The Annals of Statistics
2013, Vol. 41, No. 2, 982–1004
DOI: 10.1214/13-AOS1107

NATURAL STATISTICS FOR SPECTRAL SAMPLES

BY E. DI NARDO, P. MCCULLAGH AND D. SENATO

University of Basilicata, University of Chicago and University of Basilicata

Spectral sampling is associated with the group of unitary transformations acting on matrices in much the same way that simple random sampling is associated with the symmetric group acting on vectors. This parallel extends to symmetric functions, k-statistics and polykays. We construct spectral k-statistics as unbiased estimators of cumulants of trace powers of a suitable random matrix. Moreover we define normalized spectral polykays in such a way that when the sampling is from an infinite population they return products of free cumulants.

1. Outline. The goals of this paper are threefold.

We first introduce the notion of spectral sampling as an operation on a finite set of n real numbers $\mathbf{x} = (x_1, \dots, x_n)$ generating a random set $\mathbf{y} = (y_1, \dots, y_m)$ of $m \le n$ real numbers whose distribution is determined by $\mathbf{x}$. Spectral sampling is not the same as simple random sampling in the sense that $\mathbf{y}$ is not a subset of $\mathbf{x}$, but the parallels are unmistakable and striking. In particular, there exist symmetric functions $\mathfrak{K}_\lambda$—analogous to k-statistics and polykays—such that $E(\mathfrak{K}_\lambda(\mathbf{y}) \mid \mathbf{x}) = \mathfrak{K}_\lambda(\mathbf{x})$. In other words, the average value of $\mathfrak{K}_\lambda(\cdot)$ for spectral samples $\mathbf{y}$ taken from $\mathbf{x}$ is equal to $\mathfrak{K}_\lambda(\mathbf{x})$. The first goal is to obtain explicit expressions for these spectral k-statistics, which is done in Sections 3–5 using symbolic umbral techniques.

The second goal is to elucidate some of the concepts associated with freeness—free probability and free cumulants—in terms of spectral sampling and spectral k-statistics. For this purpose, spectral sampling may be viewed as a restriction operation $X \mapsto Y$ from a freely randomized Hermitian matrix of order n into a freely randomized Hermitian matrix of order $m \le n$, and each spectral k-statistic is class function depending only on the matrix eigenvalues. In essence, the spectral k-statistics tell us which spectral properties are preserved on average by freely randomized matrix restriction. For example, $\mathfrak{K}_{(1)}(\mathbf{x}) = \bar{\mathbf{x}}$ tells us that the eigenvalue average is preserved. Likewise, if k_2 denotes the usual sample variance with divisor $n - 1$, the second spectral statistic $\mathfrak{K}_{(2)}(\mathbf{x}) = k_2(\mathbf{x})/(n + 1)$ tells us that the eigenvalue sample variance is not preserved, but is, on average, proportional to the sample size plus one.

Finally, by considering the limit as $n \to \infty$, we show that the normalized spectral k-statistics are related to free cumulants in much the same way that polykays are related to ordinary cumulants.

Received June 2012; revised February 2013.
MSC2010 subject classifications. Primary 60B20, 46L53; secondary 62F12.
Key words and phrases. Random matrix, cumulant of traces, free cumulant, polykays.

2. Spectral sampling.

2.1. *Definition*. A random Hermitian matrix A of order n is said to be *freely randomized* if its distribution is invariant under unitary conjugation, that is, $A \sim GAG^\dagger$ for each unitary G. In particular, if H is uniformly distributed with respect to Haar measure on the group of unitary matrices of order n, $HAH^\dagger$ is freely randomized. If A is freely randomized, each leading sub-matrix is also freely randomized.

Let $\mathbf{x} = (x_1, \dots, x_n)$ be given real numbers, let $X = \operatorname{diag}(\mathbf{x})$ be the associated diagonal matrix and let $HXH^\dagger$ be the freely randomized matrix. The sample matrix Y is the leading $m \times m$ sub-matrix in the freely randomized matrix, that is, $Y = (HXH^\dagger)_{[m\times m]}$.

DEFINITION 2.1 (Spectral sample). The set of eigenvalues $\mathbf{y} = (y_1, \dots, y_m) \in \mathbb{R}^m$ of the $m \times m$ Hermitian random matrix $Y = (HXH^\dagger)_{[m\times m]}$ is called a *spectral sample* of size m from $\mathbf{x}$.

For $m = n$, the distribution is uniform with the same weight $1/n!$ on all permutations $\sigma \in \mathfrak{S}_n$; that is, $\mathbf{y}$ is a random permutation of $\mathbf{x}$. For $m < n$, however, the distribution in $\mathbb{R}^m$ is nonatomic, so the sample values $\mathbf{y}$ do not ordinarily occur among the components of $\mathbf{x}$.

If the group of unitary transformations in the preceding definition were replaced by a sub-group, the sampling distribution would be altered accordingly. The most obvious subgroups are the group of orthogonal transformations and the group of permutations $[n] \to [n]$; in each case there is an associated family of spectral functions such that $E(\mathfrak{K}_\lambda(\mathbf{y}) \mid \mathbf{x}) = \mathfrak{K}_\lambda(\mathbf{x})$. In particular, if H is a uniform random permutation, $\mathbf{y}$ is a simple random sample of size m taken from $\mathbf{x}$, and the associated spectral functions are the classical k-statistics due to Fisher [9] and the polykays due to Tukey [21].

REMARK 2.2. Within image compression [14], the random Hermitian matrix Y in Definition 2.1 is called a two-dimensional Haar transform. More generally, if X is a full matrix whose entries are the pixels ranging from 0 (black) to 255 (white), then Y contains reduced information extracted from X via the rectangular Haar matrix H. Similar transformations are employed also within classification, document analysis, hardware implementation and are known as downsampling of a vector or a matrix [18].

2.2. *Natural statistics*. For present purposes, a statistic T is a collection of functions $T_n : \mathbb{R}^n \to \mathbb{R}$ such that $T_m(\mathbf{y})$ and $T_n(\mathbf{x})$ are defined for all samples sufficiently large. For example, the usual sample variance is defined for $n \geq 2$, while the sample skewness is defined for $n \geq 3$.

DEFINITION 2.3 (Natural statistic). A statistic T is said to be *natural* if, for each $m \le n$, the average value of $T_m(\cdot)$ over random samples $\mathbf{y}$ drawn from $\mathbf{x}$ is equal to $T_n(\mathbf{x})$. In symbols,

$$E\big(T_m(\mathbf{y}) \mid \mathbf{x}\big) = T_n(\mathbf{x})$$

for each $m \le n$.

Obviously, the definition depends on what it means for $\mathbf{y}$ to be a random sample drawn from $\mathbf{x}$, that is, the choice of group in Definition 2.1. Thus, a statistic that is natural with respect to simple random sampling (a U-statistic) is not, in general, natural with respect to spectral sampling.

In Tukey [21], such functions were said to be "inherited on the average." The key point in Definition 2.3 is that a natural statistic is not a single function in isolation, but a list of functions $\{T_n : \mathbb{R}^n \to \mathbb{R}\}$. It is the property of inheritance that gives these functions a common interpretation independent of the sample size. One might be inclined to think that inheritance is no different from unbiasedness relative to a model with exchangeably distributed components. However, unbiasedness of T_n does not imply the inheritance property, nor does inheritance imply that the statistic has a limit or that its expectation exists. Unbiasedness in parametric models is a property of individual functions T_n, whereas inheritance is a property of the sequence.

For $m = n$, inheritance implies that each T_n is a symmetric function: $T_n(\mathbf{x})$ is equal to the average of the values on the permutations of $\mathbf{x}$. Tukey [21] proved that the symmetric functions

$$\begin{aligned} \tilde{\mathfrak{a}}_{r,n}(\mathbf{x}) &= \frac{1}{n}\sum_i x_i^r, \qquad \tilde{\mathfrak{a}}_{rs,n}(\mathbf{x}) = \frac{1}{(n)_2}\sum_{i\ne j} x_i^r x_j^s, \\ \tilde{\mathfrak{a}}_{rst,n}(\mathbf{x}) &= \frac{1}{(n)_3}\sum_{i\ne j\ne k} x_i^r x_j^s x_k^t, \qquad \dots \end{aligned} \tag{2.1}$$

defined, respectively, for $n \ge 1$, $n \ge 2$ and so on, are natural with respect to simple random sampling. Here and elsewhere $(n)_r = n(n-1)\cdots(n-r+1)$ denotes the descending factorial function. Ordinarily, we suppress the index n and write $\tilde{\mathfrak{a}}_{rs}(\mathbf{x})$ instead of $\tilde{\mathfrak{a}}_{rs,n}(\mathbf{x})$, the value of n being inferred from the argument $\mathbf{x} \in \mathbb{R}^n$. The unnormalized polynomials

$$\begin{aligned} \mathfrak{a}_r(\mathbf{x}) &= \sum_i x_i^r, \qquad \mathfrak{a}_{rs}(\mathbf{x}) = \sum_{i\ne j} x_i^r x_j^s, \\ \mathfrak{a}_{rsk}(\mathbf{x}) &= \sum_{i\ne j\ne t} x_i^r x_j^s x_t^k, \qquad \dots \end{aligned} \tag{2.2}$$

are the well-known *augmented symmetric functions* [19].

Every expression which is a polynomial, symmetric and inherited on the average can be written as a linear combination of the statistics in (2.1) with coefficients that do not depend on the size of the set [21]. Consequently each linear combination, with scalar coefficients independent of n, also has the inheritance property, as happens, for example, for U-statistics. The combinations that have proved to be most useful for statistical purposes are the k-statistics due to Fisher [9] and the polykays due to Tukey [21, 22], defined as follows:

$$
\begin{aligned}
&k_{(1)} = \tilde{\mathfrak{a}}_{(1)};\\
&k_{(1^2)} = \tilde{\mathfrak{a}}_{(1^2)}, \qquad k_{(2)} = \tilde{\mathfrak{a}}_{(2)} - \tilde{\mathfrak{a}}_{(1^2)};\\
&k_{(1^3)} = \tilde{\mathfrak{a}}_{(1^3)}, \qquad k_{(12)} = \tilde{\mathfrak{a}}_{(12)} - \tilde{\mathfrak{a}}_{(1^3)}, \qquad k_{(3)} = \tilde{\mathfrak{a}}_{(3)} - 3\tilde{\mathfrak{a}}_{(12)} + 2\tilde{\mathfrak{a}}_{(1^3)};\\
&k_{(1^4)} = \tilde{\mathfrak{a}}_{(1^4)}, \qquad k_{(1^2 2)} = \tilde{\mathfrak{a}}_{(1^2 2)} - \tilde{\mathfrak{a}}_{(1^4)}, \qquad k_{(13)} = \tilde{\mathfrak{a}}_{(13)} - 3\tilde{\mathfrak{a}}_{(1^2 2)} + 2\tilde{\mathfrak{a}}_{(1^4)};\\
&k_{(2^2)} = \tilde{\mathfrak{a}}_{(2^2)} - 2\tilde{\mathfrak{a}}_{(1^2 2)} + \tilde{\mathfrak{a}}_{(1^4)},\\
&k_{(4)} = \tilde{\mathfrak{a}}_{(4)} - 4\tilde{\mathfrak{a}}_{(13)} - 3\tilde{\mathfrak{a}}_{(2^2)} + 12\tilde{\mathfrak{a}}_{(1^2 2)} - 6\tilde{\mathfrak{a}}_{(1^4)}.
\end{aligned}
$$

The single index k's are the k-statistics; the multi-index k's are the polykays. For a sample of i.i.d. variables, each k-statistic is an unbiased estimator of the population cumulant, and each polykay is an unbiased estimator of cumulant products. The *degree* of each k is the sum of the subscripts. The set of natural polynomial statistics of degree i is a vector space, of dimension equal to the number of partitions of the integer i, spanned by the k's of degree i.

3. Moment symbolic method. *Univariate case*. The moment symbolic method relies on the classical umbral calculus introduced by Rota and Taylor in 1994 [16], which has been developed and refined in a series of papers starting from [7, 8]. The result is a calculus in which certain symbols represent scalar or polynomial sequences, thereby reducing the overall computational apparatus. We now review the key components.

Let R be the real or complex field whose elements are called scalars. An umbral calculus consists of a generating set $\mathcal{A} = \{\alpha, \beta, \ldots\}$, called the *alphabet*, whose elements are named *umbrae*, a polynomial ring $R[\mathcal{A}]$ and a linear functional $E\colon R[\mathcal{A}] \to R$ called evaluation. The linear functional is such that $E[1] = 1$ and

$$E\big[\alpha^i \beta^j \cdots \gamma^k\big] = E\big[\alpha^i\big] E\big[\beta^j\big] \cdots E\big[\gamma^k\big] \qquad \text{(uncorrelation property)} \tag{3.1}$$

for any set of distinct umbrae in $\mathcal{A}$ and for i, j, k nonnegative integers. To each umbra $\alpha \in \mathcal{A}$ there corresponds a sequence of scalars $a_i = E[\alpha^i]$ for $i = 0, 1, \ldots$ such that $a_0 = 1$. The scalar a_i is called the ith moment of α. Indeed any scalar random variable possessing finite moments can be represented by an umbra. A scalar sequence $\{a_i\}$ with $a_0 = 1$ is said to be represented by an umbra α if $E[\alpha^i] = a_i$ for $i = 0, 1, \ldots$.

EXAMPLE 3.1. The sequence $1, 0, 0, \ldots$ is umbrally represented by the *augmentation umbra* ε, and $1, 1, 1, \ldots$ is umbrally represented by the *unity umbra* u. These are the umbral versions of two degenerate random variables such that $P(X = 0) = 1$ and $P(Y = 1) = 1$. The sequence of moments of a unit Poisson random variable is umbrally represented by the *Bell umbra* β. This umbra plays a fundamental role in the symbolic method, as we will see later. Its ith moment is the Bell number, which is the coefficient of $z^i/i!$ in the Taylor expansion of $\exp(e^z - 1)$.

Since an umbra is a formal object, questions involving the moment problem are not taken into account. Indeed, not every umbra corresponds to a real-valued random variable.

EXAMPLE 3.2. The sequence $1, 1, 0, 0, \ldots$ is represented by the *singleton* umbra χ. Its variance $E[\chi^2] - E[\chi]^2 = -1$ is negative, so there is no real-valued random variable corresponding to χ. Nevertheless this umbra plays a fundamental role in dealing with cumulant sequences, as we will see later.

It is always possible to make the alphabet large enough so that, to each scalar sequence $\{a_i\}$, there corresponds an umbra $\alpha \in \mathcal{A}$, which is not necessarily unique. The same applies to identically distributed random variables. Two umbrae α and γ having the same moment sequence are called similar, in symbols $\alpha \equiv \gamma$, and $\mathcal{A}$ contains an unlimited supply of distinct umbrae similar to α, usually denoted by $\alpha', \alpha'', \ldots$. If the sequence $\{a_i\}$ is umbrally represented by α, then

$$\text{the sequence } \{2^i a_i\} \text{ is represented by } \alpha + \alpha = 2\alpha,$$

$$\text{the sequence } \left\{ \sum_{k=0}^{i} \binom{i}{k} a_k a_{i-k} \right\} \text{ is represented by } \alpha + \alpha'.$$

An expression such as 2α or $\alpha + \alpha'$ is an example of an umbral polynomial, that is, a polynomial $p \in R[\mathcal{A}]$ in the umbrae of $\mathcal{A}$. The support of an umbral polynomial is the set of all umbrae that occur in it. So the support of $\alpha + \alpha'$ is $\{\alpha, \alpha'\}$, and the support of 2α is $\{\alpha\}$. The formal power series

$$e^{\alpha z} = u + \sum_{i \geq 1} \alpha^i \frac{z^i}{i!} \in R[\mathcal{A}][[z]] \tag{3.2}$$

is the *generating function* of the umbra α. Moreover, each exponential formal power series

$$f(z) = 1 + \sum_{i \geq 1} a_i \frac{z^i}{i!} \in R[[z]] \tag{3.3}$$

can be umbrally represented by a formal power series (3.2) in $R[\mathcal{A}][[z]]$ [20]. In fact, if the sequence $1, a_1, a_2, \dots$ is umbrally represented by α, the action of evaluation E can be extended coefficient-wise to formal power series (3.2), so that $E[e^{\alpha z}] = f(z)$. For clarity we denote the generating function of α by $f(\alpha, z) = E[e^{\alpha z}]$. Therefore $\alpha \equiv \alpha'$ if and only if $f(\alpha, z) = f(\alpha', z)$.

The first advantage of umbral notation is the representation of operations on generating functions with operations on umbrae. For example, multiplication of exponential generating functions is umbrally represented by the sum of the corresponding umbrae, that is,

$$f(\alpha + \gamma, z) = f(\alpha, z) f(\gamma, z).$$

Therefore $f(\alpha, z)^2$ is the generating function of $\alpha + \alpha'$, which is different from the generating function $f(\alpha, 2z)$ of 2α. The sum of generating functions is represented by the auxiliary umbra $\alpha \dot{+} \gamma$, named the *disjoint sum* of two umbrae, that is,

$$f(\alpha \dot{+} \gamma, z) = f(\alpha, z) + f(\gamma, z) - 1,$$

so that $E[(\alpha \dot{+} \gamma)^i] = E[\alpha^i] + E[\gamma^i]$ for all positive integers i. Then $2f(\alpha, z) - 1$ is the generating function of $\alpha \dot{+} \alpha$ or $\alpha \dot{+} \alpha'$, and $\alpha \dot{+} \alpha \equiv \alpha \dot{+} \alpha'$.

It is also possible to compose generating functions and to represent the composition as the generating function of an umbra. First consider n uncorrelated umbrae $\alpha', \alpha'', \dots, \alpha'''$ similar to α and take their sum: the resulting umbra $\alpha' + \alpha'' + \cdots + \alpha'''$, denoted by $n.\alpha$, is called the *dot product* of the integer n and the umbra α. Its generating function is $f(n.\alpha, z) = [f(\alpha, z)]^n$ and the moments are [3]

$$E\big[(n.\alpha)^i\big] = \sum_{\lambda \vdash i} d_\lambda (n)_{l(\lambda)} a_\lambda \qquad \text{with } d_\lambda = \frac{i!}{(1!)^{r_1} r_1! (2!)^{r_2} r_2! \cdots}, \tag{3.4}$$

where λ is a partition of the integer i into $l(\lambda)$ parts, and $a_\lambda = a_1^{r_1} a_2^{r_2} \cdots$ is the moment product [8]. The right-hand side of (3.4) corresponds to $E[(X_1 + \cdots + X_n)^i]$ with $X_1, \dots, X_n$ i.i.d. with moment sequence represented by the umbra α. In (3.4), set $E[(n.\alpha)^i] = q_i(n)$, which is a polynomial of degree i in n. If the integer n is replaced by any umbra $\gamma \in \mathcal{A}$, and $(\gamma)_j = \gamma(\gamma - 1) \cdots (\gamma - j + 1)$ denotes the descending factorial polynomial, then we have $q_i(\gamma) = \sum_{\lambda \vdash i} (\gamma)_{l(\lambda)} d_\lambda a_\lambda$. The symbol $\gamma.\alpha$ such that $E[(\gamma.\alpha)^i] = E[q_i(\gamma)]$ is called the *dot-product* of the umbrae α and γ. This last equality could be rewritten by using the umbral equivalence $\simeq$ such that $p \simeq q$ iff $E[p] = E[q]$ with $p, q \in R[\mathcal{A}]$. Then we have $(\gamma.\alpha)^i \simeq q_i(\gamma)$. More generally, the umbral equivalence turns out to be useful in dealing with umbral polynomials with nondisjoint supports as we will see later. The replacement of the integer n with the umbra γ is an example of the main device employed in the symbolic method, allowing us to represent more structured moment sequences starting from (3.4). Observe that we move from the generating function $[f(\alpha, z)]^n$ to the generating function $f(\gamma.\alpha, z) = f(\gamma, \log[f(\alpha, z)])$, which is not yet the

composition of $f(\alpha, z)$ and $f(\gamma, z)$. For this purpose, the umbra α in the dot product $\gamma.\alpha$ has to be replaced by a dot product involving the Bell umbra, that is, $\beta.\alpha$. The dot product $\beta.\alpha$ is called the *α-partition umbra* with generating function $f(\beta.\alpha, z) = \exp(f(\alpha, z) - 1)$. A special property which we use later is

$$\beta.(\alpha \dot{+} \gamma) \equiv \beta.\alpha + \beta.\gamma. \tag{3.5}$$

The symbol $\gamma.(\beta.\alpha)$ has generating function which is the composition of $f(\alpha, z)$ and $f(\gamma, z)$

$$f\big(\gamma.(\beta.\alpha), z\big) = f\big(\gamma, f(\alpha, z) - 1\big). \tag{3.6}$$

Parenthesis can be avoided since $\gamma.(\beta.\alpha) \equiv (\gamma.\beta).\alpha$. The moments are

$$E\big[(\gamma.\beta.\alpha)^i\big] = \sum_{\lambda \vdash i} g_{l(\lambda)} d_\lambda a_\lambda, \tag{3.7}$$

where $\{g_i\}$ are the moments of the umbra γ [8].

EXAMPLE 3.3. The composition umbra arises naturally in connection with random sums $X_1 + \cdots + X_N$, where the X's are i.i.d., and N is distributed independently of X. The cumulant generating function of the sum is the composition $K_N(K_X(t))$ of the two generating functions. In probability theory, N is necessarily integer-valued, but there is no such constraint on the umbra γ.

Strictly connected to the composition umbra is the *compositional inverse umbra* $\alpha^{\langle -1 \rangle}$ of an umbra α, such that

$$\alpha^{\langle -1 \rangle}.\beta.\alpha \equiv \chi \equiv \alpha.\beta.\alpha^{\langle -1 \rangle}. \tag{3.8}$$

A special compositional inverse umbra is $u^{\langle -1 \rangle}$, with u the unity umbra, having generating function

$$f\big(u^{\langle -1 \rangle}, z\big) = 1 + \log(1 + z) \tag{3.9}$$

so that its ith moment is

$$E\big[\big(u^{\langle -1 \rangle}\big)^i\big] = (-1)^{i-1}(i-1)!. \tag{3.10}$$

Multivariate case. Let $\{\nu_1, \dots, \nu_m\}$ be a set of umbral monomials with support not necessarily disjoint. A vector sequence $\{g_{\mathbf{i}}\}_{\mathbf{i} \in \mathbb{N}_0^m} \in R$, with $g_{\mathbf{i}} = g_{i_1, i_2, \dots, i_m}$ and $g_{\mathbf{0}} = 1$, is represented by the m-tuple $\boldsymbol{\nu} = (\nu_1, \dots, \nu_m)$ if

$$E\big[\boldsymbol{\nu}^{\mathbf{i}}\big] = g_{\mathbf{i}} \tag{3.11}$$

for each multi-index $\mathbf{i} \in \mathbb{N}_0^m$. The elements $\{g_{\mathbf{i}}\}_{\mathbf{i} \in \mathbb{N}_0^m}$ in (3.11) are called *multivariate moments* of $\boldsymbol{\nu}$.

REMARK 3.4. Within random variables, the m-tuple $\boldsymbol{\nu} = (\nu_1, \dots, \nu_m)$ corresponds to a random vector $(X_1, \dots, X_m)$. If $\{\nu_i\}_{i=1}^m$ are uncorrelated umbrae, then $g_{\mathbf{i}} = E[\nu_1^{i_1}] \cdots E[\nu_n^{i_m}]$, and we recover the univariate symbolic method. The same happens if $\{\nu_i\}_{i=1}^m$ are umbral monomials with disjoint supports.

As done in (3.2), the generating function of $\boldsymbol{\nu}$ is the formal power series

$$e^{\nu_1 z_1 + \cdots + \nu_m z_m} = u + \sum_{k \geq 1} \sum_{|\mathbf{i}|=k} \boldsymbol{\nu}^{\mathbf{i}} \frac{\mathbf{z}^{\mathbf{i}}}{\mathbf{i}!} \in R[\mathcal{A}][[z_1, \dots, z_m]] \tag{3.12}$$

with $\mathbf{z} = (z_1, \dots, z_m)$, $|\mathbf{i}| = i_1 + \cdots + i_m$ and $\mathbf{i}! = i_1! \cdots i_m!$. If the sequence $\{g_{\mathbf{i}}\}$ is umbrally represented by $\boldsymbol{\nu}$ and has (exponential) generating function

$$f(\mathbf{z}) = 1 + \sum_{k \geq 1} \sum_{|\mathbf{i}|=k} g_{\mathbf{i}} \frac{\mathbf{z}^{\mathbf{i}}}{\mathbf{i}!}, \tag{3.13}$$

then $E[e^{\nu_1 z_1 + \cdots + \nu_m z_m}] = f(\mathbf{z})$. Taking into account (3.11), the generating function in (3.13) is denoted by $f(\boldsymbol{\nu}, \mathbf{z})$. Two umbral vectors $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ are said to be *similar*, in symbols $\boldsymbol{\nu}_1 \equiv \boldsymbol{\nu}_2$, if and only if $f(\boldsymbol{\nu}_1, \mathbf{z}) = f(\boldsymbol{\nu}_2, \mathbf{z})$, that is, $E[\boldsymbol{\nu}_1^{\mathbf{i}}] = E[\boldsymbol{\nu}_2^{\mathbf{i}}]$ for all $\mathbf{i} \in \mathbb{N}_0^m$. They are said to be *uncorrelated* if and only if $E[\boldsymbol{\nu}_1^{\mathbf{i}} \boldsymbol{\nu}_2^{\mathbf{j}}] = E[\boldsymbol{\nu}_1^{\mathbf{i}}] E[\boldsymbol{\nu}_2^{\mathbf{j}}]$ for all $\mathbf{i}, \mathbf{j} \in \mathbb{N}_0^m$.

An equation analogous to (3.4) could be given for the multivariate case, provided that integer partitions are replaced with multi-index partitions [4]. A partition of a multi-index $\mathbf{i}$ is a composition $\boldsymbol{\lambda}$, whose columns are in lexicographic order, in symbols $\boldsymbol{\lambda} \vdash \mathbf{i}$. A composition $\boldsymbol{\lambda}$ of a multi-index $\mathbf{i}$ is a matrix $\boldsymbol{\lambda} = (\lambda_{ij})$ of nonnegative integers and with no zero columns such that $\lambda_{r1} + \lambda_{r2} + \cdots + \lambda_{rk} = i_r$ for $r = 1, 2, \dots, n$. The number of columns of $\boldsymbol{\lambda}$ is the length of $\boldsymbol{\lambda}$ and denoted by $l(\boldsymbol{\lambda})$. As for integer partitions, the notation $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^{r_1}, \boldsymbol{\lambda}_2^{r_2}, \dots)$ means that in the matrix $\boldsymbol{\lambda}$ there are r_1 columns equal to $\boldsymbol{\lambda}_1$, r_2 columns equal to $\boldsymbol{\lambda}_2$ and so on, with $\boldsymbol{\lambda}_1 < \boldsymbol{\lambda}_2 < \cdots$. We set $\mathfrak{m}(\boldsymbol{\lambda}) = (r_1, r_2, \dots)$. The dot-product $n.\boldsymbol{\nu}$ of a nonnegative integer n and a m-tuple $\boldsymbol{\nu}$ is an auxiliary umbra denoting the summation $\boldsymbol{\nu}' + \boldsymbol{\nu}'' + \cdots + \boldsymbol{\nu}'''$ with $\{\boldsymbol{\nu}', \boldsymbol{\nu}'', \dots, \boldsymbol{\nu}'''\}$ a set of n uncorrelated and similar m-tuples. For $\mathbf{i} \in \mathbb{N}_0^m$ and m-tuples $\boldsymbol{\nu}$ of umbral monomials, we have

$$E\big[(n.\boldsymbol{\nu})^{\mathbf{i}}\big] = \sum_{\boldsymbol{\lambda} \vdash \mathbf{i}} \frac{\mathbf{i}!}{\mathfrak{m}(\boldsymbol{\lambda})! \boldsymbol{\lambda}!} (n)_{l(\boldsymbol{\lambda})} g_{\boldsymbol{\lambda}}, \tag{3.14}$$

where the sum is over all partitions $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^{r_1}, \boldsymbol{\lambda}_2^{r_2}, \dots)$ of the multi-index $\mathbf{i}$, $g_{\boldsymbol{\lambda}} = g_{\boldsymbol{\lambda}_1}^{r_1} g_{\boldsymbol{\lambda}_2}^{r_2} \cdots$ and $g_{\boldsymbol{\lambda}_i} = E[\boldsymbol{\nu}^{\boldsymbol{\lambda}_i}]$. The sequence in (3.14) represents moments of a sum of i.i.d. random vectors with sequence of moments $\{g_{\mathbf{i}}\}$. If we replace the integer n in (3.14) with the dot-product $\alpha.\beta$ we get the auxiliary umbra $\alpha.\beta.\boldsymbol{\nu}$ representing the sequence of moments

$$E\big[(\alpha.\beta.\boldsymbol{\nu})^{\mathbf{i}}\big] = \sum_{\boldsymbol{\lambda} \vdash \mathbf{i}} \frac{\mathbf{i}!}{\mathfrak{m}(\boldsymbol{\lambda})! \boldsymbol{\lambda}!} a_{l(\boldsymbol{\lambda})} g_{\boldsymbol{\lambda}}, \tag{3.15}$$

where the sequence $\{a_i\}$ is umbrally represented by α. In particular the generating function of the auxiliary umbra $\alpha.\beta.\boldsymbol{\nu}$ turns to be the composition of the univariate generating function $f(\alpha, z)$ and the multivariate generating function $f(\boldsymbol{\nu}, \mathbf{z})$

$$f(\alpha.\beta.\boldsymbol{\nu}, \mathbf{z}) = f\big[\alpha, f(\boldsymbol{\nu}, \mathbf{z}) - 1\big]. \tag{3.16}$$

From Example 3.3, the umbra $\alpha.\beta.\boldsymbol{\nu}$ is a generalization of a multivariate compound randomized Poisson random vector. In the next section, we show how this umbra allows us to write a formula for multivariate cumulants involving multi-index partitions. More details on the symbolic composition of multivariate formal power series can be found in [5].

4. Formal cumulants.

4.1. *Definition*. Among the sequences of numbers related to a real-valued random variable, cumulants play a central role. Whether or not the sequence $\{a_i\}$ corresponds to the moments of some distribution, we define cumulants $\{c_i\}$ by the following equation:

$$1 + \sum_{i \geq 1} a_i \frac{z^i}{i!} = \exp\left\{\sum_{i \geq 1} c_i \frac{z^i}{i!}\right\}. \tag{4.1}$$

If α is an umbra representing the sequence $\{a_i\}$, and κ_α is an umbra representing the sequence $\{c_i\}$, then by comparing (4.1) with (3.6) we have

$$\alpha \equiv u.\beta.\kappa_\alpha, \tag{4.2}$$

since $f(u, z) = \exp(z)$. The umbra κ_α is called the *α-cumulant umbra* [3] and is such that

$$f(\kappa_\alpha, z) = 1 + \log\big(f(\alpha, z)\big). \tag{4.3}$$

By comparing (4.3) with (3.6) and (3.9), we have

$$\kappa_\alpha \equiv u^{\langle -1 \rangle}.\beta.\alpha. \tag{4.4}$$

Since $u^{\langle -1 \rangle}.\beta.u \equiv u^{\langle -1 \rangle}.\beta \equiv \chi$, then equivalence (4.4) reduces to

$$\kappa_\alpha \equiv \chi.\alpha.$$

The algebraic properties of cumulants can be formalized as

$$\begin{aligned} &\textit{Homogeneity} \qquad \chi.(a\alpha) \equiv a(\chi.\alpha) \qquad \text{if } a \in R, \\ &\textit{Additivity} \quad \chi.(\alpha + \gamma) \equiv \chi.\alpha \dot{+} \chi.\gamma. \end{aligned} \tag{4.5}$$

The *semi-invariance under translation* follows from both equivalences in (4.5) by choosing as umbra α the unity umbra u.

As done in (4.1), multivariate formal cumulants $\{c_{\mathbf{i}}\}$ of a sequence of multivariate moments $\{g_{\mathbf{i}}\}$ can be defined via generating functions. Indeed, if $\{g_{\mathbf{i}}\}$ is

umbrally represented by the m-tuple $\boldsymbol{\nu}$, then the sequence $\{c_{\mathbf{i}}\}$ is umbrally represented by the m-tuple $\boldsymbol{\kappa}_{\boldsymbol{\nu}}$ such that

$$f(\boldsymbol{\nu}, \mathbf{z}) = \exp\big[f(\boldsymbol{\kappa}_{\boldsymbol{\nu}}, \mathbf{z}) - 1\big]. \tag{4.6}$$

The m-tuple $\boldsymbol{\kappa}_{\boldsymbol{\nu}}$ is named $\boldsymbol{\nu}$-cumulant. By comparing (3.16) with (4.6), the following equivalence follows:

$$\boldsymbol{\nu} \equiv u.\beta.\boldsymbol{\kappa}_{\boldsymbol{\nu}}. \tag{4.7}$$

Equivalence (4.7) can be inverted in

$$\boldsymbol{\kappa}_{\boldsymbol{\nu}} \equiv u^{\langle -1\rangle}.\beta.\boldsymbol{\nu} \qquad \text{with } f(\boldsymbol{\kappa}_{\boldsymbol{\nu}}, \mathbf{z}) = 1 + \log\big[f(\boldsymbol{\nu}, \mathbf{z})\big], \tag{4.8}$$

where $u^{\langle -1\rangle}$ is the compositional inverse of the unity umbra u. Moments of $u^{\langle -1\rangle}.\beta.\boldsymbol{\nu}$ can be computed via equation (3.15) by recalling (3.10). As before, the umbra $u^{\langle -1\rangle}.\beta$ may be replaced by the umbra χ, so that

$$\boldsymbol{\kappa}_{\boldsymbol{\nu}} \equiv \chi.\boldsymbol{\nu}. \tag{4.9}$$

Thanks to this last representation, the algebraic properties of cumulants can be formalized as

$$\begin{aligned} &\textit{Homogeneity} && \chi.(a\boldsymbol{\nu}) \equiv a(\chi.\boldsymbol{\nu}) && \text{if } a \in R, \\ &\textit{Additivity} && \chi.(\boldsymbol{\nu}_1 + \boldsymbol{\nu}_2) \equiv \chi.\boldsymbol{\nu}_1 \dot{+} \chi.\boldsymbol{\nu}_2 \\ & && && \text{if } \boldsymbol{\nu}_1 \text{ and } \boldsymbol{\nu}_1 \text{ are uncorrelated } n\text{-tuples.} \end{aligned} \tag{4.10}$$

In the additivity property (4.10), we have used the disjoint sum of two m-tuples, that is, $E[(\boldsymbol{\nu}_1 \dot{+} \boldsymbol{\nu}_2)^{\mathbf{i}}] = E[\boldsymbol{\nu}_1^{\mathbf{i}}] + E[\boldsymbol{\nu}_2^{\mathbf{i}}]$ for all $\mathbf{i} \in \mathbb{N}_0^m$. The *semi-invariance under translation* follows from both equivalences in (4.10) by choosing the m-tuple $\mathbf{u} = (u, \dots, u)$ as $\boldsymbol{\nu}$.

4.2. *Cumulants of trace powers*. Let us represent the eigenvalues of a random matrix M of dimension m by the m-tuple of umbral monomials $\boldsymbol{\nu} = (\nu_1, \dots, \nu_m)$. Cumulants of $\boldsymbol{\nu}$ can be recovered via (4.7) and (4.8). In this section we will characterize cumulants of $\mathrm{Tr}(M)$, that is, cumulants of the sequence $E[(\nu_1 + \cdots + \nu_m)^k]$ for $k = 1, 2, \dots$. Observe that

$$f(\nu_1 + \cdots + \nu_m, z) = E\big[e^{(\nu_1 + \cdots + \nu_m)z}\big] = f(\boldsymbol{\nu}, \mathbf{z}) \tag{4.11}$$

by using (3.12) and (3.13) with $\mathbf{z} = (z, \dots, z)$. Compositions of multivariate formal power series like $f(\boldsymbol{\nu}, \mathbf{z})$ in (4.11) are represented by symbols with a peculiar expression. Indeed, if $\boldsymbol{\nu}$ is a m-tuple of umbral monomials with generating function $f(\boldsymbol{\nu}, \mathbf{z})$, and $\boldsymbol{\xi}$ is a m-tuple of umbral monomials with generating function $f(\boldsymbol{\xi}, \mathbf{z})$, the m-tuple having generating function $f[\boldsymbol{\xi}, (f(\boldsymbol{\nu}, \mathbf{z}) - 1, \dots, f(\boldsymbol{\nu}, \mathbf{z}) - 1)]$ is $(\xi_1 + \cdots + \xi_m).\beta.\boldsymbol{\nu}$, that is,

$$f\big[(\xi_1 + \cdots + \xi_m).\beta.\boldsymbol{\nu}, \mathbf{z}\big] = f\big[\boldsymbol{\xi}, \big(f(\boldsymbol{\nu}, \mathbf{z}) - 1, \dots, f(\boldsymbol{\nu}, \mathbf{z}) - 1\big)\big]. \tag{4.12}$$

As in (4.2) and (4.4) for univariate and multivariate cumulants, respectively, in order to characterize cumulants of $f(\boldsymbol{\nu}, \mathbf{z})$ in (4.11), we replace the m-tuple $\boldsymbol{\xi}$ with the m-tuple $\mathbf{u} = (u, \dots, u)$ in (4.12). Denote by $\mathfrak{c}_{\boldsymbol{\nu}}$ the m-tuple such that $\boldsymbol{\nu} \equiv (u + \cdots + u).\beta.\mathfrak{c}_{\boldsymbol{\nu}}$, that is,

$$f(\boldsymbol{\nu}, \mathbf{z}) = f\big[(u + \cdots + u).\beta.\mathfrak{c}_{\boldsymbol{\nu}}, \mathbf{z}\big] \tag{4.13}$$

with $f(\boldsymbol{\nu}, \mathbf{z})$ in (4.11). If $\mathfrak{c}_{\boldsymbol{\nu}} = (\mathfrak{c}_{1,\boldsymbol{\nu}}, \dots, \mathfrak{c}_{m,\boldsymbol{\nu}})$, then $f(\mathfrak{c}_{\boldsymbol{\nu}}, \mathbf{z}) = f[\mathfrak{c}_{1,\boldsymbol{\nu}} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}}, z]$ and

$$f\big[(u + \cdots + u).\beta.\mathfrak{c}_{\boldsymbol{\nu}}, \mathbf{z}\big] = \exp\big\{m\big(f[\mathfrak{c}_{1,\boldsymbol{\nu}} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}}, z] - 1\big)\big\}.$$

DEFINITION 4.1. For fixed m, formal cumulants of the sequence $\{E[(\nu_1 + \cdots + \nu_m)^k]\}$ are umbrally represented by the umbral polynomial $\mathfrak{c}_{1,\boldsymbol{\nu}} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}}$, such that

$$\nu_1 + \cdots + \nu_m \equiv m.\beta.(\mathfrak{c}_{1,\boldsymbol{\nu}} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}}) \tag{4.14}$$

with $\mathfrak{c}_{\boldsymbol{\nu}} = (\mathfrak{c}_{1,\boldsymbol{\nu}}, \dots, \mathfrak{c}_{m,\boldsymbol{\nu}})$ given in (4.13).

In order to prove that the moments of the umbral polynomial $\mathfrak{c}_{1,\boldsymbol{\nu}} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}}$ satisfy the characterizing algebraic properties of cumulants, we need to invert (4.14).

PROPOSITION 4.2. *We have* $\mathfrak{c}_{1,\boldsymbol{\nu}} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}} \equiv \chi.\frac{1}{m}.(\nu_1 + \cdots + \nu_m)$.

PROOF. Indeed from (4.14), we have $\frac{1}{m}.(\nu_1 + \cdots + \nu_m) \equiv \beta.(\mathfrak{c}_{1,\boldsymbol{\nu}} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}})$, so $\chi.\frac{1}{m}.(\nu_1 + \cdots + \nu_m) \equiv \chi.\beta.(\mathfrak{c}_{1,\boldsymbol{\nu}} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}})$. The result follows since $\chi.\beta \equiv u$. $\square$

Thanks to Proposition 4.2, the umbral polynomial $\mathfrak{c}_{1,\boldsymbol{\nu}} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}}$ is similar to an umbra like $\chi.p$, with $p \in R[\mathcal{A}]$. So it has to satisfy the additivity and homogeneity properties like those in (4.10).

THEOREM 4.3. Additivity: *If* $\boldsymbol{\nu}_1$ *and* $\boldsymbol{\nu}_2$ *are uncorrelated* m*-tuples*, *then*

$$\mathfrak{c}_{1,\boldsymbol{\nu}_1+\boldsymbol{\nu}_2} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}_1+\boldsymbol{\nu}_2} \equiv (\mathfrak{c}_{1,\boldsymbol{\nu}_1} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}_1}) \dot{+} (\mathfrak{c}_{1,\boldsymbol{\nu}_2} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}_2});$$

Homogeneity: *if* $a \in R$, *then* $\mathfrak{c}_{1,(a\boldsymbol{\nu})} + \cdots + \mathfrak{c}_{m,(a\boldsymbol{\nu})} \equiv a(\mathfrak{c}_{1,\boldsymbol{\nu}} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}})$.

PROOF. Observe that $a.(p + q) \equiv a.p + a.q$ if $c \in R$ and p, q are umbral polynomials with disjoint supports [3]. The previous equivalence holds in particular for $a = 1/m$ and $p = (\nu_{1,1} + \cdots + \nu_{1,m})$ and $q = (\nu_{2,1} + \cdots + \nu_{2,m})$. From Proposition 4.2,

$$\mathfrak{c}_{1,\boldsymbol{\nu}_1+\boldsymbol{\nu}_2} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}_1+\boldsymbol{\nu}_2} \equiv \chi.\frac{1}{m}.(p + q) \equiv \chi.\left[\frac{1}{m}.p + \frac{1}{m}.q\right].$$

From the additivity property in (4.10), we have

$$\chi \cdot \left[\frac{1}{m} \cdot p + \frac{1}{m} \cdot q\right] \equiv \chi \cdot \frac{1}{m} \cdot p \dotplus \chi \cdot \frac{1}{m} \cdot q$$

for p and q umbral polynomials with disjoint supports. The result follows by observing that $\mathfrak{c}_{1,\boldsymbol{\nu}} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}} \equiv \chi \cdot \frac{1}{m} \cdot q$ and $\mathfrak{c}_{1,\boldsymbol{\nu}} + \cdots + \mathfrak{c}_{m,\boldsymbol{\nu}} \equiv \chi \cdot \frac{1}{m} \cdot p$. The homogeneity property follows since $b \cdot (ap) \equiv a(b \cdot p)$, for $a, b \in R$ and $p \in R[\mathcal{A}]$ [3]. The previous equivalence holds in particular for $a = 1/m$ and $p = (\nu_1 + \cdots + \nu_m)$. $\square$

The semi-invariance under translation follows since $(\mathfrak{c}_{1,\mathbf{u}} + \cdots + \mathfrak{c}_{m,\mathbf{u}}) \equiv \chi \cdot \frac{1}{m} \cdot m \equiv \chi$ whose moments are all zero except the first. The connection between multivariate cumulants of $\boldsymbol{\nu}$ and those of $\nu_1 + \cdots + \nu_m$ is given in the following proposition.

PROPOSITION 4.4. *We have* $\boldsymbol{\kappa}_{\boldsymbol{\nu}} \equiv \chi \cdot m \cdot \beta \cdot \mathfrak{c}_{\boldsymbol{\nu}}$.

PROOF. We have $(u + \cdots + u) \cdot \beta \cdot \mathfrak{c}_{\boldsymbol{\nu}} \equiv m \cdot \beta \cdot \mathfrak{c}_{\boldsymbol{\nu}}$, and from (4.13) we have $m \cdot \beta \cdot \mathfrak{c}_{\boldsymbol{\nu}} \equiv \boldsymbol{\nu}$. Therefore from (4.7), we have $u \cdot \beta \cdot \boldsymbol{\kappa}_{\boldsymbol{\nu}} \equiv m \cdot \beta \cdot \mathfrak{c}_{\boldsymbol{\nu}}$ so that $\chi \cdot u \cdot \beta \cdot \boldsymbol{\kappa}_{\boldsymbol{\nu}} \equiv \chi \cdot m \cdot \beta \cdot \mathfrak{c}_{\boldsymbol{\nu}}$. The result follows since $\chi \cdot u \cdot \beta \equiv u$ and $\chi \cdot u \cdot \beta \cdot \boldsymbol{\kappa}_{\boldsymbol{\nu}} \equiv \boldsymbol{\kappa}_{\boldsymbol{\nu}}$. $\square$

5. Spectral *k*-statistics. Tukey [21] introduced the multi-index k-statistics in connection with finite-population sampling. He showed that the multi-index k's are multiplicative in the limit as $n \to \infty$, and that they are equal to the product of Fisher's single-index k's. In the ordinary i.i.d. setting considered by Fisher, this means that each multi-index k converges to a cumulant product.

We now construct matricial polykays, indexed by an integer partition λ, as unbiased estimators of cumulant products of trace powers of a random matrix Y. Then, when the random matrix Y is defined by sub-sampling as in Section 2, that is, when a spectral sample is considered, we will prove the inheritance property by assuming that the elements of the diagonal matrix X are umbrally represented by similar and uncorrelated umbrae.

DEFINITION 5.1. The matricial polykay $\kappa_\lambda(\mathbf{y})$ of class $\lambda \vdash i$ is the symmetric polynomial in the eigenvalues $\mathbf{y} = (y_1, \dots, y_m)$ such that

$$E\big[\kappa_\lambda(\mathbf{y})\big] = \prod_{j=1}^{l(\lambda)} E\big[(\mathfrak{c}_{1,\mathbf{y}} + \cdots + \mathfrak{c}_{m,\mathbf{y}})^{\lambda_j}\big].$$

Set $l(\lambda) = r$. From Proposition 4.2 and equation (3.14), with n replaced by the umbra χ, a first expression of cumulant products of trace powers in terms of

moments of Y is

$$\begin{aligned}(5.1)\qquad &\prod_{j=1}^{r} E\big[(\mathfrak{c}_{1,\mathbf{y}} + \cdots + \mathfrak{c}_{m,\mathbf{y}})^{\lambda_j}\big] \\ &= \sum_{(\eta_1 \vdash \lambda_1, \ldots, \eta_r \vdash \lambda_r)} \prod_{j=1}^{r} \frac{(-1)^{\nu_{\eta_j}-1}}{m} d_{\eta_j} (\nu_{\eta_j} - 1)! g_{\eta_1+\cdots+\eta_r},\end{aligned}$$

where $\eta_1 + \cdots + \eta_r = (t_1, t_2, \ldots)$ is the summation of the partitions $\{\eta_1, \ldots, \eta_r\}$ and $g_{\eta_1+\cdots+\eta_r} = \prod_{j=1}^{l(\eta_1+\cdots+\eta_r)} E[\mathrm{Tr}(Y^{t_j})]$. Equation (5.1) takes into account that $E[(\chi \cdot \chi)^i] = (-1)^{i-1}(i-1)!$ for all nonnegative integers i.

A second expression which is more suitable for spectral sampling is in terms of joint moments of Y, that is, in terms of products of its trace powers. To this end, we need to work with permutations $\mathfrak{S}_i$ and with the group algebra $R[\mathcal{A}](\mathfrak{S}_i)$ on the polynomial ring $R[\mathcal{A}]$.

A permutation σ of $[i]$, or $\sigma \in \mathfrak{S}_i$, the symmetric group, can be decomposed into disjoint cycles $C(\sigma)$. In the standard representation each cycle is written with its largest element first, and the cycles are listed in increasing order of their largest element [17]. The length of the cycle $c \in C(\sigma)$ is its cardinality, denoted by $\mathfrak{l}(c)$. The number of cycles of σ is denoted by $|C(\sigma)|$. Recall that a permutation σ with r_1 1-cycles, r_2 2-cycles and so on is said to be of cycle class $\lambda = (1^{r_1}, 2^{r_2}, \ldots) \vdash i$. In particular we have $l(\lambda) = |C(\sigma)|$. The number of permutations $\sigma \in \mathfrak{S}_i$ of cycle class $\lambda = (1^{r_1}, 2^{r_2}, \ldots) \vdash i$ is usually denoted by

$$(5.2)\qquad s_\lambda = \frac{i!}{1^{r_1} r_1! 2^{r_2} r_2! \cdots}.$$

Each cycle class is a conjugacy class of the group of permutations: two elements of $\mathfrak{S}_i$ are conjugate if and only if they have the same cycle class.

Consider the group algebra $\mathbb{A}_i = R[\mathcal{A}](\mathfrak{S}_i)$. An element $f \in \mathbb{A}_i$ associates with each permutation $\sigma \in \mathfrak{S}_i$ a polynomial $f(\sigma) \in R[\mathcal{A}]$, so $\mathbb{A}_i$ is the space of $R[\mathcal{A}]$-valued functions. Multiplication in $\mathbb{A}_i$ is the convolution

$$(5.3)\qquad (f \cdot g)(\sigma) = \sum_{\rho\omega=\sigma} f(\rho) g(\omega).$$

The unitary element with respect to multiplication is the indicator function δ such that $\delta(e) = 1$, with e the identity $[i] \to [i]$, and zero otherwise. Indeed $f \cdot \delta = \delta \cdot f = f$ for $f \in \mathbb{A}_i$. If it exists, the inverse function of f in $\mathbb{A}_i$ is denoted by $f^{(-1)}$ and is such that $f^{(-1)} \cdot f = f \cdot f^{(-1)} = \delta$.

Denote by $\mu(Y)$ the function in $\mathbb{A}_i$ such that

$$(5.4)\qquad \mu(Y)(\sigma) = \prod_{c \in C(\sigma)} \mathrm{Tr}\big(Y^{\mathfrak{l}(c)}\big) \in R[\mathcal{A}]$$

for a matrix Y of order m and $\sigma \in \mathfrak{S}_i$. Evidently $\mu(Y)(\sigma)$ is a product of power sums in the eigenvalues of Y, depending only on the cycle structure. In particular we have

$$\mu(Y)(e) = \big[\mathrm{Tr}(Y)\big]^i \quad \text{and} \quad \mu(I_m)(\sigma) = m^{|C(\sigma)|}. \tag{5.5}$$

THEOREM 5.2. *Define the function $\tilde{\kappa}(\mathbf{y}) \in \mathbb{A}_i$ by*

$$\tilde{\kappa}(\mathbf{y}) = \mu(I_m)^{(-1)} \cdot \mu(Y) \tag{5.6}$$

with $\mu(Y)$ and $\mu(I_m)$ given in (5.4) and (5.5), respectively. Then

$$\mathfrak{K}_\lambda(\mathbf{y}) \simeq \big[(1!)^{r_2}(2!)^{r_3}\cdots\big]\tilde{\kappa}(\mathbf{y})(\sigma) \tag{5.7}$$

is a matricial polykay of class $\lambda = (1^{r_1}, 2^{r_2}, \dots) \vdash i$, the cycle structure of σ.

PROOF. Observe that by taking the expectation of both sides in (5.6), we have

$$E\big[\tilde{\kappa}(\mathbf{y})\big] = \mu(I_m)^{-1} \star E\big[\mu(Y)\big], \tag{5.8}$$

where $\star$ is the classical convolution on the space of R-functions on $\mathfrak{S}_i$

$$(a \star b)(\sigma) = \sum_{\rho\omega=\sigma} a(\rho)b(\omega)$$

with $E[\mu(Y)]:\sigma \in \mathfrak{S}_i \mapsto E[\mu(Y)(\sigma)] \in R$ and $E[\tilde{\kappa}(\mathbf{y})]:\sigma \in \mathfrak{S}_i \mapsto E[\tilde{\kappa}(\mathbf{y})(\sigma)] \in R$. Then the symmetric polynomial $\mathfrak{K}_\lambda(\mathbf{y})$ in (5.7) satisfies Definition 5.1 if also the function

$$E[C_{\mathbf{y}}]:\sigma \in \mathfrak{S}_i \mapsto \prod_{\sigma \in C(\sigma)} \frac{1}{(\mathfrak{l}(\sigma)-1)!} E\big[(\mathfrak{c}_{1,\mathbf{y}} + \cdots + \mathfrak{c}_{m,\mathbf{y}})^{\mathfrak{l}(\sigma)}\big]$$

is such that $E[C_{\mathbf{y}}] = \mu(I_m)^{-1} \star E[\mu(Y)]$. From (4.14), we have

$$E\big[\big(\mathrm{Tr}(Y)\big)^i\big] = \sum_{\lambda \vdash i} d_\lambda m^{l(\lambda)} \prod_{j=1}^{l(\lambda)} E\big[(\mathfrak{c}_{1,\mathbf{y}} + \cdots + \mathfrak{c}_{m,\mathbf{y}})^{\lambda_i}\big]. \tag{5.9}$$

From (5.9), by observing that

$$d_\lambda = \frac{s_\lambda}{(1!)^{r_2}(2!)^{r_3}\cdots} \tag{5.10}$$

with s_λ the number of permutations $\sigma \in \mathfrak{S}_i$ of cycle class $\lambda = (1^{r_1}, 2^{r_2}, \dots) \vdash i \le m$ given in (5.2), we have

$$E\big[\mu(Y)\big](e) = \sum_{\sigma \in \mathfrak{S}_i} m^{|C(\sigma)|} E[C_{\mathbf{y}}](\sigma) = \sum_{\sigma \in \mathfrak{S}_i} m^{|C(\sigma)|} E[C_{\mathbf{y}}](\sigma^{-1}), \tag{5.11}$$

where σ^{-1} is the inverse of σ. By using the action of $\mathfrak{S}_i$ on the group algebra $R(\mathfrak{S}_i)$ we have $E[\mu(Y)] = \mu(I_m) \star E[C_{\mathbf{y}}]$. For $i \le m$ the function $\sigma \mapsto m^{|C(\sigma)|}$ has an inverse [10], so that $E[C_{\mathbf{y}}] = \mu(I_m)^{-1} \star E[\mu(Y)]$. $\square$

Note that equation (5.8) is given in [1] as the definition of the cumulants of a random matrix. By Theorem 5.2, we have shown that $\tilde{\kappa}(\mathbf{y})$ in (5.8) are rather statistics, due also to the condition $i \leq m$ which parallels the analogous condition for Fisher's k-statistics.

The statistics $\mathfrak{K}_\lambda(\mathbf{y})$ are unbiased estimators of product of cumulants, due to Definition 5.1. The inheritance on the average is indeed strictly connected to the spectral sampling, that is, to the special structure of the matrix $Y = (HXH^\dagger)_{[m\times m]}$. When $\mathfrak{K}_\lambda(\mathbf{y})$ refers to spectral sampling, we call them spectral polykays.

DEFINITION 5.3 (Natural spectral statistics). A statistic is said to be *natural* relative to spectral sampling if, for each $m \leq n$, the average value of T_m over spectral sub-samples $\mathbf{y}$ of $\mathbf{x}$ is equal to $T_n(\mathbf{x})$,

$$E\big(T_m(\mathbf{y})|\mathbf{x}\big) = T_n(\mathbf{x}). \tag{5.12}$$

Theorem 5.5 states that the spectral k-statistics $\mathfrak{K}_\lambda(\mathbf{y})$ are natural.

We first give a proposition which moves from Lemma 7.2 in [3]. In this proposition, the evaluation operator $E[\cdot|\boldsymbol{\gamma}]$ deals the elements of the n-tuple $\boldsymbol{\gamma}$ as they were constants. A formal definition of $E[\cdot|\boldsymbol{\gamma}]$ may be found in [6].

PROPOSITION 5.4. *If $\{\gamma_1, \gamma_2, \dots, \gamma_n\}$ are uncorrelated umbrae similar to the umbra γ, then*

$$E\big[(\gamma_1 z_1 + \gamma_2 z_2 + \cdots + \gamma_n z_n)^j|\boldsymbol{\gamma}\big] = \sum_{\lambda\vdash j} d_\lambda \kappa_\lambda(\boldsymbol{\gamma}) E[\tilde{\sigma}_\lambda], \tag{5.13}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_n)$ and $\tilde{\sigma}$ is the polynomial umbra whose moments are the power sums in the indeterminates $\{z_1, z_2, \dots, z_n\}$, that is, $E[\tilde{\sigma}^i] = z_1^i + z_2^i + \cdots + z_n^i$.

PROOF. The starting point is the result of Lemma 7.2 in [3],

$$\chi\boldsymbol{\cdot}(\gamma_1 z_1 + \cdots + \gamma_n z_n) \equiv (\chi\boldsymbol{\cdot}\gamma)\tilde{\sigma}, \tag{5.14}$$

where $\{\gamma_1, \gamma_2, \dots, \gamma_n\}$ are uncorrelated umbrae similar to γ. Equivalence (5.14) may be rewritten as

$$\chi\boldsymbol{\cdot}(\gamma_1 z_1 + \cdots + \gamma_n z_n) \equiv \left(\chi\boldsymbol{\cdot}\frac{1}{n}\boldsymbol{\cdot}(\gamma_1 + \gamma_2 + \cdots + \gamma_n)\right)\tilde{\sigma} \tag{5.15}$$

as $\gamma \equiv \frac{1}{n}\boldsymbol{\cdot} n \boldsymbol{\cdot}\gamma$. Taking the dot-product with β on both sides of (5.15) gives

$$(\gamma_1 z_1 + \cdots + \gamma_n z_n) \equiv \beta\boldsymbol{\cdot}\left[\left(\chi\boldsymbol{\cdot}\frac{1}{n}\boldsymbol{\cdot}(\gamma_1 + \gamma_2 + \cdots + \gamma_n)\right)\tilde{\sigma}\right]. \tag{5.16}$$

The result follows by using (3.7) with γ replaced by the unity umbra u and by using Proposition 4.2 and Definition 5.1. In equation (3.7), the evaluation operator is intended to be replaced by $E[\cdot|\boldsymbol{\gamma}]$. $\square$

THEOREM 5.5 (Inheritance on the average). *The statistics* $\mathfrak{K}_\lambda(\mathbf{y})$ *in* (5.7) *are inherited on the average*, *that is*,

$$E\big[\mathfrak{K}_\lambda(\mathbf{y})|\mathbf{x}\big] = \mathfrak{K}_\lambda(\mathbf{x}),$$

where $\mathbf{y}$ *is a spectral random sample.*

PROOF. Since the trace is invariant under cyclic permutations, for a nonnegative integer i, we have

$$\mathrm{Tr}\big(Y^i\big) = \mathrm{Tr}\big[\big(HXH^\dagger\big)^i_{[m\times m]}\big] = \mathrm{Tr}\big[\big(XH^\dagger_{[m\times n]}H_{[m\times n]}\big)^i\big]$$

with Y given in Definition 2.1. Therefore $\mu(Y) = \mu(XB)$, with $B = H^\dagger_{[m\times n]}$ $H_{[m\times n]}$ a square matrix of dimension n independent of X. The random matrix B is an orthogonal projection on a m-dimensional subspace such that

$$\begin{aligned}\mu(B)(\sigma) &= \prod_{c\in C(\sigma)} \mathrm{Tr}\big[\big(H^\dagger_{[m\times n]}H_{[m\times n]}\big)^{l(c)}\big] \\ &= \prod_{c\in C(\sigma)} \mathrm{Tr}\big[\big(H_{[m\times n]}H^\dagger_{[m\times n]}\big)^{l(c)}\big] = \mu(I_m)(\sigma).\end{aligned} \tag{5.17}$$

For a diagonal matrix X independent of B, and by using Proposition 5.4, we have

$$E\big[\mu(XB)|\mathbf{x}\big] = \mu(I_m)\cdot\tilde{\kappa}(\mathbf{x}). \tag{5.18}$$

Indeed, if in Proposition 5.4 the umbrae $\{\gamma_1,\dots,\gamma_n\}$ are replaced by the elements of the diagonal matrix X, and the indeterminates $\{z_1,\dots,z_n\}$ by the diagonal entries of the matrix B, equation (5.13) may be updated as

$$E\big[\mathrm{Tr}(XB)^i|\mathbf{x}\big] = \sum_{\lambda\vdash i} d_\lambda\kappa_\lambda(\mathbf{x})E[\tilde{\sigma}_\lambda]. \tag{5.19}$$

Due to (5.17), we have $E[\tilde{\sigma}_\lambda] = m^{l(\lambda)}$. So again equation (5.18) follows by using the action of $\mathfrak{S}_i$ on the group algebra $R(\mathfrak{S}_i)$. The result follows from Theorem 5.2 by observing that

$$\begin{aligned}E\big[\tilde{\kappa}(\mathbf{y})|\mathbf{x}\big] &= E\big[\mu(I_m)^{-1}\cdot\mu(XB)|\mathbf{x}\big] \\ &= E\big[\mu(I_m)^{-1}\cdot\mu(I_m)\cdot\tilde{\kappa}(\mathbf{x})|\mathbf{x}\big] = \tilde{\kappa}(\mathbf{x}).\end{aligned}$$

□

REMARK 5.6. The computation of $\mu(I_m)^{-1}$ requires the solution of a system of m equations in m indeterminates $\mu(I_m)\cdot\mu(I_m)^{-1} = \mu(I_m)^{-1}\cdot\mu(I_m) = \delta$ with coefficients given by $\mu(I_m)$. This task may be performed with standard procedures in any symbolic package. A different way consists of resorting to the so-called Weingarten function on $\mathfrak{S}_i$. See [2] for the definition and the properties of the Weingarten function, which involves the characters of $\mathfrak{S}_i$ and Schur symmetric polynomials indexed by $\lambda\vdash i$.

The spectral k-statistics can be expressed on terms of power sums $S_r = \sum_{j=1}^n x_j^r$ as follows:

$$\mathfrak{K}_{(1)} = \frac{S_1}{n} = k_{(1)},$$

$$\mathfrak{K}_{(2)} = \frac{nS_2 - S_1^2}{n(n^2-1)} = \frac{k_{(2)}}{(n+1)},$$

$$\mathfrak{K}_{(1^2)} = \frac{nS_1^2 - S_2}{n(n^2-1)} = \frac{k_{(1^2)}}{(n+1)},$$

$$\mathfrak{K}_{(3)} = 2\frac{2S_1^3 - 3nS_1S_2 + n^2S_3}{n(n^2-1)(n^2-4)} = \frac{2k_{(3)}}{(n+1)(n+2)},$$

$$\mathfrak{K}_{(1,2)} = \frac{-2nS_3 + (n^2+2)S_1S_2 - nS_1^3}{n(n^2-1)(n^2-4)} = \frac{2k_{(1,2)} - nk_{(1)}k_{(2)}}{(n+1)(n+2)},$$

$$\mathfrak{K}_{(1^3)} = \frac{S_1^3(n^2-2) - 3nS_1S_2 + 4S_3}{n(n^2-1)(n^2-4)} = \frac{2k_{(1^3)} - 3k_{(1)}k_{(2)} + n(n+3)(k_{(1)})^3}{(n+1)(n+2)}.$$

The functions of degree 4 are a little more complicated,

$$\mathfrak{K}_{(4)} = 6\frac{S_4(n^3+n) - 4S_1S_3(n^2+1) + S_2^2(3-2n^2) + 10nS_1^2S_2 - 5S_1^4}{n^2(n^2-1)(n^2-4)(n^2-9)},$$

$$\begin{aligned}\mathfrak{K}_{(1,3)} = {} & 2\frac{-3nS_4(n^2+1) + S_1S_3(12+3n^2+n^4) + S_2^2(6n^2-9)}{n^2(n^2-1)(n^2-4)(n^2-9)} \\ & + \frac{-3nS_1^2S_2(n^2+1) + 2(2n^2-3)S_1^4}{n^2(n^2-1)(n^2-4)(n^2-9)},\end{aligned}$$

$$\begin{aligned}\mathfrak{K}_{(2^2)} = {} & \frac{2S_4(3n-2n^3) + 4S_1S_3(4n^2-6) + S_2^2(18+n^4-6n^2)}{n^2(n^2-1)(n^2-4)(n^2-9)} \\ & + \frac{-2nS_1^2S_2(n^2+6) + (n^2+6)S_1^4}{n^2(n^2-1)(n^2-4)(n^2-9)},\end{aligned}$$

$$\begin{aligned}\mathfrak{K}_{(1^2,2)} = {} & \frac{10nS_4 - 4S_1S_3(n^2+1) + S_2^2(n^2+6) + nS_1^2S_2(n^2+1)}{n(n^2-1)(n^2-4)(n^2-9)} \\ & + \frac{(4-n^2)S_1^4}{n(n^2-1)(n^2-4)(n^2-9)},\end{aligned}$$

$$\begin{aligned}\mathfrak{K}_{(1^4)} = {} & \frac{-30nS_4 + 4S_1S_3(4n^2-6) + S_2^2(3n^2+18) + 6nS_1^2S_2(4-n^2)}{n^2(n^2-1)(n^2-4)(n^2-9)} \\ & + \frac{(6-8n^2+n^4)S_1^4}{n^2(n^2-1)(n^2-4)(n^2-9)}.\end{aligned}$$

For comparison purposes, all of the single-index functions $\mathfrak{K}_{(r)}(\mathbf{x})$ and the k-statistics $k_{(r)}(\mathbf{x})$ for $r \geq 2$ are invariant under translation: $\mathfrak{K}_{(r)}(\mathbf{x} - \bar{\mathbf{x}}) = \mathfrak{K}_{(r)}(\mathbf{x})$. If the mean is zero, the fourth-order statistics are

$$(n)_4 k_{(4)} = n^2(n+1)S_4 - 3n(n-1)S_2^2,$$
$$\mathfrak{K}_{(4)} \propto n\big(n^2+1\big)S_4 - \big(2n^2-3\big)S_2^2,$$

showing that $\mathfrak{K}_{(4)}$ is not a simple multiple of $k_{(4)}$.

For a spectral sample, the first few conditional variances and covariances are

$$\begin{aligned}
\operatorname{var}\big(\mathfrak{K}_{(1)}(\mathbf{y}) \mid \mathbf{x}\big) &= \mathfrak{K}_{(2)}(\mathbf{x})\left(\frac{1}{m} - \frac{1}{n}\right),\\
\operatorname{cov}\big(\mathfrak{K}_{(1)}(\mathbf{y}), \mathfrak{K}_{(2)}(\mathbf{y}) \mid \mathbf{x}\big) &= 2\mathfrak{K}_{(3)}(\mathbf{x})\left(\frac{1}{m} - \frac{1}{n}\right),\\
\operatorname{var}\big(\mathfrak{K}_{(2)}(\mathbf{y}) \mid \mathbf{x}\big) &= 2\mathfrak{K}_{(2^2)}(\mathbf{x})\left(\frac{1}{m^2-1} - \frac{1}{n^2-1}\right)\\
&\quad + 2\mathfrak{K}_{(4)}(\mathbf{x})\frac{(n-m)(2m^2n^2 - 3n^2 - 3m^2 - mn + 3)}{nm(m^2-1)(n^2-1)},
\end{aligned}$$

which are similar to the covariances of the corresponding k-statistics.

We now characterize the limiting behavior of spectral polykays. To this end, we recall the notion of *free cumulant* occurring within noncommutative probability theory [13]. A noncommutative probability space is a pair (A, Φ), where A is a unital noncommutative algebra, and $\Phi : A \to \mathbb{C}$ is a unital linear functional. This gives rise to a sequence of multilinear functional $\{\Phi_i\}$ on A via $\Phi_i(a_1, \dots, a_i) = \Phi(a_1 \cdots a_i)$.

Let $\mathcal{N}C$ denote the lattice of all noncrossing partitions of $[i]$. A noncrossing partition $\pi = \{B_1, B_2, \dots, B_k\}$ of the set $[i]$ is a partition such that if $1 \leq h < l < s < k \leq i$, with $h, s \in B_n$ and $l, k \in B_{n'}$, then $n = n'$. For any noncrossing partition π and $a_1, \dots, a_i \in A$ we set

$$\Phi_\pi(a_1, \dots, a_i) = \prod_{B \in \pi} \Phi(a_{j_1} \cdots a_{j_s})$$

for $B = (j_1 < \cdots < j_s)$. Free cumulants are defined as multilinear functionals such that

$$c_\pi(a_1, \dots, a_i) = \prod_{B \in \pi} c_{|B|}(a_{j_1} \cdots a_{j_s})$$

and

$$c_i(a_1, \dots, a_i) = \sum_{\pi \in \mathcal{N}C} \mathfrak{m}(\pi, 1_i)\Phi_\pi(a_1, \dots, a_i),$$

where $\mathfrak{m}(\pi, 1_i)$ is the Moebius function on the lattice of noncrossing partitions [13]. The ith cumulants of a is $c_i = c_i(a, \dots, a)$.

By using Proposition 6.1 in [1], when m goes to infinity, the mean of the normalized spectral k-statistics $\tilde{\kappa}_\lambda^{(N)}$ corresponding to $\lambda = (1^{r_1}, 2^{r_2}, \dots) \vdash i$

$$\tilde{\kappa}_\lambda^{(N)}(\mathbf{y}) := m^{i-l(\lambda)}\tilde{\kappa}_\lambda(\mathbf{y})$$

tends toward the product of free cumulants $c_1^{r_1}c_2^{r_2}\cdots$ with $\tilde{\kappa}_\lambda(\mathbf{y}) := \tilde{\kappa}(\mathbf{y})(\sigma)$, given in (5.6), and σ a permutation of class λ.

6. Generalized spectral polykays. The notion of generalized cumulant has been discussed by McCullagh [11] and involves set partitions. In umbral terms, if π is a partition of $\{\mu_1, \mu_2, \dots, \mu_i\}$, then the generalized cumulant κ_π is defined as [3]

$$E\big[(\chi \cdot \mu)_\pi\big] = \kappa_\pi \qquad \text{with } (\chi \cdot \mu)_\pi = \prod_{B \in \Pi_i} (\chi \cdot \mu_B) \text{ and } \mu_B = \prod_{j \in B} \mu_j.$$

For example, if $i = 5$ and $\pi = \{\{\mu_1, \mu_2\}, \{\mu_3\}, \{\mu_4, \mu_5\}\}$, then

$$E\big[(\chi \cdot \mu)_\pi\big] = E\big[(\chi \cdot \mu_1\mu_2)(\chi \cdot \mu_3)(\chi \cdot \mu_4\mu_5)\big] = \kappa^{12,3,45}$$

using McCullagh's notation. Generalized k-statistics are the sample version of the generalized cumulants. The importance of generalized k-statistics stems from the following properties: the generalized k-statistics are linearly independent; every polynomial symmetric function can be expressed uniquely as a linear combination of generalized k-statistics; any polynomial symmetric function whose expectation is independent of n can be expressed as linear combination of generalized k-statistics with coefficients independent of n [12]. Due to the last property, natural statistics could be expressed as linear combinations of their generalized k-statistics with coefficients independent of n.

THEOREM 6.1. *If $\lambda \vdash i \le m$ and π is a set partition of class λ, then generalized k-statistics of spectral polykays are*

$$\tilde{l}_\pi(\mathbf{y}) \simeq \sum_{\tau \ge \pi} (-1)^{|\tau|-1}(|\tau|-1)!\tilde{\kappa}_\tau(\mathbf{y}), \tag{6.1}$$

where $\tilde{\kappa}_\tau(\mathbf{y})$ denotes the function on Π_i such that $\tilde{\kappa}(\mathbf{y})(\tau) := \tilde{\kappa}(\mathbf{y})(\sigma)$, with $\sigma \in \mathfrak{S}_i$ a permutation of the same class of τ and $\tilde{\kappa}(\mathbf{y})$ given in (5.6).

The proof of Theorem 6.1 relies on Proposition 5.4 of [3]. We do not invert equivalence (6.1) because the linear combination giving spectral polykays in terms of their generalized k-statistics is quite cumbersome; see equation (3.18) in [12]. Instead, there are alternative systems of symmetric functions that are more suitable from a computational point of view. All such systems are invertible linear functions of generalized k-statistics with coefficients independent of the sample size, and the properties given above are preserved under such transformations.

To characterize such coefficients, we first recall the Moebius inversion formula on the lattice of set partitions [15]. The set Π_i with the refinement order $\leq$ is a lattice, where $\pi \leq \tau$ if for any block in $B \in \pi$ there exists a block $B' \in \tau$ such that $B \subseteq B'$. If G is a function on Π_i and

$$F(\pi) = \sum_{\tau \geq \pi} G(\tau),$$

then the Moebius inversion formula states that

$$G(\pi) = \sum_{\tau \geq \pi} \mathfrak{m}(\pi, \tau) F(\tau), \tag{6.2}$$

where $\mathfrak{m}(\pi, \tau)$ is the so-called Moebius function. It is shown that

$$\mathfrak{m}(\pi, \tau) = (-1)^{s-t} (2!)^{r_3} (3!)^{r_4} \cdots,$$

where $r_1 + 2r_2 + \cdots = s = |\pi|$, $r_1 + r_2 + \cdots = t = |\tau|$ and $(1^{r_1}, 2^{r_2}, \ldots)$ is the partition, usually denoted by $\lambda(\pi, \tau)$, of the integer s such that r_j blocks of τ contain exactly j blocks of π.

DEFINITION 6.2. If $\lambda \vdash i \leq m$ and π is a set partition of class λ, the (transformed) generalized k-statistics of spectral polykays are $l_\tau(\mathbf{y})$ such that

$$\tilde{\kappa}_\pi(\mathbf{y}) = \sum_{\tau \geq \pi} l_\tau(\mathbf{y}), \tag{6.3}$$

where $\tilde{\kappa}_\tau(\mathbf{y})$ denotes the function on Π_i such that $\tilde{\kappa}(\mathbf{y})(\tau) := \tilde{\kappa}(\mathbf{y})(\sigma)$, with $\sigma \in \mathfrak{S}_i$ a permutation of the same class of τ and $\tilde{\kappa}(\mathbf{y})$ given in (5.6).

The linear combination in (6.3) is very simple involving coefficients all equal to 1. By using the Moebius inversion formula (6.2), from (6.3) we have

$$l_\pi(\mathbf{y}) = \sum_{\tau \geq \pi} \mathfrak{m}(\pi, \tau) \tilde{\kappa}_\tau(\mathbf{y}). \tag{6.4}$$

Since $\tilde{\kappa}(\mathbf{y})(\sigma)$ depends only on the cycle structure $C(\sigma)$, then $\tilde{\kappa}_\tau(\mathbf{y})$ depends only on the block sizes in τ. So in the sum (6.4), there are d_λ spectral k-statistics equal to $\tilde{\kappa}_\tau(\mathbf{y})$, all those having the same class λ. Therefore spectral polykays of degree i can be indexed by partitions of i. As example, by Definition 6.1, the spectral polykays up to order 4 are

$$\begin{aligned}
l_{(1)} &= \tilde{\kappa}_{(1)} \qquad (i = 1),\\
l_{(1^2)} &= \tilde{\kappa}_{(1^2)} - \tilde{\kappa}_{(2)} \qquad (i = 2),\\
l_{(1,2)} &= \tilde{\kappa}_{(1,2)} - \tilde{\kappa}_{(3)} \qquad (i = 3),\\
l_{(1^3)} &= \tilde{\kappa}_{(1^3)} - 3\tilde{\kappa}_{(1,2)} + 2\tilde{\kappa}_{(3)},\\
l_{(1,3)} &= \tilde{\kappa}_{(1,3)} - \tilde{\kappa}_{(4)} \qquad (i = 4),
\end{aligned}$$

$$
\begin{aligned}
l_{(2^2)} &= \tilde{\kappa}_{(2^2)} - \tilde{\kappa}_{(4)}, \\
l_{(1^2,2)} &= \tilde{\kappa}_{(1^2,2)} - 2\tilde{\kappa}_{(1,3)} - \tilde{\kappa}_{(2^2)} + 2\tilde{\kappa}_{(4)}, \\
l_{(1^4)} &= \tilde{\kappa}_{(1^4)} - 6\tilde{\kappa}_{(1^2,2)} + 8\tilde{\kappa}_{(1,3)} + 3\tilde{\kappa}_{(2^2)} - 6\tilde{\kappa}_{(4)}.
\end{aligned}
$$

In addition, we have $l_{(i)} = \tilde{\kappa}_{(i)}$. We take a moment to motivate this definition. Tukey [21] gives very similar equations connecting classical polykays and k-statistics. We just recall those up to order 4.

$$
\begin{aligned}
k_{(1)} &= k_{(1)} \qquad (i = 1), \\
k_{(1^2)} &= k_{(1)}k_{(1)} - \frac{1}{m}k_{(2)} \qquad (i = 2), \\
k_{(1,2)} &= k_{(1)}k_{(2)} - \frac{1}{m}k_{(3)} \qquad (i = 3), \\
k_{(1^3)} &= k_{(1)}k_{(1)}k_{(1)} - \frac{3}{m}k_{(2)}k_{(1)} + \frac{2}{m^2}k_{(3)}, \\
k_{(1,3)} &= k_{(1)}k_{(3)} - \frac{1}{m}k_{(4)} \qquad (i = 4), \\
k_{(2^2)} &= \frac{m-1}{m+1}k_{(2)}k_{(2)} - \frac{1}{m}k_{(4)}, \\
k_{(1^2,2)} &= k_{(2)}k_{(1)}k_{(1)} - \frac{2}{m}k_{(3)}k_{(1)} - \frac{m-1}{m(m+1)}k_{(2)}k_{(2)} + \frac{2}{m(m+1)}k_{(4)}, \\
k_{(1^4)} &= k_{(1)}k_{(1)}k_{(1)}k_{(1)} - \frac{6}{m}k_{(2)}k_{(1)}k_{(1)} + \frac{8}{m^2}k_{(3)}k_{(1)} \\
&\quad + \frac{3(m-1)}{m^2(m+1)}k_{(2)}k_{(2)} - \frac{6m}{m+1}k_{(4)}.
\end{aligned}
$$

The two sets of equations are very similar in structure.

The refinement order in (6.3) is inverted with respect to those connecting moments and cumulants [11]. It is the same as that employed in the change of basis between augmented symmetric functions and power sums [3] and employed by Tukey in order to show the multiplicative structure of $\tilde{\mathfrak{a}}_\lambda(\mathbf{x})$ for infinite populations.

In terms of power sums in the eigenvalues, the transformed generalized spectral polykays up to order 4 are

$$
\begin{aligned}
l_{(1,2)} &= \frac{(m+1)S_1S_2 - S_1^3 - mS_3}{m(m-1)(m+1)(m-2)}, \\
l_{(1^2,2)} &= \frac{2mS_4 + (m+3)S_1^2S_2 - (2m+2)S_1S_3 - mS_2^2 - S_1^4}{m(m-1)(m+1)(m-2)(m-3)},
\end{aligned}
$$

$$
\begin{aligned}
l_{(2^2)} &= \frac{1}{m^2(m^2-1)(m-2)(m-3)} \\
&\quad \times \{{S_1}^4 + (-3m+3+m^2){S_2}^2 \\
&\qquad + (4m-4)S_1S_3 - 2{S_1}^2S_2m + (-m^2+m)S_4\}, \\
l_{(1,3)} &= \frac{2}{(m^2-4)(m^2-1)m^2} \\
&\quad \times \{-S_4m(m^2+1) + S_1S_3(m^3+m^2+4) \\
&\qquad + S_2^2(2m^2-3) - mS_1^2S_2(3m+1) + S_1^4(2m-1)\}.
\end{aligned}
$$

The spectral statistics $l_{(1^r)}$ are the same as the corresponding polykays $k_{(1^r)}$.

THEOREM 6.3. *When m goes to infinity the mean of the normalized (transformed) generalized k-statistics*

$$l_\pi^{(N)}(\mathbf{y}) := m^{i-|\pi|}l_\pi(\mathbf{y})$$

tends to $d_\lambda c_1^{r_1}c_2^{r_2}\cdots$, with $\lambda = (1^{r_1}, 2^{r_2}, \ldots) \vdash i$, the class partition of π, and $\{c_j\}$ free cumulants.

PROOF. After multiplying both sides of (6.4) by $m^{i-|\pi|}$, we have

$$m^{i-|\pi|}l_\pi(\mathbf{y}) = \sum_{\tau\geq\pi} \mathfrak{m}(\pi,\tau)m^{i-|\pi|}\tilde{\kappa}_\tau(\mathbf{y}).$$

Since $\tau \geq \pi$, then $|\tau| \leq |\pi|$ so that

$$l_\pi^{(N)}(\mathbf{y}) = \sum_{\tau\geq\pi} \mathfrak{m}(\pi,\tau)\frac{1}{m^{|\pi|-|\tau|}}m^{i-|\tau|}\tilde{\kappa}_\tau(\mathbf{x}).$$

As m goes to infinity, for all $\tau > \pi$ having the same class partition, $m^{i-|\tau|}\tilde{\kappa}_\tau(\mathbf{y})$ tends toward the free cumulant c_τ, whereas $\frac{1}{m^{|\pi|-|\tau|}}$ goes to zero. The result follows since for $\tau = \pi$ we have $\mathfrak{m}(\pi,\tau) = 1$, and for all π having the same class partition $m^{i-|\pi|}\tilde{\kappa}_\pi(\mathbf{y})$ goes to $c_\pi = c_1^{r_1}c_2^{r_2}\cdots$. $\square$

REFERENCES

[1] CAPITAINE, M. and CASALIS, M. (2006). Cumulants for random matrices as convolutions on the symmetric group. *Probab. Theory Related Fields* **136** 19–36. MR2240781

[2] COLLINS, B. and ŚNIADY, P. (2006). Integration with respect to the Haar measure on unitary, orthogonal and symplectic group. *Comm. Math. Phys.* **264** 773–795. MR2217291

[3] DI NARDO, E., GUARINO, G. and SENATO, D. (2008). A unifying framework for k-statistics, polykays and their multivariate generalizations. *Bernoulli* **14** 440–468. MR2544096

[4] DI NARDO, E., GUARINO, G. and SENATO, D. (2009). A new method for fast computing unbiased estimators of cumulants. *Stat. Comput.* **19** 155–165. MR2486228

[5] DI NARDO, E., GUARINO, G. and SENATO, D. (2011). A new algorithm for computing the multivariate Faà di Bruno's formula. *Appl. Math. Comput.* **217** 6286–6295. MR2773373

[6] DI NARDO, E. and OLIVA, I. (2012). A new family of time–space harmonic polynomials with respect to Lévy processes. *Ann. Mat. Pura Appl.* DOI:10.1007/s10231-012-0252-3.

[7] DI NARDO, E. and SENATO, D. (2001). Umbral nature of the Poisson random variables. In *Algebraic Combinatorics and Computer Science: A Tribute to Gian-Carlo Rota* (H. Crapo and D. Senato, eds.) 245–266. Springer Italia, Milan. MR1854481

[8] DI NARDO, E. and SENATO, D. (2006). An umbral setting for cumulants and factorial moments. *European J. Combin.* **27** 394–413. MR2206475

[9] FISHER, R. A. (1929). Moments and product moments of sampling distributions. *Proc. Lond. Math. Soc.* (3) **30** 199–238.

[10] GRACZYK, P., LETAC, G. and MASSAM, H. (2003). The complex Wishart distribution and the symmetric group. *Ann. Statist.* **31** 287–309. MR1962508

[11] MCCULLAGH, P. (1984). Tensor notation and cumulants of polynomials. *Biometrika* **71** 461–476. MR0775392

[12] MCCULLAGH, P. (1987). *Tensor Methods in Statistics*. Chapman & Hall, London. MR0907286

[13] NICA, A. and SPEICHER, R. (2006). *Lectures on the Combinatorics of Free Probability*. *London Mathematical Society Lecture Note Series* **335**. Cambridge Univ. Press, Cambridge. MR2266879

[14] RAVIRAJ, P. and SANAVULLAH, M. Y. (2007). The modified 2D-Haar wavelet transformation in image compression. *Middle East J. Scientific Research* **2** 73–78.

[15] ROTA, G.-C. (1964). On the foundations of combinatorial theory. I. Theory of Möbius functions. *Z. Wahrsch. Verw. Gebiete* **2** 340–368. MR0174487

[16] ROTA, G. C. and TAYLOR, B. D. (1994). The classical umbral calculus. *SIAM J. Math. Anal.* **25** 694–711. MR1266584

[17] STANLEY, R. P. (1986). *Enumerative Combinatorics. Vol. I.* Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA. MR0847717

[18] STRANG, G. and NGUYEN, T. (1996). *Wavelets and Filter Banks*. Wellesley–Cambridge Press, Wellesley, MA. MR1411910

[19] STUART, A. and ORD, J. K. (1987). *Kendall's Advanced Theory of Statistics. Vol.* 1, 5th ed. Oxford Univ. Press, New York.

[20] TAYLOR, B. D. (1998). Difference equations via the classical umbral calculus. In *Mathematical Essays in Honor of Gian-Carlo Rota* (*Cambridge*, *MA*, 1996) (B. Sagan and R. Stanley, eds.). *Progress in Mathematics* **161** 397–411. Birkhäuser, Boston, MA. MR1627386

[21] TUKEY, J. W. (1950). Some sampling simplified. *J. Amer. Statist. Assoc.* **45** 501–519. MR0040624

[22] TUKEY, J. W. (1956). Keeping moment-like sampling computations simple. *Ann. Math. Statist.* **27** 37–54. MR0076236

E. DI NARDO
D. SENATO
DEPARTMENT OF MATHEMATICS
COMPUTER SCIENCE
AND ECONOMICS
UNIVERSITY OF BASILICATA
VIALE DELL'ATENEO LUCANO, 10
I-85100, POTENZA
ITALY
E-MAIL: elvira.dinardo@unibas.it
domenico.senato@unibas.it

P. MCCULLAGH
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5734 UNIVERSITY AVE
CHICAGO, ILLINOIS 60637
USA
E-MAIL: pmcc@galton.uchicago.edu