PCCP

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2022, **24**, 20776

Cumulant mapping as the basis of multi-dimensional spectrometry†

Leszek J. Frasinski

Received 25th May 2022,
Accepted 21st July 2022

DOI: 10.1039/d2cp02365b

rsc.li/pccp

Cumulant mapping employs a statistical reconstruction of the whole by sampling its parts. The theory developed in this work formalises and extends *ad hoc* methods of 'multi-fold' or 'multi-dimensional' covariance mapping. Explicit formulae have been derived for the expected values of up to the 6th cumulant and the variance has been calculated for up to the 4th cumulant. A method of extending these formulae to higher cumulants has been described. The formulae take into account reduced fragment detection efficiency and a background from uncorrelated events. Number of samples needed for suppressing the statistical noise to a required level can be estimated using Matlab code included in Supplemental Material. The theory can be used to assess the experimental feasibility of studying molecular fragmentations induced by femtosecond or X-ray free-electron lasers. It is also relevant for extending the conventional mass spectrometry of biomolecules to multiple dimensions.

1 Introduction and motivation

Cumulant mapping is an extension of covariance mapping[1] to more than two correlated variables. The covariance mapping technique in turn is an extension of the coincidence method[2,3] to high counting rates, where several fragmentation events may occur in an elementary sample.

From its invention covariance mapping has been used mostly in studies of ionization and fragmentation of small molecules, with some notable exceptions, such as X-ray scattering or brain studies.[4,5] Since covariance mapping requires extensive data processing, two-dimensional maps have been most practical. With continued progress in computational power an extension of covariance mapping to higher dimensions is a timely proposition.

Two recent developments have motivated this work. One is a successful application of covariance mapping to two-dimensional mass spectrometry of large biomolecules,[6,7] with good prospects for extending this technique to higher dimensions. The second development is the emergence of X-ray free-electron lasers (XFELs), which are powerful research tools for studying atomic and molecular dynamics on the femtosecond and attosecond timescales.[8] The unprecedented intensity of X-rays in the XFEL pulses induces a large number of fragmentation events, which leaves covariance mapping as the only practical method for correlating the fragments. Moreover, recent XFEL upgrades to high repetition rates, including fast data acquisition,[8] make the extension of covariance mapping to higher dimensions feasible.

2 Fragmentation scenario

The scheme for cumulant mapping is outlined in Fig. 1. A random sample of unknown objects is drawn from a Poisson distribution. The objects are fragmented, and the fragments are detected. To understand the basic principle, it is helpful to consider initially an ideal scenario where the objects are identical and always break up in the same way into distinguishable fragments. The fragments of each kind are collected in separate bins, and their number is recorded as Z, Y, X, *etc.* The sampling is repeated many times and the fragment numbers are used to reconstruct the parent objects.

2.1 Realistic conditions

The reconstruction would be trivial under the ideal conditions outlined above. In practice, however, the collected fragment numbers require extensive statistical processing for several reasons.

Firstly, each of the Z, Y, X,... random variables is usually measured at just one point on a fluctuating spectrum of mass, energy, or other quantity characterising the fragments, and normally it is not obvious in advance into which parts of the spectra the fragments may fall. Therefore it necessary to calculate the $\varkappa_n$ cumulant for each possible n-tuple of spectral points and display the reconstructed parent objects in an n-dimensional map, as shown at the bottom of Fig. 1. Moreover, the spectra of

Department of Physics, Imperial College London, London SW7 2AZ, UK.
E-mail: l.j.frasinski@imperial.ac.uk

† Electronic supplementary information (ESI) available: Matlab code that calculates the values and variances of the first four cumulants. See DOI: https://doi.org/10.1039/d2cp02365b

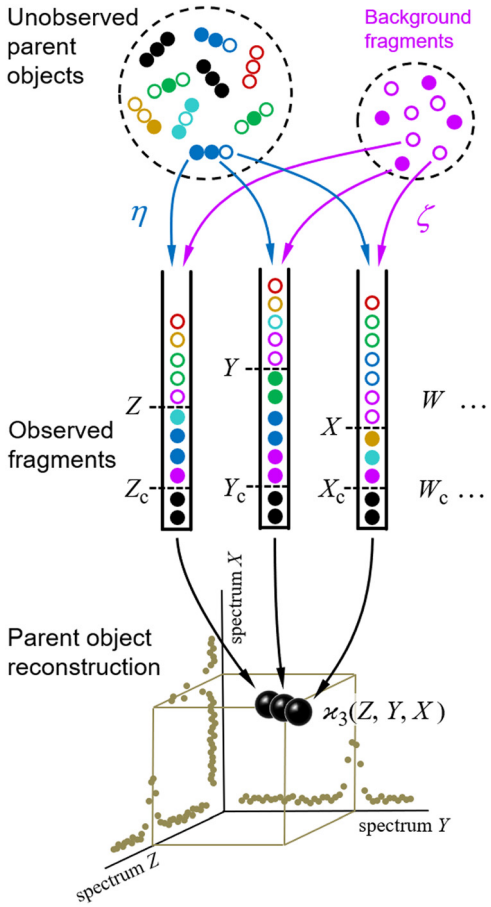

Fig. 1 The principle of reconstruction from fragments. A Poissonian sample of parent objects is fragmented and the fragments are detected with efficiency η, including some uncorrelated background in a mean proportion ζ. The detected fragments (filled circles) are counted and stored in discrete random variables Z, Y, X,..., which are processed over many samples using the cumulant formula. The cumulant value, $\varkappa_n(Z, Y, X, \ldots)$, measures the number of parent objects at a single point in an n-dimensional spectral map.

the Z, Y, X,... variables could be more than 1-dimensional, for example, they could be sourced from a position-sensitive timing detector, which effectively resolves fragments in 3D momentum space,[9] and in principle would require mapping in $3n$-dimensional space, unless the fragmentation kinematics can be used to reduce the dimensionality.

Secondly, in most experiments the fragments are detected with quantum efficiency $\eta < 100\%$. This means that some of the fragments are undetected, as indicated by empty circles in Fig. 1. (The colours mark different detection patterns to enable the reader tracing them from the sample to the bins.) Therefore, the number of tuples (Z_c, Y_c, X_c,...) of only collectively-correlated detected fragments (black filled circles) is smaller than the number of all detected fragments Z, Y, X,... (all filled circles).

And thirdly, there may be a Poissonian background of fragments (magenta circles) completely unrelated to the sample of interest. This background is characterised by parameter ζ, which is the ratio of the mean number of the background fragments to the mean number of the sample fragments. Hence, $\zeta = 0$ means no background, $\zeta = 1$ means as much background as the sample signal, *etc.*

Taking into account reduced detection efficiencies and the presence of background fragments relaxes the idealized requirements of identical parent objects and a single fragmentation pattern, which makes cumulant mapping applicable to studies of mixtures and multiplicity of fragmentation channels. The aim of this work is to estimate the cumulant value and noise when the fragments are detected under the non-ideal conditions of $\eta < 100\%$ and $\zeta > 0$.

2.2 Statistical concepts and notation

Statistics, like all natural sciences, encompasses two realms: the factual world of tangible measurements and the Platonic world of mathematical abstractions. The objects spanning both worlds are random variables, such as Z, denoted in this work with Roman capitals.

On one hand, a random variable can be measured yielding sample values, which I denote giving it an index, *e.g.* Z_i. On the other hand, the general properties of the random variables, such as probability distributions, moments, *etc.* are mathematical abstractions that cannot be known with absolute certainty. Nevertheless, we can infer these abstract properties by repetitive sampling of the random variable. I assume that in an experiment the conditions stay constant and that we draw N samples of Z, which are indexed by $i = 1, 2, 3, \ldots, N$.

The parameters describing the properties of a random variable are denoted with Greek letters, such as $\varkappa$. To infer a parameter value we can use the samples to construct an estimator denoted by a hat, *e.g.* $\widehat{\varkappa}$. For example, to estimate the first moment of variable Z, we use the sample average indicated by an overline:

$$\widehat{\varkappa_1}(Z) = \overline{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i. \tag{1}$$

Since repeating the experiment gives us a new set of samples, the sample averages and the estimators are also random variables. However, calculating their expected values (or variance, or higher-order moments) fixes them to the theoretical limit when $N \to \infty$. Angular brackets are used to denote the expected values:

$$\begin{aligned}\varkappa_1(Z) = \langle\widehat{\varkappa_1}\rangle = \langle\overline{Z}\rangle &= \left\langle \frac{1}{N}\sum_{i=1}^{N} Z_i \right\rangle \\ &= \frac{1}{N}\sum_{i=1}^{N}\langle Z_i\rangle = \frac{1}{N}N\langle Z\rangle = \langle Z\rangle.\end{aligned} \tag{2}$$

The bulk of this work is dedicated to such calculations.

3 Fragment correlations

In general, $\widehat{\varkappa_n}$ stands for an estimator of collective correlations among n fragments. When $n = 1$ the problem is degenerate and the best we can do is to estimate the mean number of only one

kind of a fragment, Z, using the sample average according to eqn (1).

3.1 Covariance

When $n = 2$ the appropriate estimator is the sample covariance of the two fragments Z and Y:

$$\widehat{\varkappa_2}(Z, Y) = \overline{(Z - \overline{Z})(Y - \overline{Y})}.$$

It is worth noting that that this estimator is biased. In principle, the bias can be removed by using Bessel's correction factor $N/(N-1)$ each time a degree of freedom of the sample has been used to calculate an inner average. In practice however, the bias is insignificant for $N \gg 1$ and can be ignored where appropriate.

Calculating the expected value of this estimator leads to the well known formula for covariance:

$$\begin{aligned}\varkappa_2(Z, Y) &= \langle\widehat{\varkappa_2}\rangle = \left\langle\overline{(Z - \overline{Z})(Y - \overline{Y})}\right\rangle \\ &= \langle(Z - \overline{Z})(Y - \overline{Y})\rangle = \langle(Z - \langle Z\rangle)(Y - \langle Y\rangle)\rangle \\ &= \langle ZY\rangle - \langle Z\rangle\langle Y\rangle = \mathrm{cov}(Z, Y).\end{aligned}$$

Introducing mean-centered variables

$$z_0 = Z - \langle Z\rangle, \quad y_0 = Y - \langle Y\rangle, \tag{3}$$

gives us a compact version of the formula:

$$\varkappa_2(Z, Y) = \langle z_0 y_0\rangle. \tag{4}$$

(Symbols z, y, x, *etc.* are reserved for later use.)

3.2 The problem of more than two fragments

When there are three fragments, an extension of eqn (4) has been proposed:[10]

$$\varkappa_3(Z, Y, X) = \langle z_0 y_0 x_0\rangle, \tag{5}$$

and the suitability of this '3-fold covariance' formula has been demonstrated experimentally[10,11] and theoretically.[12]

One may expect that this method of extending the covariance formula works for four fragments:

$$\varkappa_4^{\mathrm{trial}}(Z, Y, X, W) = \langle z_0 y_0 x_0 w_0\rangle.$$

Unfortunately, this trial for the formula is unsuitable.[13] The reason is that if we have only pairwise correlations, *e.g.* Z with Y and X with W, then

$$\varkappa_4^{\mathrm{trial}} = \langle z_0 y_0\rangle\langle x_0 w_0\rangle \neq 0,$$

but we want $\varkappa_4 = 0$ because there is no collective correlation among all four fragments.

3.3 The solution

To find the correct formula for $n \geq 4$, I start with listing the desired properties of $\varkappa_n = \varkappa_n(Z, Y, X, \ldots)$:

- $\varkappa_n \neq 0$ only if all arguments are collectively correlated;
- $\varkappa_n$ has units of the product of all arguments;
- $\varkappa_n$ is linear in the arguments;
- $\varkappa_n$ is invariant under interchange of any two arguments.

It turns out that these properties uniquely determine the formula. For example, the reader is invited to check that the following formula has the desired properties:

$$\begin{aligned}\varkappa_4(Z, Y, X, W) = &\langle z_0 y_0 x_0 w_0\rangle \\ &- (\langle z_0 y_0\rangle\langle x_0 w_0\rangle + \langle z_0 x_0\rangle\langle y_0 w_0\rangle + \langle z_0 w_0\rangle\langle y_0 x_0\rangle),\end{aligned}$$

and that other products of expected values, such as $\langle z_0\rangle\langle y_0 x_0 w_0\rangle$ cannot contribute to the formula because

$$\langle z_0\rangle = \langle Z - \langle Z\rangle\rangle = \langle Z\rangle - \langle Z\rangle = 0.$$

The formula for $\varkappa_4$ can be simplified by writing

$$\varkappa_4(Z, Y, X, W) = \langle z_0 y_0 x_0 w_0\rangle - \sum^{3}\langle z_0 y_0\rangle\langle x_0 w_0\rangle, \tag{6}$$

where $\sum^{3}$ denotes a sum over all ways of paring the four variables.

Similarly,

$$\varkappa_5(Z, Y, X, W, V) = \langle z_0 y_0 x_0 w_0 v_0\rangle - \sum^{10}\langle z_0 y_0\rangle\langle x_0 w_0 v_0\rangle, \tag{7}$$

and

$$\begin{aligned}\varkappa_6(Z, Y, X, W, V, U) = &\langle z_0 y_0 x_0 w_0 v_0 u_0\rangle \\ &- \sum^{15}\langle z_0 y_0\rangle\langle x_0 w_0 v_0 u_0\rangle \\ &- \sum^{10}\langle z_0 y_0 x_0\rangle\langle w_0 v_0 u_0\rangle \\ &- \sum^{15}\langle z_0 y_0\rangle\langle x_0 w_0\rangle\langle v_0 u_0\rangle.\end{aligned} \tag{8}$$

Formulae for collective correlations among more variables can be constructed in a similar manner.

3.4 Cumulants

In statistics the $\varkappa_n$ parameter that measure the collective correlations of n random variables is known as the n-variate joint cumulant of the first order[14] (chapters 3, 12 and 13). In this work I shall shorten this name and call it simply the nth cumulant.

Cumulants are useful in seemingly disjoint areas of science, such as light–matter interactions,[15] quantum theory of multi-electron correlations,[16] bond breaking of diatomic molecules,[17] neural network theory,[18] financial data analysis,[19] and gravitational interaction of dark matter.[20] In physics multivariate cumulants are also known as the Ursell functions.[21]

3.5 Physical interpretation of $\varkappa_n$

So far, the cumulants have been introduced as parameters describing multivariate distributions. Now we want to apply

them to the problem of recovering the parent objects from their fragments.

Let us suppose that we repetitively gather a sample of objects in a random manner, so the number of objects S in the sample follows the Poisson distribution:

$$S \sim \mathrm{Pois}(\lambda) \equiv P(S = k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}, \tag{9}$$

where $P(S = k)$ is the probability of having exactly k objects in the sample and parameter λ is the expected number of the objects in a sample. Next, we fragment the objects and detect only the fragments. And from the detected fragments we want to infer the identity of the undetected parent objects.

To understand why cumulants are useful in this task, we first consider ideal conditions: there is only one way of fragmenting the parent, we detect every fragment, and there is no background of fragments from other processes. Such fragmentation process can be written as

$$S \to (Z, Y, X, \ldots),$$

where $(Z, Y, X, \ldots)$ is a tuple containing n fragments. In such a simple process the number of fragments of each kind matches the number of parent objects. Hence, the random variables are equal:

$$Z = Y = X = \ldots = S$$

and

$$z_0 = y_0 = y_0 = \ldots = s_0,$$

which reduces the expected values in eqn (2) and (4)–(8) to the central moments of the Poisson distribution:

$$\langle z_0 y_0 x_0 \ldots \rangle = \langle s_0^n \rangle = \langle (S - \langle S \rangle)^n \rangle = \mu_n.$$

Since these moments are known polynomials of λ[14] (Section 5.9), the calculation of cumulants is straightforward:

$$\varkappa_1(S) = \langle S \rangle = \lambda, \tag{10a}$$

$$\varkappa_2(S, S) = \mu_2 = \lambda, \tag{10b}$$

$$\varkappa_3(S, S, S) = \mu_3 = \lambda, \tag{10c}$$

$$\varkappa_4(S, S, S, S) = \mu_4 - 3\mu_2{}^2 = (\lambda + 3\lambda^2) - 3\lambda^2 = \lambda. \tag{10d}$$

Note how the sum in eqn (6) cancels out the higher powers of λ leaving just the linear term. In fact this is the general property of the Poisson-distribution cumulants[14] (Example 3.10):

$$\varkappa_n(S, S, S, \ldots) = \lambda.$$

We conclude that under the ideal fragmentation scenario cumulant mapping of fragments gives us a statistical estimate of the mean number of parent objects.

4 Uncorrelated background

Ideal conditions are rarely met in practice. We should estimate how a reduced detection efficiency and a background from uncorrelated fragments affect cumulant mapping. Both effects influence the statistics of the cumulant estimator in a similar manner; those fragments that are detected but have undetected siblings effectively contribute to the uncorrelated background, which is depicted in Fig. 1 using non-black filled circles.

When $\eta < 100\%$ or $\zeta > 0$, then the random variables S, Z, Y,... are only partially correlated. To find the formula for $\varkappa_n$ we need to separate the correlated and uncorrelated parts of these variables.

4.1 Correlated and uncorrelated parts

While the parent objects in the sample still follow the Poisson distribution given by eqn (9), the reduced detection efficiency effectively combines the parent distribution with a binomial distribution of the partial detection giving another Poisson distribution:

$$Z \sim \mathrm{Pois}(\lambda) * \mathrm{Binom}(\eta_Z) \to \mathrm{Pois}(\eta_Z \lambda).$$

When a Poissonian background from other, uncorrelated fragments is added, the compound probability distribution continues to be Poissonian with a modified expected value:

$$Z \sim \mathrm{Pois}(\eta_Z \lambda) \boxplus \mathrm{Pois}(\eta_Z \zeta_Z \lambda) \to \mathrm{Pois}(\eta_Z(1 + \zeta_Z)\lambda).$$

Similarly, $Y \sim \mathrm{Pois}(\eta_Y(1 + \zeta_Y)\lambda)$, *etc.*

Since the binomial sampling of each of the Z, Y, X,... fragments is independent, their joint detection efficiency is a product of the individual efficiencies. Therefore the probability distribution of the detected fragments Z correlated with all the other detected fragments is given by

$$Z_\mathrm{c} \sim \mathrm{Pois}(\theta_n \lambda),$$

where

$$\theta_n = \eta_Z \eta_Y \eta_X \ldots.$$

By the same reasoning $Y_\mathrm{c} \sim \mathrm{Pois}(\theta_n \lambda)$, $X_\mathrm{c} \sim \mathrm{Pois}(\theta_n \lambda)$, *etc.* Moreover, the correlated parts are present in each kind of fragment to the same extent, therefore

$$Z_\mathrm{c} = Y_\mathrm{c} = X_\mathrm{c} = \ldots = S_\mathrm{c} \sim \mathrm{Pois}(\theta_n \lambda). \tag{11}$$

Since the number of detected fragments is the sum of the correlated and uncorrelated parts:

$$Z = Z_\mathrm{c} + Z_\mathrm{u}, \quad Y = Y_\mathrm{c} + Y_\mathrm{u}, \quad X = X_\mathrm{c} + X_\mathrm{u}, \ldots \tag{12}$$

and a sum of Poisson distributions is a Poisson distribution, the distributions of uncorrelated parts can be found as follows:

$$\begin{aligned} Z_\mathrm{u} = Z - Z_\mathrm{c} &\sim \mathrm{Pois}(\eta_Z(1 + \zeta_Z)\lambda - \theta_n \lambda) \\ &= \mathrm{Pois}((\eta_Z(1 + \zeta_Z) - \theta_n)\lambda), \\ Y_\mathrm{u} &\sim \mathrm{Pois}((\eta_Y(1 + \zeta_Y) - \theta_n)\lambda), \\ X_\mathrm{u} &\sim \mathrm{Pois}((\eta_X(1 + \zeta_X) - \theta_n)\lambda), \ldots. \end{aligned}$$

Further calculations are significantly simplified when mean-centered variables are introduced for the correlated and

uncorrelated parts:

$$\begin{aligned} s &= S_c - \langle S_c \rangle, \\ z &= Z_u - \langle Z_u \rangle, \\ y &= Y_u - \langle Y_u \rangle, \\ x &= X_u - \langle X_u \rangle, \ldots . \end{aligned} \tag{13}$$

Using eqn (3), (12), and (11) we obtain

$$\begin{aligned} z_0 &= s + z, \\ y_0 &= s + y, \\ x_0 &= s + x, \ldots . \end{aligned} \tag{14}$$

One useful implication of eqn (13) is that the expected values of the mean-centered variables vanish:

$$\langle s \rangle = \langle z \rangle = \langle y \rangle = \langle x \rangle = \ldots = 0. \tag{15}$$

The second useful property is that they can be regarded as independent, which is exactly true if only one or two detectable fragments are produced. For three or more fragments it may happen that one fragment that is not detected relegates the other fragments to the background in spite of their correlation. It can be shown that these residual correlations do not affect the expected values of cumulants, nor the variance of the first and second cumulant.

4.2 Expected values of cumulants

The first cumulant is unusual because formally it cannot distinguish between the sample fragments and the uncorrelated background. To deal with this ambiguity I shall redefine the first cumulant given by eqn (2), so it measures only the fragments coming from the sample:

$$\varkappa_1(Z) = \langle Z_c \rangle = \langle S_c \rangle = \eta_Z \lambda = \theta_1 \lambda. \tag{16}$$

This definition not only makes $\varkappa_1$ consistent with the higher cumulants but also gives it the meaning of a signal that is separate from the background. As illustrated in Fig. 2, when in spectral analysis the sample fragments form a peak riding on a broad background given by $\langle Z_u \rangle = \eta_Z \zeta_Z \lambda$, then $\langle Z_c \rangle = \eta_Z \lambda$ is just the peak height.

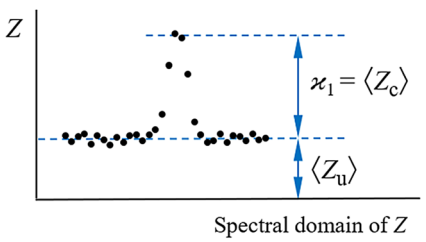

Fig. 2 How to separate the first cumulant from the uncorrelated background. Cumulant $\varkappa_1$ measures only the correlated fragments Z_c that form a peak on a spectrum. The peak rides on a background of uncorrelated fragments Z_u. This definition of $\varkappa_1$ keeps the background parameter $\zeta_Z = \langle Z_u \rangle / \langle Z_c \rangle$ consistent with the higher cumulants.

The higher cumulants can be calculated using the mean-centered variables, $z_0 = s + z$, $y_0 = s + y$, *etc.* For example,

$$\begin{aligned} \varkappa_2(Z, Y) &\overset{4}{=} \langle z_0 y_0 \rangle \overset{14}{=} \langle (s + z)(s + y) \rangle \\ &= \langle s^2 \rangle + \langle s \rangle (\langle z \rangle + \langle y \rangle) + \langle z \rangle \langle y \rangle \\ &\overset{15}{=} \langle s^2 \rangle \overset{13,11}{=} \theta_2 \lambda, \end{aligned} \tag{17}$$

where the numbers above equality signs refer to the equations used.

We notice that the formulae for all higher cumulants can be derived in a similar manner. When expanding $\langle z_0 y_0 x_0 \ldots \rangle$, most of the terms vanish because of eqn (15), and we are left only with polynomials of moments of s. These polynomials are the same as in eqn (10) except now $\varkappa_n$ follows Pois($\theta_n \lambda$) rather then Pois(λ). Therefore, we obtain a general result:

$$\varkappa_n(Z, Y, X, \ldots) = \theta_n \lambda. \tag{18}$$

The simplicity of this result is remarkable. Despite extensive data processing required to estimate cumulants, their meaning is simple: cumulants reconstruct objects from partially detected fragments disregarding any background of other fragments.

4.3 Estimators of cumulants

To construct cumulant estimators, the expected values in eqn (4)–(8) should be replaced with sample averages. The simplest action is to use eqn (1) everywhere. If, however, unbiased estimators are desired, factor $1/N$ should be replaced with $1/(N-1)$ whenever a degree of freedom has already been used, for example

$$\begin{aligned} \widehat{\varkappa_2}(Z, Y) &= \overline{(Z - \overline{Z})(Y - \overline{Y})} \\ &= \frac{1}{N-1} \sum_{i=1}^{N} \left(Z_i - \frac{1}{N} \sum_{i=j}^{N} Z_j \right) \left(Y_i - \frac{1}{N} \sum_{i=j}^{N} Y_j \right). \end{aligned} \tag{19}$$

These estimators can be plotted as 2-dimensional maps[1] or slices of higher-dimensional maps.[10]

4.4 Noise of estimators

Due to the finite number of samples collected, cumulant estimators are noisy. When assessing the feasibility of an experiment involving a cumulant map, the expected noise on the map is of primary concern. It is known that with increasing dimensionality of the map, the noise-to-signal ratio (N/S) increases.[10] There are two sources of this deterioration. Firstly, each time the map dimensionality is increased, the signal decreases because it is multiplied by the detection efficiency according to eqn (18). And secondly, the higher the cumulant, the more subtraction of lower correlations is needed, which contributes more noise from the subtrahends.

These effects can be quantified by calculating the variance of the cumulant estimator, $\mathrm{var}(\widehat{\varkappa_n})$, and finding the noise-to-signal ratio:

$$\mathrm{N/S} = \sigma_n / \varkappa_n, \quad \text{where } \sigma_n = \sqrt{\mathrm{var}(\widehat{\varkappa_n})}. \tag{20}$$

Since the calculations of the variance are lengthy, they are relegated to the appendix. As usual, we find that the standard deviation $\sigma_n \propto 1/\sqrt{N}$, therefore, once we know σ_n we can estimate the number of samples, N, needed for the required noise-to-signal ratio and assess the experimental feasibility.

5 Summary of analytical results

The values and variances of the cumulants are quite complicated functions of the counting rate, λ, the detection efficiency, η, and the relative background, ζ. Rather than inspecting the analytical formulae, it is more informative to plot the results for some chosen argument ranges. Matlab code that calculates the values and variances of cumulants up to the 4th one is provided separately.† When reading and using the code, it is helpful to refer to the equations written in normal mathematical notation.

The equations give cumulant estimates $\widehat{\varkappa_n}$ constructed from the samples according to eqn (1), the cumulant values $\varkappa_n$, and the cumulant variances $\mathrm{var}(\widehat{\varkappa_n})$. The auxiliary quantities are the central moments of the correlated parts, $\langle s^n\rangle$, and of the uncorrelated parts, $\langle z^2\rangle$, $\langle y^2\rangle$, $\langle x^2\rangle$, and $\langle w^2\rangle$ (see eqn (13)).

5.1 1D spectrum

Note that the expected value of the first cumulant depends only on the correlated part, which is justified in the discussion of eqn (16). Hence

$$\begin{aligned}\widehat{\varkappa_1} &= \overline{Z} = \overline{Z_\mathrm{c}} + \overline{Z_\mathrm{u}},\\ \varkappa_1 &= \langle\overline{Z_\mathrm{c}}\rangle = \langle S_\mathrm{c}\rangle = \theta_1\lambda,\\ \mathrm{var}(\widehat{\varkappa_1}) &= \frac{1}{N}\big(\langle s^2\rangle + \langle z^2\rangle\big),\end{aligned}$$

where

$$\begin{aligned}\langle s^2\rangle &= \theta_1\lambda,\\ \langle z^2\rangle &= ((1+\zeta_Z)\eta_Z - \theta_1)\lambda,\\ \theta_1 &= \eta_Z.\end{aligned}$$

5.2 2D covariance map

The second cumulant is commonly known as covariance.

$$\begin{aligned}\widehat{\varkappa_2} &= \overline{(Z-\overline{Z})(Y-\overline{Y})},\\ \varkappa_2 &= \langle\widehat{\varkappa_2}\rangle = \langle s^2\rangle = \theta_2\lambda,\\ \mathrm{var}(\widehat{\varkappa_2}) &\approx \frac{1}{N}\big(\langle s^4\rangle - \langle s^2\rangle^2 + \langle s^2\rangle\big(\langle z^2\rangle + \langle y^2\rangle\big)\\ &\quad + \langle z^2\rangle\langle y^2\rangle\big),\end{aligned}$$

where

$$\begin{aligned}\langle s^2\rangle &= \theta_2\lambda,\\ \langle s^4\rangle &= \theta_2\lambda + 3\theta_2^2\lambda^2,\\ \langle z^2\rangle &= ((1+\zeta_Z)\eta_Z - \theta_2)\lambda,\\ \langle y^2\rangle &= ((1+\zeta_Y)\eta_Y - \theta_2)\lambda,\\ \theta_2 &= \eta_Z\eta_Y.\end{aligned}$$

5.3 3D cumulant map

The third cumulant is sometimes called 3-fold covariance or 3-dimensional covariance.

$$\begin{aligned}\widehat{\varkappa_3} &= \overline{(Z-\overline{Z})(Y-\overline{Y})(X-\overline{X})},\\ \varkappa_3 &= \langle\widehat{\varkappa_3}\rangle = \langle s^3\rangle = \theta_3\lambda,\\ \mathrm{var}(\widehat{\varkappa_3}) &\approx \frac{1}{N}\big(\langle s^6\rangle - \langle s^3\rangle^2 + \langle s^4\rangle\sum^3\langle z^2\rangle\\ &\quad + \langle s^2\rangle\sum^3\langle z^2\rangle\langle y^2\rangle + \langle z^2\rangle\langle y^2\rangle\langle x^2\rangle\big),\end{aligned}$$

where

$$\begin{aligned}\langle s^2\rangle &= \theta_3\lambda,\\ \langle s^3\rangle &= \theta_3\lambda,\\ \langle s^4\rangle &= \theta_3\lambda + 3\theta_3^2\lambda^2,\\ \langle s^6\rangle &= \theta_3\lambda + 25\theta_3^2\lambda^2 + 15\theta_3^3\lambda^3,\\ \langle z^2\rangle &= ((1+\zeta_Z)\eta_Z - \theta_3)\lambda,\\ \langle y^2\rangle &= ((1+\zeta_Y)\eta_Y - \theta_3)\lambda,\\ \langle x^2\rangle &= ((1+\zeta_X)\eta_X - \theta_3)\lambda,\\ \theta_3 &= \eta_Z\eta_Y\eta_X.\end{aligned}$$

5.4 4D cumulant map

The fourth cumulant formulae are the main result of this work.

$$\begin{aligned}\widehat{\varkappa_4} &= \overline{(Z-\overline{Z})(Y-\overline{Y})(X-\overline{X})(W-\overline{W})}\\ &\quad - \sum^3\overline{(Z-\overline{Z})(Y-\overline{Y})}\;\overline{(X-\overline{X})(W-\overline{W})},\\ \varkappa_4 &= \langle\widehat{\varkappa_4}\rangle = \langle s^4\rangle - 3\langle s^2\rangle^2 = \theta_4\lambda,\\ \mathrm{var}(\widehat{\varkappa_4}) &\approx \frac{1}{N}(\langle s^8\rangle - \langle s^4\rangle^2 + 48\langle s^4\rangle\langle s^2\rangle^2 - 12\langle s^6\rangle\langle s^2\rangle - 36\langle s^2\rangle^4\\ &\quad + \big(\langle s^6\rangle - 6\langle s^4\rangle\langle s^2\rangle + 9\langle s^2\rangle^3\big)\sum^4\langle z^2\rangle\\ &\quad + \big(\langle s^4\rangle - \langle s^2\rangle^2\big)\sum^6\langle z^2\rangle\langle y^2\rangle\\ &\quad + \langle s^2\rangle\sum^4\langle z^2\rangle\langle y^2\rangle\langle x^2\rangle + \langle z^2\rangle\langle y^2\rangle\langle x^2\rangle\langle x^2\rangle),\end{aligned}$$

where

$$\langle s^2\rangle = \theta_4\lambda,$$
$$\langle s^4\rangle = \theta_4\lambda + 3\theta_4^2\lambda^2,$$
$$\langle s^6\rangle = \theta_4\lambda + 25\theta_4^2\lambda^2 + 15\theta_4^3\lambda^3,$$
$$\langle s^8\rangle = \theta_4\lambda + 119\theta_4^2\lambda^2 + 490\theta_4^3\lambda^3 + 105\theta_4^4\lambda^4,$$
$$\langle z^2\rangle = ((1+\zeta_Z)\eta_Z - \theta_4)\lambda,$$
$$\langle y^2\rangle = ((1+\zeta_Y)\eta_Y - \theta_4)\lambda,$$
$$\langle x^2\rangle = ((1+\zeta_X)\eta_X - \theta_4)\lambda,$$
$$\langle w^2\rangle = ((1+\zeta_W)\eta_W - \theta_4)\lambda,$$
$$\theta_4 = \eta_Z\eta_Y\eta_X\eta_W.$$

6 Discussion of the results

The figures in this section have been drawn using the Matlab code provided.† For a detailed inspection of the results, it is recommended to run the code and vary the figure options, such as rotate the 3D plots or change the argument ranges.

6.1 Ideal conditions

It is instructive first to look at the results under the ideal conditions of full detection efficiency and no background. Substituting η = 100% and ζ = 0 into the equations in Section 5 we find that there is no contribution from the uncorrelated parts and the expressions for the noise-to-signal ratio (N/S) calculated from eqn (20) are relatively simple functions of λ. These functions are plotted in Fig. 3 for N = 1.

With increasing counting rate, *i.e.* increasing λ, the noise of the first cumulant approaches zero as $1/\sqrt{\lambda}$, which reflects the well-known fact that a high counting rate is always advantageous in collecting 1D spectra.

The N/S of the second cumulant (covariance) approaches a constant value of $\sqrt{2}$ with an increasing counting rate. This tells us that for covariance mapping there is little advantage in increasing λ beyond 1 or 2, unless we want to accommodate weak and strong features on the same map. (In fact a high counting rate exacerbates map distortions due to fluctuations in experimental conditions, which induce common-mode fragment correlations.[22,23])

Cumulants higher than the second one have N/S increasing at high counting rates due to the higher powers of λ present in the expressions for $\mathrm{var}(\widehat{\varkappa_n})$. Therefore, for $n \geq 3$ there is an optimal counting rate at low values of λ as the minima of the green and orange curves show in Fig. 3.

6.2 Reduced detection efficiency

To assess how a reduced detection efficiency affects the noise, we plot N/S as a function of η and $\eta\lambda$, as shown in Fig. 4. The reason for choosing the latter argument rather than just λ is that $\eta\lambda$ is the mean number of only the detected fragments, which is what is observed experimentally on 1D spectra. For simplicity, it is assumed that η is the same for every fragment.

As expected, the noise of the 2nd and higher cumulants substantially increases when the detection efficiency is very low. However, when the detection efficiency is reduced only moderately, to around 50%, the increase in the noise is also moderate, even for the 4th cumulant. Since 50% detection efficiency is

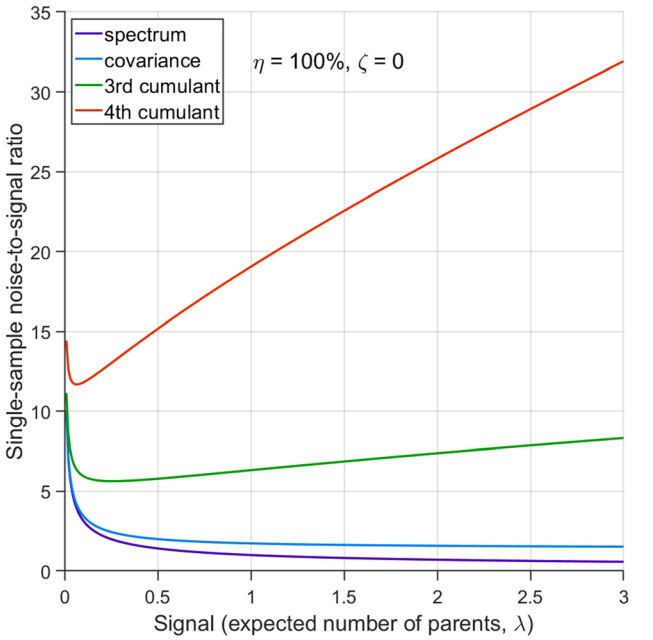

Fig. 3 Cumulant noise as the function of the number of parent objects under the ideal conditions. Unlike for the first and second cumulants, the noise for the higher cumulants is minimal at low counting rates.

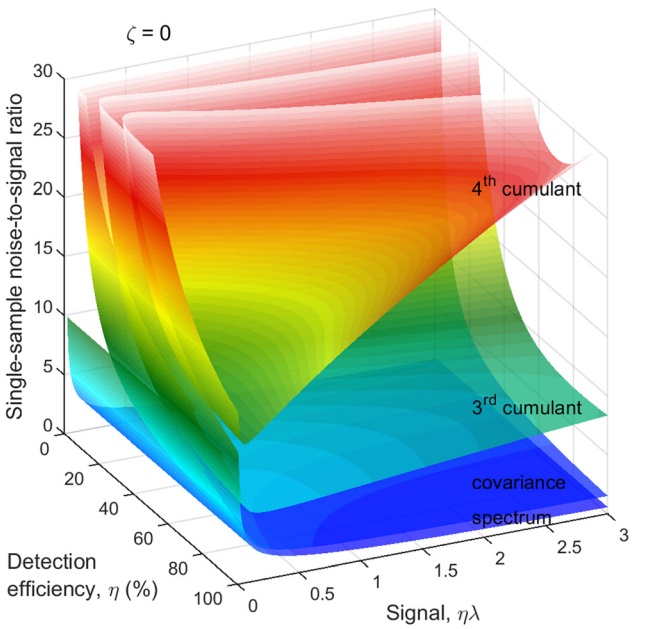

Fig. 4 Cumulant noise as in Fig. 3 but resolved for η. When detection efficiency is reduced to about 50%, there is only a modest increase in the noise.

within the reach of modern particle detectors, it makes cumulant mapping a feasible proposition.

6.3 Background fragments

The next correction to the ideal conditions worth considering is a background of uncorrelated fragments (magenta circles in Fig. 1). This is done using a realistic value of η = 50% and plotting N/S as a function of ζ and $\eta\lambda$ in Fig. 5. This choice of arguments means that they are proportional, respectively, to the relative background level and the height of a peak on a 1D spectrum, as shown in Fig. 2.

For simplicity, we assume the same background level, ζ, for each kind of fragment, which makes the noise of higher cumulants to grow faster with increasing ζ than the lower ones because of the higher powers of ζ present in the expressions for $\mathrm{var}(\widehat{\kappa_n})$. In some experiments this may be an over-pessimistic assumption because some of the Z, Y, X, or W fragments may experience little or no background at all. The code provided† accepts ζ and η tailored to each kind of fragment.

The optimum counting rate is broadly the same as for no background shown in Fig. 4. With an increasing background level, however, the optima for the higher cumulants shift to even lower counting rates.

6.4 Number of samples needed

When planning an experiment, the calculated noise is used to estimate the number of samples, N, needed to suppress N/S to an acceptable level. We can use eqn (20) to calculate N for a fixed noise level, *e.g.* N/S = 0.1. Taking η = 50%, the result is shown in Fig. 6 on a logarithmic scale. The need to reduce the counting rate is clearly visible for the higher cumulants, especially at high background levels. The optimal $\eta\lambda$ for the 3rd and 4th cumulants is below 0.05 at $\zeta > 5$, which means that the observed fragments should be detected in less than 1 in 20 single-sample spectra. Such a low counting rate is comparable to the requirement of coincidence experiments. Unlike coincidences, however, cumulant mapping can accommodate higher counting rates if necessary, albeit at an increased noise.

6.5 Practical implications

Fig. 6 tells us that at least on the order of 10^5 samples will be needed to obtain a good 4th cumulant map. If we want to complete data collections in 15–20 minutes, then the sampling rate should be at least 100 Hz, and 1 kHz or more is desirable.

Such sampling rates are now routinely available from femtosecond lasers and becoming available from XFELs.[8] For example, the LCLS-II XFEL will be operating at up to 1 MHz repetition rate, enabling researchers to probe over 10^9 samples in a single experimental run. In principle, such a large number of samples makes it possible to build clear cumulant maps of even higher order than the 4th one. In practice, however, the data acquisition speed is likely to be the limiting factor, since every single-sample spectrum needs to be recorded.

Cumulant mapping can significantly enhance the conventional mass spectrometry, whose main application is in identifying large biomolecules. The conventional approach is to obtain a high-quality mass spectrum of fragments and search for a match in a large database of molecular spectra. Therefore, the development of commercial spectrometers is driven towards the high mass resolution at the expense of the repetition rate, which is normally below 1 Hz.

Recently covariance mapping has been successfully applied to analyse mass spectra obtained from a commercial spectrometer.[6,7]

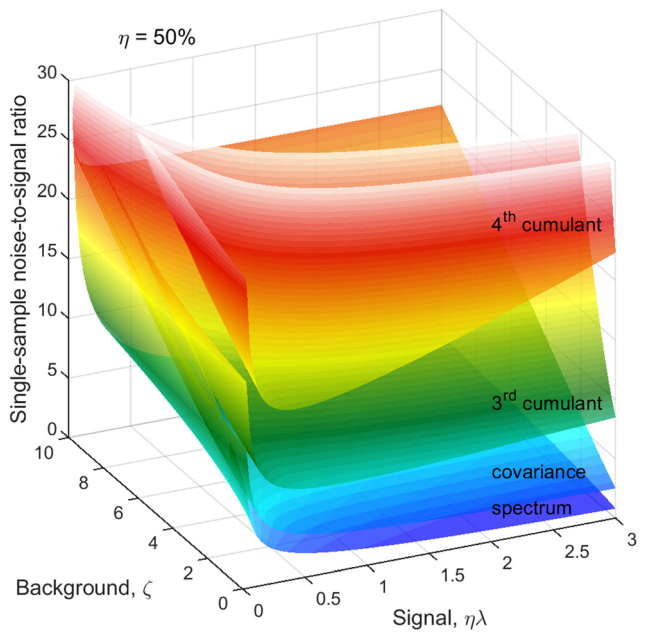

Fig. 5 Cumulant noise as in Fig. 3 but at a reduced η and resolved for ζ. The noise increases with increasing background, especially for the higher cumulants and higher counting rates.

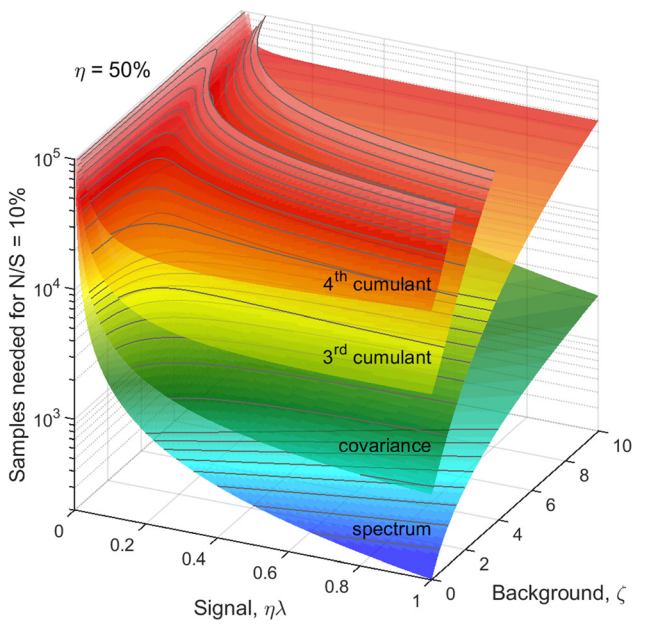

Fig. 6 Number of samples needed for a fixed noise-to-signal ratio. In realistic experimental conditions about 10^5 samples are needed to build a clear 4th cumulant map.

Rather than relying on the high mass resolution, the technique resolves the spectra in the second dimension and partially reconstructs the parent objects on a 2D map. Such a reconstruction allows some parent identifications that would be impossible using only a 1D spectrum of any quality. Clearly, this technique can utilise higher-cumulant mapping to perform a more complete parent reconstruction. Since many samples are needed to build cumulant maps, the spectrometer should operate at high repetition rates, for example, by employing the time-of-flight technique.

The volume of a multi-dimensional cumulant map can be very large and cumbersome to explore. In the simplest approach, cross-sections and projections of the map can be used to visualise the reconstructed molecules.[10] If the locations of the reconstructed molecules are to be found, computational methods of artificial intelligence can be used to discover and identify them.

Cumulant mapping spectrometry can be combined with laser-induced fragmentation. On one hand, such combination allows us to elucidate the dynamics of molecular ionization and fragmentation,[9] on the other hand, it makes it possible to tune the fragmentation to specific bonds in large biomolecules.[24]

7 Conclusions and outlook

Cumulant mapping forms a firm theoretical basis for the concept of 'multi-fold covariance'. The derived formulae enable the experimentalist to assess quantitatively the feasibility of studying multiple correlations in a fragmentation experiment. The key requirements are detection efficiency of around 50%, and a sufficient number of samples, which in practice translates to sampling at high repetition rates.

Studies of molecular fragmentations induced by femtosecond or X-ray lasers are obvious areas for applying cumulant mapping. The technique can be used to extend the conventional mass spectrometry to multiple dimensions and substantially enhance its selectivity. Extension to particles other than electrons or ions should be straightforward. In particular, photons in a wide spectral range from near infrared to hard gamma rays are the promising candidates.

Since cumulant mapping relies on the Poisson distribution of the samples, in principle, it is applicable to any repetitive Poissonian process. For example, the neuronal spike trains closely follow Poisson point processes.[25,26] It could be speculated that cumulants represent the high-level brain functions emerging from correlations in low-level neuronal activity.

Conflicts of interest

There are no conflicts to declare.

Appendix: variance calculations

Variance of $\widehat{\varkappa_n}$ is needed to estimate the number of samples that have to be collected to measure the cumulant with given noise-to-signal ratio. Since only an estimate is needed, I will use approximations in deriving some of the formulae.

Useful formulae

The expected value of random variables, or functions of random variables, is distributive over summation:

$$\langle A + B\rangle = \langle A\rangle + \langle B\rangle. \tag{21}$$

This is generally not true for multiplication, unless A and B are independent:

$$\langle AB\rangle = \langle A\rangle\langle B\rangle + \mathrm{cov}(A, B). \tag{22}$$

Their variances and the covariance can be expressed in terms of expected values:

$$\mathrm{var}(A) = \langle(A - \langle A\rangle)^2\rangle = \langle A^2\rangle - \langle A\rangle^2, \tag{23}$$

$$\mathrm{cov}(A, B) = \langle(A - \langle A\rangle)(B - \langle B\rangle)\rangle = \langle AB\rangle - \langle A\rangle\langle B\rangle. \tag{24}$$

An expected value of a sample average is equal to the expected value of the whole population:

$$\langle\overline{A}\rangle = \langle A\rangle, \tag{25}$$

but the variance and covariance are scaled down by the number of samples[14] (eqn (10.7) and (10.22)):

$$\mathrm{var}(\overline{A}) = \frac{1}{N}(\langle A^2\rangle - \langle A\rangle^2), \tag{26}$$

$$\mathrm{cov}(\overline{A}, \overline{B}) = \frac{1}{N}(\langle AB\rangle - \langle A\rangle\langle B\rangle). \tag{27}$$

Since covariance is a linear function, the covariance of a random variable A with the sum of several other random variables $X_1, X_2, X_3, \ldots$ is the sum of the covariances:

$$\mathrm{cov}\left(A, \sum_i X_i\right) = \sum_i \mathrm{cov}(A, X_i). \tag{28}$$

Whereas variance is a nonlinear function, it can be expanded as follows:

$$\begin{aligned}\mathrm{var}\left(\sum_i X_i\right) &= \mathrm{cov}\left(\sum_i X_i, \sum_j X_j\right) = \sum_i\sum_j \mathrm{cov}(X_i, X_j)\\ &= \sum_i \mathrm{var}(X_i) + 2\sum_i\sum_{j<i} \mathrm{cov}(X_i, X_j).\end{aligned} \tag{29}$$

The variance and covariance of products of random functions can be calculated using the delta method. If $F(\boldsymbol{X})$ and $G(\boldsymbol{X})$ are functions of a random vector $\boldsymbol{X} = [X_1, X_2, X_3, \ldots]$, then the Taylor expansion of F and G gives[14] (eqn (10.12) and (10.13))

$$\begin{aligned}\mathrm{var}(F(\boldsymbol{X})) \approx &\sum_i \left.\frac{\partial F}{\partial X_i}\right|_\xi^2 \mathrm{var}(X_i)\\ &+ \sum_i\sum_{j\neq i} \left.\frac{\partial F}{\partial X_i}\right|_\xi \left.\frac{\partial F}{\partial X_j}\right|_\xi \mathrm{cov}(X_i, X_j),\end{aligned} \tag{30}$$

$$\operatorname{cov}(F(\boldsymbol{X}), G(\boldsymbol{X})) \approx \sum_i \left.\frac{\partial F}{\partial X_i}\right|_{\boldsymbol{\xi}} \left.\frac{\partial G}{\partial X_i}\right|_{\boldsymbol{\xi}} \operatorname{var}(X_i) + \sum_i \sum_{j\neq i} \left.\frac{\partial F}{\partial X_i}\right|_{\boldsymbol{\xi}} \left.\frac{\partial G}{\partial X_j}\right|_{\boldsymbol{\xi}} \operatorname{cov}(X_i, X_j), \tag{31}$$

where the partial derivatives are evaluated at a particular value of $\boldsymbol{X}$, which in this work is set to its expected value $\boldsymbol{\xi} = \langle \boldsymbol{X} \rangle$.

To calculate the variance of the product of two random variables, I make the following choice for eqn (30):

$$F = AB, \quad \boldsymbol{X} = [A, B], \quad \boldsymbol{\xi} = [\langle A \rangle, \langle B \rangle],$$

which allows me to calculate the partial derivatives:

$$\left.\frac{\partial F}{\partial A}\right|_{\boldsymbol{\xi}} = \langle B \rangle, \quad \left.\frac{\partial F}{\partial B}\right|_{\boldsymbol{\xi}} = \langle A \rangle.$$

Substituting these intermediate results into eqn (30) gives us the required formula:

$$\operatorname{var}(AB) \approx \langle A \rangle^2 \operatorname{var}(B) + \langle B \rangle^2 \operatorname{var}(A) + 2\operatorname{cov}(A, B), \tag{32}$$

including the special case when $B = A$:

$$\operatorname{var}(A^2) \approx 4\langle A \rangle^2 \operatorname{var}(A). \tag{33}$$

In a similar way if I choose the following for eqn (31):

$$F = AB, \quad G = CD, \quad \boldsymbol{X} = [A, B, C, D], \quad \boldsymbol{\xi} = [\langle A \rangle, \langle B \rangle, \langle C \rangle, \langle D \rangle],$$

and calculate the other two non-zero partial derivatives

$$\left.\frac{\partial G}{\partial C}\right|_{\boldsymbol{\xi}} = \langle D \rangle, \quad \left.\frac{\partial G}{\partial D}\right|_{\boldsymbol{\xi}} = \langle C \rangle,$$

I obtain the formula for the covariance of products:

$$\operatorname{cov}(AB, CD) \approx \langle A \rangle \langle C \rangle \operatorname{cov}(B, D) + \langle A \rangle \langle D \rangle \operatorname{cov}(B, C) + \langle B \rangle \langle C \rangle \operatorname{cov}(A, D) + \langle B \rangle \langle D \rangle \operatorname{cov}(A, C). \tag{34}$$

When estimating experimental cumulants we use sample averages, $\overline{\boldsymbol{X}}$, to center random variables. In theoretical estimations the expected values, $\langle \boldsymbol{X} \rangle$, can be used instead, which significantly simplifies the calculations. These two versions of the cumulant estimators are asymptotically equal when the number of samples tends to infinity:

$$\widehat{\varkappa_n}(\boldsymbol{X} - \overline{\boldsymbol{X}}) \simeq \widehat{\varkappa_n}(\boldsymbol{X} - \langle \boldsymbol{X} \rangle). \tag{35}$$

This approximation makes no difference when calculating the expected cumulant values, or the variance of the first and the second cumulants. For the third and higher cumulants the experimental variance is a little lower than the theoretical one. This is because the sample average contributes additional negative terms in powers of $1/N$ expansion to the variance[14] (Example 10.3). By using the right hand side of eqn (35), I effectively estimate deviations from the true, unknown value of the cumulant, rather than the noise on the map.

Another approximation I use, also affecting only the variance of the third and higher cumulants, is the neglect of the residual correlations that are discussed following eqn (15).

1D spectrum

The calculation of the variance quoted in Section 5.1 is straightforward:

$$\operatorname{var}(\widehat{\varkappa_1}) \overset{1}{=} \operatorname{var}(\overline{Z}) \overset{23,26}{=} \frac{1}{N}\langle (Z - \langle Z \rangle)^2 \rangle \overset{3,14}{=} \frac{1}{N}\langle (s+z)^2 \rangle \overset{15}{=} \frac{1}{N}\left(\langle s^2 \rangle + \langle z^2 \rangle\right),$$

where the numbers above equality signs refer to the equations used. The last equality relies on the independence of s and z, which is always possible to satisfy having only one variable Z.

2D covariance map

To start variance calculations we write the estimator of the second cumulant in terms of mean-centered variables:

$$\widehat{\varkappa_2} \overset{35}{\simeq} \overline{(Z - \langle Z \rangle)(Y - \langle Y \rangle)} \overset{3,14}{=} \overline{(s+z)(s+y)} = \overline{s^2} + \overline{s(z+y)} + \overline{zy},$$

and expand the variance of the sum:

$$\operatorname{var}(\widehat{\varkappa_2}) \overset{29}{=} \operatorname{var}\left(\overline{s^2}\right) + \operatorname{var}\left(\overline{s(z+y)}\right) + \operatorname{var}(\overline{zy}) + \text{covariance terms}.$$

We calculate the variance terms one by one:

$$\operatorname{var}\left(\overline{s^2}\right) \overset{26}{=} \frac{1}{N}\left(\langle s^4 \rangle - \langle s^2 \rangle^2\right),$$

$$\operatorname{var}\left(\overline{s(z+y)}\right) \overset{26}{=} \frac{1}{N}\left(\langle s^2(z+y)^2 \rangle - \langle s(z+y) \rangle^2\right) \overset{21,22}{=} \frac{1}{N}\Big(\langle s^2 \rangle\left(\langle z^2 \rangle + \langle y^2 \rangle + 2\langle z \rangle \langle y \rangle\right) - \left(\langle s \rangle(\langle z \rangle + \langle y \rangle)\right)^2\Big) \overset{15}{=} \frac{1}{N}\left(\langle s^2 \rangle\left(\langle z^2 \rangle + \langle y^2 \rangle\right)\right),$$

$$\operatorname{var}(\overline{zy}) \overset{26}{=} \frac{1}{N}\left(\langle z^2 y^2 \rangle - \langle zy \rangle^2\right) \overset{22,15}{=} \frac{1}{N}\langle z^2 \rangle \langle y^2 \rangle.$$

Note a useful simplification of eqn (15): a term vanishes if it has a factor linear in the expected value of a mean-centered variable. For this reason all the covariance terms are zero, for example

$$\operatorname{cov}\left(\overline{s^2}, \overline{s(z+y)}\right) \overset{27}{=} \frac{1}{N}\left(\langle s^2 s(z+y) \rangle - \langle s^2 \rangle \langle s(z+y) \rangle\right) \overset{21,22}{=} \frac{1}{N}\left(\langle s^3 \rangle(\langle z \rangle + \langle y \rangle) - \langle s^2 \rangle \langle s \rangle(\langle z \rangle + \langle y \rangle)\right) \overset{15}{=} 0.$$

Gathering all the variance terms we obtain the formula for $\operatorname{var}(\widehat{\varkappa_2})$ given in Section 5.2.

3D cumulant map

The calculation of the third cumulant estimator is similar to the second one:

$$\widehat{\varkappa_3} \overset{35}{\simeq} \overline{(Z - \langle Z \rangle)(Y - \langle Y \rangle)(X - \langle X \rangle)} \overset{3,14}{=} \overline{(s+z)(s+y)(s+x)} = \overline{s^3} + \overline{s^2(z+y+x)} + \overline{s(zy+zx+yx)} + \overline{zyx}.$$

Expanding the variance of the sum we find that the covariance terms vanish again and we need to calculate only variances:

$$\mathrm{var}\left(\overline{s^3}\right) \overset{26}{=} \frac{1}{N}\left(\langle s^6\rangle - \langle s^3\rangle^2\right),$$

$$\mathrm{var}\left(\overline{s^2(z+y+x)}\right) \overset{26}{=} \frac{1}{N}\left(\langle s^4(z+y+x)^2\rangle - \langle s(z+y+x)\rangle^2\right)$$

$$\overset{22}{=} \frac{1}{N}\left(\langle s^4\rangle\left\langle\left(\sum^3 z\right)^2\right\rangle - \langle s\rangle^2\left\langle\sum^3 z\right\rangle^2\right)$$

$$\overset{21,22,15}{=} \frac{1}{N}\langle s^4\rangle\sum^3\langle z^2\rangle,$$

$$\mathrm{var}\left(\overline{s(zy+zx+yx)}\right) \overset{26}{=} \frac{1}{N}\left(\langle s^2(zy+zx+yx)^2\rangle - \langle s(zy+zx+yx)\rangle^2\right)$$

$$\overset{22}{=} \frac{1}{N}\left(\langle s^2\rangle\left\langle\left(\sum^3 zy\right)^2\right\rangle - \langle s\rangle^2\left\langle\sum^3 zy\right\rangle^2\right)$$

$$\overset{21,22,15}{=} \frac{1}{N}\langle s^2\rangle\sum^3\langle z^2\rangle\langle y^2\rangle,$$

$$\mathrm{var}(\overline{zyx}) \overset{26}{=} \frac{1}{N}\left(\langle z^2y^2x^2\rangle - \langle zyx\rangle^2\right) \overset{22,15}{=} \frac{1}{N}\langle z^2\rangle\langle y^2\rangle\langle x^2\rangle.$$

Gathering all the variance terms we obtain the formula for $\mathrm{var}(\widehat{\kappa_3})$ given in Section 5.3.

4D cumulant map

The calculation for the fourth cumulant is more laborious because not only the formula for the estimator is longer but also some of the covariance terms in the variance expansion do not vanish. While using more advanced tools, such as *k*-statistics,[14] may shorten the calculations, I shall continue with the same method as before:

$$\begin{aligned}\widehat{\kappa_4} &\overset{35}{\cong} \overline{(Z-\langle Z\rangle)(Y-\langle Y\rangle)(X-\langle X\rangle)(W-\langle W\rangle)}\\ &\quad - \sum^3 \widehat{\kappa_2}(Z,Y)\,\widehat{\kappa_2}(X,W)\\ &\overset{3,14}{=} \overline{(s+z)(s+y)(s+x)(s+w)}\\ &\quad - \sum^3 \left(\overline{s^2} + \overline{s(z+y)} + \overline{zy}\right)\left(\overline{s^2} + \overline{s(x+w)} + \overline{xw}\right)\\ &= \overline{s^4} + \sum^4 \overline{s^3z} + \sum^6 \overline{s^2zy} + \sum^4 \overline{szyx} + \overline{zyxw}\\ &\quad - 3\overline{s^2}^2 - 3\overline{s^2}\sum^4 \overline{sz} - \overline{s^2}\sum^6 \overline{zy}\\ &\quad - 2\sum^6 \overline{sz}\;\overline{sy} - \sum^{12} \overline{sz}\;\overline{yx} - \sum^3 \overline{zy}\;\overline{xw}.\end{aligned}$$

To calculate the variance of this estimator we apply eqn (29) and evaluate the variance of the first two terms:

$$\mathrm{var}\left(\overline{s^4}\right) \overset{26}{=} \frac{1}{N}\left(\langle s^8\rangle - \langle s^4\rangle^2\right),$$

$$\mathrm{var}\left(\sum^4 \overline{s^3z}\right) \overset{29,15}{=} \sum^4 \mathrm{var}\left(\overline{s^3z}\right) + 0$$

$$\overset{26,15}{=} \frac{1}{N}\left(\sum^4 \langle s^6z^2\rangle - 0^2\right)$$

$$\overset{22}{=} \frac{1}{N}\langle s^6\rangle\sum^4\langle z^2\rangle,$$

and continue applying eqn (29), (26), (22), and (15) to the next three terms:

$$\mathrm{var}\left(\sum^6 \overline{s^2zy}\right) = \frac{1}{N}\langle s^4\rangle\sum^6\langle z^2\rangle\langle y^2\rangle,$$

$$\mathrm{var}\left(\sum^4 \overline{szyx}\right) = \frac{1}{N}\langle s^2\rangle\sum^4\langle z^2\rangle\langle y^2\rangle\langle x^2\rangle,$$

$$\mathrm{var}(\overline{zyxw}) = \frac{1}{N}\langle z^2\rangle\langle y^2\rangle\langle x^2\rangle\langle w^2\rangle.$$

To calculate the variance of the next three terms we can use eqn (32) and (33) for the variance of a product of two variables:

$$\mathrm{var}(-3\overline{s^2}^2) = 9\,\mathrm{var}(\overline{s^2}\;\overline{s^2}) \overset{33}{\approx} 36\langle\overline{s^2}\rangle^2\mathrm{var}\left(\overline{s^2}\right)$$

$$\overset{25,26}{=} \frac{36}{N}\langle s^2\rangle^2\left(\langle s^4\rangle - \langle s^2\rangle^2\right),$$

$$\mathrm{var}\left(-3\overline{s^2}\sum^4\overline{sz}\right) \overset{32,15}{\approx} 9\langle\overline{s^2}\rangle^2\mathrm{var}\left(\sum^4\overline{sz}\right) + 0 + 2\times 0$$

$$\overset{25,29,15}{=} 9\langle s^2\rangle^2\left(\sum^4 \mathrm{var}(\overline{sz}) + 0\right)$$

$$\overset{26,15}{=} \frac{9}{N}\langle s^2\rangle^2\sum^4\left(\langle s^2z^2\rangle - 0^2\right)$$

$$\overset{22}{=} \frac{9}{N}\langle s^2\rangle^3\sum^4\langle z^2\rangle,$$

and in a very similar way we find

$$\mathrm{var}\left(-\overline{s^2}\sum^6\overline{zy}\right) \approx \frac{1}{N}\langle s^2\rangle^2\sum^6\langle z^2\rangle\langle y^2\rangle.$$

When we proceed to calculate the variances of the last line of $\widehat{\kappa_4}$, we notice that all terms in eqn (32) vanish. Therefore all remaining variance terms are zero.

To calculate the covariance terms of $\widehat{\kappa_4}$, we refer to eqn (27) and note that the covariance vanishes unless both *A* and *B* have

the same subset of the z, y, x, w variables. Therefore, the only non-zero covariance terms are:

$$\begin{aligned}
2\,\mathrm{cov}\left(\overline{s^4}, -3\overline{s^2}^2\right) &= -6\mathrm{cov}\left(\overline{s^4}, \overline{s^2}\;\overline{s^2}\right) \\
&\overset{34,15}{\approx} -6\times 2\langle\overline{s^2}\rangle^2\mathrm{cov}\left(\overline{s^4}, \overline{s^2}\right) + 0 \\
&\overset{25,27}{=} -\frac{12}{N}\langle s^2\rangle(\langle s^6\rangle - \langle s^4\rangle\langle s^2\rangle), \\
2\,\mathrm{cov}\left(\sum^4\overline{s^3z}, -3\overline{s^2}\sum^4\overline{sz}\right) &\overset{28,15}{=} -6\sum^4\mathrm{cov}\left(\overline{s^3z}, \overline{s^2}\overline{sz}\right) + 0 \\
&\overset{34,15}{\approx} -6\sum^4(\langle\overline{s^2}\rangle\mathrm{cov}\left(\overline{s^3z}, \overline{sz}\right) + 0) \\
&\overset{25,27,15}{=} -\frac{6}{N}\langle s^4\rangle\langle s^2\rangle\sum^4\langle z^2\rangle,
\end{aligned}$$

and in a very similar way

$$2\,\mathrm{cov}\left(\sum^6\overline{s^2zy}, -\overline{s^2}\sum^6\overline{zy}\right) \approx -\frac{2}{N}\langle s^2\rangle^2\sum^6\langle z^2\rangle\langle y^2\rangle.$$

Gathering all the variance and covariance terms gives us the formula for $\mathrm{var}(\widehat{\varkappa_4})$ given in Section 5.4.

References

1 L. J. Frasinski, K. Codling and P. A. Hatherly, *Science*, 1989, **246**, 1029–1031.
2 J. H. D. Eland, F. S. Wort and R. N. Royds, *J. Electron Spectrosc. Relat. Phenom.*, 1986, **41**, 297–309.
3 L. J. Frasinski, M. Stankiewicz, K. J. Randall, P. A. Hatherly and K. Codling, *J. Phys. B: Atom. Mol. Phys.*, 1986, **19**, L819.
4 L. J. Frasinski, *J. Phys. B: At., Mol. Opt. Phys.*, 2016, **49**, 152004.
5 C. Vallance, D. Heathcote and J. W. L. Lee, *J. Phys. Chem. A*, 2021, **125**, 1117–1133.
6 T. Driver, B. Cooper, R. Ayers, R. Pipkorn, S. Patchkovskii, V. Averbukh, D. R. Klug, J. P. Marangos, L. J. Frasinski and M. Edelson-Averbukh, *Phys. Rev. X*, 2020, **10**, 041004.
7 T. Driver, V. Averbukh, L. J. Frasinski, J. P. Marangos and M. Edelson-Averbukh, *Anal. Chem.*, 2021, **93**, 10779–10788.
8 N. Huang, H. Deng, B. Liu, D. Wang and Z. Zhao, *Innovation*, 2021, **2**, 100097.
9 F. Allum, C. Cheng, A. J. Howard, P. H. Bucksbaum, M. Brouard, T. Weinacht and R. Forbes, *J. Phys. Chem. Lett.*, 2021, **12**, 8302–8308.
10 L. J. Frasinski, P. A. Hatherly and K. Codling, *Phys. Lett. A*, 1991, **156**, 227–232.
11 W. A. Bryan, W. R. Newell, J. H. Sanderson and A. J. Langley, *Phys. Rev. A: At., Mol., Opt. Phys.*, 2006, **74**, 053409.
12 J. Mikosch and S. Patchkovskii, *J. Mod. Opt.*, 2013, **60**, 1426–1438.
13 V. Zhaunerchyk, L. J. Frasinski, J. H. D. Eland and R. Feifel, *Phys. Rev. A: At., Mol., Opt. Phys.*, 2014, **89**, 053418.
14 A. Stuart and K. Ord, *Kendall's advanced theory of statistics. Vol. 1, Distribution theory*, Hodder Arnold, London, 6th edn, 1994.
15 M. Sánchez-Barquilla, R. Silva and J. Feist, *J. Chem. Phys.*, 2020, **152**, 034108.
16 W. Kutzelnigg and D. Mukherjee, *J. Chem. Phys.*, 1999, **110**, 2800–2809.
17 O. Brea, M. El Khatib, C. Angeli, G. L. Bendazzoli, S. Evangelisti and T. Leininger, *J. Chem. Theory Comput.*, 2013, **9**, 5286–5295.
18 M. Helias and D. Dahmen, *Statistical field theory for neural networks*, Springer, 2020.
19 K. Domino, *Phys. A*, 2020, **558**, 124995.
20 C. Uhlemann, *J. Cosmol. Astropart. Phys.*, 2018, 030.
21 H. D. Ursell, *Math. Proc. Cambridge Philos. Soc.*, 1927, **23**, 685–697.
22 L. J. Frasinski, V. Zhaunerchyk, M. Mucke, R. J. Squibb, M. Siano, J. H. D. Eland, P. Linusson, P. Vd Meulen, P. Salén and R. D. Thomas, *et al.*, *Phys. Rev. Lett.*, 2013, **111**, 073002.
23 O. Kornilov, M. Eckstein, M. Rosenblatt, C. P. Schulz, K. Motomura, A. Rouzée, J. Klei, L. Foucar, M. Siano and A. Lübcke, *et al.*, *J. Phys. B: At., Mol. Opt. Phys.*, 2013, **46**, 164028.
24 R. K. Ayers, PhD thesis, Imperial College London, 2022.
25 C. Gardella, O. Marre and T. Mora, *Neural Comput.*, 2019, **31**, 233–269.
26 A. T. Campo, *Comput. Struct. Biotechnol. J.*, 2020, **18**, 2699–2708.