectrum can be accurately estimated in the range of multipoles of interest.

The beam b_ℓ is decomposed into a pixelization-induced part $b_\ell^{\rm pix}$ and a smoothing part $b_\ell^{\rm pix}$. In the previous sections, we have assumed that pixelization distortions were negligible, i.e. $b_\ell^{\rm pix} \approx 1$ for all ℓ, but this assertion is true for a mode ℓ only if the pixels of $\boldsymbol{x}$ are small compared with $180/\ell^\circ$. We considered the HEALPIX pixelization of the sphere, where a map at resolution $N_{\rm side}$ has $12N_{\rm side}^2$ equal-area pixels. In terms of its effect on the power spectrum, pixelizing a signal at resolution $N_{\rm side}$ is well approximated by smoothing with a Gaussian kernel of full width at half-maximum (FWHM) $41.7/N_{\rm side}{}^\circ$. This beam, shown in Fig. A1, smoothly decays as ℓ increases, and imposes an effective band-limit of $\ell_{\rm max} = 7N_{\rm side}$ on the pixelized map. However, it is well known that the accessible multipoles for a map at HEALPIX resolution $N_{\rm side}$ lie within $\ell \in [0, 2N_{\rm side}]$. Higher multipoles are not accessible because the integrals in the spherical harmonics and Legendre transforms are not accurately approximated by matrix multiplications at this resolution. It is essential to smooth the initial map before degrading it in order to avoid a mismatch between the band-limits, while insuring that the power spectrum can be reconstructed up to $\ell_{\rm max} = 2N_{\rm side}$. We investigated this issue and found that band-limiting the map at $\ell_{\rm max} = 4N_{\rm side}$ gives optimal performance. This can be realized by a Gaussian smoothing of FWHM $60/N_{\rm side}{}^\circ$, as illustrated in Fig. A1.

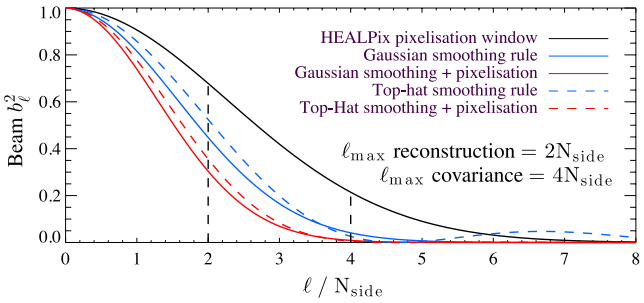


Figure A1. Beams that apply to the power spectrum estimates of pixelized, smoothed maps at HEALPIX resolution $N_{\rm side}$. The beam must be accounted for in the construction of the model covariance matrices and inverted when comparing the estimates with theory.

However, the smoothing procedure introduces information from outside the mask into the map, hence biasing the estimates (Aurich & Lustig 2011; Copi et al. 2011; Feeney, Peiris & Pontzen 2011). This effect depends on the resolution of the map and the shape of the mask, and can significantly affect the power spectrum on a wide range of scales. To avoid this smoothing-induced contamination, the mask must be extended. A systematic procedure is to smooth the complement of the mask, i.e. $1 - \boldsymbol{m}$, and keep the pixels below a certain threshold (0.01 in this work). But because the Gaussian kernel is band-limited in harmonic space but not compact in pixel space, the smoothing will always leak contamination signal into the data. Also, depending on the threshold, the mask extension might be large and thus significantly increase the variance of the estimates. This can be resolved by using a top-hat smoothing of diameter $90/N_{\rm side}{}^\circ$, which band-limits the data at $\ell_{\rm max} = 4N_{\rm side}$ similar to the Gaussian smoothing. The top-hat kernel is compact and the mask extension is smaller and requires no threshold. However, its extent in harmonic space is infinite (in other words it is not band-limited), as shown in Fig. A1, which can introduce further contamination in the map due to approximated smoothing. This can be avoided by performing the top-hat smoothing at high resolution with a large band-limit. Alternatively, the smoothing can easily be performed in pixel space through explicit convolution since the kernel has a simple shape and the procedure can be parallelized.

In conclusion, the power spectrum of a high-resolution signal can be estimated from a map at lower resolution $N_{\rm side}$ provided that the band-limits, the pixelization and smoothing-induced biases are correctly handled. Before degradation, the initial map is smoothed with a Gaussian kernel of FWHM $60/N_{\rm side}{}^\circ$ or a top-hat kernel of diameter $90/N_{\rm side}{}^\circ$, and the mask must be extended accordingly to minimize the contamination due to smoothing. The estimation can then be performed on the low-resolution map in the range $\ell \in [0, 4N_{\rm side}]$ and the signal covariance matrix created with a prior $b_\ell^2 C_\ell$ up to $\ell_{\rm max} = 4N_{\rm side}$ in order to incorporate all the information in the data. Dividing by the beam b_ℓ^2 leads to accurate unbiased band-power estimates of C_ℓ in the range $\ell \in [0, 2N_{\rm side}]$.

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019

APPENDIX B: KARHUNEN–LOÈVE COMPRESSION DEMYSTIFIED

The Karhunen–Loève (KL) transform (Vogeley & Szalay 1996; Tegmark et al. 1997, 1998, 2002a) refers to finding a basis in which the transformed data pixels are statistically orthogonal with respect to a prior on the power spectrum. The transformation matrix $\mathbf{B}$ satisfies

$$\tilde{\boldsymbol{x}} = \mathbf{B}\tilde{\boldsymbol{y}}, \tag{B1}$$

such that

$$\langle \tilde{y}_n \tilde{y}^*_{n'} \rangle = \lambda_n \delta_{nn'}, \tag{B2}$$

and the columns $\boldsymbol{B}_n$ are the eigen-vectors of the pixel–pixel covariance matrix, i.e.

$$\tilde{\mathbf{C}}\boldsymbol{B}_n = \lambda_n \boldsymbol{B}_n. \tag{B3}$$

Hence $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{y}}$ retain the same information and KL compression is equivalent to a principal component analysis performed through singular value decomposition (SVD) of the pixel–pixel covariance matrix. However, this approach is only optimal if the noise covariance is diagonal. Anisotropic noise requires a so-called pre-whitening operation, changing the eigen-problem to be solved into

$$\tilde{\mathbf{S}}\boldsymbol{B}_n = \lambda_n \tilde{\mathbf{N}}\boldsymbol{B}_n, \tag{B4}$$

where the eigen-vectors now diagonalize both the signal and the noise covariance matrices. If the transformation is renormalized such that $\boldsymbol{B}_n^\dagger \mathbf{N} \boldsymbol{B}_n = 1$ before transforming the map $\boldsymbol{x}$, the coefficients λ_n can be interpreted as signal-to-noise ratios, i.e.

$$\langle \tilde{y}_n \tilde{y}^*_{n'} \rangle = \delta_{nn'}(1 + \lambda_n). \tag{B5}$$

The power spectrum is invariant under rotation in the isotropic case, it can be estimated from $\boldsymbol{y}$ provided that the matrices $\tilde{\mathbf{C}}$ and $\mathbf{P}^\ell$ are in the same coordinate system. In the KL basis, the most informative contributions to the spectrum explicitly come from the first KL modes, as they correspond to high signal-to-noise ratio modes. Hence KL compression is often used to remove the noisiest modes and speed up power spectrum estimation, in particular for the matrix inversion in the QML estimator. However, this gain is usually offset by the cost of the preliminary SVD required to compute the KL transformation matrix.

Moreover, although removing the noisiest modes leaves the PCL and QML estimates unbiased, it increases their variance and potentially impacts the estimates on a wide range of scales. In particular, assuming a constant diagonal noise, the KL modes calculated from a full sky covariance matrix **ssS** directly relate to the theory spectrum. The largest modes correspond to the largest C_ℓ's with multiplicity $2\ell + 1$. On the cut sky, this degeneracy is broken and each mode uniquely relates to a linear combination of C_ℓ's with finite support peaking at the previous full sky value (thus conserving the order of the modes). Small masks increase the scope of this combination, and removing small KL modes can thus impact a wide range of multipoles. Logically, considering the CMB spectrum with a typical CMB mask shows that the lowest KL modes, the noisiest, relate to high ℓ multipoles, and removing them leaves the low ℓ power spectrum estimates unchanged. However, for a galaxy survey the theory spectrum is flatter and the mask larger, and the noisiest modes then relate to a wide range of multipoles. Therefore, removing them also impacts the largest scales. Fortunately this effect is reduced when performing the estimation in bins, and the KL compression has a negligible impact on the quality and the variance estimates.

We did not use KL compression in the context of our analyses because the gain in computer time is marginal. We were able to run the QML estimator at $N_{\rm side} = 64$ without any approximation.

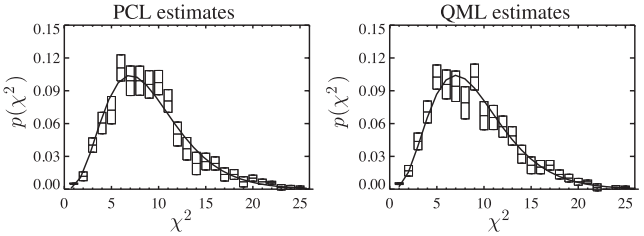


Figure C1. Histograms of the χ^2 values of the PCL (left) and QML (right) estimates of the 600 CMASS mock catalogues. We used multipole bins of size $\Delta\ell = 11$, and the theory prediction from CAMB SOURCES. The boxes show the Poisson errors due to sample variance. The histograms were normalized and compared with the theoretical χ^2 probability distributions (thick lines). Their good agreement demonstrates the validity of the χ^2 given by equation 39.

APPENDIX C: χ^2 FOR BAND-POWER ESTIMATES

To compare power spectrum measurements with theory predictions, we used a χ^2 defined as

$$\chi^2 = \sum_{bb'} (\hat{C}_b - C_b)\mathbf{V}^{-1}_{bb'}(\hat{C}_{b'} - C_{b'}) \tag{C1}$$

where $\hat{C}_b$ denotes the (PCL or QML) band-power estimates and $C_{b'}$ the theory band-powers, constructed from the theory power spectrum C_ℓ (calculated from CAMB_SOURCES) using the window functions $W_{b\ell}$ defined in equations (17) and (18). The covariance matrix $\mathbf{V}$ is calculated using equation (11) and is equal to the inverse of the Fisher matrix when the QML estimator is used. Note that the shot noise $\frac{1}{G}$ must be subtracted from the autospectrum estimates $\hat{C}_b$.

It is well known that observed power spectra $\mathcal{C}_\ell$ as calculated by equation (1) are only described by Gaussian statistics at high-ℓ when the central limit theorem applies. For low-ℓ estimates, one must resort to alternative likelihood functions (see, e.g. Jaffe et al. 1999; Hamimeche & Lewis 2009). However, when estimating band-powers rather than individual multipoles, these effects can be neglected, and a Gaussian likelihood can be used to compare the estimates with theory band-powers. To illustrate this point, Fig. C1 shows the χ^2 values for the PCL and QML estimates of the 600 CMASS mock catalogues with bin size $\Delta\ell = 11$ in the range [2,120], corresponding to 11 degrees of freedom. The normalized histograms follow the theoretical χ^2 distributions, demonstrating the validity of a Gaussian likelihood, as expected when using band-powers.

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

Downloaded from https://academic.oup.com/mnras/article-abstract/435/3/1857/1019739 by Imperial College London Library user on 19 September 2019