LSS Bispectral Estimation with MODAL

David Gwilym Baker[1]

[1]DAMTP, University of Cambridge

Abstract and Background

This poster presents the MODAL bispectral estimator for LSS (Large Scale Structure) datasets. It is an efficient, and highly accurate estimator with a multitude of use cases. Here we draw particular note to MODAL's ability in:

- Reconstructing given theoretical templates.
- Producing fully generic Non-Gaussian (NG) inflationary field perturbations at bispectral level.
- Extracting highly accurate f_{NL} parameters from simulated and observed datasets.

More widely MODAL is being developed for estimation in a number of different domains, such as the CMB and for projected-LSS. MODAL was first introduced in [1]. The LSS component of MODAL that I have worked on has been implemented in C++ with full support for MPI and Threaded functionalities in hybrid and is highly-scalable on HPC cluster systems.

With new surveys such as Euclid and LSST allowing us to probe cosmological observables to greater precision, we start to look to LSS for frontier constraints on cosmologies. The ability to accurately and quickly extract statistics at bispectral level from LSS datasets (which are naturally nonlinear) and beyond is critical to adequately advantage this new data. Whether this be in the context of primordial non-gaussianity (PNG) constraints, cosmological parameter constraints, or for astrophysical applications at smaller scales.

Introduction to MODAL for LSS

The MODAL methodology is predicated on an efficient compression of bispectral statistics into a functional basis expansion with the requisite symmetries:

$$B_{\mathrm{SN}}(k_1,k_2,k_3) = \sum_{i=0}^{i=n_{\mathrm{modes}}} c_i^{\mathcal{Q}} \cdot \mathcal{Q}_i = \sum_{j=0}^{j=n_{\mathrm{modes}}} c_j^{\mathcal{R}} \cdot \mathcal{R}_j \quad , \quad B_{\mathrm{SN}} = \left(\sqrt{\frac{\Pi_i k_i}{\Pi_i P(k_i)}}\right) B$$

Where $c \in \{\alpha, \beta\}$ are expansion coefficients dependent on whether we compress a theoretical template (α's) or a data extracted bispectrum (β's), and P/B the relevant power/bispectrum. A signal-to-noise (SN) weighting is used throughout our estimation pipeline to reduce dynamic range related error accumulation.

Our Q_n are polynomial functions from a class that span $\mathcal{V}_B$:

$$\mathcal{V}_B : \quad \left\{ \begin{array}{ccc} k_1 + k_2 \geq k_3 & + & (2\ \mathrm{perms}) \\ k_{\mathrm{min}} \leq k_i \leq k_{\mathrm{max}} & , & i = \{1,2,3\} \end{array} \right\}$$

For numerical reasons, we choose these Q_n to be constructed from Shifted Legendre Polynomials (SLP), which are complete on $[0,1]\times[0,1]\times[0,1] \supset \mathcal{V}_B$; a rescaling from $k_{\mathrm{max}} \to 1$ is required w.l.o.g. Our Q_n have to obey bispectral symmetries, hence:

$$Q_n(k_1,k_2,k_3) = q_{\{r}(k_1) q_s(k_2) q_{t\}}(k_3) \quad , \quad q_m(x) \in P_m(2x-1) = \mathrm{SLP}$$

With $n \to \{r,s,t\}$ a mapping from modes to polynomial orders.

The key advantage of the MODAL methodology is that compression of B into $\mathcal{Q}_n$ is constructed *separable* in $\{k_i\}$. For datasets with N points along each direction, the typical FFT method scales as $\mathcal{O}(N^6)$ whereas MODAL scales as $\mathcal{O}(n_{\mathrm{modes}} \cdot N^3)$[2].

For convenience we also introduce an orthonormal basis, constructed out of linear combinations of $\mathcal{Q}_n$ on the bispectral domain: $\mathcal{R}_n := \sum_{a=0}^{a=n_{\mathrm{modes}}} \lambda_{na} \mathcal{Q}_a$. For this transition we first need to determine our inner product space for the $\mathcal{Q}_n$:

$$\langle \mathcal{Q}_a, \mathcal{Q}_b \rangle_{\mathcal{V}_B} = \Gamma_{ab} \leftrightarrow \mathrm{CholeskyDecomp}(\Gamma) = \lambda^{-1}$$

The final step is to determine expressions for the coefficients $\{\alpha, \beta\}$ for each use case, this is covered in other sections.

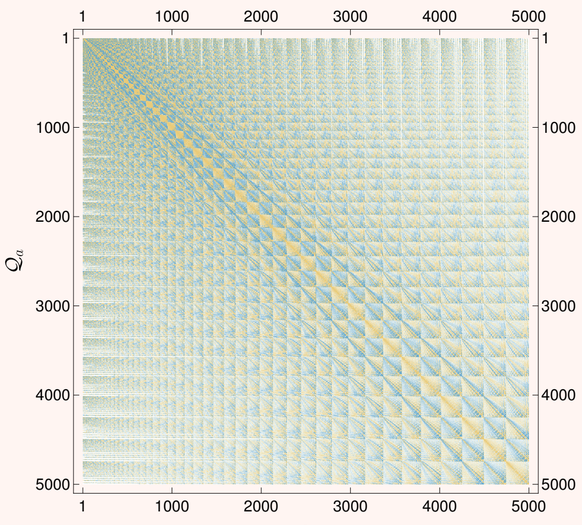

Figure 1. An analytic visualisation of the inner product space defined by $\langle \mathcal{Q}_a, \mathcal{Q}_b \rangle_{\mathcal{V}_B}$ for 5000 modes, computed analytically. Note near orthogonality.

Estimation Fidelity Probes

We wish to quantify how good our MODAL bispectral estimation is. For two functions on $\mathcal{V}_B$, F_1 and F_2, we define *shape* and *amplitude* correlators:

$$\mathcal{S} := \frac{\langle F_1, F_2 \rangle}{\sqrt{\langle F_1, F_1 \rangle \langle F_2, F_2 \rangle}} \quad , \quad \mathcal{A}(F_1, F_2) := \sqrt{\frac{\langle F_1, F_1 \rangle}{\langle F_2, F_2 \rangle}}$$

We also define *total* correlator, $\mathcal{T}$, and the usual 'f_{nl}' estimate in our framework:

$$1 - \mathcal{T}(F_1, F_2) := \left[\frac{\langle F_2 - F_1, F_2 - F_1 \rangle}{\langle F_2, F_2 \rangle}\right]^{1/2} \quad , \quad f_{nl} := \langle F_1, F_2 \rangle / \langle F_2, F_2 \rangle$$

Note that these quantities are not always argument symmetric, usually we choose F_2 to be the (theoretical) template we wish to compare some MODAL reconstruction against.

Theoretical Validation

In order to compute the coefficients needed for a MODAL expansion of some bispectral template F we need to compute:

$$\mathcal{Z}_b^{\alpha} := \langle F, \mathcal{Q}_b \rangle_{\mathcal{V}_B} = \sum_{a=0}^{a=n_{\mathrm{modes}}} \alpha_a \langle \mathcal{Q}_a, \mathcal{Q}_b \rangle_{\mathcal{V}_B} = \sum_{a=0}^{a=n_{\mathrm{modes}}} \alpha_a \Gamma_{ab}$$

Both $\mathcal{Q}$ and F naturally live on $\mathcal{V}_B$ and thus this integral can be computed without further analysis. Typically the integral over $\mathcal{V}_B$ is analytically intractable and thus numerical integration techniques need be used; for numerical reasons the exact implementation turns out to be rather important, but that is beyond the scope of this poster.

Given a $\mathcal{Z}_b^{\alpha}$, we can solve for α_b with a simple linear solve (or inversion):

$$\mathcal{Z}_b^{\alpha} = \sum_{a=0}^{a=n_{\mathrm{modes}}} \alpha_a \Gamma_{ab} \xrightarrow[\mathrm{Solve}]{\mathrm{Linear}} \alpha_b \implies B_{\mathrm{F}} \sim \sum_{a=0}^{a=n_{\mathrm{modes}}} \alpha_a^{\mathcal{Q}} \mathcal{Q}_a(k_1,k_2,k_3)$$

Using a reconstruction grid size of 512 points/dimension, we highlight the f_{NL} reconstruction % error for the three well-known inflationary bispectral shapes, as a function of # terms included in our MODAL expansion:

Shape:	Local	Equilateral	Orthogonal	∼ CPU hours
10 modes	3E-1	7E-3	2E-1	0.2
50 modes	1E-3	1E-4	4E-3:	0.4
100 modes	4E-4	3E-4	1E-3	0.7
500 modes	2E-6	4E-5	7E-5	2.7

Table 1. MODAL reconstruction fidelity for theoretical templates. CPU hours refers to extraction component of compute. The system used 2 MPI ranks + 16 OpenMP threads.

Note that the profile is not monotonically improving; a 'flaw' resultant from looking at f_{NL} with MODAL.

Although requiring more advanced numerical stability optimisations, MODAL estimation in the theoretical domain has been performed in our pipeline across a wide variety of shapes with up to 10000 modes. In practise there is no apparent constraint on desired accuracy bar a compute time balance.

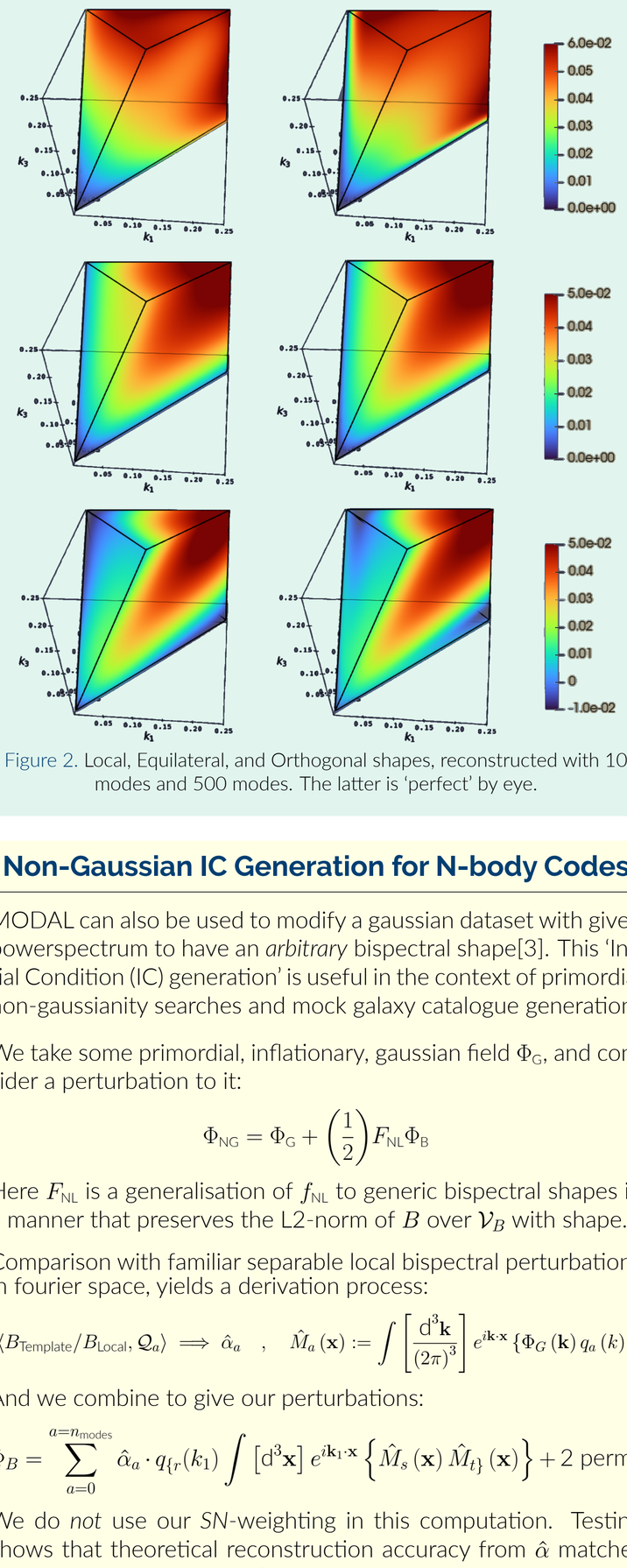

Figure 2. Local, Equilateral, and Orthogonal shapes, reconstructed with 10 modes and 500 modes. The latter is 'perfect' by eye.

Non-Gaussian IC Generation for N-body Codes

MODAL can also be used to modify a gaussian dataset with given powerspectrum to have an *arbitrary* bispectral shape[3]. This 'Initial Condition (IC) generation' is useful in the context of primordial non-gaussianity searches and mock galaxy catalogue generation.

We take some primordial, inflationary, gaussian field Φ_{G}, and consider a perturbation to it:

$$\Phi_{\mathrm{NG}} = \Phi_{\mathrm{G}} + \left(\frac{1}{2}\right) F_{\mathrm{NL}} \Phi_{\mathrm{B}}$$

Here F_{NL} is a generalisation of f_{NL} to generic bispectral shapes in a manner that preserves the L2-norm of B over $\mathcal{V}_B$ with shape.

Comparison with familiar separable local bispectral perturbations in fourier space, yields a derivation process:

$$\langle B_{\mathrm{Template}}/B_{\mathrm{Local}}, \mathcal{Q}_a \rangle \implies \hat{\alpha}_a \quad , \quad \hat{M}_a(\mathbf{x}) := \int \left[\frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^3}\right] e^{i\mathbf{k}\cdot\mathbf{x}} \{\Phi_G(\mathbf{k}) q_a(k)\}$$

And we combine to give our perturbations:

$$\Phi_B = \sum_{a=0}^{a=n_{\mathrm{modes}}} \hat{\alpha}_a \cdot q_{\{r}(k_1) \int \left[\mathrm{d}^3\mathbf{x}\right] e^{i\mathbf{k}_1\cdot\mathbf{x}} \left\{\hat{M}_s(\mathbf{x}) \hat{M}_{t\}}(\mathbf{x})\right\} + 2\ \mathrm{perms}$$

We do *not* use our *SN*-weighting in this computation. Testing shows that theoretical reconstruction accuracy from $\hat{\alpha}$ matches what we ultimately extract from our fields Φ_{NG} very closely, quantifying expectations of bispectrum encoding in our IC with n_{modes}.

Simulated Data Validation

Although the above tools are useful, clearly the MODAL application with the most promise for cosmological inference is that where MODAL is used to extract bispectra from simulated and real datasets. In order to do this, we derive an integral equality:

$$\langle F_1, F_2 \rangle_{\mathcal{V}_B} = \frac{1}{8\pi^2} \int \left[\frac{d^3k_1}{(2\pi)^3} \frac{d^3k_2}{(2\pi)^3} \frac{d^3k_3}{(2\pi)^3}\right] \left\{\frac{(2\pi)^6 \delta_D(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3) F_1 F_2}{k_1 k_2 k_3}\right\}$$

The maximum likelihood estimator for our data (δ) bispectrum, $\hat{B}_{\mathrm{data}}$, in the limit of weak non-gaussianity can be shown to be:

$$\hat{B}_{\mathrm{data}} = \delta_{\mathbf{k}_1}\delta_{\mathbf{k}_2}\delta_{\mathbf{k}_3} - 3\langle \delta_{\mathbf{k}_1}\delta_{\mathbf{k}_2} \rangle \delta_{\mathbf{k}_3} \quad , \quad \hat{B}^{\mathrm{SN}}_{\mathrm{data}} = \left(\sqrt{\frac{k_1 k_2 k_3}{P(k_1)P(k_2)P(k_3)}}\right) \hat{B}_{\mathrm{data}}$$

Our desired inner product is $\langle \hat{B}^{\mathrm{SN}}_{\mathrm{data}}, \mathcal{Q}_b \rangle_{\mathcal{V}_B}$; we thus leverage our separable mode expansion and define:

$$M_a(\mathbf{x}) := \int \left[\frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^3}\right] e^{i\mathbf{k}\cdot\mathbf{x}} \left\{\frac{\delta_{\mathbf{k}} q_a(k)}{\sqrt{k P_\delta(k)}}\right\}$$

Our inner product is then:

$$\mathcal{Z}_b^{\beta} := \langle \hat{B}^{\mathrm{SN}}_{\mathrm{data}}, \mathcal{Q}_b \rangle_{\mathcal{V}_B} = (2\pi)^3 \int \left[\mathrm{d}^3\mathbf{x}\right] \left\{M_r M_s M_t - 3\langle M_{\{r} M_s \rangle M_{t\}}\right\}$$

If we assume that the 'linear term' on the RHS is negligible (true for simulations, sometimes true for real data), this simplifies our compute significantly. As per the α pipeline, we can linear solve to obtain $\beta^{\mathcal{Q}}$ or obtain $\beta^{\mathcal{R}}$ directly via a λ transform.

To validate full consistency of our pipeline, we can generate an ensemble of Gaussian Random Field (GRF) initial data, imbue each with a set arbitrary bispectrum using MODAL, then analyse the resultant δ_{NG} with our β pipeline estimator. The extent to which we are able to recover an input f_{NL} gives us some insight as to the intrinsic fidelity of our MODAL estimator for simulation and observable datasets - there will naturally be other noise sources and sample size related constraints in those cases, indep. of MODAL.

Below we demonstrate a MODAL estimated reconstruction of a *Tree Level* gravitational bispectrum imbued onto an ensemble of GRF initial data as a function of grid size. 500 modes are used for IC generation and data bispectrum extraction.

Grid Size	$\mathcal{T}$ error (samples)	∼ CPU hours/IC	∼ CPU hours/βs
128 (72 proc)	1.5 (100)	0.1	0.0
256 (72 proc)	0.38 (100)	1.1	0.0
512 (288 proc)	0.22 (100)	11.6	0.5

Table 2. Tree level bispectrum IC generated and extracted from simulated data with MODAL using 500 modes. CPU hours refers to GRF+non-Gaussian IC generation, and bispectral extraction for the ensemble. N.B. there *is* imperfect parallelisation scaling.

$\mathcal{T}$ is a *very* stringent test of performance; it roughly corresponds to the average error in the estimated bispectrum v.s. IC template target at any $\{k_1,k_2,k_3\}$ position in $\mathcal{V}_B$. The error in $\mathcal{T}$ always exceeds that in f_{NL}.

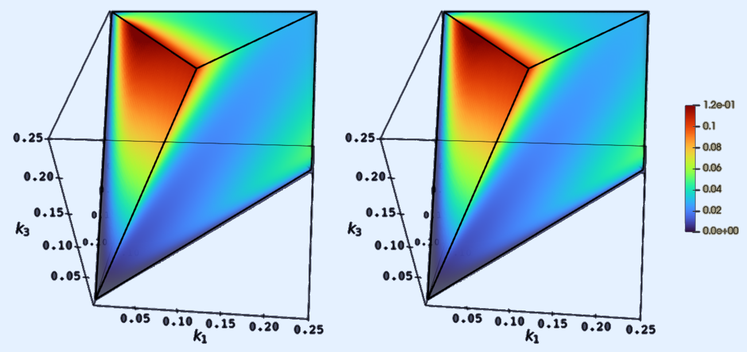

Figure 3. The theoretical tree bispectrum (left) compared against the average extracted from 100 IC generations on a 512 grid (right).

Conclusions and Future work

The sections above clearly demonstrate a versatile and effective compressed MODAL pipeline for generation and extraction of bispectral statistics in LSS datasets. Our extraction pipeline is comfortably fast enough for use cases with *and* without vast HPC resources. Our IC pipeline is fast in comparison to the N-body codes that they naturally supplement.

In coming months we intend to use our high fidelity MODAL estimator to evaluate mock catalogues for upcoming surveys, probe improved methods of IC setting in N-body simulations, and for the extraction of EFT parameters. In future we intend to publish a public version of the code(s) for the wider cosmological community to use in their analyses.

Acknowledgements

With thanks to Professor Paul Shellard, Alexander Miranthis, and Petar Suman.

This work was performed using the DiRAC Data Intensive service at Leicester, operated by the University of Leicester IT Services, which forms part of the STFC DiRAC HPC Facility (www.dirac.ac.uk). The equipment was funded by BEIS capital funding via STFC capital grants ST/K000373/1 and ST/R002363/1 and STFC DiRAC Operations grant ST/R001014/1. DiRAC is part of the National e-Infrastructure.

References

[1] J. R. Fergusson, M. Liguori, and E. P. S. Shellard.
General CMB and primordial bispectrum estimation: Mode expansion, map making, and measures of F_{NL}.
Phys. Rev. D, 82(2):023502, July 2010.

[2] J. R. Fergusson, D. M. Regan, and E. P. S. Shellard.
Rapid Separable Analysis of Higher Order Correlators in Large Scale Structure.
Phys. Rev. D, 86:063511, 2012.

[3] D. M. Regan, M. M. Schmittfull, E. P. S. Shellard, and J. R. Fergusson.
Universal non-Gaussian initial conditions for N-body simulations.
Phys. Rev. D, 86(12):123524, December 2012.