Galaxy Clustering Working Group
Description of the Likelihood Fitting Work Package

Ariel G. Sánchez & Will J. Percival

October 8, 2015

1 General description of the work package

The Likelihood Fitting Work Package (LFWP) must produce likelihood modules to derive cosmological measurements from two-point clustering measurements, including those made in both configuration and Fourier space. They may constrain the information used from these measurements, for example extracting only the BAO signal, or they may use the full information available. The modules to do this must be statistically rigorous, and include allowances for all known sources of systematic errors. Ultimately, the Euclid consortium will wish to present a singel set of constraints, so we expect to have one or two final "Euclid product algorithms".

This document presents a brief outline of the main issues that need to be addressed in order to achieve the goal of the WP, including a prioritized list of tasks in near term (1 year), mid term (3 years), and long term (5 years or more).

2 Short-term tasks: a roadmap for covariance matrix estimation

The likelihood of the measured tow-point statistics is likely to be close to Gaussian in form, with the primary uncertainty being knowing the covariance matrix $\mathbf{C}$. Given the long lead-time for creating large sets of simulations, a short-term task for the LFWP is to develop a roadmap for the estimation of the covariance matrix. This task must be performed in coordination with other WP and working groups.

In most clustering analyses, $\mathbf{C}$ is estimated directly from a set of mock catalogues. Recent studies have provided a description of the dependence of the noise in the estimated covariance matrix on the number of mock catalogues used (Taylor et al. 2013), its propagation to the derived parameter uncertainties (Dodelson & Schneider 2013) and the correct way to include this additional uncertainty in the obtained cosmological constraints (Percival 2014). The results from these studies show that a large number of mock catalogues are necessary in order to keep this additional uncertainty under control. For a large-volume survey like Euclid, this requirement might be extremely demanding or even infeasible. Thus, Euclid cannot simply use the methods adopted in the past, so a detailed study of the estimation of the covariance matrix is a high priority task for the LFWP.

The questions to be answered in the short term include:

- **What are the requirements in terms of number of N-body simulations and mock catalogues?** The formulae of Dodelson & Schneider (2013) can be used to estimate the required number of mocks to maintain the additional uncertainty of the derived cosmological constraints, $\sigma_{\rm extra}$, below a given threshold. These limits range from a few thousands mocks for $\sigma_{\rm extra} = 0.1\sigma_{\rm ideal}$ to tens of thousands for a challenging limit of $\sigma_{\rm extra} = 0.01\sigma_{\rm ideal}$. These limits must be carefully defined for different clustering statistics and analysis configurations.

- **What are the requirements in terms of size and accuracy of N-body simulations and mock catalogues?** What resolution is required and what size of mocks is required. Is it necessary to simulate the full survey? This questions is clearly linked to both the simulations WG and the previous item.
- **Are approximate N-body methods accurate enough for covariance matrix estimates?** Approximate methods such as PINOCCHIO (Monaco et al. 2002, 2013), COLA (Tassev et al. 2013), PATCHY (Kitaura 2014) or EZMOCKS (Chuang et al. 2015) can be used to construct mock catalogs with significantly less computational resources than N-body simulations. A detailed comparison of the covariance matrices derived from these methods with that of full N-body simulations must be performed.
- **What schemes could be implemented to minimize the impact of the uncertainties on C?** We must test the applicability of techniques to reduce the impact of the noise in the covariance matrix such as shrinkage (Pope & Szapudi 2008), covariance tapering (Paz & Sánchez 2015).
- **Theoretical models of the covariance matrix**. We must develop and test models of **C**, taking into account the effects of cosmic variance, shot noise, non-linear clustering evolution, redshift-space distortions and bias. A hybrid approach by calibrating uncertainties in models against mock catalogues could also help to significantly reduce the required number of simulations.
- **Do we need to estimate C for different models?** In the standard likelihood calculation, the covariance depends on the model being tested. We must asses the effect of varying the covariance matrix when exploring a given parameter space, as opposed to the usual approach in which a fixed **C** is assumed. If the covariance matrix must be varied, what are the requirements in terms of the range of cosmological models we need mocks for?

3 Mid-term tasks: the likelihood function

Although not at the same level of priority, on a mid-term time scale the LFWP must address issues related to the correct characterization of the likelihood function and the posterior parameter distributions. These tasks require to address the following questions:

- **What is the correct shape of the likelihood function?** Clustering analyses usually assume a Gaussian likelihood function. Instead, it is possible to use the true form with a covariance matrix at a fixed fiducial cosmology (Hamimeche & Lewis 2008). This assumption might introduce systematic biases in the obtained parameter constraints (e.g. Kalus et al. 2015).
- **How do we combine measurements from different methods?** The analysis of Anderson et al. (2014) showed that, although they are strongly correlated, the combination of information from $\xi(r)$ and $P(k)$ can help to improve the obtained constraints, as they have different noise properties. We must explore the optimal combination of two-point statistics to maximize the obtained cosmological information.
- **How do we include information from higher order statistics?** What is the min/max scale to be included given issues with calculating covariances? How do we include other information from non-2-pt statistics?
- **How do we correctly combine systematic and statistical errors?** Cosmology-independent systematic errors like the lack of knowledge of Zodiacal background must be correctly combined with cosmology-dependent errors such as cosmic variance.

4 Long-term tasks: Euclid likelihood modules

On a longer time scale the LFWP must produce the likelihood modules to derive cosmological constraints from Euclid clustering measurements. These modules must be fast, able to work for multiple models, and correctly account for the effect of the uncertainties in the covariance matrix, systematic errors, marginalising over nuisance parameters.

Although the construction of the full Euclid modules for likelihood calculations is a long-term task, this tool would also be useful for the design of the analyses that will be applied to the survey. Assuming a model for an observable (e.g. the correlation function or the power spectrum) and its covariance matrix, we could generate mock clustering measurements that can be plugged into the likelihood modules and used to compute forecasts for Euclid beyond the simplified Fisher matrix approach. These forecasts could then be used to compare the performance of different analysis configurations, such as the number of redshift bins, galaxy sample selection, etc. to optimise their design.

The questions related to the construction of the likelihood modules include:

- **Should we base these modules on existing packages such as** CosmoMC **or** Monte Python **or build an entirely new code?** CosmoMC is the most commonly used tool for cosmological studies as it includes modules for the analysis of Planck and several additional data sets. Releasing CosmoMC-compatible modules would simplify the use of Euclid clustering products outside the Consortium. We must evaluate if CosmoMC satisfies our needs or if we need to develop a new code.
- **Should we use MCMC, nested sampling, or base our analysis on other methods (e.g. Hamiltonian Monte Carlo)?** These alternatives must be evaluated, exploring ways to quantify and minimise the errors from each methodology.
- **What is the best way to characterize the resulting parameter posterior distributions for different models?** Is it sufficient to present single parameter marginalsed measurements, with 2D (or 3D) plots showing primary degeneracies, or do we need a more complicated scheme?

References

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

Chuang C.-H., Kitaura F.-S., Prada F., Zhao C., Yepes G., 2015, MNRAS, 446, 2621

Dodelson S., Schneider M. D., 2013, Phys. Rev. D, 88, 063537

Hamimeche S., Lewis A., 2008, Phys. Rev. D, 77, 103013

Kalus B., Percival W. J., Samushia L., 2015, arXiv:1504.03979

Kitaura F.-S., Yepes G., Prada F., 2014, MNRAS, 439, L21

Koda J., Blake C., Beutler F., Kazin E., Marin F., 2015, arXiv:1507.05329

Monaco P., Theuns T., Taffoni G., 2002, MNRAS, 331, 587

Paz D.J., Sánchez, A. G. 2015, arXiv:1508.03162

Percival W. J. et al., 2014, MNRAS, 439, 2531

Tassev S., Zaldarriaga M., Eisenstein D. J., 2013, JCAP, 6, 36

Taylor A., Joachimi B., Kitching T., 2013, MNRAS, 432, 1928