Separating the Inseparable

Oliver H. E. Philcox[1,2,3,*] and et al.
[1]Department of Physics, Columbia University, New York, NY 10027, USA
[2]Department of Physics, Stanford University, Stanford, CA 94305, USA
[3]Simons Society of Fellows, Simons Foundation, New York, NY 10010, USA

Overview of ML x Inflation project.

I. PRIMORDIAL BISPECTRA

New physics generates higher-order statistics at the end of inflation. This is best described by thinking about the statistics of the primordial curvature perturbation, ζ. This has power spectrum P_ζ:

$$\langle\zeta(\mathbf{k})\zeta(\mathbf{k}')\rangle = (2\pi)^3\delta_{\rm D}\left(\mathbf{k}+\mathbf{k}'\right)P_\zeta(k), \qquad P_\zeta(k) = \frac{\Delta^2_\zeta}{k^3}\left(\frac{k}{k_{\rm pivot}}\right)^{n_s-1}, \tag{1}$$

where $\Delta^2_\zeta \equiv 2\pi^2 A_s$ is the power spectrum amplitude, $k_{\rm pivot} = 0.05{\rm Mpc}^{-1}$ (usually) and $n_s \approx 0.96$ is the slope. New physics is well described by the bispectrum of the curvature perturbation ζ:

$$\langle\zeta(\mathbf{k}_1)\zeta(\mathbf{k}_2)\zeta(\mathbf{k}_3)\rangle_c = (2\pi)^3\delta_{\rm D}\left(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3\right)B_\zeta(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3). \tag{2}$$

This usually satisfies certain properties, in particular:

- It is real. $B^*_\zeta = B_\zeta$.
- It is isotropic. $B_\zeta(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) = B_\zeta(k_1,k_2,k_3)$, where $k_i \equiv |\mathbf{k}_i|$.
- It is fully symmetric: $B_\zeta(k_1,k_2,k_3) = B_\zeta(k_2,k_1,k_3)$ *et cetera.*
- It is (usually) approximately scale-invariant.

$$B_\zeta(\alpha k_1,\alpha k_2,\alpha k_3) = \alpha^{2(n_s-4)}B_\zeta(k_1,k_2,k_3), \tag{3}$$

  where α is any constant. If $n_s \approx 1$ then $B_\zeta \sim k^{-6}$. This property can be broken in some models of inflation.
- It is normalized. We typically split the bispectrum into a *template* and an *amplitude*, as

$$B_\zeta(k_1,k_2,k_3) = f^X_{\rm NL}B_X(k_1,k_2,k_3), \tag{4}$$

  where B_X encodes the model of interest (and one could sum over X if there are multiple interesting templates). The amplitude $f^X_{\rm NL}$ is defined via

$$B_\zeta(k_{\rm pivot},k_{\rm pivot},k_{\rm pivot}) = \frac{18}{5}f^X_{\rm NL}\frac{\Delta^4_\zeta}{k^6_{\rm pivot}}, \tag{5}$$

  i.e. it measures the amplitude of the equilateral bispectrum at some given wavenumber relative to the power spectrum squared. (If $n_s = 1$, it is the amplitude of $k^6B_\zeta(k,k,k)$ on all scales).

The last property is typically used to define a dimensionless *shape function*. Here, we'll define this as

$$B_X(k_1,k_2,k_3) = \frac{18}{5}\left[P_\zeta(k_1)P_\zeta(k_2)P_\zeta(k_3)\right]^{2/3}S_X(k_1,k_2,k_3), \tag{6}$$

with $S_X(k_{\rm pivot},k_{\rm pivot},k_{\rm pivot}) = 1$. It's common to define this using the simpler $n_s = 1$ limit (whence $B_X(k_1,k_2,k_3) = \frac{18}{5}S_X(k_1,k_2,k_3)/(k_1k_2k_3)^2$), but we'll keep the general form here. Note that many inflation calculations assume $n_s = 1$ (the de Sitter limit).

---

* ohep2@cantab.ac.uk

II. CMB BISPECTRA

The CMB bispectrum generated by a primordial signal is given by

$$\langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} \rangle_c = f_{\rm NL}^X \int_{\mathbf{k}_1\mathbf{k}_2\mathbf{k}_3} (2\pi)^3 \delta_{\rm D}\left(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3\right) B_X(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) h_{\ell_1 m_1}(\mathbf{k}_1) h_{\ell_2 m_2}(\mathbf{k}_2) h_{\ell_3 m_3}(\mathbf{k}_3), \tag{7}$$

where $\int_{\mathbf{k}} \equiv (2\pi)^{-3}\int d^3\mathbf{k}$, and $h_{\ell m}(\mathbf{k})$ is some transfer function given explicitly by

$$h_{\ell m}(\mathbf{k}) = 4\pi i^\ell \int_{\mathbf{k}} \mathcal{T}_\ell(k) Y^*_{\ell m}(\hat{\mathbf{k}}) \tag{8}$$

where $Y_{\ell m}$ is a spherical harmonic and $\mathcal{T}_\ell$ is the radiation transfer function from CLASS. The exact form of this isn't important here.

A common task in primordial CMB cosmology is to estimate $f_{\rm NL}^X$ given a template B_X. To do this, one typically uses a matched filter, essentially computing:

$$\widehat{f}_{\rm NL}^X \sim \sum_{\ell_i m_i} \frac{\partial \langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} \rangle_c}{\partial f_{\rm NL}^X} \tilde{a}^*_{\ell_1 m_1} \tilde{a}^*_{\ell_2 m_2} \tilde{a}^*_{\ell_3 m_3}, \tag{9}$$

projecting the model onto the (suitably filtered) data $\tilde{a}_{\ell m}$ and being lazy with summation indices.

Let's see how this can be computed in practice. From our model:

$$\widehat{f}_{\rm NL}^X \sim \int d\mathbf{r} \prod_{i=1}^3 \left[\sum_{\ell_i m_i} \int_{\mathbf{k}_i} h_{\ell_i m_i}(\mathbf{k}) \tilde{a}^*_{\ell_i m_i} e^{i\mathbf{k}_i\cdot\mathbf{r}}\right] B_X(k_1,k_2,k_3), \tag{10}$$

where we have used a computational trick to write $(2\pi)^3\delta_{\rm D}\left(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3\right) = \int d\mathbf{r}\, e^{i(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3)\cdot\mathbf{r}}$ which will help us later. This is a *mess* to compute. We have three sums over ℓ, m (scaling as $\ell_{\rm max}^6$), three integrals over all the **k**-modes of interest, *and* an integral over **r**.

Here, our lives are made much easier by assuming that the bispectrum is *separable*. In this case,

$$B_X(k_1,k_2,k_3) = f_1(k_1) f_2(k_2) f_3(k_3). \tag{11}$$

and we can rewrite our estimator as

$$\widehat{f}_{\rm NL}^X \sim \int d\mathbf{r} \prod_{i=1}^3 \left[\sum_{\ell_i m_i} \int_{\mathbf{k}_i} h_{\ell_i m_i}(\mathbf{k}) \tilde{a}^*_{\ell_i m_i} e^{i\mathbf{k}_i\cdot\mathbf{r}} f_i(k_i)\right]. \tag{12}$$

This is a *huge* amount easier to compute. Now we take our data, do a few spherical harmonic transforms and a numerical integral (all of which are possible in $\mathcal{O}(\ell_{\rm max}^2)$ time) and spit out a function of **r**-only (the square bracket). Denoting this by $F(\mathbf{r})$, the final estimator is just

$$\widehat{f}_{\rm NL}^X \sim \int d\mathbf{r} F_1(\mathbf{r}) F_2(\mathbf{r}) F_3(\mathbf{r}) \tag{13}$$

which is just a 3D sum. Because of the separability in (11), we are able to estimate $f_{\rm NL}$ efficiently. This is known as the KSW estimator. Without the separability, we cannot perform the computation directly, and thus cannot measure $f_{\rm NL}^X$. (There are some work-arounds, such as measuring a *binned* bispectrum like one does in LSS, or projecting onto some polynomial coefficients, known as a *modal* decomposition. These are interesting but have their limitations).

III. PROBLEM STATEMENT

The above discussion motivates the following. What if we could split an arbitrary bispectrum up into a sum of separable components? Let's imagine we have some precomputed bispectrum template B_X. This might be analytic, or it might be tabulated numerically (say for $k \in [2, 2000]/\chi_\star$, where $\chi_\star$ is the distance to the CMB, $\approx 14\,000\,\mathrm{Mpc}$). Can we find a separable approximation? This would look like:

$$B_X(k_1,k_2,k_3) \approx B_X^{\rm fact}(k_1,k_2,k_3) \equiv \sum_{n=1}^N f_1^{(n)}(k_1) f_2^{(n)}(k_2) f_3^{(n)}(k_3) + 2 \text{ perms.}, \tag{14}$$

where we have included the permutations explicitly to ensure the symmetries, and we sum over N terms. If the template of interest *is* separable (e.g., the local and equilateral templates below), the decomposition is exact using only a small number of terms. In general, one searches for an *approximate* decomposition, trying to define a set of functions $\{f_i^{(n)}\}$, where the above form is a good approximation to the full $B_X(k_1, k_2, k_3)$. In some simple examples, f_n could be polynomials; however, many physical templates of interest have more complex dependencies, featuring, for example $1/(k_1+k_2+k_3)$, $1/(k_1+k_2-k_3)$ or $\sin[\ln k_1/(k_2+k_3)]$. For such shapes, we probably require non-analytic f_n functions e.g., some form of neural networks.

The full problem can be described as follows.

- **Start**: We have a precomputed bispectrum $B_X(k_1, k_2, k_3)$ for a range of values of k. In many simple cases, we can set $n_s = 1$ and work with the shape function, S_X, instead of B_X which has simpler symmetries (it is symmetric, satisfies $S(k, k, k) = 1$ and scale invariant).

- **Problem**: Find a set of functions $\{f_i^{(n)}(k)\}$ for which the factorized bispectrum $f_i^{(n)}$ closely approximates the true bispectrum B_X. This can be done at fixed N. In practice, we would want to search for the best approximation at $N = 1$ (for example), then try increasing N gradually until we get an approximation of appropriate accuracy.

- **Distance Metric**: We need a measure of distance between the factorized and true bispectra. This is given in terms of the following inner product:

$$\langle S|S'\rangle = \int_0^1 dx \int_{1-x}^1 dy\, S(1, x, y)S'(1, x, y), \tag{15}$$

  assuming $n_s = 1$, for two shapes S, S'. Outside the simple $n_s = 1$ scale-invariant limit, one can use the more general form (which is proportional to the former in the aforementioned limit)

$$\langle B|B'\rangle = \int_{\text{triangle}} dk_1 dk_2 dk_3 \frac{B(k_1, k_2, k_3)B'(k_1, k_2, k_3)}{P_\zeta(k_1)P_\zeta(k_2)P_\zeta(k_3)}, \tag{16}$$

  where the integral is taken over the triangle domain specified by $|k_1 - k_2| \leq k_3 \leq k_1 + k_2$. We define a (log) loss function as

$$\mathcal{L}(B^{\text{fact}}|B) = \langle B - B^{\text{fact}}|B - B^{\text{fact}}\rangle, \tag{17}$$

  which is convex. Note that $\langle B_X|B_X\rangle^{-1/2}$ is proportional to the theoretical error bar on f_{NL}^X from an ideal 3D experiment – our loss function quantifies how well the approximated template captures the signal-to-noise of the former.

- **Tools**: Any functional representation of $f_i^{(n)}(k)$, whether it is a spline, symbolic regression, neural networks, FFTs of neural networks or beyond.

- **Outcome**: Find some set of functions which has $\mathcal{L}(B^{\text{fact}}|B) \ll \langle B|B\rangle$, such that our approximation is accurate. Then we can compute the constraints on these using known methods.

IV. EXAMPLES

Here, I give a couple of notable bispectrum examples (always assuming $n_s = 1$). First the local bispectrum tells us about multi-field inflation and is defined by its shape function:

$$S_{\text{loc}}(k_1, k_2, k_3) = \frac{1}{3}\frac{k_1^2}{k_2 k_3} + 2 \text{ perms} \tag{18}$$

The equilateral shape probes particle dynamics and is given by

$$S_{\text{eq}}(k_1, k_2, k_3) = \left(\frac{k_1}{k_2} + \ 5 \text{ perms.}\right) - \left(\frac{k_1^2}{k_2 k_3} + 2 \text{ perms.}\right) - 2. \tag{19}$$

These are clearly explicitly sum-separable with a small number of terms. There's also interesting *folded* shapes, which involve, e.g. $1/(k_1+k_2-k_3)$ and peak at $k_3 \approx k_1+k_2$. These can tell us about non-standard vacuum states in inflation.

There's some interesting examples from the *cosmological collider* approach. These use symmetries of inflationary space-time to derive three-point (and beyond) correlation function. This is a good template for the "low-speed" collider (2307.01751):

$$S_\alpha(k_1,k_2,k_3) = S_{\rm eq}(k_1,k_2,k_3) + \frac{1}{3}\frac{k_1^2}{k_2k_3}\left[1+\left(\alpha\frac{k_1^2}{k_2k_3}\right)\right]^{-1} + 2\text{ perms.} \tag{20}$$

for $\alpha \lesssim 1$. A separable form *does* exist, but it requires some contour integration tricks to obtain and has a lot of terms (infinite, in fact). For massive particles, there's some funky oscillations, e.g. (1610.06559)

$$S_{\rm clock}(k_1,k_2,k_3) \propto \Theta_H(k_1+k_2-\alpha_0k_3)\left(\frac{k_3}{k_1+k_2}\right)^{1/2}\sin\left(\mu\ln\frac{k_1+k_2}{2k_3}+\delta\right) + 2\text{ perms.} \tag{21}$$

where $\alpha_0 \sim 5$ restricts to squeezed shapes ($k_3 \ll k_1+k_2$) via the Heaviside function Θ_H, $\mu > 0$ encodes the mass of the field and δ is some phase. For slightly less massive fields, there's shapes like (1610.06559)

$$S_{\rm int}(k_1,k_2,k_3) \propto \Theta_H(k_1+k_2-\alpha_0k_3)\frac{k_1^2+k_2^2+k_3^2}{(k_1+k_2+k_3)^{7/2-3\nu}}(k_1k_2k_3)^{1/2-\nu} \tag{22}$$

for $0<\nu<3/2$, which are not immediately separable. This scales as $(k_3/k_1)^{1/2-\nu}$ in the squeezed limit.

Some interesting shapes are given by the recent numerical approaches to inflationary computation, e.g. the CosmoFlow code. These cannot be computed analytically. It would be *great* to be able to analyze these, since they're a more principled approach than the templates above (which are computed only in certain limits and are approximate at best).