Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **503,** 5061–5084 (2021)
Advance Access publication 2021 March 13

doi:10.1093/mnras/stab709

Mitigating contamination in LSS surveys: a comparison of methods

Noah Weaverdyck[1,2,★] and Dragan Huterer[1]

[1] *Department of Physics, University of Michigan, 450 Church St, Ann Arbor, MI 48109-1040, USA*
[2] *Leinweber Center for Theoretical Physics, University of Michigan, 450 Church St, Ann Arbor, MI 48109-1040, USA*

Accepted 2021 March 5. Received 2021 March 3; in original form 2020 September 25

ABSTRACT

Future large-scale structure surveys will measure the locations and shapes of billions of galaxies. The precision of such catalogues will require meticulous treatment of systematic contamination of the observed fields. We compare several existing methods for removing such systematics from galaxy clustering measurements. We show how all the methods, including the popular pseudo-C_ℓ Mode Projection and Template Subtraction methods, can be interpreted under a common regression framework and use this to suggest improved estimators. We show how methods designed to mitigate systematics in the power spectrum can be used to produce clean maps, which are necessary for cosmological analyses beyond the power spectrum, and we extend current methods to treat the next-order multiplicative contamination in observed maps and power spectra, which reduced power spectrum errors from $\Delta\chi^2_{C_\ell} \simeq 10$ to $\simeq 1$ in simulated analyses. Two new mitigation methods are proposed, which incorporate desirable features of current state-of-the-art methods while being simpler to implement. Investigating the performance of all the methods on a common set of simulated measurements from Year 5 of the Dark Energy Survey, we test their robustness to various analysis cases. Our proposed methods produce improved maps and power spectra when compared to current methods, while requiring almost no user tuning. We end with recommendations for systematics mitigation in future surveys, and note that the methods presented are generally applicable beyond the galaxy distribution to any field with spatial systematics.

Key words: methods: data analysis – methods: statistical – surveys – cosmology: observations – large-scale structure of Universe.

1 INTRODUCTION

Over the past 40 yr, cosmological surveys have produced increasingly detailed maps of the large-scale structure (LSS) in the Universe (de Lapparent, Geller & Huchra 1986; Lumsden et al. 1992; York et al. 2000; Colless et al. 2001; Drinkwater et al. 2010; Dawson et al. 2013; de Jong et al. 2015; Abbott et al. 2018; Aihara et al. 2018). These observations have proven crucial for testing our understanding of gravity and cosmological structure formation, and helped to constrain cosmological parameters to the percent level (e.g. Anderson et al. 2014; Alam et al. 2017; Abbott et al. 2018). Recent observations from DES have for the first time imposed strong constraints on dark energy using an LSS survey alone, independently of the cosmic microwave background (Abbott et al. 2019). Upcoming ground-based missions like DESI (DESI Collaboration et al. 2016), and the Rubin Observatory's LSST (Collaboration 2012), along with space-based missions like SPHEREx (Doré et al. 2014), Euclid (Amendola et al. 2018), and RST (formerly WFIRST) (Spergel et al. 2013) will truly herald the age of precision cosmology, mapping up to ∼20 billion galaxies across the sky and bringing unprecedented precision to measurements of the dark energy equation of state and modified gravity. Such statistical precision makes the control of systematic errors in these data sets of paramount importance to avoid biasing cosmological analyses.

Cosmological information is extracted from LSS observations in multiple ways. The most common approach is to calculate the two-point correlation function or its Fourier counterpart, the power spectrum, to characterize the spatial distribution of galaxies (galaxy clustering) or their shapes (weak lensing). To date, these have been used in cosmological analyses to great success (Hauser & Peebles 1973; Peebles & Hauser 1974; Davis & Peebles 1983; Peacock & Nicholson 1991; Saunders, Rowan-Robinson & Lawrence 1992; Baugh & Efstathiou 1993; Fisher et al. 1993a, b; Feldman, Kaiser & Peacock 1994; Baugh 1996; Dodelson & Gaztañaga 2000; Eisenstein & Zaldarriaga 2001; Huterer, Knox & Nichol 2001; Miller & Batuski 2001; Peacock et al. 2001; Percival et al. 2001; Connolly et al. 2002; Dodelson et al. 2002; Tegmark et al. 2004; Blake 2019)

The two-point function contains all available information when the field it characterizes is Gaussian, but non-linear gravitational collapse induces non-Gaussianity at late times and small scales. Therefore there is considerable cosmological information that is inaccessible to the two-point function. This has led to growing interest in using complementary statistical representations of LSS observations, such as higher order N-point functions (Peebles & Grot(Peebles & Groth 1975; Cooray & Hu 2001; Feldman et al. 2001; Feldman et al. 2001; Scoccimarro et al. 2001; Scoccimarro et al. 2001; Verde et al. 2002; Sefusatti et al. 2006; Marín et al. 2013; Gil-Marín et al. 2015; Slepian & Eisenstein 2015; Slepian et al. 2015) et al. 2015; Slepian & Eisenstein 2015; Slepian et al. 2015), statistics of peaks (Jain & Van Waerbeke 2000; Marian et al. 2011; Liu et al. 2015) and voids (White (White 1979; Fry 1986; Biswas, Alizadeh & Wandelt 2010; Bos et al.

[★] E-mail: nweaverd@umich.edu

Published by Oxford University Press on behalf of Royal Astronomical Society

2012; Leclercq et al. 2015; Nadathur & Hotchkiss 2015; Pisani et al. 2015), density-split statistics (Friedrich et al. 2018), marked power spectra (Sheth (Sheth 2005; White & Padmanabhan 2009; White 2016; Philcox, Massara & Spergel 2020), Minkowski functionals (Gott, Weinberg & Me(Gott, Weinberg & Melott 1987; Hikage, Komatsu & Matsubara 2006; Kratochvil et al. 2012; Munshi et al. 2012; Petri et al. 2015; Mawdsley et al. 2020), wavelet transforms (Allys et al. 2020; Cheng et al. 2020) and more. These methods rely on the accurate mapping of the underlying cosmological fields from which they are derived, so there is increasing need for tools to mitigate systematic contamination at the levels of both the map and the two-point functions.

Here, we consider a very general class of systematics that describes an arbitrary spatial modulation of the observed field. Such generic sources of error are one of the most serious contaminants in our quest to probe cosmology with future surveys. For definiteness, we focus on the case of galaxy clustering, where the systematic error corresponds to a modulation of the galaxy selection function in redshift or across the observing footprint. However, the methods we test and develop in this paper are general enough to apply to any real or complex field for which there exist maps of potential contaminants (e.g. shear or Sunyaev-Zeldovich-effect fields).

Spatially varying systematics in LSS maps may be caused by a large variety of physical effects. These include observing conditions and dust extinction (both of which effectively create a position-dependent 'screen,' obscuring background galaxies), bright objects and star–galaxy separation (which can eclipse, change the shape, or be confused for galaxies close to them on the sky), and variations in sensitivity of the detector (which include potentially time- and position-dependent variations in the focal plane), or imaging pipeline. In all of these cases, failure to fully account for variability in the selection function will result in residual artifacts – *calibration errors* – in the final data product and potentially bias results (Huterer, Cunha & Fang 2013; Shafer & Huterer 2015; Weaverdyck, Muir & Huterer 2018). The presence of calibration errors is evidenced by a number of surveys (Vogeley 1998; Scranton et al. 2002; Goto, Szapudi & Granett 2012; Ho et al. 2012, 2013; Pullen & Hirata 2013; Giannantonio et al. 2014; Agarwal, Ho & Shandera 2014a; Agarwal et al. 2014b) which have shown a significant excess of power at large scales where calibration errors are thought to be most prevalent. Recent observations (e.g. from the Dark Energy Survey Leistedt et al. 2016) demonstrate, however, that such contamination is by no means limited to large scales alone. In addition to adding power, calibration errors induce a multiplicative effect, coupling different scales and thus affecting all scales in the survey, including those smaller than the typical size of the calibration systematic itself (Huterer et al. 2013; Shafer & Huterer 2015). Much recent work (Ross et al. 2011; Ho et al. 2012; Leistedt et al. 2013; Pullen & Hirata 2013; Leistedt & Peiris 2014; Agarwal et al. 2014b; Morrison & Hildebrandt 2015; Rykoff, Rozo & Keisler 2015; Awan et al. 2016; Delubac et al. 2016; Kalus et al. 2016; Prakash et al. 2016; Suchyta et al. 2016; Bautista et al. 2018; Awan & Gawiser 2019; Kalus et al. 2019; Kitanidis et al. 2020; Kong et al. 2020; Rezaie et al. 2020; Ross et al. 2020; Wagoner, Rozo & Fang 2020) has focused on mitigating these systematics in order to probe the underlying cosmology.

The simplest strategy to ameliorate the effects of calibration errors is to simply mask scales or data points suspected having large levels of contamination. More sophisticated strategies include using maps of suspected contaminants – so-called 'templates' – to correct the observations. An alternative and complementary approach is to forward-model many possible realizations of the cosmic initial conditions (e.g. Jasche & Kitaura 2010; Jasche & Wandelt 2013; Kitaura 2013; Wang et al. 2014; Jasche, Leclercq & Wandelt 2015; Wang et al. 2016; Modi et al. 2019; Porqueres et al. 2019). One then evolves these initial conditions in time (while adding realizations of non-linearities, bias, and observational/instrument systematics), and performs joint inference of cosmology, the initial conditions and late-time 'true' fields, given observations. Another forward-modelling approach involves the injection of false images into observations in order to sample the selection function (Suchyta et al. 2016; Kong et al. 2020). While such forward approaches are powerful and very general, they also require extensive computational resources and are complicated to implement. In contrast, using templates to clean contaminated observations and directly infer the underlying fields is straightforward to implement and can be readily incorporated into ongoing or completed analyses. They have been the dominant approach in the community thus far, and so these are the methods we focus on here.

In this paper, we revisit and extend state-of-the-art LSS systematics-cleaning strategies. We interpret them through a regression framework to highlight commonalities and differences of the methods, as well as some tacit assumptions. In doing so, we show that the common pseudo-C_ℓ Mode Projection (MP) method is equivalent to linear regression. We use this framework to propose straightforward extensions that leverage the extensive body of literature and tools that have been developed for regression analyses. We rigorously test the performance of several existing methods, plus new ones that we propose, on a common set of simulated observations from current and future surveys.

We study performance using an ensemble of simulated galaxy overdensity maps, such that we can assess both the accuracy and precision of each method. We provide a library of templates and a contaminated overdensity map as input to each cleaning method, which then produces an estimate of the true overdensity map and power spectrum that we assess for accuracy. We repeat the process over a large number of sky realizations and for various configurations of templates to asses the precision and robustness of each method. A schematic outline of this process is shown in Fig. 1.

The paper is organized as follows. In Section 2, we describe in detail our general model for contamination, which encompasses a wide range of systematics due to foregrounds or instrument calibration errors. In Section 3, we describe several existing methods for systematics mitigation; in Section 4, we reinterpret the methods through a common framework to facilitate comparison, and in Section 5 we use this to map several aspects on to well-known techniques in statistics and propose two new mitigation methods. In Section 6, we describe the fiducial synthetic surveys on which we test the efficacy of the methods that we study. Section 7 shows the results of these performance comparisons, while Section 9 has our conclusions. Several appendices show important but more technical and detailed aspects of the investigation.

2 CONTAMINATION MODEL

We first introduce the model for contamination of the observed LSS fields. It is very general, encompassing most known sources of real-world contamination. We can model the observed number density map as a combination of the true galaxy number density map ($N_{\rm true}(\hat{\boldsymbol{n}})$) modulated by a direction-dependent screen $(1 + f_{\rm sys}(\hat{\boldsymbol{n}}))$, plus an additive contamination term $N_{\rm add}(\hat{\boldsymbol{n}})$:

$$N_{\rm obs}(\hat{\boldsymbol{n}}) = (1 + f_{\rm sys}(\hat{\boldsymbol{n}}))N_{\rm true}(\hat{\boldsymbol{n}}) + N_{\rm add}(\hat{\boldsymbol{n}}). \tag{1}$$

We will primarily address multiplicative contamination as this characterizes most known LSS contaminants; one exception is a

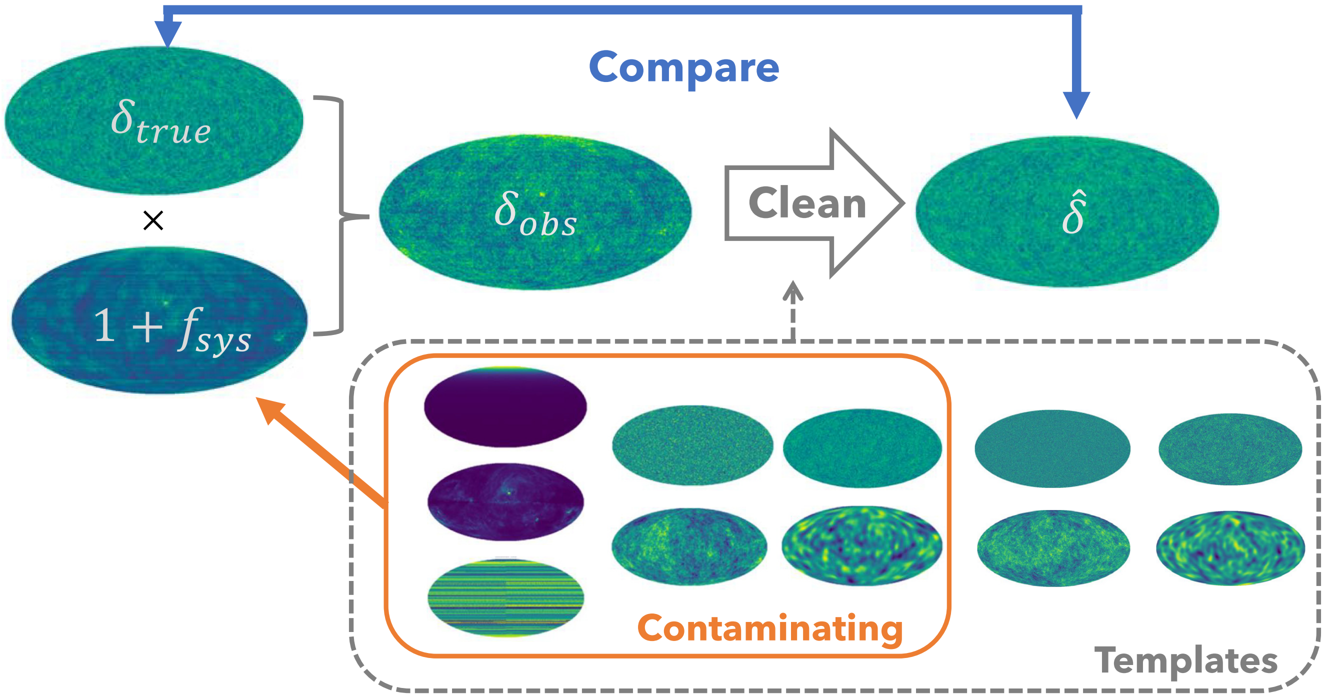

Figure 1. Analysis procedure for a single map. A set of templates is generated (dashed box) along with a true overdensity map $\delta_{\rm true}$. A subset of the templates (orange box) contaminates the true overdensity map to generate the observed overdensity field $\delta_{\rm obs}$. We generate an estimated signal map $\hat{\delta}$ using one of the cleaning methods, and compare it to the truth, either at a map-level or power spectrum level. This is repeated for many realizations of the signal map and the performance of each cleaning method is assessed.

contaminating population of objects such as stars, which we discuss briefly in Section 4.2. Therefore, we take $N_{\rm add}(\hat{\boldsymbol{n}}) \to 0$ for simplicity and focus on the first term $f_{\rm sys}(\hat{\mathbf{n}})$, which fully characterizes the systematic modulation of the true field such that pixels with $f_{\rm sys}(\hat{\boldsymbol{n}}) = 0$ are free of contamination. Using $N = \bar{N}(1+\delta)$ and defining the ratio of true to observed mean number density as $\gamma = \bar{N}_{\rm true}/\bar{N}_{\rm obs}$, the observed overdensity can be written as

$$\delta_{\rm obs}(\hat{\boldsymbol{n}}) = \gamma(\delta(\hat{\boldsymbol{n}})+1)(f_{\rm sys}(\hat{\boldsymbol{n}})+1) - 1. \tag{2}$$

Here, γ enforces the constraint that $\langle\delta_{\rm obs}\rangle_{\rm pix} = 0$ across the survey footprint, even though this is not necessarily true for the *true* overdensity field δ. This is due to the fact that we can only access the observed mean number density $\bar{N}_{\rm obs}$, which differs from the true mean both because of systematic contamination and because of sample variance from a limited survey footprint (see Section 4.2 for details).

This model for contamination is similar to the one used in (Huterer et al. 2013; Shafer & Huterer 2015; Muir & Huterer 2016; Weaverdyck et al. 2018) to assess the impacts of *residual* calibration errors that remain in the data after cleaning. Here we focus on the methods used to perform such cleaning, and so use the screen model to describe contamination more generally.

We extend the screening formalism by considering that the total systematic modulation is comprised of $N_{\rm sys}$ individual systematics, each of which acts as its own screen. Thus, we have

$$\begin{aligned} 1 + f_{\rm sys} &= \textstyle\prod_{i=1}^{N_{\rm sys}}(1+f_i) \\ &\simeq 1 + \textstyle\sum_{i=1}^{N_{\rm sys}} f_i + \sum_{j\neq k}^{N_{\rm sys}} f_j f_k + [\text{higher order terms}], \end{aligned} \tag{3}$$

where we have suppressed $\hat{\boldsymbol{n}}$ in the notation for convenience both here and in what follows. Note that even if a systematic individually contributes to $f_{\rm sys}$ linearly, there exist interaction terms with other systematics up to order $N_{\rm sys}$. Here, and in general, $f_i \equiv f_i(\hat{\mathbf{n}})$ is a column vector with each element corresponding to a pixel, unless otherwise noted.

3 BACKGROUND: EXISTING MITIGATION METHODS

The principal goal of this paper is to compare various proposed systematics mitigation methods. The methods that we test are all designed to use maps that trace potential contamination in order to mitigate the impact of systematics, i.e. they assume that the systematic $f_i(\hat{\boldsymbol{n}})$ is a function of some tracer $t_i(\hat{\boldsymbol{n}})$. We refer to these tracer maps as *templates*, and examples include maps of stellar density, extinction, or summary statistics of observing conditions (e.g. mean g-band seeing) in each region of the sky throughout the duration of the survey (see Leistedt et al. 2016 for a detailed description of the process for creating templates from multiepoch observational data for the Dark Energy Survey). Sources of error for which we have no templates (e.g. shot noise) are implicitly subsumed into the overdensity field.

We will investigate how effectiveness depends on analysis choices and suggest improvements where possible. We start with three principal methods that have been applied in the literature: the Dark Energy Survey Year 1 method (henceforth DES-Y1), the Template Subtraction (TS) method, and the MP method. While at face value the algorithms associated with these methods seem quite different, we demonstrate that they can be translated into a common mathematical framework of linear regression. Doing so allows us to distill commonalities and differences between the methods, as well to identify simplifications and extensions to them. We include three additional methods based on these insights.

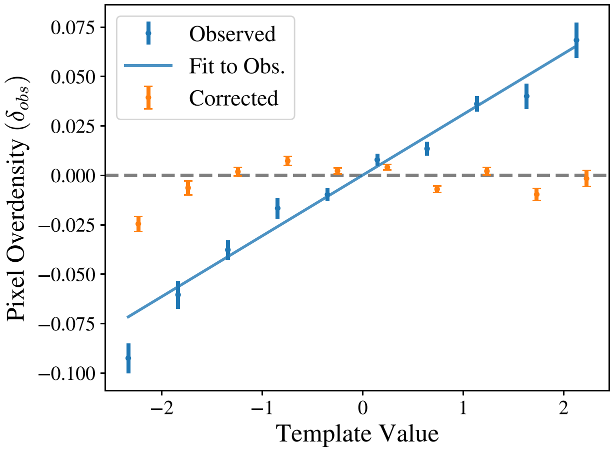

Figure 2. Illustration of the DES-Y1 cleaning method, showing the total observed pixel overdensity ($\delta_{\rm obs}$) as a function of a *template's* pixel overdensity, in 10 evenly spaced bins. Given the estimated covariance matrix (diagonals shown by blue error bars), the best-fit trend (blue line) can be calculated and used to reweight the observed map, producing a corrected map whose dependence on the template is removed (orange points, with corresponding standard errors on the pixel means). The process is then iterated for other templates until a satisfactory threshold is reached; see the text for details.

For all the methods, we will work with maps that are divided into pixels in HEALPix[1] (Gorski et al. 2005) format, which summarize the mean galaxy overdensity or template values within each pixel (see Section 6 for details). Furthermore, while we work in the context of cleaning galaxy overdensity fields, the methods are applicable more generally to corrections of any field for which we have templates of potential contamination, and so we denote the true signal more generally as s and the observed field as $d_{\rm obs}$. In our application, these correspond to the true and observed galaxy overdensity fields, $\delta_{\rm true}$ and $\delta_{\rm obs}$. In the sections that follow, we use $\hat{x}$ to denote an estimate of x, and $\tilde{C}_\ell^{xx}$ to indicate a realization-specific measurement of the power spectrum, as compared to its theoretical mean C_ℓ^{xx}.

3.1 Dark energy Survey Y1 method

The method used to derive galaxy weights for the Year-1 DES release is one of the more sophisticated mitigation methods applied to date. It is described in detail in Elvin-Poole et al. (2018), but we review its main features here. Hereafter referred to as 'DES-Y1,' it builds on the method first proposed as the 'Weights' method in Ross et al. (2011), wherein one-dimensional relationships between observed galaxy densities and systematic templates (there called 'survey property maps') are removed by iteratively applying multiplicative factors ('weights') to galaxies. Fig. 2 shows one example of how the observed overdensity varies with a template. Multiplicative weights are applied to galaxies to de-trend the data, shifting the blue line to lie atop the dashed line. This method is explicitly a regression method, with versions employing linear fits (Ross et al. (Laurent et al. 2017; Ross et al. 2017; Ata et al. 2018), splines (Hernández-Monteagudo et al. 2014) or higher order polynomials (Nicola, Refregier & Amara 2016) as fitting functions for the 1D relationships.

Here, we describe the version we adopt, which closely follows the implementation in Elvin-Poole et al. (2018) used on the DES-Y1 data. For each template t_i, we group pixels into 10 evenly spaced bins based on their template values, and independent of location on the sky (e.g. all pixels with a mean i-band seeing value within 10 per cent of the max would be grouped). We then find the mean galaxy overdensity over the pixels in each bin.[2] A 10 × 10 covariance matrix of these bin means is estimated by performing the same bin-averaging process on a set of 400 uncontaminated mock maps, generated with a fiducial power spectrum for the overdensity field (we assume the true overdensity power spectrum to generate these mocks).

Next, we use `scipy.optimize` and the estimated covariance to find the parameters $\{m_i, b_i\}$ of the best-fitting line[3] of the binned overdensity to each binned template i:

$$\frac{\langle N_{\rm obs}\rangle_j}{\bar{N}_{\rm obs}} = m_i \langle t_i\rangle_j + b_i \tag{4}$$

where $\langle\,\cdot\,\rangle_j$ indicates the average pixel value in bin j of the given template. See Fig. 2 (blue points and trend) for an illustration.

The template with the most significant fit is used to reweight the number density in each pixel as $N'_{\rm obs}(\hat{\boldsymbol{n}}) = N_{\rm obs}(\hat{\boldsymbol{n}})/(\hat{m}_i t(\hat{\boldsymbol{n}}) + \hat{b}_i)$, where the significance metric is defined below. Having removed the effect of the dominant systematic, the whole process is repeated: for each template, the pixels are assigned to bins and averaged, the new best-fitting parameters are computed from equation (4), and the trend from the most significant template is removed from the data. The process stops when all templates are below a predefined significance threshold.

In general, the more contamination from template i, the stronger relationship the relationship with the observed galaxy density. However, some level of correlation is expected just by chance, and this depends on the spatial clustering of each template. The DES-Y1 method addresses this in two ways: (1) by using a different covariance matrix for the observed overdensity for each template as described above, and (2) by having a template-specific significance threshold, calibrated on mocks. Specifically, the significance statistic used is $\Delta\chi_i^2/[\Delta\chi_i^2]_{68}$, where $\Delta\chi_i^2$ is the improvement in χ^2 for the binned fit on template i, compared to a null hypothesis of $m_i = b_i = 0$. It is normalized to the 68th percentile of the same quantity measured on uncontaminated signal mocks ($[\Delta\chi_i^2]_{68}$). We use the stopping criterion $\Delta\chi_i^2/[\Delta\chi_i^2]_{68} < \Delta\chi^2_{\rm threshold} = 2$, but find that our results change little when changing this threshold between 1 and 4 (see Appendix F).

There are a number of required parameter choices in the DES-Y1-type method. These include the criterion for selecting the most significant template,[4] the significance threshold that determines when to stop weighting, the prior power spectrum for generating mocks, and choices associated with binning (e.g. number of bins, equally spaced versus equally filled, etc.). Here, we use the fiducial choices from Elvin-Poole et al. (2018), and investigate some of the effects of these choices in Appendix F.

[1] http://healpix.sourceforge.net

[2] In Elvin-Poole et al. (2018), extreme regions are removed by eye: each template is inspected and bins that exhibit an average fluctuation in number density of > 20 per cent are masked, as are regions where visual inspection suggests a deviation from non-monotonic behaviour (see their fig. 3). We neglect this step, as it is difficult to automate robustly and in our tests we found that it did not alter our results.

[3] Elvin-Poole et al. (2018) also use linear fits for almost all templates, with only a couple exceptions. As noted in Section 4.2, even if a template is thought to contaminate non-linearly, the relationship can usually be made linear through an appropriate transformation of the template.

[4] E.g. one could consider an R^2 statistic, the commonly used F-statistic, Akaike or Bayesian information criteria, etc.

3.2 Template subtraction

The TS method uses the cross-power of templates with the observed sky to estimate contamination of each template at each angular scale. Contamination is subtracted directly from the two-point clustering statistics. The method was proposed in Ho et al. (2012) and Ross et al. (2011) where it was called the 'cross-correlation' technique, and we review it here.

TS assumes the observed overdensity d_{obs} is a linear combination of the true galaxy overdensity s and individual template overdensities t_i:

$$d_{\text{obs}} = s + \sum_{i=1}^{N_{\text{tpl}}} \alpha_i t_i. \tag{5}$$

Any systematics or noise not accounted for by templates are subsumed into the signal s. In Ho et al. (2012), d_{obs} and t_i are taken to be in multipole space, such that $\langle ss\rangle \rightarrow \langle s_{\ell m} s_{\ell m}\rangle = C_\ell^{ss}$ (where $\langle \cdot \rangle$ is the ensemble average over many sky realizations), and $\alpha \rightarrow \alpha_\ell$ is a function of ℓ. The companion paper of Ross et al. (2011) works in configuration space, so in their version of TS the data vectors are in pixel-space (they also take some additional steps; see footnote 7).

We will work in harmonic space and so follow Ho et al. (2012), but we will keep the notation general until dealing with the two-point functions where we will explicitly work with power spectra. The treatment for configuration space is largely identical. To apply the method, one would simply substitute the correlation function for the power spectrum $C_\ell^{ij} \rightarrow w^{ij}(\theta)$ and $\alpha_\ell \rightarrow \alpha(\theta)$. See e.g. Crocce et al. (2016) for an application of TS to the correlation function.

If we consider just a single contaminant for simplicity ($N_{\text{tpl}} = 1$), and assume that it is uncorrelated with the underlying galaxy field, then from equation (5) the two-point function of the observed field is

$$\langle d_{\text{obs}} d_{\text{obs}}\rangle = \langle ss\rangle + \alpha^2 \langle tt\rangle. \tag{6}$$

Then, on average,

$$\langle t d_{\text{obs}}\rangle / \langle tt\rangle = C_\ell^{td} / C_\ell^{tt} = \alpha_\ell \tag{7}$$

and the contamination at each multipole can be estimated as

$$\hat{\alpha}_\ell = \tilde{C}_\ell^{td} / C_\ell^{tt}, \tag{8}$$

where the tilde in $\tilde{C}_\ell$ indicates the power spectrum that is measured from the observed sky realization, and $\tilde{C}_\ell^{tt} = C_\ell^{tt}$ since we take templates to be fixed.

An estimate of the power spectrum can then be found to be

$$\hat{C}_\ell^{ss} = \left(\tilde{C}_\ell^{dd} - \hat{\alpha}_\ell^2 C_\ell^{tt}\right)\left(1 - \frac{1}{2\ell+1}\right)^{-1}. \tag{9}$$

Here, $[1 - 1/(2\ell+1)]^{-1} = [(2\ell+1)/(2\ell)]$ is a factor found by Elsner, Leistedt & Peiris (2016) that is needed to debias the estimator.[5] The bias arises because the process is too aggressive – any chance correlation between template and the true signal is also removed, resulting in an underestimate of the true clustering power.

The TS method is easily generalized to multiple templates by extending the dimensionality of terms as

$$\alpha_\ell \text{ (scalar)} \rightarrow \boldsymbol{\alpha}_\ell \ (N_{\text{tpl}})$$
$$C_\ell^{tt} \text{ (scalar)} \rightarrow \boldsymbol{C}_\ell^{TT} \ (N_{\text{tpl}} \times N_{\text{tpl}})$$

and equations (8) and (9) become

$$\hat{\boldsymbol{\alpha}}_\ell = [\tilde{\boldsymbol{C}}_\ell^{TT}]^{-1} [\tilde{\boldsymbol{C}}_\ell^{Td}], \tag{10}$$

$$\hat{C}_\ell^{ss} = \left(\tilde{C}_\ell^{dd} - \hat{\boldsymbol{\alpha}}^\dagger \boldsymbol{C}_\ell^{TT} \hat{\boldsymbol{\alpha}}\right)\left(\frac{2\ell+1}{2\ell+1-N_{\text{tpl}}}\right). \tag{11}$$

For the cut-sky equations, we refer the reader to Elsner et al. (2016).

While previous work on TS has focused on the cleaned power spectrum, an estimate of cleaned overdensity field itself is also of interest for cosmological study, as it contains more information than just its power spectrum.

A map estimate from the TS method can be produced as

$$\hat{s}^{TS}(\hat{\boldsymbol{n}}) = \sum_{\ell=1}^{\infty} \sum_{m=-\ell}^{\ell} \hat{s}_{\ell m}^{TS}\, Y_{\ell m}(\hat{\boldsymbol{n}}), \tag{12}$$

where the harmonic coefficients of the map are given by

$$\hat{s}_{\ell m}^{TS} = (d_{\text{obs}})_{\ell m} - \sum_{i=1}^{N_{\text{tpl}}} (t_{\ell m})_i (\hat{\alpha}_\ell)_i \tag{13}$$

and the (biased) power spectrum of the cleaned map is equivalent to the first factor in equation (11).

3.3 Mode projection

MP (Kalus et al. 2016; Percival 2018; Alonso, Sanchez & Slosar 2019; Kalus et al. 2019; Nicola et al. 2019) assumes the same contamination model as TS, given by equation (5). The original formulation (Rybicki & Press 1992; Leistedt et al. 2013) cleans the map-level systematics by assigning infinite variance to contaminating templates. This procedure desensitizes the power spectrum estimate to the templates and is equivalent to marginalizing over the contamination amplitude of each template (Leistedt et al. 2013).

In particular, it updates the map-level covariance matrix C as follows

$$\begin{aligned} \boldsymbol{C}' &= \left[\boldsymbol{C} + \sum_k^{N_{\text{tpl}}} \lim_{\beta\rightarrow\infty} (\beta_k t_k t_k^\dagger)\right] \\ &= \lim_{\beta\rightarrow\infty} \left[\boldsymbol{C} + \beta T T^\dagger\right] \end{aligned} \tag{14}$$

where t_k are the individual template maps, which can represent either real spin-0 or complex spin-2 fields (Alonso et al. 2019), and which can be assembled into a matrix T, with t_k as the k^{th} column. In previous works with MP, the maps have been represented in pixel space, but in principle the operations can also be performed in harmonic space, e.g. representing a spin-0 field by its complex harmonic coefficients. For clarity and continuity, we will assume the maps are N_{pix}-length vectors in what follows, as opposed to their multipole transforms. There are some benefits to performing MP in harmonic space, however, which we explore in Section 4.

The main challenge with the original formulation of MP is that it requires the construction and inversion of a covariance matrix for the whole map, which is often intractable. To remedy this, Elsner, Leistedt & Peiris (2017) extended MP to the popular (albeit sub-optimal) pseudo-C_ℓ estimator. In practice, this is achieved by computing the pseudo-C_ℓs of the overdensity field after first applying a filter F, where

$$\begin{aligned} \boldsymbol{F} &= \lim_{\beta\rightarrow\infty} \left(I + \beta T T^\dagger\right)^{-1} \\ &= I - T(T^\dagger T)^{-1} T^\dagger, \end{aligned} \tag{15}$$

[5] In the case of the correlation function, the bias cannot be written in a signal-independent fashion, and so requires a prior signal power spectrum or simulations to estimate.

where the second expression follows from the Sherman–Morrison–Woodbury formula. It is easy to see that $T(T^\dagger T)^{-1}T^\dagger$ is a projection matrix, projecting an $N_{\rm pix}$-dimensional map on to a $N_{\rm tpl}$-dimensional subspace. The filter thus removes any components of the observed map within the subspace spanned by the templates (hence the alternate name of Mode *De*projection).

Taking the case of a single template map t for simplicity, F then takes the form $(I - (tt^\dagger)/(t^\dagger t))$, resulting in a filtered overdensity map

$$\hat{s} = \boldsymbol{F}\, d_{\rm obs} \tag{16}$$

$$= \left[I - t(t^\dagger t)^{-1}t^\dagger\right] d_{\rm obs} \tag{17}$$

$$= d_{\rm obs} - t\hat{\alpha}_{\rm mp} \tag{18}$$

where

$$\hat{\alpha}_{\rm mp} = (t^\dagger d_{\rm obs})/(t^\dagger t) = \tilde{\sigma}^2_{\rm td}/\sigma^2_{\rm tt}, \tag{19}$$

and $\tilde{\sigma}^2_{\rm td}$ is a measure of the covariance of maps t and d. Note that this is very similar to the TS estimate in equation (8), but here the covariances are taken over the whole footprint, rather than for a single mode ℓ. We can make the connection even more explicit by noting that in the full-sky case,

$$\tilde{\sigma}^2_{\rm td} = \frac{1}{4\pi}\sum_{\ell=0}^{\infty}(2\ell+1)\tilde{C}_\ell^{\rm td} \tag{20}$$

While Elsner et al. (2017) introduce this filtered map only as a means to compute the power spectrum, it can be used on its own as an estimate for the cleaned overdensity field. However, as with TS, the power spectrum of this cleaned *map* is a biased estimate of the true power spectrum, as some of the signal is removed in the cleaning process:

$$\langle C_\ell^{\hat{s}\hat{s}}\rangle = \langle (d_{\rm obs} - t\hat{\alpha}_{\rm mp})^\dagger (d_{\rm obs} - t\hat{\alpha}_{\rm mp})\rangle \tag{21}$$

$$= C_\ell^{ss} - \frac{C_\ell^{tt}}{4\pi(\sigma^2_{tt})^2}\left(2C_\ell^{ss}\sigma^2_{tt} - \tfrac{1}{4\pi}\textstyle\sum_{\ell'}(2\ell'+1)C_{\ell'}^{ss}C_{\ell'}^{tt}\right). \tag{22}$$

In the full sky case, the power spectrum estimate can be debiased analytically (Elsner et al. 2017):

$$\hat{C}_\ell^{ss} = \textstyle\sum_{\ell'}\left[(I+B)^{-1}\right]_{\ell\ell'} C_{\ell'}^{\hat{s}\hat{s}}, \tag{23}$$

where

$$B_{\ell\ell'} = \frac{C_\ell^{tt}}{4\pi\left(\sigma^2_{tt}\right)^2}\left(-2\sigma^2_{tt}\delta_{\ell\ell'} + \frac{2\ell'+1}{4\pi}C_{\ell'}^{tt}\right) \tag{24}$$

and $\delta_{\ell\ell'}$ is the Kronecker delta. In the presence of a mask, one can debias via iteration or assuming a prior power spectrum (Elsner et al. 2017). As we work in the full-sky case, we debias analytically via equation (23), though we do not expect an iterative or prior-based debiasing to significantly alter our conclusions.

The procedure outlined above easily generalizes to multiple maps by extending the dimensionality of the terms:

$$\begin{aligned}
\alpha\ (\text{scalar}) &\rightarrow \alpha\ (N_{\rm tpl}),\\
t\ (N_{\rm pix}) &\rightarrow \boldsymbol{T}\ (N_{\rm pix}\times N_{\rm tpl}),\\
\sigma^2_{\rm td}\ (\text{scalar}) &\rightarrow \sigma^2_{\boldsymbol{T}d}\ (N_{\rm tpl}),\\
C_\ell^{\rm tt}\ (\text{scalar}) &\rightarrow \mathbf{C}_\ell^{\boldsymbol{TT}}\ (N_{\rm tpl}\times N_{\rm tpl}),\\
\sigma^2_{\rm tt}\ (\text{scalar}) &\rightarrow \sigma^2_{\boldsymbol{TT}}\ (N_{\rm tpl}\times N_{\rm tpl}).
\end{aligned} \tag{25}$$

Hereafter, we will use ‘MP’ to refer to the pseudo-C_ℓ MP method described above, due to the popularity of the pseudo-C_ℓ power spectrum estimator and the adoption of this version into `NaMaster`[6](Alonso et al. 2019), in anticipation of LSST. We again refer the reader to Elsner et al. (2017) for the modifications necessary to account for the mask, and specifically to their equation (21) for the multi-template version of the debiasing matrix, which we use to correct for all MP power spectrum estimates (see Alonso et al. (2019) for the equivalent formulae for spin-2 fields).

4 PLACING INTO A COMMON MATHEMATICAL FRAMEWORK

To facilitate a comparison of the methods, it is useful to place them into a common mathematical framework. In this section, we show how all three methods presented so far can be interpreted through a regression analysis lens, and in doing so help identify different assumptions within each method and possible avenues for improvement. Moreover, we can leverage the powerful suite of tools that have already been developed and tested for regression to the task of systematics removal, facilitating and accelerating the process.

4.1 Connections to regression

We have purposefully formulated the methods (e.g. equations 19 and 8) in a manner designed to make the connections between MP and TS apparent. TS is equivalent to running the MP algorithm, but with each original template $(t_i(\hat{\boldsymbol{n}}))$ decomposed into a set of independent templates $(t_\ell^i(\hat{\boldsymbol{n}}))$, where

$$t_\ell^i(\hat{\boldsymbol{n}}) = \sum_{m=-\ell}^{\ell} t_{\ell m}^i Y_{\ell m}(\hat{\boldsymbol{n}}). \tag{26}$$

Fig. 3 shows this schematically. In other words, (pseudo-C_ℓ) MP can be considered a special case of TS, where the contamination is assumed to be independent of scale *and* the full template map is used to estimate such contamination. It has been pointed out before in the context of 3D clustering estimates that TS and MP can be related if they use equivalent templates (Kalus et al. 2016).

Casting the two methods into this form allows us to make the connection to standard linear regression wherein a measured response $\boldsymbol{y}$ is assumed to be a linear combination of predictors given by the $\boldsymbol{\alpha}$ and a noise term $\boldsymbol{\epsilon}$:

$$\boldsymbol{y} = \boldsymbol{X\alpha} + \boldsymbol{\epsilon}. \tag{27}$$

$\boldsymbol{X}$ is a $n \times p$ matrix, where p is the number of predictors (potentially including a column of ones – the intercept term), $\boldsymbol{\alpha}$ a vector of length p, and $\mathbf{y}$ and $\boldsymbol{\epsilon}$ vectors of length n.

Perhaps the most common regression method, Ordinary Least Squares (OLS), finds the vector $\hat{\alpha}$ that minimizes the squared residuals:

$$\hat{\boldsymbol{\alpha}} = \text{argmin}_\alpha ||\boldsymbol{y} - \boldsymbol{X\alpha}||^2 \tag{28}$$

$$= (\boldsymbol{X}^\dagger\boldsymbol{X})^{-1}\boldsymbol{X}^\dagger\boldsymbol{y}, \tag{29}$$

where the second expression follows if X is full column-rank (i.e. the number of observations exceeds the degrees of freedom from the predictors). This is equivalent to the maximum-likelihood solution if one assumes the noise of each element, ϵ_i, is independent and identically Gaussian distributed,

$$P(\boldsymbol{y}|\boldsymbol{X}\alpha) \sim \mathcal{N}(0, I\sigma_\epsilon). \tag{30}$$

[6] https://github.com/LSSTDESC/NaMaster

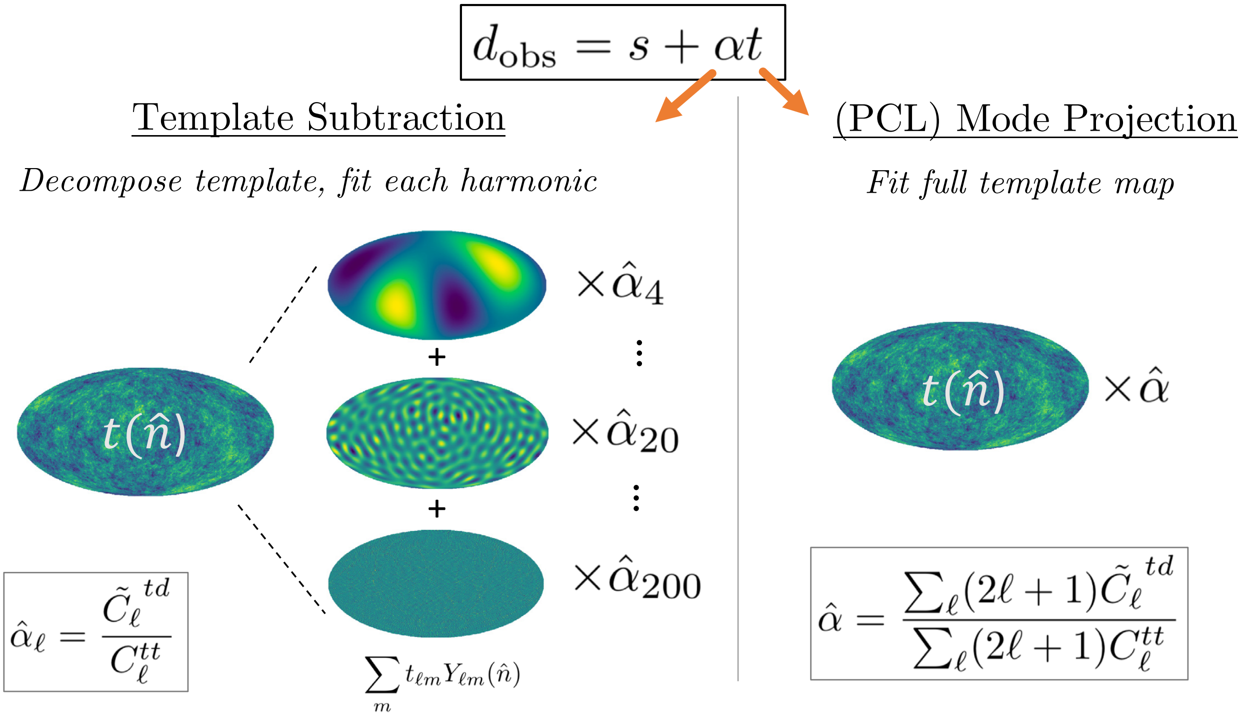
$d_{\mathrm{obs}} = s + \alpha t$
Template Subtraction
Decompose template, fit each harmonic
(PCL) Mode Projection
Fit full template map
$t(\hat{n})$
$\times\hat{\alpha}_4$
+
$\times\hat{\alpha}_{20}$
+
$\times\hat{\alpha}_{200}$
$\sum_m t_{\ell m} Y_{\ell m}(\hat{n})$
$\hat{\alpha}_\ell = \dfrac{\tilde{C}_\ell^{td}}{C_\ell^{tt}}$
$t(\hat{n})$
$\times\hat{\alpha}$
$\hat{\alpha} = \dfrac{\sum_\ell (2\ell+1)\tilde{C}_\ell^{td}}{\sum_\ell (2\ell+1) C_\ell^{tt}}$

Figure 3. Schematic illustration of the difference between the TS and (pseudo-C_ℓ) MP methods. TS allows templates to have different levels of contamination at each scale. This is analogous to performing MP, but first decomposing each template map into a series of derived templates, each corresponding to a different harmonic ℓ. See Section 4.1 for details.

such that the log-likelihood goes as $\mathcal{L} \propto |\boldsymbol{y} - \boldsymbol{X\alpha}|^2$. Even if the assumption of Gaussianity is violated, by the Gauss-Markov theorem equation (28) still corresponds to the unbiased estimator with minimum variance if the errors $\boldsymbol{\epsilon}$ are uncorrelated and have equal variance.

We can write equation (27) in terms of the OLS estimates as

$$\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\epsilon}} = \boldsymbol{X}(\boldsymbol{X}^\dagger\boldsymbol{X})^{-1}\boldsymbol{X}^\dagger\boldsymbol{y} + \hat{\boldsymbol{\epsilon}} \tag{31}$$

where the residuals are defined as

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}. \tag{32}$$

The quantities of interest in the typical regression problem are the coefficients $\boldsymbol{\alpha}$ or the predicted response $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\alpha}}$, with the goal of understanding the influence of predictors or to predict future observations, and hence the residuals are largely used to assess whether the basic OLS assumptions hold. However, comparing equation (32) to equations (13) and (18), we see that both MP and TS can be interpreted as OLS regression methods where the observed overdensity signal is regressed on to the templates, and the reconstructed overdensity signal $\hat{s}$ and power spectrum $C_\ell^{\hat{s}\hat{s}}$ correspond to the map and power spectrum of the residuals $\hat{\boldsymbol{\epsilon}}$.

MP uses the full map footprint, with each pixel corresponding to a single observation, for a total of N_{tpl} fit coefficients. In contrast, TS can be interpreted as performing multiple OLS regressions in parallel on smaller subspaces – one at each multipole in our case – for a total of $N_\ell \times N_{\mathrm{tpl}}$ fit coefficients (see Fig. 3).

We can write the TS amplitudes computed by equation (10) in OLS form as

$$\hat{\alpha}_\ell = (\boldsymbol{T}_\ell^\dagger\boldsymbol{T}_\ell)^{-1}\boldsymbol{T}_\ell^\dagger d_\ell, \tag{33}$$

where T_ℓ is a $(2\ell+1) \times N_{\mathrm{tpl}}$ matrix, with each column corresponding to a template, consisting of all the harmonic coefficients for a fixed ℓ:

$$\boldsymbol{T}_\ell = \begin{pmatrix} t^1_{\ell,-\ell} & t^2_{\ell,-\ell} & \cdots & t^{N_{\mathrm{tpl}}}_{\ell,-\ell} \\ t^1_{\ell,-\ell+1} & t^2_{\ell,-\ell+1} & \cdots & t^{N_{\mathrm{tpl}}}_{\ell,-\ell+1} \\ \vdots & \vdots & \ddots & \vdots \\ t^1_{\ell,\ell} & t^2_{\ell,\ell} & \cdots & t^{N_{\mathrm{tpl}}}_{\ell,\ell} \end{pmatrix}. \tag{34}$$

In cases where the multipoles (or angular scales) are binned, the number of fit coefficients is reduced to $N_{\mathrm{bins}} \times N_{\mathrm{tpl}}$, which reduces the variance of the contamination estimate. Indeed, MP corresponds to a limiting case, where the modes of each template are averaged with equal weight before fitting. However, in principle one could apply weights differently across scales, and as we will show, this can produce improved coefficient estimates. Alternatively, one could fit individual modes as in TS but combine at the coefficient level – potentially useful if certain scales are of particular interest for a given analysis.[7]

An immediate consequence of the OLS interpretation of these methods is in making explicit the assumptions that MP and TS are making about the underlying density field – they are exactly the 'OLS' assumptions for the error term ϵ in the regression model: independent, Gaussian and of equal variance, in whatever basis the map is represented. These assumptions hold well for TS, which performs a separate regression at each multipole ℓ. In this case, the assumed OLS 'noise' terms are the set of harmonic coefficients of the map

[7] In their real-space analysis of SDSS galaxies, Ross et al. (2011) seem to implement a version of this. They use TS to produce fit coefficients for a large number of scales and templates, but ultimately select one coefficient to apply to all scales for each template. However, it is unclear how they compute the single summary coefficient.

($s_{\ell m}$) at that multipole, which have $\mathrm{Cov}[s_{\ell m_1}, s_{\ell m_2}] = C_\ell^{ss}\delta_{m_1 m_2}$. For MP, these assumptions are violated, as the covariance matrix between overdensity pixels is not diagonal, $\mathrm{Cov}[s(\hat{\boldsymbol{n}}_i), s(\hat{\boldsymbol{n}}_j)] \neq \sigma^2_{\mathrm{sig}}\delta_{ij}$.

Since the primary contribution to the 'noise' of the OLS fit is the clustering signal itself, we can diagonalize it by performing MP in multipole space, with the maps d, s, and t_i becoming complex column vectors comprised of the map spherical harmonic coefficients. The noise of the observed overdensity $d_{\ell m}$ is then $\mathrm{Cov}[s_{\ell_1 m_1}, s_{\ell_2 m_2}] = C_\ell^{ss}\delta_{\ell_1\ell_2}\delta_{m_1 m_2}$. While diagonal, this varies strongly with ℓ and therefore violates the assumption of equal variance, a property known as 'heteroskedasticity' in the statistics literature.

However, once the noise is diagonal, we can improve the MP estimate of $\hat{\boldsymbol{\alpha}}$ by weighting the observed data and template modes by (a prior inferred) $1/\sqrt{C_\ell^{ss}}$:

$$\hat{\alpha} = \frac{\sum_{\ell=0}^{\infty}(2\ell+1)\tilde{C}_\ell^{td}/C_\ell^{ss}}{\sum_{\ell=0}^{\infty}(2\ell+1)\tilde{C}_\ell^{tt}/C_\ell^{ss}}. \quad (35)$$

This is equivalent to a weighted least-squares approach and recovers the maximum-likelihood estimate of $\hat{\boldsymbol{\alpha}}$, eschewing the erroneous assumption of a flat signal power spectrum. This of course only works in the ideal full-sky case, but in principle it should not be difficult to extend to a masked sky, e.g. using a predicted cut-sky C_ℓ^{ss} computed using the standard coupling matrix from the mask (e.g. Hivon et al. 2002; Elsner et al. 2017) along with the cut-sky harmonics of the templates and datavector, or appropriate binning of modes. This can be viewed as a form of 'prewhitening' the data, which accounts for the off-diagonal pixel covariance in the likelihood through an appropriate transform. We explore the potential improvement from such prewhitening in Appendix B, finding that it improves cleaning, but is subdominant to differences between cleaning methods and higher order corrections we discuss below.

Finally, we note that both the TS bias from Elsner et al. (2016), as well as the pseudo-C_ℓ MP bias from Elsner et al. (2017) result trivially when interpreting them through the OLS lens, in which the variance of observed residuals is well known to be biased low:

$$\langle\hat{\epsilon}^\dagger\hat{\epsilon}\rangle = \left(\frac{N_{\mathrm{data}} - p}{N_{\mathrm{data}}}\right)\epsilon^\dagger\epsilon. \quad (36)$$

For TS, the regression at each harmonic has $N_{\mathrm{data}} = 2\ell + 1$ and number of predictors $p = N_{\mathrm{tpl}}$, leading exactly to the debiasing terms for the signal power estimate in equations (9) and (11). The debiasing terms for MP in equation (22) are more complicated and dependent on the signal and template clustering, but if we take both C_ℓ^{ss} and C_ℓ^{tt} to be independent of ℓ, equation (22) reduces to

$$\langle\hat{s}\hat{s}\rangle = C^{ss}\left(1 - \frac{1}{\sum_{\ell'=0}^{\ell_{\max}}(2\ell'+1)}\right), \quad (37)$$

where $p = N_{\mathrm{tpl}} = 1$ and $N_{\mathrm{data}} = \sum_{\ell'=0}^{\ell_{\max}}(2\ell'+1) = (\ell_{\max}+1)^2$ is the total number of Fourier modes in the map. This is in keeping with the interpretation of Elsner et al. (2017), wherein each template removes one degree of freedom from the number of observed Fourier modes. This interpretation can help to assess the risk of overfitting based on the size of the template library.

By making connections between current methods and linear regression explicit, we not only facilitate their interpretation, but can more easily identify the tacit assumptions within these methods, as well as readily improve upon them, drawing on the large body of research into the statistical properties of various regression approaches.

4.2 Additive versus multiplicative treatment

The systematic contamination described in equation (2) results in both additive and multiplicative contributions to the overdensity field. One way in which the methods described here differ is whether or not they ignore the multiplicative contributions. These take the form $\delta(\hat{\boldsymbol{n}})f_{\mathrm{sys}}(\hat{\boldsymbol{n}})$ and, if unaddressed, can bias cosmological constraints in upcoming surveys (Shafer & Huterer 2015). Here, we show how so-called 'additive methods' like MP (or any other regression method) and be readily adapted to account for these multiplicative terms and so lead to improved map and power spectrum estimates.

For clarity, we reformulate equation (2) in the general notation of Section 3 for direct comparison with the additive methods, taking $\delta_{\mathrm{obs}} \to d_{\mathrm{obs}}$ and $\delta \to s$, such that

$$d_{\mathrm{obs}} = \gamma(1+s)(1+f_{\mathrm{sys}}) - 1 \quad (38)$$

where again we have suppressed the pixel index. $\gamma = \bar{N}_{\mathrm{true}}/\langle N_{\mathrm{obs}}\rangle_{\mathrm{pix}}$ accounts for the so-called integral constraint, wherein the mean *observed* number density is used to compute the overdensity field, rather than the true full-sky mean density (see Appendix D for a more in depth look at the impact of this monopole term).

To compare with the additive methods, it is convenient to define a zero-centred systematic as

$$f_i' \equiv \frac{f_i - \bar{f}_i}{1 + \bar{f}_i}, \quad (39)$$

and write equation (38) in an equivalent but zero-centred form:

$$d_{\mathrm{obs}} = \gamma'(1+s)\left(1 + f_{\mathrm{sys}}'\right) - 1, \quad (40)$$

with the new prefactor

$$\gamma' \equiv \gamma(1+\bar{f}_{\mathrm{sys}}) = \left(1 + \langle s' f_{\mathrm{sys}}'\rangle_{\mathrm{pix}} + s_0\right)^{-1} \quad (41)$$

ensuring that the monopole in d_{obs} is zero, and having the property that $\langle\gamma'\rangle \approx 1$.[8] Here, $s_0 \equiv \langle s\rangle_{\mathrm{pix}}$ characterizes the global overdensity in which the footprint resides, and $s' \equiv s - s_0$ is the deviation from that local overdensity.

Thus the observed overdensity field contaminated with a generic systematic f_i can be equivalently written as contamination from a *zero-centred* systematic with a rescaled amplitude, f_i'.

Expanding equation (40), we have

$$d_{\mathrm{obs}} = s + \gamma' f_{\mathrm{sys}}' + \gamma' s f_{\mathrm{sys}}' + (\gamma' - 1)(s+1), \quad (42)$$

Comparing to additive models like MP and TS, which take

$$d_{\mathrm{obs}}^{\mathrm{add}} = s + f_{\mathrm{sys}}' = s + \sum_{i=1}^{N_{\mathrm{tpl}}} f_i', \quad (43)$$

we see that they assume that $\gamma' = 1$ (no mean local overdensity and a vanishing correlation between signal and systematics over the footprint), as well also that $s f_{\mathrm{sys}}' = 0$ *for every pixel*, a much stronger assumption. Despite these assumptions, additive estimates of the total contamination are unbiased, provided the templates T span the space of the true contamination:

$$\langle\hat{f}_{\mathrm{sys}}\rangle = \langle T\hat{\boldsymbol{\alpha}}\rangle = \langle T(T^\dagger T)T^\dagger d_{\mathrm{obs}}\rangle \quad (44)$$

$$\approx T(T^\dagger T)T^\dagger f_{\mathrm{sys}}' = f_{\mathrm{sys}}'. \quad (45)$$

[8] Here the approximation stems from making the assumption $\langle x^{-1}\rangle \approx \langle x\rangle^{-1}$, which holds very well for the cases we are studying where the mean is taken over a footprint with $N_{\mathrm{pix}} \gtrsim 10^5$ and shot noise is subdominant.

Intuitively this makes sense, since in the ensemble average the multiplicative term sf will vanish.

From equation (40), we can then make an improved estimate of the signal map as

$$\hat{s} = \frac{1+d_{\rm obs}}{1+\hat{f}_{\rm sys}} - 1, \qquad (46)$$

$$= \frac{d_{\rm obs}-\hat{f}_{\rm sys}}{1+\hat{f}_{\rm sys}} \qquad (47)$$

where the second form makes clear that this is a simple rescaling of the additive signal estimate, $d_{\rm obs} - \hat{f}_{\rm sys}$. Therefore, *in a model with multiplicative contamination, signal estimates from additive methods can be improved by weighting the estimated signal map by* $1/(1 + \hat{f}_{\rm sys})$. Such reweighting should be avoided for contaminants that are thought to contribute additively to the number density (such as stellar contamination), as these modify γ and result in an additive contribution to the overdensity, but no direction-dependent multiplicative terms (see e.g. Crocce et al. (2016), Nicola et al. (2019)).

To explicitly close the loop on the aforementioned methods, the DES-Y1 method performs a series of 1-D regressions and iteratively weights the observed overdensity in a manner equivalent to equation (46) for each template, whereas MP estimates contamination via a single $N_{\rm tpl}$-dimensional regression, with a signal estimate that can be improved via equation (47).[9] Applying the multiplicative correction makes MP equivalent to the Weights model where the coefficients are derived from a simultaneous multiple regression on all the templates (such as in Bautista et al. (2018), Ross et al. (2020)), but with an additional correction to debias the inferred two-point function. Thus a pixelized weights map for MP can be produced[10] as

$$w(\hat{\boldsymbol{n}}) = (\hat{s}(\hat{\boldsymbol{n}}) + 1)/(d_{\rm obs}(\hat{\mathbf{n}}) + 1), \qquad (49)$$

Fig. 4 illustrates the effect of the multiplicative terms – as well as the impact of neglecting them – on the residuals of a map with a single linear, multiplicative contaminant. The diagonal, dotted line shows the expected relation that would be precisely followed by a purely additive contaminant. A multiplicative contamination adds significant scatter around this relation, shown as the grey points. This scatter remains when the contamination is cleaned with an additive method (orange), but is effectively removed when the multiplicative component is taken into account (blue). Fig. 5 shows how errors on the estimated overdensity field are drastically reduced when applying the multiplicative correction of equation (47) to a realistic use case with multiple contaminating systematics (see Section 6 for details of implementation).

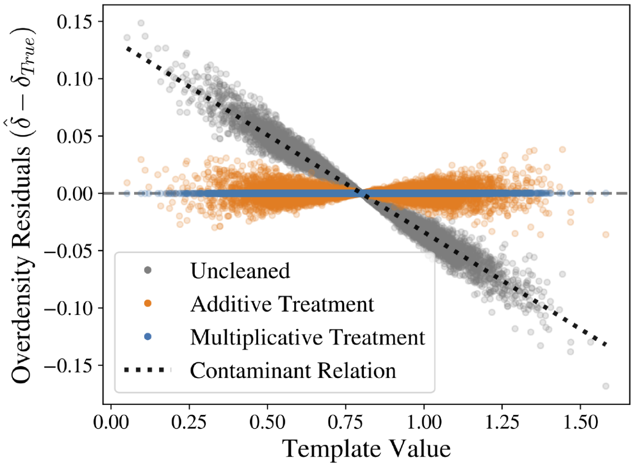

Figure 4. The error in estimates of the overdensity δ in a toy Gaussian map when contaminated with a single template. Gray points indicate the pixel-based difference between the observed, uncleaned overdensity and the true overdensity when the contamination is multiplicative (additive contamination would lie directly along the dotted line). Orange points are the result when erroneously assuming the contamination is only additive. Blue points are the result when correctly treating the multiplicative component.

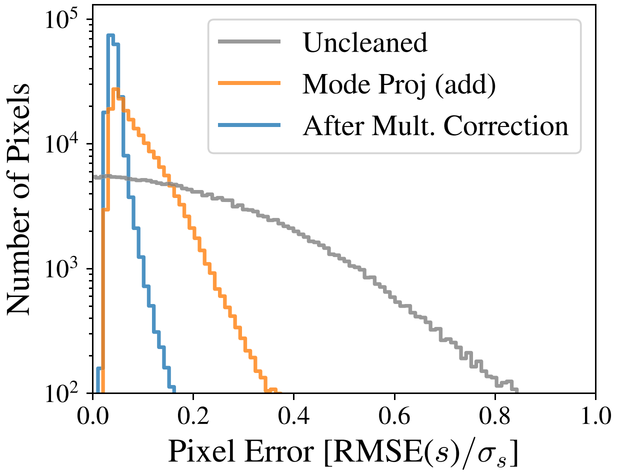

Figure 5. Distribution of pixel errors before cleaning (grey), after cleaning with MP but *before* multiplicative correction (orange), and *after* multiplicative correction (blue). The errors have been calculated as the RMSE of each pixel across 100 cleaned mocks in our fiducial configuration of a DES-like survey as described in Section 6, and have been normalized to the expected dispersion from the true overdensity field.

[9] As noted in Section 2 even linear contaminants will have interaction terms up to order $N_{\rm tpl}$, such that in principle, for equation (44) to fully capture $f_{\rm sys}$, additional templates up to $t_i t_j t_k ... t_{N_{\rm tpl}}$ would need to be included in the template library. A more precise and efficient approach would be to not add any interaction templates, but instead combine the base systematic estimates as

$$\hat{f}_{\rm sys,alt} = \prod_{i=1}^{N_{\rm tpl}} (1 + \hat{f}_i) = \prod_{i=1}^{N_{\rm tpl}} (1 + \hat{\alpha}_i t_i), \qquad (48)$$

where recall t_i corresponds to the ith template and ith column of T, and $\hat{\alpha}_i$ the ith element of $\hat{\alpha}$. This is closer to the treatment of the DES-Y1 method, wherein weightings for each $\hat{f}_i$ are applied in series and thus cumulatively.

In practice, we find that equation (44) is a very good approximation since $\sigma^2_{\rm sys} \lesssim \mathcal{O}(10^{-2})$, so the non-linear interaction contributions to $f_{\rm sys}$ *due to each systematic acting as its own multiplicative screen* are fairly negligible, i.e. $f'_{\rm sys} \approx \left(\sum_{i=1}^{N_{\rm tpl}} f'_i\right)$, as long as the templates sufficiently capture the form of contamination: $f'_i = \alpha_i t_i$. Of course this latter condition is a basic requirement of all of the methods we describe here, one that can and should be verified through standard residual plots and other regression diagnostic techniques to ensure an appropriate contamination model for each template. Methods that incorporate template selection criteria, such as the proposed Elastic Net, can help to satisfy this by allowing a large number of templates to be included in order to address potential higher order terms with little penalty.

[10] This can be released on its own or, as with the DES-Y1 data release, as an additional column at the catalogues level. cf. https://des.ncsa.illinois.edu/releases/y1a1/key-catalogs/key-redmagic).

4.3 Multiplicative effect on likelihood

While the multiplicative term vanishes in the ensemble average, resulting in the same ensemble pixel mean as the additive-only prediction ($\langle d_{\rm obs}\rangle = f'_{\rm sys}$), the pixel variance is modulated:

$$\begin{aligned}\mathrm{Var}[d_{\mathrm{obs}\,i}] &\approx \langle [s\gamma'(1+f'_{\rm sys})]^2\rangle \\ &\approx \langle s^2\rangle(1+f'_{\rm sys})^2 \\ &= \sigma^2_{\rm sig}(1+f'_{\rm sys})^2,\end{aligned} \tag{50}$$

where for large $N_{\rm pix} \gtrsim 10^5$, $\langle\gamma'\rangle \approx \langle\gamma'^2\rangle \approx 1$. The corresponding covariance between pixels is

$$\mathrm{Cov}(d_{\rm obs}, d_{\mathrm{obs}\,j}) \approx \left\langle (\gamma')^2 \left[s_i(1+f_{\mathrm{sys}}{}'_i)\right]\left[s_j(1+f_{\mathrm{sys}}{}'_j)\right]\right\rangle. \tag{51}$$

This is the source of the systematic-dependent scatter in Fig. 4, which will result in biased two-point statistics from additive methods. Because the contamination estimate is unbiased, the correction of equation (47) almost fully suppresses this variance, but the multiplicative terms also impact the likelihood when performing the regression. In pixel-space a simple fix would be to iterate: use an initial estimate of $\langle \hat{f}_{\rm sys}\rangle$ with equation (50) to apply inverse variance weights to the maps before making a second estimate of $\langle \hat{f}_{\rm sys}\rangle$. In practice these are 'errors on the errors' and so the impacts will be subdominant to the multiplicative correction to the datavector itself.

5 APPLICATIONS

We can use the insights of the previous sections to propose two additional methods, as well as to estimate the errors on the cleaned map. We now describe these in turn.

5.1 Iterative forward selection

We include an iterative Forward Selection method that incorporates some of the main features of the DES-Y1 method, but adopts some of the simplifying assumptions of MP. The result is greatly simplified and easier to implement than the full DES-Y1 method.

We keep the core of the template selection algorithm, but modify the fit procedure and significance criterion to eliminate the need to generate mocks. We do this by adopting the same implicit assumptions of MP: that pixels are uncorrelated and have equal variance. This allows for an analytical solution for the best-fitting parameters $\theta = \{m_i, b_i\}$ and their covariance Cov_θ for each template, which we obtain using `numpy.polyfit`.[11] We then adopt a simplified significance criterion of $\Delta\chi^2_{\rm FS} = \theta^T[\mathrm{Cov}_\theta]^{-1}\theta$, and use the same stopping threshold as the DES-Y1 method.[12]

This Iterative Forward Selection method is a fast and simple method that incorporates some of the key aspects of the DES-Y1 method, the iterative weighting and template selection, while avoiding the most computationally expensive parts, the generation of mocks. We expect some loss of precision by not including a covariance matrix in the fitting step, but on the other hand to gain some precision by not having to bin pixels, so this method can help to benchmark the importance of including the covariance matrix in a DES-Y1-like method.

5.2 Elastic net

We also propose a method that closely mimics MP but incorporates template selection, thereby reducing the impact of overfitting when the template library is large.[13] Having shown that MP is equivalent to linear regression, we adopt a regression method specifically designed to automatically select predictors based on the data.

This selection is accomplished by modifying the *Loss* function that is optimized when fitting, which is equivalent to applying a prior to the template coefficients and finding their maximum a posteriori (MAP) estimate. Specifically, instead of finding $\hat{\alpha}$ that minimizes the square of the residuals ($||d_{\rm obs} - T\alpha||^2$), we instead minimize

$$\mathrm{Loss} = \frac{1}{2N_{\rm pix}}||d_{\rm obs} - T\alpha||_2^2 + \lambda_1||\alpha||_1 + \frac{\lambda_2}{2}||\alpha||_2^2, \tag{52}$$

where

$$||\alpha||_1 = \sum_i^{N_{\rm tpl}} |\alpha_i| \tag{53}$$

is the L1-norm of α, and

$$||\alpha||_2 = \left(\sum_i^{N_{\rm tpl}} |\alpha_i^\dagger \alpha_i|\right)^{1/2} \tag{54}$$

is the usual vector L2-norm of α. Here, λ_1 and λ_2 are hyperparameters that are tuned from the data, which we now discuss in turn:

(i) The L1-norm term incentivizes sparsity in α by penalizing non-zero coefficients of templates, thus naturally performing template *selection*. This is useful because the number of templates in modern surveys can be enormous – e.g. Leistedt & Peiris (2014) produce ∼3700 templates for their analysis of SDSS quasars – and so it is common to pre-select only a handful to use, for fear of removing true signal. Since we don't know a priori which templates are contaminating, the incorporation of an automated selection scheme enables a more agnostic, data-centric approach to cleaning a large library of templates, while mitigating the risk of overfitting. The use of this penalty term in isolation (i.e. setting $\lambda_2 = 0$) is often called the Least Absolute Shrinkage and Selection Operator (Foster & George 1994), and has a Bayesian interpretation of applying a zero-centred Laplace prior on the elements of $\boldsymbol{\alpha}$, with a width $\propto 1/\lambda_1$ (see e.g. Starck et al. (2013) for a discussion). L1 priors to induce sparsity have been used in a variety of astrophysical problems, such as for source separation in cosmic microwave background analyses (Bobin et al. 2007, 2013; Wagner-Carena et al. 2019) or in reconstructing mass maps from weak lensing data (Leonard, Lanusse & Starck 2014; Lanusse et al. 2016; Jeffrey et al. 2018).

(ii) The L2-norm term helps address collinearity (i.e. correlation) between template maps which, when present, can cause the matrix $T^\dagger T$ to be ill-conditioned and the variance of contamination estimates to be large. When it is the only additional penalty term (i.e. $\lambda_1 = 0$), this is often called Ridge Regression, or Tikhonov Regularization. It is straightforward to show that, from a Baysian perspective, this

[11] To estimate Cov_θ, `numpy.polyfit` assumes a diagonal Gaussian covariance of the pixels, scaled so that the best-fit model has reduced χ^2 of $\chi^2_{\rm red} = \chi^2/(N_{\rm pix} - 2) = 1$

[12] As with the DES-Y1 method, this method can suffer from a lack of convergence when the threshold is low, where chance correlations between the signal realization and templates result in a loop of the same series of templates being repeatedly reweighted. We adopted a limit of 10 × $N_{\rm tpl}$ reweightings for each signal realization before breaking the loop and using the resulting signal estimate as is. This occurred occasionally and at very low thresholds, with no discernible effect on the estimated maps or power spectra.

[13] See Leistedt & Peiris (2014) for an alternative approach that preselects templates for projection using a χ^2 threshold.

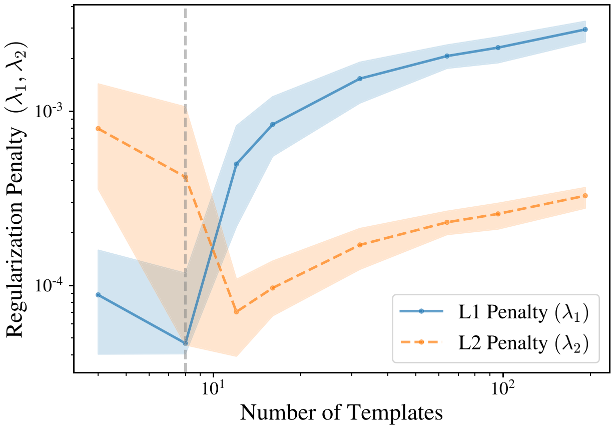

Figure 6. Best-fitting L1 and L2 penalty coefficients in the regularization technique described in Section 5.2, as a function of the number of templates used for cleaning, $N_{\rm tpl}$ (new signal and template maps are generated at each value of $N_{\rm tpl}$). In all cases, 12 templates are contaminating the observed data (vertical dashed line). The importance of the L1 penalty, facilitating template selection, becomes increasingly important as more templates are included for cleaning. Lines and shaded region indicate the median and central 68 per cent probability mass of 50 mocks at each $N_{\rm tpl}$ for the central bin of our fiducial DES-like survey. Here, $\rho_{\rm tpl} = 0.2$ within template groups, though plots for other $\rho_{\rm tpl}$ look similar. See Section 6 for details of implementation.

method is equivalent to placing a zero-centred Gaussian prior on the elements of α, with a width $\propto 1/\lambda_2$.

Since each penalty term addresses a different issue with standard regression, it is not uncommon to combine them, as proposed by Zou & Hastie (2005), in a method known as the 'Elastic Net'. We use the `scikit-learn` (Pedregosa et al. 2011) implementation, `ElasticNetCV`, with a hyperparameter space of $\lambda_1/(\lambda_1 + \lambda_2) \in \{0.1, 0.5, 0.9\}$ and 100 values of $(\lambda_1 + \lambda_2)$ spanning three orders of magnitude, which are automatically determined from the input data (the default setting). We use fivefold cross-validation to determine the best λ_1 and λ_2, trained on a random selection of 30 per cent of the input map pixels.[14]

In this five-fold cross-validation scheme, the training sample (30 per cent of the map) is itself partitioned into five equal subsamples. For each combination of hyperparameters, one subsample is withheld for validation, while the other four are used to train the model by minimizing equation (52). The mean squared error (MSE) of the validation sample is then computed and stored (i.e. the first term in equation 52). One of the four training subsamples is then withheld as the new validation set, and the process is repeated until each of the five subsamples has been used exactly once for validation, with their average MSE used to characterize the goodness-of-fit for the given set of hyperparameters λ_1 and λ_2.

Setting $\lambda_1 = \lambda_2 = 0$ reduces to OLS regression and hence to the pseudo-C_ℓ MP method, while sampling extreme values for the relative weight of the L1 versus L2 penalty allows for the effective use of only one of the penalty terms, if preferred by the data. The use of cross-validation on a subset of the map allows the data to dictate which model is most appropriate, with minimal risk of overfitting. We illustrate the utility of this in Fig. 6, which shows how the cross-validation scheme naturally increases the L1 penalty when fitting for more (*un*contaminating) templates. We found that the L2 penalty became increasingly important when the correlation between templates increased beyond $\rho_{\rm tpl} \gtrsim 0.9$.

5.3 Map errors

We can use the regression framework to gain insight into how errors in the estimated overdensity map are distributed across pixels. This aids the propagation of map errors in cross-correlation studies and summary statistics beyond the two-point functions, as well as helps to identify regions that may benefit from masking.

For simplicity, we assume additive contamination and correction and ignore higher order terms:

$$d_{\rm add} = s + f_{\rm sys} = s + T\alpha. \tag{55}$$

The estimated contamination amplitude is then

$$\hat{\alpha}_{\rm mp} = (T^\dagger T)^{-1} T^\dagger d_{\rm add} \tag{56}$$

$$= \alpha + (T^\dagger T)^{-1} T^\dagger s \tag{57}$$

such that our signal estimate is

$$\hat{s}_{\rm mp} = d_{\rm add} - T\hat{\alpha}_{\rm mp} \tag{58}$$

$$= s - T(T^\dagger T)^{-1} T^\dagger s \tag{59}$$

$$\equiv (I - H)s, \tag{60}$$

where the matrix $H \equiv T(T^\dagger T)^{-1} T^\dagger$ is often called the 'Hat' or 'Projection' matrix in the statistics literature. Then

$${\rm Var}[(\hat{s}_{\rm mp} - s)_i] = {\rm Var}[(Hs)_i] = [H {\rm Var}[s] H^\dagger]_{ii}. \tag{61}$$

If we make the assumption that the signal covariance is diagonal, then ${\rm Var}[s] \approx \sigma^2_{\rm sig} I$ and

$${\rm Var}[(\hat{s}_{\rm mp} - s)_i] \approx \sigma^2_{\rm sig} (HH^\dagger)_{ii} = \sigma^2_{\rm sig} H_{ii}, \tag{62}$$

where we have used the fact that H is both Hermitian and idempotent so that $HH^\dagger = HH = H$.

Despite a number of simplifying assumptions and the fact that some of the methods only fit for some of the templates, we find that with the exception of TS, H_{ii} is a remarkably good predictor[15] of how the errors in the overdensity estimates are distributed for all the methods. The errors arise from removing real signal during the cleaning process, with H_{ii} as a measure of how susceptible pixel i is to such overcorrection. This also indicates that while to first order all correlation with templates is removed from the estimated overdensity field, the templates remain imprinted on the map through their absence; there is missing signal in precisely their spatial configuration.

Intuitively, H_{ii} as a distance measure of pixel i from the centre of mass of other pixels in the $N_{\rm tpl}$-dimensional space spanned by the templates. This is sometimes referred to as 'leverage', as pixels with higher H_{ii} have larger impact when performing a regression.[16] This

[14] We performed the cross-validation procedure on a subset rather than the full footprint as further protection against overfitting, but this is likely overly cautious and subsequent tests showed little difference in performance between training on 30 per cent as compared to the full footprint.

[15] Note that $H_{ii}{}^{\rm I}$ only requires the diagonal elements of H, which are far more tractable to calculate than the full $N_{\rm pix} \times N_{\rm pix}$ matrix.

[16] This phenomenon is very familiar from the simple case of fitting a 1D line to a scatter of 2D points $\{x, y\}$, where the best-fit line is 'pulled' preferentially to points that lie farther from $\bar{x}$.

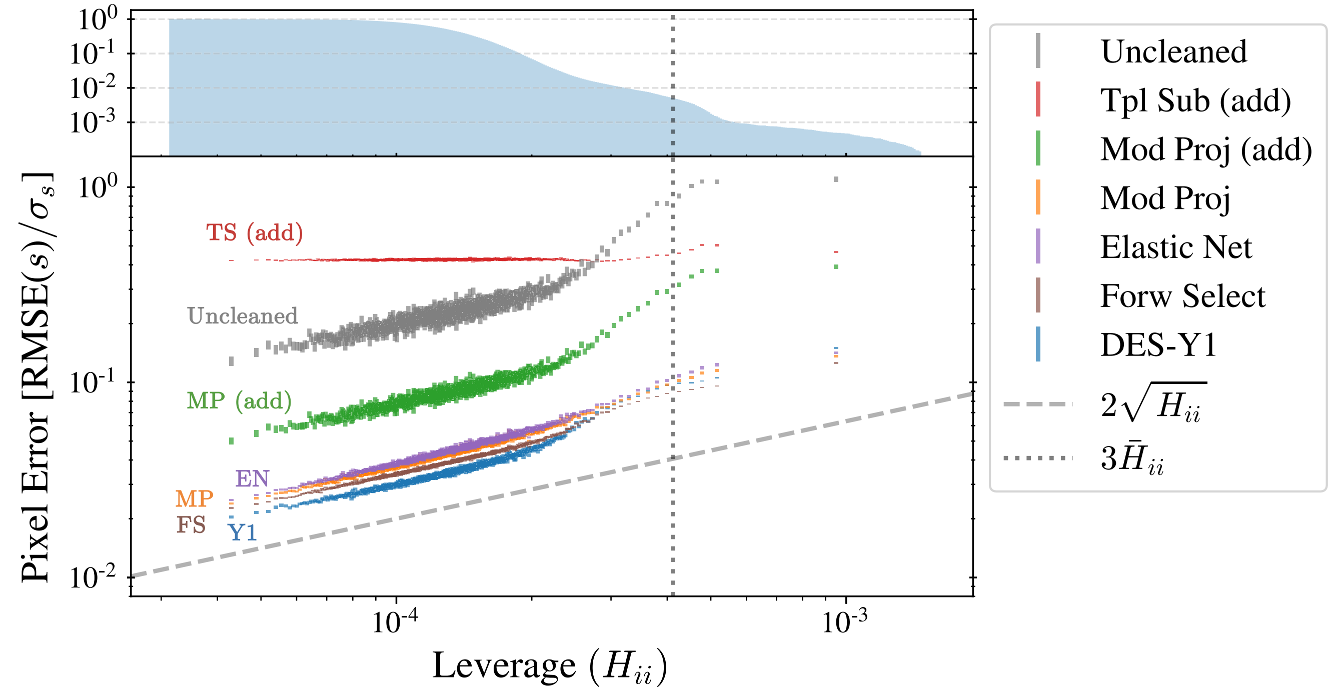

Figure 7. RMSE of pixel overdensity estimates, normalized to expected dispersion from the true overdensity due to cosmic variance, versus pixel leverage for 100 signal realizations. The vertical axis shows the standard error across pixels in 1000 equal-sized bins (in this case 197 pixels per bin at `Nside`=128). The error in both observed and estimated overdensity scales as roughly $\propto H_{ii}^{1/2}$ for all methods (dashed line, to guide the eye). The dotted vertical line indicates a commonly used threshold of 3 times the mean leverage across pixels to identify pixels that may have an undue impact on regression fit parameters. The histogram in the top panel indicates the number of pixels at a given leverage. $N_{\rm tpl} = 27$, $N_{\rm sys} = 11$, and $\sigma^2_{\rm sys} = 0.01$.

can be seen by observing that the estimated systematic field can be written as

$$\hat{f}_{\rm sys} = H d_{\rm obs} \tag{63}$$

such that the leverage

$$H_{ii} = \frac{\partial \hat{f}^{(i)}_{\rm sys}}{\partial d^{(i)}_{\rm obs}} \tag{64}$$

encodes the sensitivity of the contamination estimate to an observed over- or underdensity at pixel i. Because pixels with high leverage can have an outsized effect on the estimated contamination, we expect leverage to be a useful tool for identifying potentially problematic pixels that should be masked before cleaning, in addition to providing error estimates for those pixels that remain.

It is straightforward to derive the mean leverage value as

$$\bar{H}_{ii} \leq N_{\rm tpl}/N_{\rm pix}, \tag{65}$$

with the equality holding if T is full rank, since

$$\sum_i^{N_{\rm pix}} H_{ii} \;=\; {\rm Tr}(H) \leq N_{\rm tpl}, \tag{66}$$

providing a basis on which to determine extreme leverage values.

The main panel of Fig. 7 shows the root mean squared error (RMSE $= \sqrt{\langle(\hat{s}-s)^2\rangle}$) of each pixel computed over 100 cleaned DES-like mock maps plotted against leverage H_{ii} from 27 cleaning templates, 11 of which are contaminating. Pixels are grouped into 1000 bins of 197 pixels, according to their leverage value, and we show the mean and standard error of the RMSE for each bin. We see that pixels with low leverage value have much smaller error in the estimated overdensity map, and that the error goes roughly as $\propto H_{ii}^{1/2}$ (diagonal dashed line), as predicted by equation (62). The TS method is an exception to this trend likely because the regression happens in a different space, at each harmonic separately, and so does not relate cleanly to the pixel leverage.[17]

The top panel of Fig. 7 shows the fraction of map pixels below a given leverage (note the log scale), with the vertical dotted line indicating $3 \times \bar{H}_{ii}$, which is one of two common thresholds used in statistics to flag points that may bias a regression analysis ($2 \times \bar{H}_{ii}$ being the other). Here, 0.5 per cent of map pixels exceed $3 \times \bar{H}_{ii}$; these pixels potentially merit further inspection or masking, as they are particularly prone to biasing the regression. The trend of the uncleaned data may be surprising, but as noted in Section 4.2, because of the integral constraint, $d_{\rm obs}$ is insensitive to a monopole in $f_{\rm sys}$ and so as long as templates approximately trace the true contamination, overdensities near the mean of the templates (i.e. low H^{ii}) will be most accurately measured, even if contamination is greater than at other points in the map (see Appendix D).

A complementary statistic is the ‘Cook’s distance’ (Cook 1977, 1979) for each pixel, which uses H_{ii} and $\hat{s}_i$ to provide a measure of the total change in the $\hat{s}$ map if pixel i were to be masked (assuming additive contamination and correction). Along with the leverage, we expect this to be a useful tool when performing template-based mitigation of spatial systematics and for mask creation. We leave further investigation of these as diagnostic tools, as well as generalization to the multiplicative case, to a later work.

[17] In principle, one could construct the analogous leverage quantity $H_{\ell m} = t_{\ell m}[\mathbf{C}^{\mathbf{TT}}_\ell]^{-1} t^\dagger_{\ell m}$ in harmonic space for the analysis of errors in $\hat{s}_{\ell m}$, which may be useful for cross correlation analyses in harmonic space.

We next describe the fiducial survey on which we test the performance of foreground-cleaning methods.

6 EVALUATING PERFORMANCE

Our analysis is fully synthetic, with the procedure depicted in Fig. 1. We compare the cleaning methods described, including results for both the standard additive MP case (denoted 'MP (add.)') as well as one with the multiplicative correction from equation (47) (denoted simply 'MP'). For the Elastic Net, we only show results that include the multiplicative correction.

We only consider full-sky maps in this paper. Extension to partial-sky surveys should be fairly straightforward, requiring the usual correction of cut-sky power spectra, but this applies equally across the full-sky spectra estimated with each method here and so we do not expect it to qualitatively change the main results.

6.1 Templates

We first describe the fiducial set of templates that we use, for both contamination and cleaning purposes. We adopt several classes of templates in order to span a range of possible contaminants and their spectral behaviour. In most cases, we use multiple templates of the same class by generating Gaussian realizations of maps from the same theoretical power spectrum. The classes of template we use are as follows:

(i) $C_\ell \propto (\ell+1)^0$ (white noise);
(ii) $C_\ell \propto (\ell+1)^{-1}$;
(iii) $C_\ell \propto (\ell+1)^{-2}$;
(iv) $C_\ell \propto \exp[-(\ell/10)^2]$;
(v) a 'Cat-scratch' map, with 128 horizontal stripes to model a basic scanning pattern and/or differences in depth due to overlapping tiles;
(vi) a 2D Gaussian 'spot' map; and
(vii) a $E(B-V)$ extinction map, with dependence on latitude removed.

The last three correspond to static maps which do not change throughout the analysis. We use the full-sky $E(B-V)$ map[18] from Planck (Abergel et al. 2014), but since this is dominated by emission near the galactic plane, which LSS surveys typically avoid, we reweight the map to remove its major latitudinal dependence.

We normalize the individual templates to the same overall variance, and construct a total systematic map as a product of some or all of the individual template maps:

$$1 + f_{\rm sys} = \prod_{i=1}^{N_{\rm sys}} (1 + \alpha_i t_i) \tag{67}$$

Note that this model can generally encompass contamination to any polynomial order simply by including templates that are products of others (e.g. $t_{\rm new} \equiv t_i^2$), and incrementing $N_{\rm sys}$ accordingly. Similarly, non-linear contamination can often be made linear through an appropriate transformation of the template map.[19]This total systematic map is then scaled to a desired overall map variance $\sigma^2_{\rm sys}$, thus determining the overall contamination field $f_{\rm sys}$. We use a fiducial level of contamination of $\sigma^2_{\rm sys} = 0.01$, as we found this to produce fluctuations similar to those seen in the DES-Y1 data (Elvin-Poole et al. 2018); this corresponds to an RMS error on δ of ~ 10 per cent. Changing the level of contamination $\sigma^2_{\rm sys}$ did not significantly alter our results.

We perform the contamination and cleaning procedure shown in Fig. 1 on each redshift bin and for each cleaning method over many sky realizations, and plot the mean and central 68 per cent probability mass of the relevant quality statistic. We use the same set of templates and total systematic map for across redshift bins and sky realizations, but generate a new set for each unique combination of parameter choices (e.g. level of cross-correlation between templates, number of templates used, etc.) in order to minimize any effects from specific template realizations.

We use `CLASS` (Lesgourgues 2011) to compute theoretical galaxy clustering power spectra for a mock LSS survey, including contributions from redshift-space and Doppler distortions and lensing. We found gravitational potential terms to contribute $\lesssim 1$ per cent to the resultant C_ℓ for $\ell > 7$ but increased computation time by an order of magnitude, so we neglect them. Since we find the cleaning procedures are not strongly sensitive to the signal power spectrum, this should not impact our results. We then use `Healpy` (Zonca et al. 2019) to generate full-sky Gaussian realizations of LSS overdensity ($\delta \equiv \delta\rho/\rho$) maps for each redshift bin with `NSIDE` $= 128$. We compare the impact of using lognormal maps in Appendix A, finding it does not change our results.

6.2 Cosmological model and simulated survey

We assume a standard Lambda cold dark matter cosmological model with one species of massive neutrino and parameter values from best-fit Planck 2018: $\{\Omega_{\rm c}, \Omega_{\rm b}, h, n_{\rm s}, \sigma_8, m_\nu/{\rm eV}\} = \{0.26499, 0.04938, 0.6732, 0.96605, 0.8120, 0.06\}$. Given the precise parameter constraints from current probes, the dependence of our results on cosmological parameters is expected to be very minimal. In contrast, the choice of the parameter set to be *determined* from the survey may be highly dependent on the residual systematics.

In general for comparing the methods, the exact form of the galaxy power spectra is not very consequential, so we use a fiducial survey comparable to the completed Y5 Dark Energy Survey, for which a realistic level of contamination can be estimated based on existing data. We assume the number density distribution of galaxies to be in the form

$$\frac{{\rm d}n}{{\rm d}z} \propto \left(\frac{z}{z_0}\right)^\alpha \exp\left[-(z/z_0)^\beta\right], \tag{68}$$

where $z_0 = 0.55$, $\alpha = 2.65$, and $\beta = 3.34$. We assume five redshift bins centred at redshifts $\{0.225, 0.375, 0.525, 0.675, 0.825\}$, with galaxy bias of $\{1.4, 1.6, 1.6, 1.95, 2\}$, respectively, and containing galaxies with Gaussian redshift dispersion of $\sigma_z = 0.05$. These values were chosen to closely approximate the `REDMAGIC` redshift distribution given in Elvin-Poole et al. (2018).

We choose to work primarily in harmonic space. Therefore, starting with some map with overdensity $\delta \equiv \delta N/N$, where N is the galaxy count over some patch, the expansion in spherical harmonics gives

$$\delta(\hat{\boldsymbol{n}}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\hat{\boldsymbol{n}}), \tag{69}$$

and the angular power spectrum is given by

$$C_\ell = \sum_{m=-\ell}^{\ell} \frac{|a_{\ell m}|^2}{2\ell+1}. \tag{70}$$

[18] https://wiki.cosmos.esa.int/planckpla/index.php/CMB_and_astrophysical_component_maps#The_.5Bmath.5DE.28B-V.29.5B.2Fmath.5D_map_for_extra-galactic_studies

[19] E.g. Elvin-Poole et al. (2018) fit linear models to the square root of exposure time and sky brightness, based on how how they contribute to the depth map.

Because we are working in the full-sky limit, all well-known estimators of power return the same result, so here we make use of the `anafast` and `alm2cl` functions in `Healpy`. To more accurately account for the cosmological impact of the cleaning methods on data from a DES Y5-like survey, we divide the assumed sample variance $\sigma^2_{C_\ell}$ by a factor of $f_{\rm sky} = 0.116$.

We add shot noise to the theoretical power spectrum as $C_\ell \rightarrow C_\ell + \bar{n}^{-1}$, with $\bar{n} = 1.5 \times 10^8$, but this is negligible at the large scales we work with ($\ell \le 350$). We are primarily interested in studying the systematic impacts of cleaning (or not) using spatial templates, so it is reasonable to focus on cases where the signal-to-noise is large (i.e. shot noise is negligible).[20]

7 SIMULATION RESULTS

To compare methods, we compare the fidelity of the cleaned data products to the truth, either at the map level or at the level of the power spectrum, rather than look for cosmological-parameter biases. We do this for a few reasons: (1) the map and power spectrum are more general, being independent of (but easily mapped to) any specific cosmological model one wants to test, or summary statistic one wants to use; (2) while we primarily study applications to galaxy clustering data here, the methods themselves are quite general and can easily be applied to other data sets for which one has tracers of potential contamination, such as shear or convergence maps; and (3) galaxy clustering *alone* leads to relatively weak cosmological constraints and is rarely used on its own to constrain cosmology.

We therefore limit ourselves to investigating biases in data space and leave the investigation of impacts on cosmological constraints to a later work when weak lensing data can be incorporated in a more realistic fashion. At this stage, the test bed is sufficiently representative to compare foreground-cleaning methods in a manner to inform future LSS analyses.

7.1 Characterizing performance

We first study the impacts of the different methods on the estimated maps and power spectra for a single configuration and compare the residual biases of each. For this fiducial comparison, we generate 50 mocks for each redshift bin and contaminate them with 11 systematics, two from each of the four Gaussian classes, plus the three static templates. We construct a template library that contains the contaminating templates, plus four additional realizations from each Gaussian class, for a total of 27 cleaning templates. Each method uses this library to produce estimates of the overdensity field and power spectra.

We show map residuals of each cleaning method for the lowest redshift bin in Fig. 8, where the residuals are binned into deciles of the true overdensity. Results for other redshift bins are similar. From left- to right-hand side, in approximate order of performance, the figure shows the TS method (red), MP without (green) and with (orange) multiplicative correction, the Elastic Net method (purple), Forward Selection (brown), and the DES-Y1 method (blue).

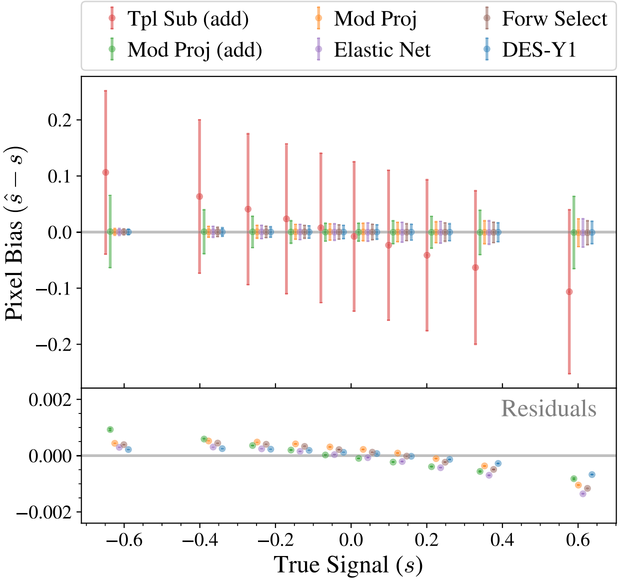

Figure 8. Error in overdensity estimates for different cleaning methods, binned in deciles of the true overdensity and with points offset for clarity. The top plot includes error bars indicating the standard deviation of pixel errors in each bin, while the bottom plot is a zoomed-in version to better display how the means deviate from zero. Overcorrection at the map-level is only significant for TS, which underestimates the magnitude of both peaks and voids, while other methods are very close to unbiased. See the text for details.

The overcorrection of TS is evident, with density fluctuations consistently *under*-estimated (i.e. peaks and voids are both less extreme than they should be). The other methods are all very close to unbiased with respect to the true overdensity field, with bias of the mean $\lesssim 0.001$ for each bin. The multiplicative methods show significantly reduced within-bin scatter (i.e. smaller error bars) compared to the additive ones – the additive TS and MP methods (leftmost, red and green) have typical errors in the overdensity of $\sigma_{\rm s} \sim 0.1$ and $\sigma_{\rm s} \sim 0.01 - 0.05$, respectively, compared to the errors of $\sigma_{\rm s} \sim 0.005 - 0.02$ for the multiplicative methods. This suggests that applying the multiplicative correction results in significantly improved map estimates, making them excellent candidates for map-based analyses, such as as counts-in-cells or density-split statistics.

While the signal estimates are unbiased (with the exception of TS), the errors of the additive methods increase near extremes of the density field. This is similar to the result in Fig. 5, which showed larger errors at extreme template values, in part for the same reasons. Both Figs 7 and 8 indicate a clear stratification of the methods, with the methods that fail to treat the multiplicative component of contamination showing significantly larger error.

We also compare the maps in harmonic space. The left-hand panel of Fig. 9 shows the per-multipole performance of the cleaning algorithms as $(1 - C_\ell^{\hat{s}s}/C_\ell^{ss})$ versus the multipole ℓ, where

$$C_\ell^{s\hat{s}} = \langle s_{\ell m} \hat{s}^*_{\ell m} \rangle. \tag{71}$$

This quantifies the fractional *missing* cross-power between the true and estimated maps, such that a perfect reconstruction corresponds to 0, and pure noise corresponds to 1 (note the log scale). This conveys the approximate level of error expected when using cleaned maps for cross-correlation studies.

All of the methods that treat the multiplicative contamination perform significantly better than the additive methods. The corrected

[20] Shot noise may have the effect of (1) rendering the the regression residuals more diagonal in pixel-space (or flattening them in harmonic space), which could actually improve the regression procedure, and/or (2) introduce significant skewness in the distribution. We would expect the impacts of these to be similar to those of prewhitening the data or using lognormal mocks, and so based on our results in Appendices A and B, we do not expect shot noise to significantly impact on our findings.

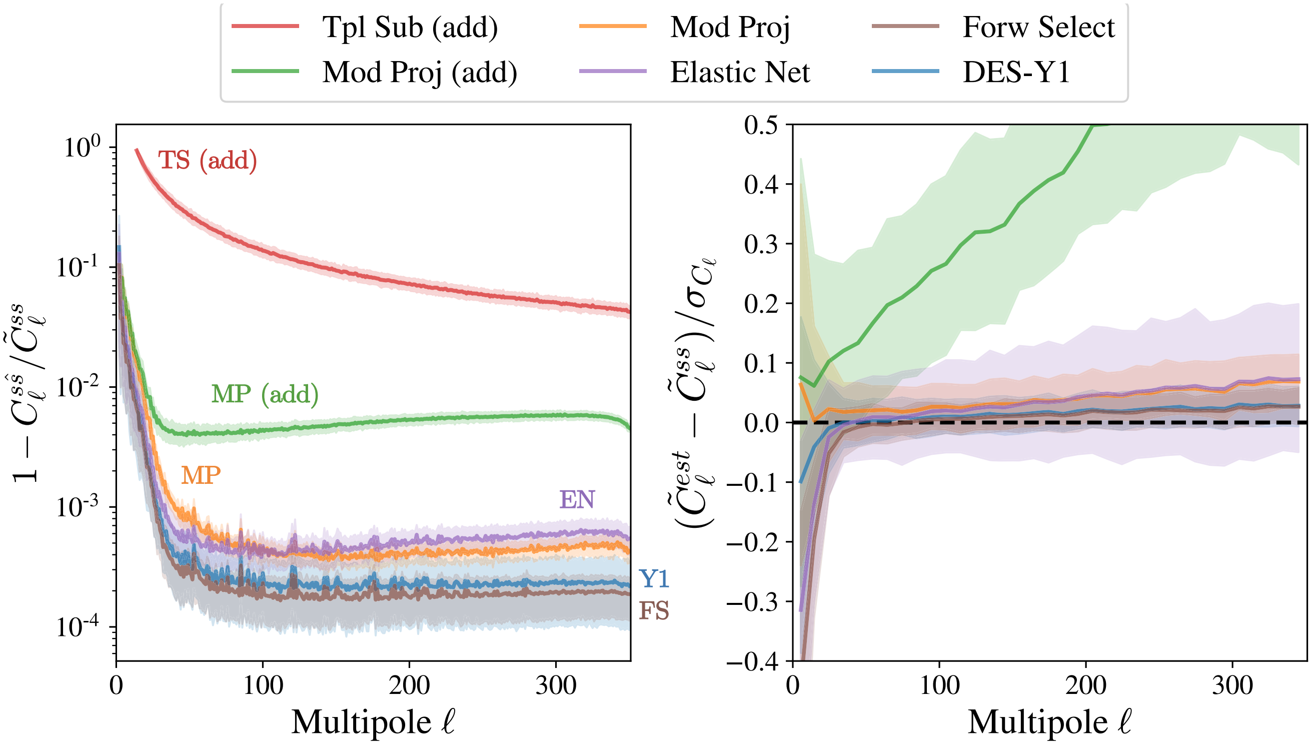

Figure 9. Left-hand panel: Error in map reconstruction for each method as a function of multipole ℓ in a DES-like survey, shown as the deficit in correlation at each multipole between the true and cleaned maps ($1 - C_\ell^{s\hat{s}}/C_\ell^{ss}$). A perfect reconstruction corresponds to 0, whereas pure noise corresponds to 1. For all methods (except perhaps TS), the cleaned map is a good approximation of the true map for cross-correlation purposes, especially at scales $\ell \gtrsim 30$. Right-hand panel: Error in power spectrum estimation shown as the residual angular power relative to sample variance $(\tilde{C}_\ell^{\rm est} - \tilde{C}_\ell^{ss})/\sigma_{C_\ell}$ in bins of $\Delta\ell = 10$. Solid lines indicate means of cleaning performed on 50 signal realizations of each bin and shaded regions indicate the central 68 per cent probability mass of the 250 total realizations. The multiplicative correction applied to MP removes most of the bias of the method (green to orange). Here, we use 27 templates, of which 11 are contaminating the data.

MP and Elastic Net, and the DES-Y1 method all have excellent performance at $\ell \gtrsim 30$ or scales below about 0.2 degrees on the sky, showing $\lesssim 0.1$ per cent error. Maps cleaned with these methods should therefore be excellent candidates for cross-correlation studies. Even the additive MP method performs quite well with error of $\lesssim 1$ per cent in this case, and as such it may be adequate for many studies.

In the right-hand panel of Fig. 9 we show the error in the power spectrum estimate as the difference between the estimated (after cleaning) and true angular power in bins of $\Delta\ell = 10$ and normalized to sample variance $(\tilde{C}_\ell^{\rm est} - \tilde{C}_\ell^{ss})/\sigma_{C_\ell}$, where for σ_{C_ℓ} we use the standard Gaussian approximation for cosmic variance, scaled by $1/f_{\rm sky}$. Unlike $C_\ell^{s\hat{s}}$, this quantity is insensitive to phase-differences between the true and reconstructed maps of the map.

To lowest order, all of the methods work well, and the residual biases are below cosmic variance for the large angles studied here (note that systematic shifts will become more significant with larger multipole bins). For MP, the performance is satisfactory only once it is corrected for the multiplicative bias via equation (46), which both reduces bias and uncertainty in the estimated power spectra. We do not show TS on the right for clarity – its mean traces the mean for the additive MP method, but the dispersion is very large, exceeding the plot limits.

The Elastic Net and Forward Selection methods show a similar deficit at large scales as does MP before it is debiased via equation (47). This is because the power spectra of the clustering signal and most of the cleaning templates peak at low ℓ, such that more power is removed from large scales. The contribution from the signal power spectrum to this effect (i.e. heteroskedasticity) is mitigated for the DES-Y1 method, which uses the signal covariance. In practice, biases exhibited by any of the methods for the signal power spectrum could be estimated and removed by running on realistic contaminated mocks.

7.2 Susceptibility to overfitting

Any template-fitting model faces a challenge to neither underfit nor overfit the data. In the case of underfitting, residual contamination will be left over in the map and inferred to be signal. In the case of overfitting, a portion of the signal will be inadvertently removed from the map, having been mistaken for systematics. Additionally, increasing the number of fitted templates increases the variance of the estimated power spectrum, which will increase the error of $\tilde{C}_\ell^{\rm est}$ in a mean-squared sense (Elsner et al. 2017).

MP and TS address the risk of overfitting by estimating how much signal power is lost from over-correction given the template library and scaling the power spectrum accordingly (equations 11 and 23). In contrast, the DES-Y1 and Forward Selection methods use thresholds to limit the templates used for cleaning to only those that are most significant, an approach that was also implemented in the *Extended* MP method of Leistedt & Peiris (2014) for the QML power spectrum estimator (though as shown by Elsner et al. (2016), this comes at the cost of an unknown bias in the power spectrum). As described in Section 5.2, the Elastic Net reduces overfitting by adding a prior on the template coefficients to reduce the number of templates used.

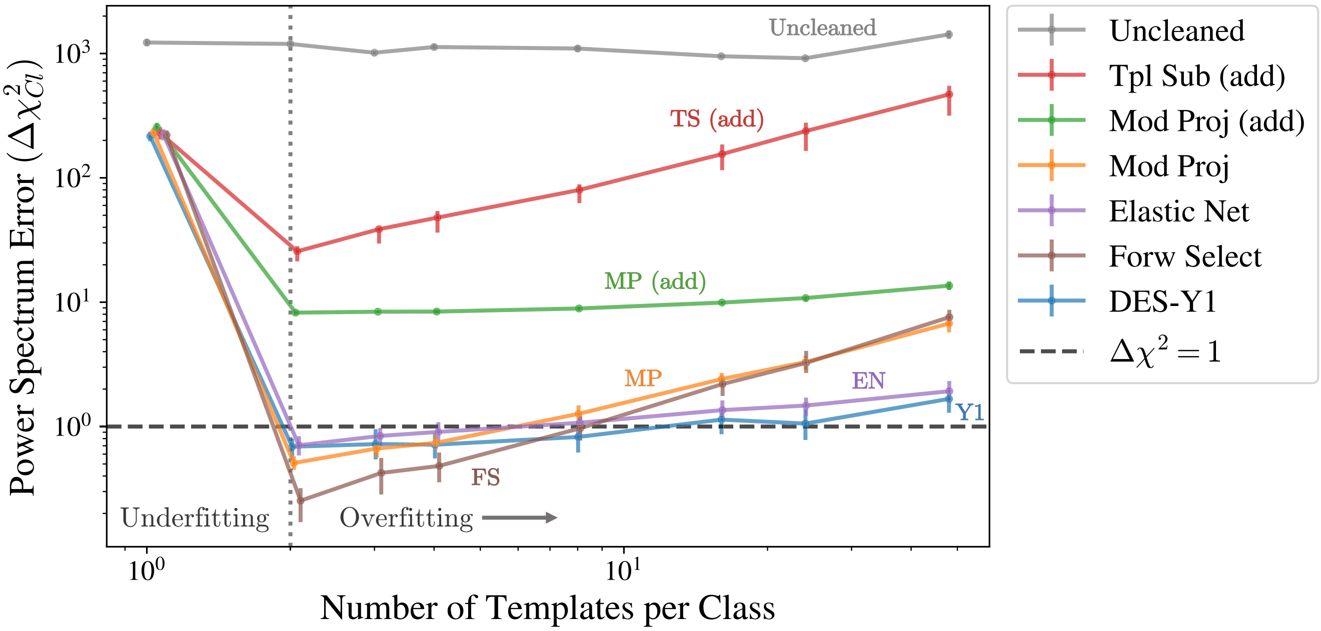

Figure 10. Bias in the angular power spectrum, $\Delta\chi^2_{C_\ell}$, as a function of the number of templates fit to the map. We consider Gaussian templates which have a correlation of $\rho_{\rm tpl} = 0.2$ within each of the four template classes, as defined in Section 6.1. We generate two template realizations per class with which to contaminate each signal map ($N_{\rm sys} = 2$, denoted by the vertical dotted line). The templates used to perform the cleaning vary from one to 48 for each type of template spectrum, for a total of four to 196 templates, with new realizations generated for each $N_{\rm tpl}$ (this is the source of the noise in the 'Uncleaned' line). The TS, MP, and Forward Selection methods are all mildly susceptible to overfitting – signaled by the increase in $\Delta\chi^2_{C_\ell}$ for $N_{\rm tpl} > 2$ – though only TS to a degree where it overcomes the penalty for neglecting a contaminating template ($N_{\rm tpl} = 1$). For the *additive* MP method, $\Delta\chi^2_{C_\ell}$ is dominated by the bias from not addressing the multiplicative contribution to the power spectrum (see Fig. 9, right-hand panel), while the other methods are dominated by increased variance from chance correlations. The bias from failing to correct for the multiplicative term dominates even when fitting for ~200 templates. The DES-Y1 and Elastic Net display a lesser dependence on $N_{\rm tpl}$, and so are more robust to overfitting. See Section 7.2 for details.

While each of the methods addresses overfitting in its own way, the library of templates fed to them has in most cases already been narrowed from a much larger set of possible templates through decisions made by researchers. For example, almost all modern surveys observe any given patch of sky multiple times, resulting in multiple values for each observing condition for every pixel. To produce a scalar template map requires compressing these values into a summary statistic and, as it isn't known a priori which statistic will best capture systematic contamination of the data, multiple statistics may be computed, each corresponding to its own template (see e.g. Leistedt et al. 2016). If just one statistic (such as the mean) is chosen as representative as is often done, there is the very real risk of discarding potential templates that more accurately capture the contamination, resulting in residual contamination, or underfitting.

Therefore, one of the key performance metrics for these methods is their ability to handle increasing numbers of non-contaminating templates without degrading map or power spectrum estimates, and so simultaneously mitigate the risks of under and overfitting.

To characterize the error in the reconstructed angular power spectrum, we use the sum of squared errors between the true and reconstructed power spectra, normalized by sample variance:

$$\Delta\chi^2_{C_\ell} = \sum_{z\rm bins} \sum_{\ell=\ell_{\rm min}}^{350} \frac{\left(\tilde{C}^{\rm est}_\ell(z) - \tilde{C}^{ss}_\ell(z)\right)^2}{\sigma^2_{C^{ss}_\ell(z)}}, \qquad (72)$$

where $\ell_{\rm min} = 2$, except for TS where $\ell_{\rm min} = {\rm Ceil}[(N_{\rm tpl} - 1)/2]$, since with $N_{\rm tpl}$ templates, all signal is removed for $\ell \le (N_{\rm tpl} - 1)/2$.

In Fig. 10, we show $\Delta\chi^2_{C_\ell}$ as a function of the number of templates used to clean the maps (Fig. E1 shows the same plot for map-level statistics, which demonstrate very similar behaviour to $\Delta\chi^2_{C_\ell}$). We generate two template realizations per class with which to contaminate each signal map, and vary the number of templates used to perform the cleaning from one to 24 for *each template class*. The true contaminants are always 'selected first', such that $N_{\rm tpl} = N_{\rm sys}$ = 2 represents correctly fitting for the two contaminating templates from each class (vertical dotted line), whereas $N_{\rm tpl} > 2$ indicates the penalty for overfitting of non-contaminating templates. The error bars come from many signal realizations for the same template maps, and different template and signal map realizations are used for each value of $N_{\rm tpl}$.

Fig. 10 demonstrates that all methods are susceptible to overfitting, as indicated by the fact that $\Delta\chi^2_{C_\ell}$ increases for $N_{\rm tpl} > 2$, but that some are more susceptible than others. TS and additive MP are the worst-performing methods with $\Delta\chi^2_{C_\ell} \gtrsim 10$ for all cases, with TS showing a strong dependence on $N_{\rm tpl}$. Multiplicative MP and Forward Selection display approximately the same $\Delta\chi^2_{C_\ell} \tilde{\propto} N_{\rm tpl}$ scaling as TS, whereas The Elastic Net and DES-Y1 methods show a much weaker scaling, indicating that they are much more robust to a larger number of templates.

The trend for *additive* MP method indicates the importance of the multiplicative correction. Here, the error in the power spectrum does not scale with $N_{\rm tpl}$ as strongly as that of TS or the multiplicative MP method because it is dominated by the bias from not addressing the multiplicative contribution to the power spectrum (see Fig. 9, right-hand panel), not the increased variance from a larger number of templates. The bias from failing to correct for the multiplicative term dominates the additive MP error even when overfitting by ~200 templates (or equivalently, roughly 19 templates to quadratic order).

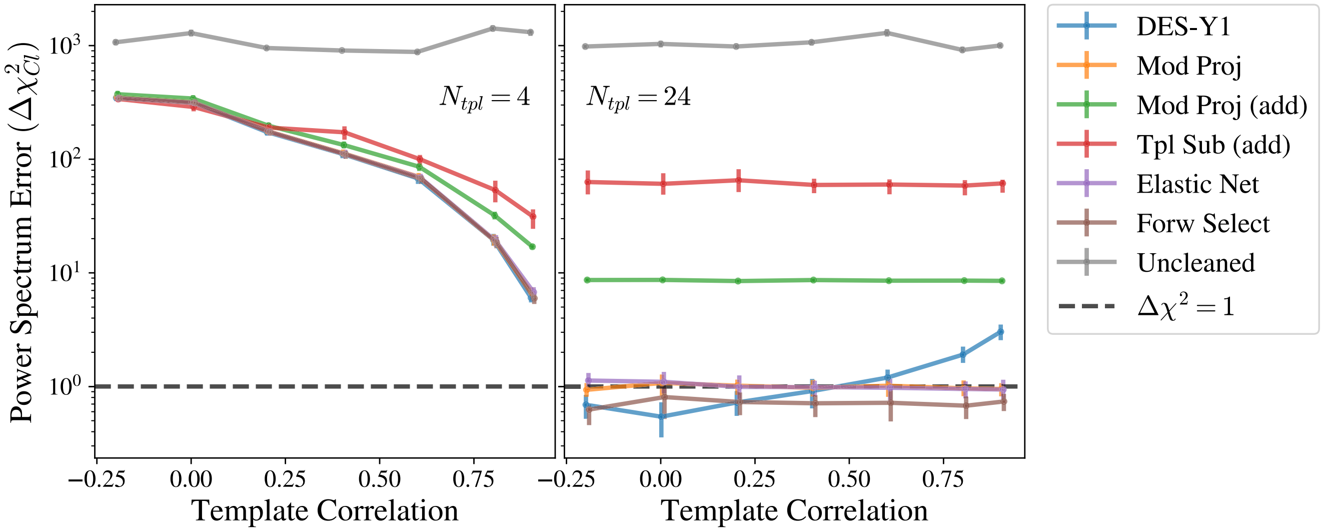

Figure 11. Error in the angular power spectrum, $\Delta\chi^2_{C_\ell}$, as a function of the level of cross-correlation imposed between the templates within the same class. We assumed contamination from two realizations from each of the four classes (i.e. $N_{\rm sys} = 8$). The left-hand panel assumes cleaning with only one of the contaminating templates from each class, while in the right-hand panel we clean for four templates from each class, including the contaminating ones. Note that in the case where template correlation $\rho_{\rm tpl} \rightarrow 1$, the two templates are identical and it is equivalent to cleaning only for one contaminating templates, an ideal scenario. In the right-hand panel, we see that while the DES-Y1 outperforms others when templates are completely orthogonal, it suffers as the level of correlation between templates increases. The Elastic Net method mitigates this problem.

Were the plot to continue to the right, we would expect the error to begin to scale similarly to the other MP and TS methods.[21]

Another key point is that for all cases except TS, the penalty for overfitting is dwarfed by the penalty for neglecting contaminating templates ($N_{\rm tpl} = 1$ on the x-axis). This suggests that the researchers should err on the side of overfitting, rather than risk removing contaminating templates from the cleaning library. This is especially true if using a method that is more robust to overfitting, such as the Elastic Net or DES-Y1 method. In sum, the DES-Y1, Forward Selection, MP with multiplicative correction, and Elastic Net methods all perform very well relative to the uncleaned case, with the Elastic Net and DES-Y1 methods being most robust to overfitting and achieving the best performance with $\Delta\chi^2_{C_\ell} \simeq 1$ even when $N_{\rm tpl} \gg N_{\rm sys}$.

7.3 Impact of correlated templates

Real templates often have groups of templates that are highly similar to one another in their spectral behaviour and/or in their correlation to one another, which we have modeled here as different template classes. The same tracer/property measured in different wavelength bands, or different summary statistics (e.g. the mean versus median) for the same tracer in a multiepoch survey are both common examples that can result in very similar templates. We wish to investigate the impact of selecting a non-optimal template for cleaning, which only partially describes the true systematic. This could be either through the choice of a non-optimal summary statistic, or through the a priori choice of a ‘representative’ template from a group of similar templates in order to mitigate the risk of overfitting, as is commonly done in current surveys.

We test this by cleaning with sets of templates that have varying levels of within-class correlation. For each template class (corresponding to one of the spectra listed in Section 6.1) we use `Healpy.synfast` to generate template realizations with off-diagonal covariance terms between templates *i* and *j* of

$$C^{ij}_\ell = \begin{cases} \rho_{\rm tpl}\sqrt{C^{ii}_\ell C^{jj}_\ell}, & \text{if } i \text{ and } j \text{ in same class} \\ 0, & \text{if } i \text{ and } j \text{ in different classes} \end{cases} .$$

We only use the first four classes from Section 6.1, which are defined by their spectrum and from which we can generate multiple Gaussian realizations with defined levels of cross-correlation.

Fig. 11 shows the performance of the methods when the within-class correlation between templates is varied. We again consider the case of two contaminating systematics from each of the four Gaussian template classes. The left-hand panel shows the case where for each class we have chosen only one of the templates to clean with, deeming it ’representative’ of the template group. As within-class correlation between the systematics increases, the cleaning templates are more representative and can increasingly remove more of the unaccounted for contamination. At $\rho_{\rm tpl} = 0.9$, the multiplicative methods are able to reduce the error to $\Delta\chi^2_{C_\ell} \sim 6$ compared to $\Delta\chi^2_{C_\ell} \sim 300$ for the uncorrelated case.

Despite the additional freedom of the TS method to fit multipoles independently, it does not do a better job than the other methods of correcting for the ’unknown’ systematics. The multiplicative methods have almost identical performance, with the dominant contributions to residual errors in the power spectrum resulting from the unaccounted for systematics and, to a lesser extent, failing to treat the multiplicative term of the contamination.

The right-hand panel in Fig. 11 illustrates the other approach of including many possible templates rather than preselecting a few: we use six cleaning templates from each class: the two true systematics, plus four more that are uncontaminating, for a total of 24. We find that with the exception of the DES-Y1 method, performance

[21] The multiplicative bias is not the dominant contribution for TS because its *effective* number of templates is much larger, since it performs N_ℓ regressions for each template.

of the methods is largely independent of the correlation between templates.[22]

Comparing the panels, even if using a high threshold of similarity of $\rho_{\rm tpl} = 0.9$ to discard templates, significantly more error is introduced through neglecting a contaminating template than through overfitting, so it is better to not preselect templates solely on the basis of similarity to others and instead err on the side of too many templates rather than too few. TS is the one exception to this, where each additional template results in N_ℓ additional fits. While the additional freedom does not substantially protect against unknown systematics, it does result in a much steeper penalty for overfitting from higher $N_{\rm tpl}$.

7.4 Extensions

By interpreting current LSS systematics cleaning methods in the context of regression, we have facilitated their comparison and interpretation, as well as motivated several possible extensions to them. We have explored some of these extensions in this work, such as the Elastic Net method in Section 5.2, and the use of the leverage statistic to predict overdensity errors and aid mask creation, but with the extensive body of regression methods, there are many more that we must leave to future work. For example, one promising avenue for regression methods that use a threshold for template selection would be to motivate that threshold by controlling the ratio of Type I (false correction) to Type II (false omission) errors in the selection process via the False Discovery Rate (Benjamini & Hochberg 1995), based on the relative impact of each type of error on the analysis.

We have noted individually multiple cases where the assumptions made by the methods do not hold and how they might be improved. A full treatment of these effects is beyond the scope of this paper and would include the full non-Gaussian likelihood of $P(d_{\rm obs}|\hat{f}_{\rm sys})$, including contributions from systematics, but as we show in Appendices A and B, the corrections from these are minor compared to the methodological differences and the improvements we suggest. Generalized linear models may be a promising compromise for future mitigation routines, preserving off-the-shelf implementation and diagnostic tools, while providing greater specificity for the likelihood and relaxing some of the tacit assumptions of MP and OLS regression.

The methods presented here are general enough to be applicable in any situation where one has an external prediction (template) for systematic contamination of observational data, and is equally applicable to spin-2 fields. The insights gained can be used to further extend linear models like the ones in this work, or inform the formulation of non-linear contamination models, non-parametric methods, or machine learning approaches such as that of Rezaie et al. (2020).

8 SUMMARY OF METHODS

Here, we summarize our findings about the performance of systematic-cleaning methods.

(i) DES-Y1 method: The most complicated method of the ones we studied, the DES-Y1 method resulted in some of the lowest biases in the cleaned maps. It usefully includes prior information about the covariance between pixels in the fitting procedure, albeit in a coarse way. However, it is also somewhat complicated to implement, as it requires a large number of parameter choices on the part of the researcher (binning number and procedure, significance statistic and threshold, power spectrum prior) and the generation of realistic mocks. We observed some degradation of its performance as the correlation between templates increased. It is one of the two methods most robust to overfitting when using a large library of templates that are not actually contaminating the data (the other being Elastic Net).

(ii) MP: The standard pseudo-C_ℓ MP method, as introduced in Elsner et al. (2017) and implemented in `NaMaster`(Alonso et al. 2019). We showed that it is equivalent to removing the result of an OLS regression of the observed data on to the template maps (thus providing a map estimate), with an additional step to debias the power spectrum. This removes most of the contamination present, but can be simply adapted to, and significantly improved by, treating the multiplicative component of contamination instead of just the additive term. We demonstrate how to do this in Section 4.2. In all cases we studied, the error from not correcting the multiplicative term dominated over error induced from overfitting – as Fig. 10 illustrates, in the *ideal* case where our templates exactly matched the systematics, not treating the multiplicative term introduced as much error as using ∼30 times more templates than systematics in the cleaning procedure.

(iii) TS: Equivalent to performing an individual OLS regression at each multipole, resulting in large variance and significant loss of signal from overfitting. As a result, it does not reconstruct maps well and generally performs most poorly in all of our tests. However, our implementation is a limiting case, where each harmonic from each template is allowed to contaminate independently, in contrast to MP where all modes contaminate identically. The work here should make it straightforward to construct a hybrid method where all modes contribute identically like in MP (as is physically motivated) and hence have small variance, but where certain modes are prioritized for cleaning, based on the analysis case.

(iv) Iterative forward selection: This is a method we propose, which is a much simpler version of the DES-Y1 method that requires only a single tunable parameter (a significance threshold) and no mocks. We found that it produces excellent results and is robust to correlation between templates, but is not as robust to overfitting, displaying the same dependence of roughly $\Delta\chi^2_{C_\ell} \propto N_{\rm tpl}$ as the MP and TS methods.

(v) Cross-validated elastic net: A cleaning method we introduce, which we find has the best overall performance, being consistently low error and robust to overfitting. It is equivalent to MP, but with the amplitude of contamination for each template having a mixed Gaussian/Laplace prior applied to encourage sparsity and thus automatically select the important templates. The 'priors' are not strictly such in a Bayesian sense, as their strengths are determined by the data through cross-validation. It is easy to implement using out-of-the-box software and does not require a user-defined prior for the power spectrum or debiasing step, providing the best balance of performance, ease of implementation, interpretability and robustness.

9 CONCLUSIONS

In this paper, we carried out a broad comparison of methods used to remove astrophysical, atmospheric, and instrumental systematic errors that affect galaxy-clustering measurements. We have gener-

[22] We found this to be true for both map-level and 2-pt reconstruction statistics, though we only show the latter here. It is not obvious from the outset that this would be the case – Forward Selection methods are often criticized for being less reliable when predictors are correlated, though this is in the context of the more typical regression scenario where it is the predictors themselves that are of interest, as opposed to the residuals which is our focus here. The source of the dependence of the DES-Y1 results on $\rho_{\rm tpl}$ is not entirely clear, but our investigations found it to be mildly impacted by both binning choices and the total monopole of systematic maps.

alized previous work by (1) showing how different methods can be interpreted under a common regression framework, (2) jointly assessing the robustness of methods on simulated data, (3) investigating the reconstruction fidelity of LSS *map(s)*, rather than just their clustering statistics, as the maps are useful points of departure for numerous other analyses (e.g. summary statistics beyond the power spectrum, cross-correlations, searches for signatures of dark matter or exotic new physics); and (4) proposing improvements to current methods, as well as new, hybrid and efficient methods for the systematics cleaning.

We employed a simple and general model for systematics, given in equation (1), which allows for spatially varying multiplicative and additive systematic errors with a range of clustering properties to any generic cosmological field. Equipped with that model, we defined a testing procedure that attempts to mimic real-world conditions for LSS surveys, where the true galaxy map is contaminated with an unknown set of systematics and a set of known templates is used to model and correct for the contamination. Given our methodology (pictorially described in Fig. 1) and a set of assumptions about the fiducial DES Y5-like survey used to generate the maps, we studied the performance of the systematics-cleaning methods under different conditions.

We showed that both TS and MP, while developed independently, can be interpreted through a regression framework where the signal of interest corresponds to the noise term of a regression model. This allowed us to straightforwardly apply known statistical results and techniques to these methods. We used this to adapt additive methods to account for multiplicative errors (Fig. 5), and identify potentially highly contaminated map pixels as a function of their 'leverage' (Fig. 7), while opening up avenues for further improvement. One such avenue we touched on was to optimize MP (or other regression methods) by prewhitening the maps in harmonic space. Recognizing that the noise of the regression is the clustering signal itself, we proposed that the maps could be efficiently and optimally inverse-variance weighted in harmonic space, where the clustering signal is diagonal. This is equivalent to accounting for the off-diagonal pixel covariance in the pixel-based regression methods, which is rarely done for tractability reasons (but see Wagoner et al. 2020 for one approach). We found this to improve results (Fig. B1), but be subdominant to the multiplicative correction and differences between the cleaning methods.

We introduced two new methods for cleaning: (1) the 'Forward Selection' method, which is a greatly simplified version of the DES-Y1 method that achieves similar performance albeit being less robust to a large number of templates; and (2) the 'Elastic Net' method, a simple out-of-the-box method that implements MP, but which automatically selects important templates. We found that the Elastic Net method is very robust, with strong performance even when there is a large number of templates (Fig. 10) or templates are highly correlated (Fig. 11); both are cases where other methods display weaknesses. This method is very easy to implement, and we recommend it for future surveys.

On the whole, we found that all of the methods perform quite well, dramatically improving the chi-squared difference between the cleaned and true (uncontaminated) angular power spectrum. At the map level, TS was the only method that did not significantly reduce the RMS overdensity error across pixels (Figs 7 and E1), and so we do not recommend the version implemented here for map reconstruction. Once we adopted only the algorithms that take into account both additive and multiplicative errors, all of the methods improved $\Delta\chi^2_{C_\ell}$ by three orders of magnitude relative to the uncleaned case. Moreover, overfitting did not lead to large degradation in the reconstructed power spectra (see Fig. 10), which is encouraging. Finally, we found that the performance of the various systematics-cleaning methods is very weakly dependent on the level of cross-correlation between the template maps used for the cleaning, with the DES-Y1 method being mildly more susceptible.

We end with several recommendations based on this work as follows:

(I) Current and future cleaning methods should account for multiplicative contamination. 'Weights' methods like the DES-Y1 method already do this and other methods like (Pseudo-C_ℓ) MP can easily do so via equations (46)–(47).

(II) Cleaning methods based on a single OLS regression are equivalent to (Pseudo-C_ℓ) MP and so should debias inferred two-point functions accordingly. For more complicated methods where the bias cannot be determined analytically, it can be characterized and removed through performing cleaning on mock catalogues.

(III) Analyses should err on the side of overfitting rather than underfitting for templates, as the error from the former tends to be small provided templates do not contain any more information about the true density field than would occur by chance. Researchers should avoid arbitrarily removing templates from the library prior to cleaning based solely on their similarity to other templates. Larger template libraries result in increased variance of the map and power spectrum estimators, especially with the very large number of templates that will be available to future surveys. Therefore:

(IV) In scenarios where a very large template library is available, the data itself should be used to select a subset for cleaning. Among the methods that we studied, this is accomplished by either a DES-Y1 type method or the Elastic Net with cross-validation. Both show good robustness, and the latter is simple to implement with common software. The theoretical connections we have made between methods should make alternative template selection routines such as those in Leistedt et al. (2016) and Rezaie et al. (2020) simple to adapt and implement.

(V) The cleaning methods used thus far can – and should – be viewed in the context of regression, with the estimated overdensity field corresponding to the regression residuals. Researchers should make use of the powerful suite of existing tools and diagnostic measures to assess the validity of regression models when cleaning LSS data (e.g. leverage for outlier detection, Q–Q plots, partial regression/residual plots) and to aid mask creation. This is applicable to all methods studied in this paper.

ACKNOWLEDGEMENTS

We thank Boris Leistedt and Jessie Muir for helpful comments and encouragement. The authors have been supported by DOE under Contract No. DE-FG02-95ER40899. DH has also been supported by NASA under Contract No. 14-ATP14-0005, and N W has been supported through a Leinweber Graduate Fellowship.

DATA AVAILABILITY

The data sets generated for this study are available from the corresponding author upon request.

REFERENCES

Abbott T. M. C. et al., 2018, Phys. Rev. D, 98, 043526
Abbott T. et al., 2019, Phys. Rev. Lett., 122, 171301
Abergel A. et al., 2014, Astron. Astrophys., 571, A11

Agarwal N., Ho S., Shandera S., 2014a, JCAP, 1402, 038

Agarwal N. et al., 2014b, J. Cosmol. Astropart. Phys., 1404, 007

Aihara H. et al., 2018, Publ. Astron. Soc. Japan, 70, S4

Alam S. et al., 2017, MNRAS, 470, 2617

Allys E., Marchand T., Cardoso J. F., Villaescusa-Navarro F., Ho S., Mallat S., 2020, Phys. Rev. D, 102, 103506

Alonso D., Sanchez J., Slosar A., 2019, MNRAS, 484, 4127

Amendola L. et al., 2018, Living Rev. Rel., 21

Anderson L. et al., 2014, MNRAS, 441, 24

Ata M. et al., 2018, MNRAS, 473, 4773

Awan H., Gawiser E., 2019, ApJ, 890, 32

Awan H. et al., 2016, ApJ, 829, 50

Baugh C. M., 1996, MNRAS, 280, 267

Baugh C. M., Efstathiou G., 1993, MNRAS, 265, 145

Bautista J. E. et al., 2018, ApJ, 863, 110

Benjamini Y., Hochberg Y., 1995, J. R. Astron. Soc B, 57, 289

Biswas R., Alizadeh E., Wandelt B. D., 2010, Phys. Rev. D, 82, 023002

Blake C., 2019, MNRAS, 489, 153

Bobin J., Starck J., Fadili J., Moudden Y., 2007, IEEE Trans. Image Process., 16, 2662

Bobin J., Starck J.-L., Sureau F., Basak S., 2013, A&A, 550, A73

Bos E. G. P., van de Weygaert R., Dolag K., Pettorino V., 2012, MNRAS, 426, 440

Cheng S., Ting Y.-S., Ménard B., Bruna J., 2020, MNRAS, 499, 5902

Coles P., Jones B., 1991, MNRAS, 248, 1

LSST Dark Energy Science Collaboration, 2012, preprint (arXiv:1211.0310)

Colless M. et al., 2001, MNRAS, 328, 1039

Connolly A. J. et al., 2002, ApJ, 579, 42

Cook R. D., 1977, Technometrics, 19, 15

Cook R. D., 1979, J. Am. Stat. Assoc., 74, 169

Cooray A., Hu W., 2001, ApJ, 548, 7

Crocce M. et al., 2016, MNRAS, 455, 4301

Davis M., Peebles P. J. E., 1983, ApJ, 267, 465

Dawson K. S. et al., 2013, AJ, 145, 10

de Jong J. T. A. et al., 2015, A&A, 582, A62

de Lapparent V., Geller M. J., Huchra J. P., 1986, ApJ, 302, L1

Delubac T. et al., 2016, MNRAS, 465, 1831

DESI Collaboration et al., 2016, preprint (arXiv:1611.00036)

Dodelson S., Gaztañaga E., 2000, MNRAS, 312, 774

Dodelson S. et al., 2002, ApJ, 572, 140

Doré O. et al., 2014, preprint (arXiv:1412.4872)

Drinkwater M. J., Jurek R. J., Blake C., Woods D., Pimbblet K. A. et al., 2010, MNRAS, 401, 1429

Eisenstein D. J., Zaldarriaga M., 2001, ApJ, 546, 2

Elsner F., Leistedt B., Peiris H. V., 2016, MNRAS, 456, 2095

Elsner F., Leistedt B., Peiris H. V., 2017, MNRAS, 465, 1847

Elvin-Poole J. et al., 2018, Phys. Rev. D, 98, 042006

Feldman H. A., Kaiser N., Peacock J. A., 1994, ApJ, 426, 23

Feldman H. A., Frieman J. A., Fry J. N., Scoccimarro R., 2001, Phys. Rev. Lett., 86, 1434

Fisher K. B., Davis M., Strauss M. A., Yahil A., Huchra J. P., 1993a, ApJ, 402, 42

Fisher K. B., Davis M., Strauss M. A., Yahil A., Huchra J. P., 1993b, ApJ, 402, 42

Foster D. P., George E. I., 1994, Ann. Statist., 22, 1947

Friedrich O. et al., 2018, Phys. Rev. D, 98

Fry J. N., 1986, ApJ, 306, 358

Giannantonio T., Ross A. J., Percival W. J., Crittenden R., Bacher D. et al., 2014, Phys. Rev. D, 89, 023511

Gil-Marín H., Noreña J., Verde L., Percival W. J., Wagner C., Manera M., Schneider D. P., 2015, MNRAS, 451, 539

Gorski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelman M., 2005, ApJ, 622, 759

Goto T., Szapudi I., Granett B. R., 2012, MNRAS, 422, L77

Gott J., Richard I., Weinberg D. H., Melott A. L., 1987, ApJ, 319, 1

Hauser M. G., Peebles P. J. E., 1973, ApJ, 185, 757

Hernández-Monteagudo C. et al., 2014, MNRAS, 438, 1724

Hikage C., Komatsu E., Matsubara T., 2006, ApJ, 653, 11

Hilbert S., Hartlap J., Schneider P., 2011, A&A, 536, A85

Hivon E., Gorski K. M., Netterfield C. B., Crill B. P., Prunet S., Hansen F., 2002, ApJ, 567, 2

Ho S. et al., 2012, ApJ, 761, 14

Ho S. et al., 2013, J. Cosmol. Astropart. Phys., 05, 040

Huterer D., Knox L., Nichol R. C., 2001, ApJ, 555, 547

Huterer D., Cunha C. E., Fang W., 2013, MNRAS, 432, 2945

Jain B., Van Waerbeke L., 2000, ApJ, 530, L1

Jasche J., Kitaura F. S., 2010, MNRAS, 407, 29

Jasche J., Wandelt B. D., 2013, MNRAS, 432, 894

Jasche J., Leclercq F., Wandelt B. D., 2015, J. Cosmol. Astropart. Phys, 1501, 036

Jeffrey N. et al., 2018, MNRAS, 479, 2871

Kalus B., Percival W. J., Bacon D. J., Samushia L., 2016, MNRAS, 463, 467

Kalus B., Percival W. J., Bacon D. J., Mueller E. M., Samushia L., Verde L., Ross A. J., Bernal J. L., 2019, MNRAS, 482, 453

Kitanidis E. et al., 2020, MNRAS, 496, 2262

Kitaura F.-S., 2013, MNRAS, 429, 84

Kong H. et al., 2020, MNRAS, 499, 3943

Kratochvil J. M., Lim E. A., Wang S., Haiman Z., May M., Huffenberger K., 2012, Phys. Rev. D, 85, 103513

Lanusse F., Starck J.-L., Leonard A., Pires S., 2016, A&A, 591, A2

Laurent P. et al., 2017, J. Cosmol. Astropart. Phys, 1707, 017

Leclercq F., Jasche J., Sutter P. M., Hamaus N., Wandelt B., 2015, J. Cosmol. Astropart. Phys, 2015, 047

Leistedt B., Peiris H. V., 2014, MNRAS, 444, 2

Leistedt B., Peiris H. V., Mortlock D. J., Benoit-Lévy A., Pontzen A., 2013, MNRAS, 435, 1857

Leistedt B. et al., 2016, ApJS, 226, 24

Leonard A., Lanusse F., Starck J.-L., 2014, MNRAS, 440, 1281–1294

Lesgourgues J., 2011, preprint (arXiv:1104.2932)

Liu J., Petri A., Haiman Z., Hui L., Kratochvil J. M., May M., 2015, Phys. Rev. D, 91, 063507

Lumsden S. L., Nichol R. C., Collins C. A., Guzzo L., 1992, MNRAS, 258, 1

Marian L., Hilbert S., Smith R. E., Schneider P., Desjacques V., 2011, ApJ, 728, L13

Marín F. A. et al., 2013, MNRAS, 432, 2654

Mawdsley B. et al., 2020, MNRAS, 493, 5662

Miller C. J., Batuski D. J., 2001, ApJ, 551, 635

Modi C., White M., Slosar A., Castorina E., 2019, JCAP, 1911, 023

Morrison C. B., Hildebrandt H., 2015, MNRAS, 454, 3121–3133

Muir J., Huterer D., 2016, Phys. Rev. D, 94, 043503

Munshi D., van Waerbeke L., Smidt J., Coles P., 2012, MNRAS, 419, 536

Nadathur S., Hotchkiss S., 2015, MNRAS, 454, 2228

Nicola A., Refregier A., Amara A., 2016, Phys. Rev. D, 94, 083517

Nicola A. et al., 2019, J. Cosmol. Astropart. Phys, 03, 044

Peacock J. A., Nicholson D., 1991, MNRAS, 253, 307

Peacock J. A. et al., 2001, Nature, 410, 169

Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825

Peebles P. J. E., Groth E. J., 1975, ApJ, 196, 1

Peebles P. J. E., Hauser M. G., 1974, ApJS, 28, 19

Percival W. J., 2018, preprint (arXiv:1810.04263)

Percival W. J. et al., 2001, MNRAS, 327, 1297

Petri A., Liu J., Haiman Z., May M., Hui L., Kratochvil J. M., 2015, Phys. Rev. D, 91, 103511

Philcox O. H. E., Massara E., Spergel D. N., 2020, Phys. Rev. D, 102, 043516

Pisani A., Sutter P. M., Hamaus N., Alizadeh E., Biswas R., Wandelt B. D., Hirata C. M., 2015, Phys. Rev. D, 92, 083531

Porqueres N., Kodi Ramanah D., Jasche J., Lavaux G., 2019, A&A, 624, A115

Prakash A. et al., 2016, ApJS, 224, 34

Pullen A. R., Hirata C. M., 2013, Publ. Astron. Soc. Pac., 125, 705

Rezaie M., Seo H., Ross A., Bunescu R. C., 2020, MNRAS, 495, 1613

Ross A. J. et al., 2011, MNRAS, 417, 1350

Ross A. J. et al., 2017, MNRAS, 464, 1168

Ross A. J. et al., 2020, MNRAS, 498, 2354

Rybicki G. B., Press W. H., 1992, ApJ, 398, 169

Rykoff E. S., Rozo E., Keisler R., 2015, preprint (arXiv:1509.00870)

Saunders W., Rowan-Robinson M., Lawrence A., 1992, MNRAS, 258, 134

Scoccimarro R., Feldman H. A., Fry J. N., Frieman J. A., 2001, ApJ, 546, 652

Scranton R. et al., 2002, ApJ, 579, 48

Sefusatti E., Crocce M., Pueblas S., Scoccimarro R., 2006, Phys. Rev. D, 74, 023522

Shafer D. L., Huterer D., 2015, MNRAS, 447, 2961

Sheth R. K., 2005, MNRAS, 364, 796

Slepian Z., Eisenstein D. J., 2015, MNRAS, 454, 4142

Slepian Z. et al., 2015, MNRAS, 468, 1070

Spergel D. et al., 2013, preprint (arXiv:1305.5422)

Starck J.-L., Donoho D. L., Fadili M. J., Rassat A., 2013, A&A, 552, A133

Suchyta E. et al., 2016, MNRAS, 457, 786

Taruya A., Takada M., Hamana T., Kayo I., Futamase T., 2002, ApJ, 571, 638–653

Tegmark M. et al., 2004, ApJ, 606, 702

Verde L. et al., 2002, MNRAS, 335, 432

Vogeley M. S., 1998, Toward High-Precision Measures of Large-Scale Structure. Kluwer, Dordrecht, p. 395

Wagner-Carena S., Hopkins M., Rivero A. D., Dvorkin C., 2019, MNRAS, 494, 1507

Wagoner E. L., Rozo E., Fang X., 2020, MNRAS, preprint (arXiv:2009.10854)

Wang H., Mo H. J., Yang X., Jing Y. P., Lin W. P., 2014, ApJ, 794, 94

Wang H. et al., 2016, ApJ, 831, 164

Weaverdyck N., Muir J., Huterer D., 2018, Phys. Rev. D, 97, 043515

White S. D. M., 1979, MNRAS, 186, 145

White M., 2016, J. Cosmol. Astropart. Phys., 2016, 057

White M., Padmanabhan N., 2009, MNRAS, 395, 2381

Xavier H. S., Abdalla F. B., Joachimi B., 2016, MNRAS, 459, 3693

York D. G. et al., 2000, AJ, 120, 1579

Zonca A., Singer L., Lenz D., Reinecke M., Rosset C., Hivon E., Gorski K., 2019, J. Open Source Soft., 4, 1298

Zou H., Hastie T., 2005, J. R. Statistical Soc. B , 67, 301

APPENDIX A: LOGNORMAL VERSUS GAUSSIAN SIGNAL MAPS

While the methods presented here are quite general for any case where systematic contamination can be traced using a template, we have specifically worked in the context of galaxy clustering. In this case, the signal map s that we are attempting to model is the galaxy overdensity δ, which is subject to the constraint $\delta > -1$ (as is the case for any overdensity statistic). Thus our assumption that s is Gaussian breaks down at low redshift and at small scales, when $|\delta|$ can be large.

It is well known that galaxy and shear overdensities are better approximated by a lognormal distribution (see e.g. Coles & Jones 1991; Taruya et al. 2002; Hilbert, Hartlap & Schneider 2011; Xavier, Abdalla & Joachimi 2016), so we run the methods on a series of lognormal maps to see if the relative performance of the methods changes.

We generate 100 Gaussian signal realizations $s_G(\hat{\boldsymbol{n}})$ of the lowest redshift bin of our fiducial DES survey, for which the cosmological signal will be most non-Gaussian. We generate lognormal versions of these maps by first computing the transformation that achieves zero-mean lognormal overdensity field *in the ensemble* (Hilbert et al. 2011), then centring and scaling so that each realization of the lognormal field has the same mean and variance as its Gaussian counterpart. The two steps correspond to the mathematical operations:

(i) $s'_{LN}(\hat{\boldsymbol{n}}) = \mathrm{e}^{s_G(\hat{\boldsymbol{n}})} - \mathrm{e}^{\mathrm{Var}[s_G(\hat{\boldsymbol{n}})]/2}$

(ii) $s_{LN}(\hat{\boldsymbol{n}}) = \sqrt{\frac{\mathrm{Var}[s_G(\hat{\boldsymbol{n}})]}{\mathrm{Var}[s'_{LN}(\hat{\boldsymbol{n}})]}}\left(s'_{LN}(\hat{\boldsymbol{n}}) - \bar{s}'_{LN}(\hat{\boldsymbol{n}})\right).$

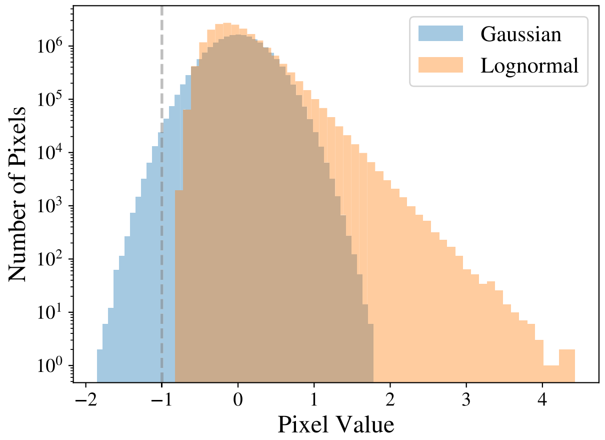

Figure A1. Distribution of pixel overdensities across all 100 realizations of the lognormal (orange) and Gaussian (blue) maps of the galaxy overdensity in the lowest redshift bin of our fiducial DES-like survey. The Gaussian maps contain pixels with $s < -1$, which is non-physical in cases like this where s corresponds to an overdensity.

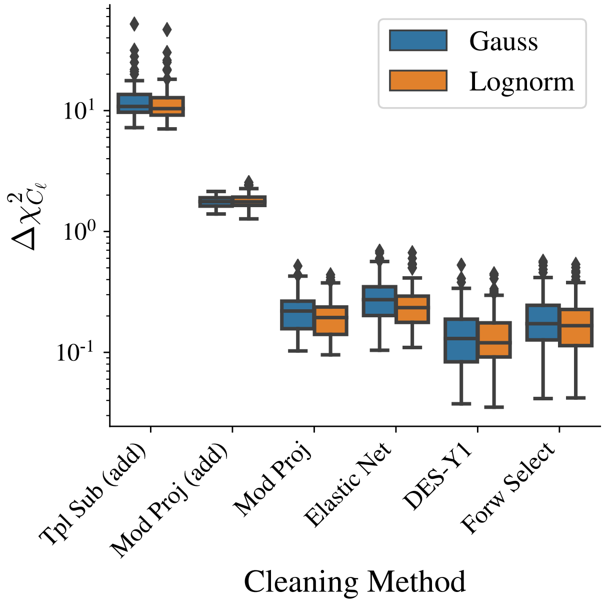

Figure A2. Box plot showing the performance of each cleaning method when using Gaussian (blue, left) versus lognormal (orange, right) signal maps, as measured by $\Delta\chi^2_{C_\ell}$ of the power spectrum. Filled boxes show the 25-50-75 per cent quartiles, with whiskers encompassing the rest of the distribution out to 1.5 × the inter-quartile range. Points beyond this range are indicated by diamonds. Regardless of whether lognormal or Gaussian maps are used, the relative performance of the methods to one another is largely unchanged, and the Gaussian approximation is negligible compared to neglecting the multiplicative correction of Section 4.2.

The resulting lognormal realizations are then of the form

$$s_{\mathrm{LN}}(\hat{\boldsymbol{n}}) = \lambda_1 \mathrm{e}^{s_G}(\hat{\boldsymbol{n}}) - \lambda_0, \qquad \text{(A1)}$$

with scale and shift parameters of $\lambda_1 = 0.9123 \pm 0.0017$ and $\lambda_0 = 0.9697 \pm 0.0017$, respectively, for our lowest redshift bin, which is the most non-Gaussian.

Fig. A1 shows the distribution of pixel overdensities across all realizations of the lognormal and Gaussian signal maps. It is clear that the Gaussian maps contain many pixels with $s < -1$, which is non-physical for our case, where s corresponds to an overdensity. The lognormal maps avoid this problem and are a better approximation of the true overdensity distribution. As we have shown, most of the cleaning methods can be viewed under a regression framework wherein the signal distribution is assumed to be Gaussian, so we investigate whether our comparison of methods changes when using a more realistic lognormal distribution.

Fig. A2 shows the error in the power spectrum reconstruction, given by the $\Delta\chi^2_{C_\ell}$ statistic, for the different methods. We find that while there is some overall shift, using the lognormal signal maps does not change the relative behaviour of the methods; none of them display a unique susceptibility to the assumption of Gaussianity in the signal maps.

APPENDIX B: EFFECT OF PREWHITENING

In their derivation of the bias on the estimated power spectrum after (pseudo-C_ℓ) MP, Elsner et al. (2017) assume that the map d has been decorrelated ('prewhitened') before projecting out the templates. This is quite difficult to do in practice, as it requires the inversion of an $N_{\rm pix} \times N_{\rm pix}$ matrix, the same problem with QML estimators for the power spectrum. Indeed, one of the assumptions of pseudo-C_ℓ estimation is that pixels are uncorrelated (though individual pixels are weighted by an estimate of their inverse noise variance and by the mask, see e.g. Alonso et al. (2019).)

As shown in Section 4.1, however, the dominant 'noise' in our observations is actually our true clustering signal, so a true 'prewhitening' step should more appropriately inverse weight the data by the expected clustering variance. This can be done efficiently in harmonic space when there is no mask, as the clustering signal is diagonal, circumventing the need to invert a large covariance matrix.

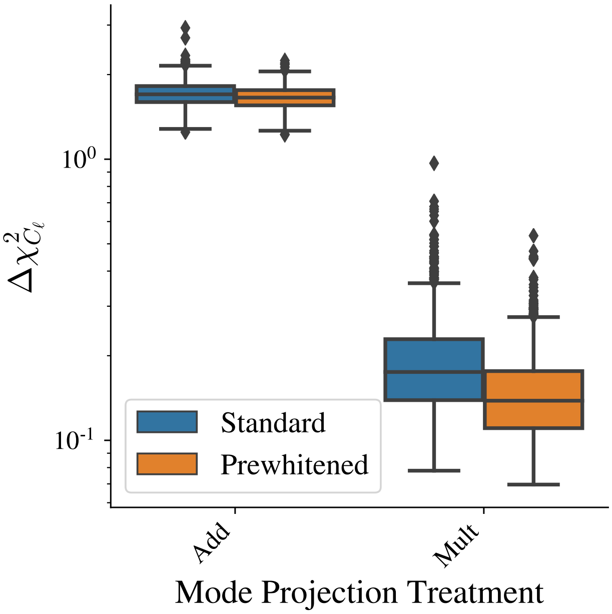

Figure B1. Impact of prewhitening before cleaning with the multiplicative and additive versions of the MP method on 1000 realizations for our fiducial contamination model. The standard MP method assumes a flat power spectrum for the target signal, resulting in a suboptimal estimate of contamination. This can be improved through 'prewhitening' the data vector and templates using a prior power spectrum, which can be shown to be equivalent to a standard weighted regression procedure in harmonic space. There is clear but modest improvement from the standard case (blue) to the nearly optimal, prewhitened case (orange), with the most improvement seen for realizations that have large error. This can be seen by the preferential reduction of extreme points at the high end of the box plots in the prewhitened case (note the log scale).

We can define prewhitened data vectors for our observed overdensity field and templates as

$$(d_{\rm obs})'_{\ell m} = (d_{\rm obs})_{\ell m}/\sqrt{C^{ss}_\ell}, \tag{B1}$$

$$(t_i)'_{\ell m} = (t_i)_{\ell m}/\sqrt{C^{ss}_\ell}, \tag{B2}$$

which results in coefficient estimates of

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{T}'^{\dagger}\boldsymbol{T}')^{-1}\boldsymbol{T}'^{\dagger}\boldsymbol{d}'_{\rm obs}, \tag{B3}$$

where T' is a $N_{\ell m} \times N_{\rm tpl}$ matrix with complex entries defined in equation (B2). We can compute the amplitudes directly with

$$\hat{\alpha} = \frac{\sum_{\ell=0}^{\ell_{\max}}(2\ell+1)\tilde{C}_\ell^{td}/C^{ss}_\ell}{\sum_{\ell=0}^{\ell_{\max}}(2\ell+1)\tilde{C}_\ell^{tt}/C^{ss}_\ell}. \tag{B4}$$

We found that prewhitening improved $\Delta\chi^2_{C_\ell}$ by a mean of ∼0.05 with dispersion 0.08 across the mocks, with similar shifts regardless of whether the multiplicative correction was applied or not. Fig. B1 shows the improvement from the standard case (blue) to the prewhitened case (orange) for both additive and multiplicative MP. While we do not show it, we found that the benefit of prewhitening increased for realizations that had worse power spectrum estimates (higher $\Delta\chi^2_{C_\ell}$), in effect catching and mitigating particularly bad realizations.

In practice, one would either assume a prior power spectrum for prewhitening or compute it iteratively, just as one does for the MP debiasing step, so this could easily be incorporated into existing MP routines such as `NaMaster`. As noted in Section 4.1, since MP is equivalent to regression, this improvement also quantifies the expected level of improvement that would come from accounting for the covariance between pixels in pixel-based regression methods.

Analyses on real data will of course be complicated by the mask, which correlates different multipoles, but this can be addressed by suitable binning of the multipoles. Indeed, the standard pseudo-C_ℓ MP assumes a flat power spectrum and so can be thought of as the limiting case of using only a single bin across multipoles with equal weighting, such that even a rough estimate of the signal power spectrum should offer improvement.

The other methods tested here should benefit similarly from prewhitening, with the possible exception of the DES-Y1 method, which already incorporates an estimate of the covariance of s (which accounts for much of the methods' complexity). The Forward Selection method we presented may be particularly impacted, since the estimated covariance of the fit parameters is underestimated when the pixel covariance is neglected, and this is used for the significance criterion for selecting a template. This could be one reason why the Forward Selection method sometimes failed to reduce all templates to below a significance of $\Delta\chi^2/\Delta\chi^2_0 = 2$ – such a threshold was artificially low compared to what would be expected from random variation.

As noted in equation (51), the prewhitening step in equation (B2) should optimally include contributions from the systematics as well. However as this represents minor perturbations to the major prewhitening correction above and is hence a small 'error on the error', the effects should be small. This is consistent with Elvin-Poole et al. (2018), who found negligible impact on their method from neglecting the additional systematics contribution to their estimated covariance matrices.

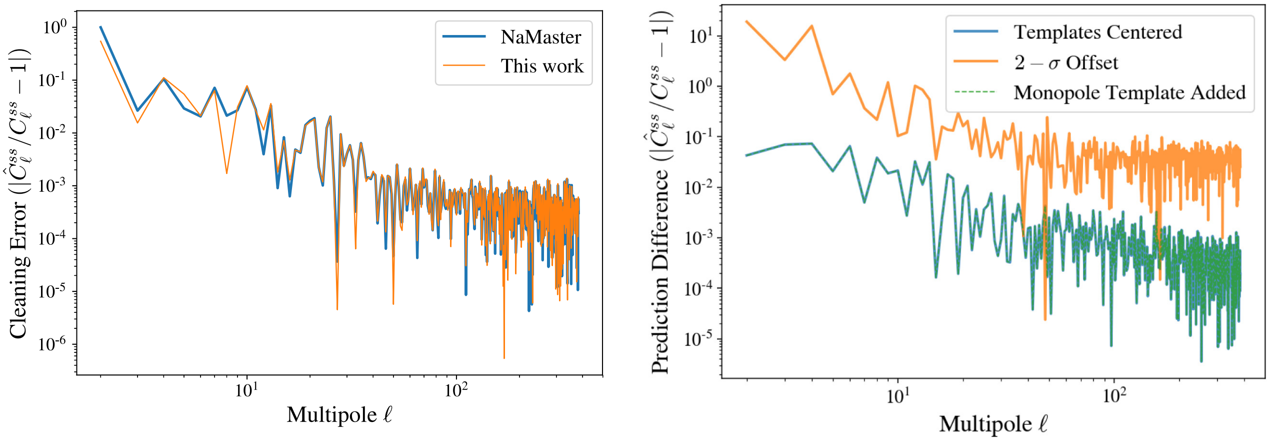

Figure B2. Validation tests of the MP map-cleaning procedure. Left-hand panel: Comparison of MP performance on an additive-only contaminated map (as assumed by the MP method), using `NaMaster` (blue) and our own implementation (orange). The agreement between the two is very good. Right-hand panel: Impact of not precentreing cleaning templates in `NaMaster`. The blue curve indicates the standard use case, where contamination is additive and completely described by the templates, which have been individually centred at zero. If templates are instead centred at another value (here we add a constant $2\sigma_{\mathrm{tpl},\,i}$ offset to each template, where $\sigma_{\mathrm{tpl},\,i}$ is the standard deviation of values in template map i. Adding a monopole template completely mitigates the bias from non-centred templates.

APPENDIX C: COMPARISON WITH `NaMaster`

We have used our own implementation of the MP method and have tested it against that of `NaMaster`, finding good agreement. `NaMaster` computes the power spectra given a set of templates and observations, but does not produce map estimates, so we compare the two implementations using the cleaned power spectrum only. The left-hand panel of Fig. B2 shows the relative error of the estimated power spectrum when cleaned using NaMaster versus our own implementation, using the exact same contaminated map and templates and we find good agreement (this held true for all realizations tested). There is very slight disagreement at larger scales (low ℓ), which may be numerical artifacts from the `Master` (Hivon et al. 2002) algorithm implemented to account for mode coupling on a cut sky being applied to full-sky input maps. Regardless, the deviations between the two are small for $\ell > 2$.

APPENDIX D: ACCOUNTING FOR THE MONOPOLE

It is worth saying a few words about the monopole term, both as in terms of prediction and as it relates to regression.

First, the *overdensity* residuals do not correspond to the *number* density residuals. Even with a perfect reconstruction $\hat{s} = s$, the true number density will be unknown up to a factor of γ,

$$N_{\mathrm{true}} = \gamma \langle N_{\mathrm{obs}} \rangle_{\mathrm{pix}} (s + 1), \tag{D1}$$

and as such the estimated number density could be quite different from the truth. Fig. 4 shows a somewhat unintuitive consequence of this. The single systematic that contaminates the field has the form $f_{\mathrm{sys}} \propto -t$, so that it only obscures galaxies from view ($f_{\mathrm{sys}} \leq 0$). At $t = 0$, there is no contamination and so $N_{\mathrm{obs}} = N_{\mathrm{true}}$, however as the figure shows the *over*-density residuals are quite large. This is because the *mean* number density is significantly underestimated, so pixels with no obscuration are preferentially (and wrongly) estimated to reside in overdense regions.[23]

Secondly, a net monopole in f_{sys} corresponds to the intercept in the regression methods (a column of ones in T). In OLS regression (or pseudo-C_ℓ MP), the fit is guaranteed to go through the centre of mass of the points, $(\bar{t}, \bar{d}_{\mathrm{obs}})$, such that including a monopole is unnecessary with such methods if working with overdensities and zero-centred templates. In such cases, the 'projection' of the monopole has already been done by subtracting the mean from the density and template maps (consider equation 17 with a template of all 1s). We showed in equations (38)–(44) how how this also holds in the multiplicative case.

In realistic situations, there is high susceptibility to human error if a monopole term is not included – previously zero-centred maps can easily shift through template transformations, mask adjustments, and the application of a mask to mocks, resulting in wildly biased contamination estimates that may be difficult to detect. For example, it is easy to pass templates that are not zero-centred to current pseudo-C_ℓ MP methods such as implemented in `NaMaster` and receive highly biased spectra without warning (see right-hand panel of Fig. B2).

The DES-Y1 and Forward Selection methods both already include an intercept term, in keeping with the original formulation of the DES-Y1 method, though in practice it should be very close to zero.

We therefore opt to include a monopole term in our Elastic Net method, as this ensures the method is robust and generalizes the process beyond overdensities to non-zero mean fields, and it will naturally be ignored as a template if it does not contribute information.

APPENDIX E: MAP ERROR WHEN VARYING NUMBER OF TEMPLATES

Fig. E1 shows the RMSE of the estimated overdensity map when varying the number of templates used for cleaning (see Fig. 10 for details). Typical map errors are small, with $\mathrm{RMSE}_s \lesssim 0.03$ for most of the cases studied and trends similar to Fig. 10. TS shows particularly bad map reconstruction due to the fact that it fully removes the largest scale modes which have a small contribution to $\Delta\chi^2_{C_\ell}$ but a

[23] In other words, $\gamma > 1$, so from equation (38), $\langle d_{\mathrm{obs}} |_{f_{\mathrm{sys}}=0} \rangle_{\mathrm{pix}} > 0$.

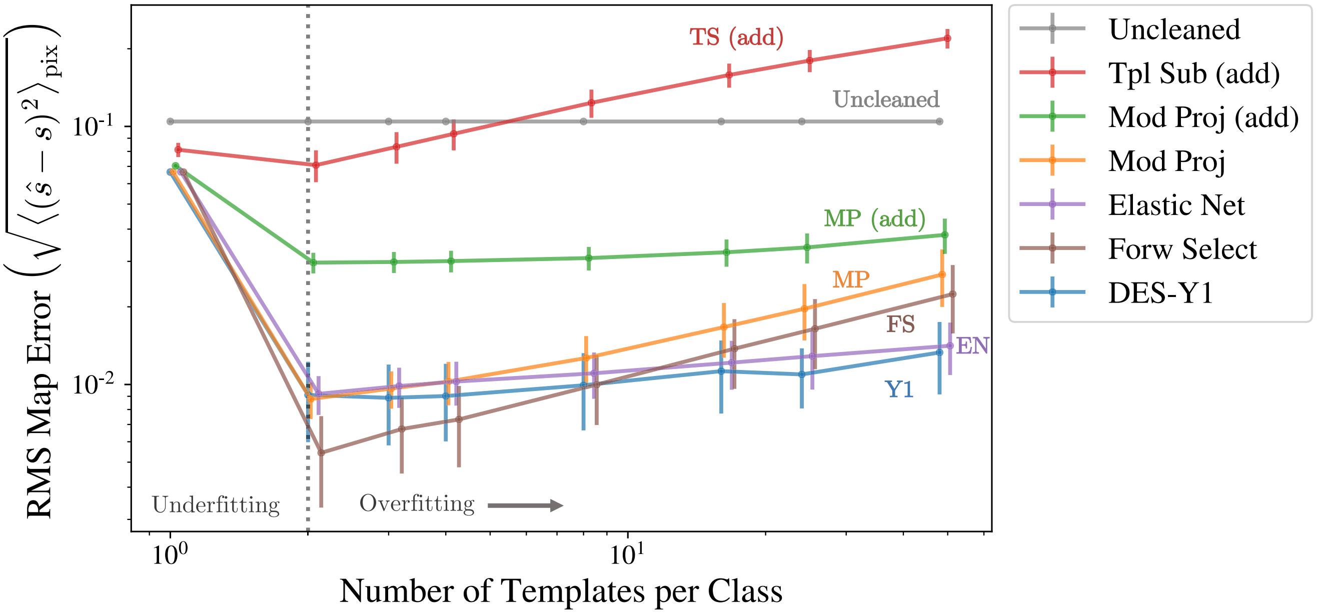

Figure E1. Same as Fig. 10 but showing the RMSE in the estimated overdensity map for each method, rather than error in the power spectrum. Trends are very similar. See Section 7.2 for details.

large contribution to the RMSE because of inducing a bias on a large number of pixels.

APPENDIX F: IMPACT OF $\Delta\chi^2/\Delta\chi^2_0$ ON DES-Y1 ANALYSIS

Here, we investigate the effect of $\Delta\chi^2/\Delta\chi^2_0$ and $\sigma^2_{\rm sys}$ on the efficacy of the DES-Y1 method, as described in Section 3.1. We describe the reconstruction quality with the residual chi-squared between the cleaned and true model, $\Delta\chi^2_{C_\ell}$.

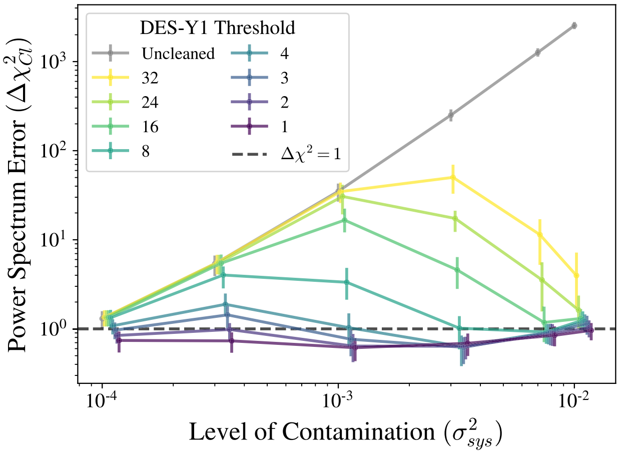

Figure F1. Dependence of the power spectrum error ($\Delta\chi^2_{C_\ell}$) on the level of contamination $\sigma^2_{\rm sys}$ (x-axis), and on the stopping criterion $\Delta\chi^2/\Delta\chi^2_0$ used for the DES-Y1 method (colors). Points are offset for clarity. For comparison, the variance across pixels from the true overdensity in each bin ranges from $\sigma^2_{\rm sig}\in[0.075, 0.122]$ for the five redshift bins of our fiducial survey, corresponding to factors of 7.5 – 1220 times larger than $\sigma^2_{\rm sys}$ for the points shown.

Fig. F1 shows how $\Delta\chi^2/\Delta\chi^2_0$ affects the reconstruction quality for the DES-Y1 method, as a function of the level of contamination parametrized by the systematic-error variance $\sigma^2_{\rm sys}$. We find little reduction in error by lowering the significance threshold below $\Delta\chi^2_{\rm threshold}=4$.

At our fiducial level of contamination ($\sigma^2_{\rm sys}=10^{-2}$), almost all contaminating templates exceed the highest threshold displayed of $\Delta\chi^2/\Delta\chi^2_0=32$ and so are corrected for. The larger the contamination, the more precisely its form can be determined, so as the level of contamination decreases, some contaminated templates are left uncorrected for. This results in the somewhat counter-intuitive turnover in the error for a given threshold level. We found that the lowest threshold of $\Delta\chi^2/\Delta\chi^2_0=1$ consistently outperformed higher thresholds, despite the risk of overfitting, in agreement with our results in Section 7.2, which showed that the extra power from residual contamination is likely more pernicious than the excess removal of power due to overfitting.

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