The Journal of Open Source Software

Triumvirate: A Python/C++ package for three-point clustering measurements

Mike Shengbo Wang[1¶], **Florian Beutler**[1], **and Naonori S. Sugiyama**[2]

1 Institute for Astronomy, University of Edinburgh, Royal Observatory Edinburgh, Blackford Hill, Edinburgh EH9 3HJ, United Kingdom **2** National Astronomical Observatory of Japan, Mitaka, Tokyo 181-8588, Japan ¶ Corresponding author

DOI: 10.21105/joss.05571

Software

- Review
- Repository
- Archive

Editor: Ivelina Momcheva

Reviewers:

- @alfonso-veropalumbo
- @wcoulton

Submitted: 07 April 2023
Published: 07 November 2023

Summary

`Triumvirate` is a Python/C++ package for measuring the three-point clustering statistics in large-scale structure (LSS) cosmological analyses. Given a catalogue of discrete particles (such as galaxies) with their spatial coordinates, it computes estimators of the multipoles of the three-point correlation function, also known as the bispectrum in Fourier space, in the tri-polar spherical harmonic (TripoSH) decomposition proposed by Sugiyama et al. (2019). The objective of `Triumvirate` is to provide efficient end-to-end measurement of clustering statistics which can be fed into downstream galaxy survey analyses to constrain and test cosmological models. To this end, it builds upon the original algorithms in the `hitomi` code[1] developed by Sugiyama et al. (2018, 2019), and supplies a user-friendly interface with flexible input/output (I/O) of catalogue data and measurement results, with the built program configurable through external parameter files and tracked through enhanced logging and warning/exception handling. For completeness and complementarity, methods for measuring two-point clustering statistics are also included in the package.

Statement of need

The analysis of higher-order clustering statistics is a key pursuit of the current and forthcoming generations of large galaxy surveys such as the Dark Energy Spectroscopic Instrument (DESI)[2] (DESI Collaboration et al., 2016) and *Euclid*[3] (Euclid Consortium et al., 2011). Although the matter density fluctuations in the Universe have been observed to be almost Gaussian on large scales, primordial non-Gaussianity in the initial conditions of structure formation and late-time non-linear gravitational dynamics can both leave potentially detectable signals in the higher-order moments of the galaxy distribution on very large and small scales (Bernardeau et al., 2002). Therefore any measurement of the three-point clustering statistics, the leading non-Gaussian moment, offers a promising probe of both early-Universe and gravitational physics. In addition, it complements two-point clustering statistics in constraining cosmological models by breaking down certain parameter degeneracies (Sefusatti et al., 2006).

In contrast to the two-point statistic analysis which has become standard in recent galaxy surveys (e.g. BOSS Collaboration et al., 2017; eBOSS Collaboration et al., 2021), three-point clustering statistics have more degrees of freedom and thus can be compressed in a greater number of ways with different choices of the coordinate system. For the TripoSH decomposition mentioned above, there is a need for a computational program that is easy to use, versatile and suited for the large data sets expected from modern galaxy surveys, and `Triumvirate` is designed to meet that demand. More specifically, it can compute:

[1] github.com/naonori/hitomi
[2] desi.lbl.gov
[3] sci.esa.int/euclid, euclid-ec.org

- three-point clustering statistics, namely multipoles of the bispectrum in Fourier space and of the three-point correlation function (3PCF) in configuration space (Sugiyama et al., 2019);
- two-point clustering statistics, namely multipoles of the power spectrum and two-point correlation function (2PCF) (Sugiyama et al., 2018);
- for both three- and two-point statistics, the local plane-parallel estimator for a pair of survey-like data and random catalogues, and the global plane-parallel estimator for a cubic-box simulation (Feldman et al., 1994; Sugiyama et al., 2019; Yamamoto et al., 2006);
- for both three- and two-point statistics, the configuration-space window function multipoles, which are used to convolve theoretical models derived in Fourier space through the Hankel transform (Sugiyama et al., 2019; Wilson et al., 2016).

For the global plane-parallel estimators, the simulation box is placed at the spatial infinity (or equivalently the observer is), so that the line of sight to each particle can be treated as the same and taken to be along the z-axis. For the local plane-parallel estimators, the observer is placed at the origin in the survey coordinates, and the line of sight is chosen to point towards one of the particles in a triplet or pair for three- or two-point clustering measurements respectively.

The geometry of the survey leaves an imprint on the clustering statistics, where in Fourier space the effect is a convolution with the survey window function. This convolution mixes different multipoles of the underlying clustering statistics and the survey window, and the precise convolution formula (i.e. the number of multipoles to include in modelling) needed to achieve a given level of convergence depends on the precise survey geometry including any sample weights applied. Therefore the functionality to measure the window function is an integral part of this program.

These functionalities are essential to cosmological inference pipelines, and can help validate any analytical covariance matrix predictions against sample estimates. Since precise covariance matrix estimates usually require clustering measurements repeated over a large number of simulated mock catalogues, computational efficiency is an important objective.

Finally, `Triumvirate` also enables comparison studies between alternative compressed statistics of three-point clustering, which may have different constraining power on different cosmological parameters. There are existing software packages for some of these alternative approaches:

- `pylians`[4] (Villaescusa-Navarro, 2018) computes the bispectrum with the Scoccimarro estimator (Scoccimarro, 2015) for triangle configurations parametrised by two wavenumbers and the angle between the corresponding wavevectors;
- `nbodykit`[5] (Hand et al., 2018) computes the isotropised 3PCF with a pair-counting algorithm (Slepian & Eisenstein, 2015), although in principle this can be generalised to anisotropic 3PCF (Slepian & Eisenstein, 2018), or be implemented using FFTs (Slepian & Eisenstein, 2016);
- Philcox (2021)[6] advocates a windowless cubic estimator for the bispectrum in the Soccimarro decomposition, which can be evaluated using FFTs. However, this approach requires the inversion of a Fisher matrix obtained from a suite of Monte Carlo realisations.

As these programs use a different decomposition of three-point clustering statistics and focus on either configuration- or Fourier-space statistics only, `Triumvirate` fulfills complementary needs in current galaxy clustering analyses.

[4] pylians3.readthedocs.io
[5] nbodykit.readthedocs.io
[6] github.com/oliverphilcox/Spectra-Without-Windows

Implementation

Direct calculation of Fourier modes of density field fluctuations as a sum over a large number of particles is computationally infeasible, but the TripoSH estimators can be cast in a form amenable to fast Fourier transforms (FFTs), which can be utilised to speed up evaluations. In our numerical scheme, the particles are assigned to regular mesh grids with appropriate weighting, which is a combination of spherical harmonics and weights from the input catalogues. Fourier-space fields are obtained by FFTs over the cubic mesh grids, and clustering statistics are formed by multiplying the discretely sampled and transformed fields grid-by-grid, before binning in spherical shells. For the three-point statistics, the shot noise components are effectively two-point statistics and can be calculated the same way.

`Triumvirate` supports mesh assignment schemes up to order 4, namely the piecewise cubic spline (PCS) scheme, where order p refers to the number of grid points a particle is assigned to. Since mesh assignment convolves the underlying field with a sampling window, the transformed Fourier-space field should be compensated by dividing out the sampling window (Hockney & Eastwood, 1988). For power spectrum measurements only where no inverse FFT is involved the interlacing technique can be used to reduce the amount of aliasing (Sefusatti et al., 2016), an artifact of discrete Fourier transform where the sampled Fourier mode at each wavenumber receives contributions from other modes, with the effect increasingly prominent as the Nyquist wavenumber is approached; without interlacing, the corrections from Jing (2005) (see eq. 20 therein) are adopted instead. For all clustering statistics, increasing the mesh assignment order and/or the number of grid cells can help reduce aliasing (at the expense of speed and memory).

Features

Besides code refactoring, `Triumvirate` has many value-added features in comparison with the predecessor `himoti`:

- The frontend is written in Python for interactivity and convenience, with Cython binding the C++ backend. Although the user will typically use the Python interface, the C++ code can also be compiled and executed independently.
- Measurement pipelines can be configured through external parameter files (in the YAML format for the Python program), cleanly separating user inputs from the program itself. Alternatively, measurement parameters can be set for individual Python methods without the use of a parameter file.
- The reading of catalogue data is implemented via `astropy.io` (Astropy Collaboration et al., 2022) and `nbodykit` (Hand et al., 2018), with flexible support for different file formats such as text and `fits` files.
- Numerical algorithms are parallelised with OpenMP, with for-loops over catalogue particles and mesh grid cells distributed amongst multiple CPU threads.
- Mesh assignment schemes from order $p = 1$ to 4 are supported: nearest grid point (NGP), cloud-in-cell (CIC), triangular-shape cloud (TSC) and piecewise cubic spline (PCS).
- Interlacing is supported for power spectrum measurements.
- Two normalisation choices are implemented for all clustering statistics, one as a sum of catalogue particles (Feldman et al., 1994, eq. 2.4.1) and another as a sum over the mesh grid (Sugiyama et al., 2019, eq. 37).
- A customised logger is provided for runtime tracking, with enhanced handling of warnings and exceptions for parameter and data I/O.

Performance

When a large number of grid cells, N_{mesh}, are used to sample the density fields from a catalogue of N_{part} particles on a mesh with $N_{\text{mesh}} \gg N_{\text{part}}$, the dominant operations are FFTs with complexity $\mathcal{O}\left(N_{\text{mesh}} \ln N_{\text{mesh}}\right)$. Therefore the complexity for three-point clustering measurements is $\mathcal{O}\left(N_{\text{bin}}^2 N_{\text{mesh}} \ln N_{\text{mesh}}\right)$, where N_{bin} is the number of coordinate bins.

It is worth noting that in `Triumvirate`, the spherical harmonic weights are applied to individual particles rather than the mesh grids. This should result in more accurate results at the expense of memory usage, as multiple meshes need to be stored for spherical harmonics of different degrees and orders. We estimate the minimum memory usage for bispectrum measurements to be $11M$ and $9M$ respectively for local and global plane-parallel estimators, where $M = 16N_{\text{mesh}}$ bytes (roughly $1.5 \times 10^{-8} N_{\text{mesh}}$ gibibytes[7]); for local and global plane-parallel 3PCF estimators, the figures are $10M$ and $9M$ respectively.

In the table below, we show the wall time and peak memory usage for bispectrum and 3PCF measurements of a few select multipoles and grid numbers with $N_{\text{bin}} = 20$, using a single core on one AMD EPYC 7H12 processor with base frequency 2.60 GHz. With multithreading enabled, the run time is reduced (see the last column in the table). Here 'lpp' and 'gpp' denote local and global plane-parallel approximations respectively. For the global plane-parallel estimates, the catalogue used is a cubic box containing $N_{\text{part}} = 8 \times 10^6$ particles; for the local plane-parallel estimates, the data and random catalogues contain $N_{\text{part}} = 6.6 \times 10^5$ and 1.3×10^7 particles respectively. Since both the bispectrum and 3PCF are computed with FFTs, the computation time and memory usage for them are roughly the same, with minor differences due to the slightly different number of mesh grids needed for evaluation.

Multipole/N_{mesh}	128^3	256^3	512^3	512^3 (32 threads)
$B_{000}^{(\text{lpp})}$	96 s, 1.8 GiB	215 s, 4.4 GiB	1247 s, 25 GiB	85 s, 25 GiB
$B_{000}^{(\text{gpp})}$	42 s, 0.7 GiB	172 s, 2.9 GiB	1185 s, 21 GiB	59 s, 21 GiB
$B_{202}^{(\text{lpp})}$	320 s, 1.8 GiB	1030 s, 4.4 GiB	6449 s, 25 GiB	267 s, 25 GiB
$B_{202}^{(\text{gpp})}$	42 s, 0.7 GiB	176 s, 2.9 GiB	1187 s, 21 GiB	60 s, 21 GiB
$\zeta_{000}^{(\text{lpp})}$	90 s, 1.8 GiB	211 s, 4.4 GiB	1403 s, 21 GiB	83 s, 21 GiB
$\zeta_{000}^{(\text{gpp})}$	43 s, 0.7 GiB	178 s, 2.9 GiB	1226 s, 19 GiB	55 s, 19 GiB
$\zeta_{202}^{(\text{lpp})}$	267 s, 1.8 GiB	964 s, 4.4 GiB	6377 s, 21 GiB	266 s, 21 GiB
$\zeta_{202}^{(\text{gpp})}$	43 s, 0.7 GiB	177 s, 2.9 GiB	1241 s, 19 GiB	57 s, 19 GiB

Future work

`Triumvirate` will be routinely maintained and updated depending on user feedback. One extension of interest is the inclusion of other three-point clustering estimators with different coordinate systems and compression choices, and the functionality to transform between them. The ability to measure clustering statistics from a density field already sampled on a mesh grid may also be useful. In addition, porting the code to graphic processing units (GPUs) can bring further parallelisation that can enhance the performance of the code.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 853291).

[7]Note that 1 gibibyte (GiB) is 2^{30} bytes, as opposed to 1 gigabyte (GB) which is 10^9 bytes. GiB is the preferred unit by job schedulers such as Slurm for computer clusters.

FB is a Royal Society University Research Fellow.

We thank the reviewers for their valuable feedback and suggestions, which have improved the functionality and documentation of this code.

References

Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., Earl, N., Starkman, N., Bradley, L., Shupe, D. L., Patil, A. A., Corrales, L., Brasseur, C. E., Nöthe, M., Donath, A., Tollerud, E., Morris, B. M., Ginsburg, A., Vaher, E., Weaver, B. A., Tocknell, J., Jamieson, W., … Astropy Project Contributors. (2022). The Astropy Project: Sustaining and Growing a Community-oriented Open-source Project and the Latest Major Release (v5.0) of the Core Package. *Astrophys. J.*, *935*(2), 167. https://doi.org/10.3847/1538-4357/ac7c74

Bernardeau, F., Colombi, S., Gaztañaga, E., & Scoccimarro, R. (2002). Large-scale structure of the universe and cosmological perturbation theory. *Phys. Rep.*, *367*, 1–248. https://doi.org/10.1016/s0370-1573(02)00135-7

BOSS Collaboration, Alam, S., Ata, M., Bailey, S., Beutler, F., Bizyaev, D., Blazek, J. A., Bolton, A. S., Brownstein, J. R., Burden, A., Chuang, C.-H., Comparat, J., Cuesta, A. J., Dawson, K. S., Eisenstein, D. J., Escoffier, S., Gil-Marín, H., Grieb, J. N., Hand, N., … Zhao, G.-B. (2017). The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: Cosmological analysis of the DR12 galaxy sample. *Mon. Not. R. Astron. Soc.*, *470*(3), 2617–2652. https://doi.org/10.1093/mnras/stx721

DESI Collaboration, Aghamousa, A., Aguilar, J., Ahlen, S., Alam, S., Allen, L. E., Allende Prieto, C., Annis, J., Bailey, S., Balland, C., Ballester, O., Baltay, C., Beaufore, L., Bebek, C., Beers, T. C., Bell, E. F., Bernal, J. L., Besuner, R., Beutler, F., … Zu, Y. (2016). The DESI Experiment Part I: Science,Targeting, and Survey Design. *arXiv e-Prints*. https://doi.org/10.48550/arXiv.1611.00036

eBOSS Collaboration, Alam, S., Aubert, M., Avila, S., Balland, C., Bautista, J. E., Bershady, M. A., Bizyaev, D., Blanton, M. R., Bolton, A. S., Bovy, J., Brinkmann, J., Brownstein, J. R., Burtin, E., Chabanier, S., Chapman, M. J., Choi, P. D., Chuang, C.-H., Comparat, J., … Zheng, Z. (2021). Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Cosmological implications from two decades of spectroscopic surveys at the Apache Point Observatory. *Phys. Rev. D*, *103*, 083533. https://doi.org/10.1103/PhysRevD.103.083533

Euclid Consortium, Laureijs, R., Amiaux, J., Arduini, S., Augueres, J.-L., Brinchmann, J., Cole, R., Cropper, M., Dabin, C., Duvet, L., Ealet, A., Garilli, B., Gondoin, P., Guzzo, L., Hoar, J., Hoekstra, H., Holmes, R., Kitching, T., Maciaszek, T., … Zucca, E. (2011). Euclid Definition Study Report. *arXiv e-Prints*. https://doi.org/10.48550/arXiv.1110.3193

Feldman, H. A., Kaiser, N., & Peacock, J. A. (1994). Power-Spectrum Analysis of Three-dimensional Redshift Surveys. *426*, 23. https://doi.org/10.1086/174036

Hand, N., Feng, Y., Beutler, F., Li, Y., Modi, C., Seljak, U., & Slepian, Z. (2018). nbodykit: An Open-source, Massively Parallel Toolkit for Large-scale Structure. *Astron. J.*, *156*(4), 160. https://doi.org/10.3847/1538-3881/aadae0

Hockney, R. W., & Eastwood, J. W. (1988). *Computer Simulation Using Particles* (1st ed.). CRC Press. https://doi.org/10.1201/9780367806934

Jing, Y. P. (2005). Correcting for the Alias Effect When Measuring the Power Spectrum Using a Fast Fourier Transform. *Astrophys. J.*, *620*(2), 559. https://doi.org/10.1086/427087

Philcox, O. H. E. (2021). Cosmology without window functions. II. Cubic estimators for the galaxy bispectrum. *Phys. Rev. D*, *104*(12), 123529. https://doi.org/10.1103/physrevd.104.123529

Scoccimarro, R. (2015). Fast estimators for redshift-space clustering. *Phys. Rev. D*, *92*, 083532. https://doi.org/10.1103/PhysRevD.92.083532

Sefusatti, E., Crocce, M., Pueblas, S., & Scoccimarro, R. (2006). Cosmology and the bispectrum. *Phys. Rev. D*, *74*, 023522. https://doi.org/10.1103/PhysRevD.74.023522

Sefusatti, E., Crocce, M., Scoccimarro, R., & Couchman, H. M. P. (2016). Accurate estimators of correlation functions in Fourier space. *Mon. Not. R. Astron. Soc.*, *460*(4), 3624–3636. https://doi.org/10.1093/mnras/stw1229

Slepian, Z., & Eisenstein, D. J. (2015). Computing the three-point correlation function of galaxies in $\mathcal{O}(N^2)$ time. *Mon. Not. R. Astron. Soc.*, *454*(4), 4142–4158. https://doi.org/10.1093/mnras/stv2119

Slepian, Z., & Eisenstein, D. J. (2016). Accelerating the two-point and three-point galaxy correlation functions using fourier transforms. *Mon. Not. R. Astron. Soc.*, *455*(1), L31–L35. https://doi.org/10.1093/mnrasl/slv133

Slepian, Z., & Eisenstein, D. J. (2018). A practical computational method for the anisotropic redshift-space three-point correlation function. *Mon. Not. R. Astron. Soc.*, *478*(2), 1468–1483. https://doi.org/10.1093/mnras/sty1063

Sugiyama, N. S., Saito, S., Beutler, F., & Seo, H.-J. (2019). A complete FFT-based decomposition formalism for the redshift-space bispectrum. *Mon. Not. R. Astron. Soc.*, *484*(1), 364–384. https://doi.org/10.1093/mnras/sty3249

Sugiyama, N. S., Shiraishi, M., & Okumura, T. (2018). Limits on statistical anisotropy from BOSS DR12 galaxies using bipolar spherical harmonics. *Mon. Not. R. Astron. Soc.*, *473*(2), 2737–2752. https://doi.org/10.1093/mnras/stx2333

Villaescusa-Navarro, F. (2018). *Pylians: Python libraries for the analysis of numerical simulations*. Astrophysics Source Code Library [ascl:1811.008].

Wilson, M. J., Peacock, J. A., Taylor, A. N., & de la Torre, S. (2016). Rapid modelling of the redshift-space power spectrum multipoles for a masked density field. *Mon. Not. R. Astron. Soc.*, *464*(3), 3121–3130. https://doi.org/10.1093/mnras/stw2576

Yamamoto, K., Nakamichi, M., Kamino, A., Bassett, B. A., & Nishioka, H. (2006). A Measurement of the Quadrupole Power Spectrum in the Clustering of the 2dF QSO Survey. *Publ. Astron. Soc. Jpn.*, *58*(1), 93–102. https://doi.org/10.1093/pasj/58.1.93