Methodol Comput Appl Probab (2012) 14:253–262
DOI 10.1007/s11009-010-9202-y

On a Generalized Chu–Vandermonde Identity

Stefano Favaro · Igor Prünster · Stephen G. Walker

Received: 11 February 2010 / Accepted: 12 October 2010 /
Published online: 2 November 2010

Abstract In the present paper we introduce a generalization of the well–known Chu–Vandermonde identity. In particular, by inductive reasoning, the identity is extended to a multivariate setup in terms of the fourth Lauricella function. The main interest in such generalizations derives from the species diversity estimation and, in particular, prediction problems in Genomics and Ecology within a Bayesian nonparametric framework.

Keywords Bayesian nonparametrics · Chu–Vandermonde identity · Multivariate convolution · Lauricella function · Prediction · Species diversity

AMS 2000 Subject Classifications 05A19 · 33C20 · 62E15 · 62P10

S. Favaro and I. Prünster are partially supported by MIUR, grant 2008MK3AFZ, and Piedemont Region.

S. Favaro
Dipartimento di Statistica e Matematica Applicata,
Università degli Studi di Torino and Collegio Carlo Alberto,
Corso Unione Sovietica 218/bis, 10134,
Torino, Italy
e-mail: stefano.favaro@unito.it

I. Prünster (✉)
Dipartimento di Statistica e Matematica Applicata,
Università degli Studi di Torino, ICER and Collegio Carlo Alberto,
Corso Unione Sovietica 218/bis, 10134,
Torino, Italy
e-mail: igor@econ.unito.it

S. G. Walker
Institute of Mathematics, Statistics and Actuarial Science,
University of Kent, Canterbury, CT27NZ, UK
e-mail: S.G.Walker@kent.ac.uk

1 Introduction

Among elegant results implied by the binomial theorem, one of the most attractive and widely known results is Vandermonde's identity, named after Alexandre–Théophile Vandermonde:

$$\binom{n+m}{q} = \sum_{q_1=0}^{q} \binom{n}{q_1}\binom{m}{q-q_1} \tag{1}$$

for $n, m, q \in \mathbb{N}_0$. Combinatorially, we can think of this identity as related to the following illustrative example: a group of people consists of n left-handed and m right-handed persons, and we are trying to establish how many combinations exist such that there are exactly q women in the group. We can categorize each possible arrangement into one of $r+1$ categories. The $r+1$ categories are indexed from 0 to r, and an arrangement falls under category q_1 if there are exactly q_1 left-handed women, and the remaining women $(q-q_1)$ are right-handed. In particular, the $\binom{n}{q_1}\binom{m}{q-q_1}$ part merely counts how many arrangements fall under category q_1. The sum adds up all possible arrangements which fall under one of the categories. From a probabilistic point of view, the Vandermonde identity is related to the hypergeometric probability distribution. In particular, when both sides of Eq. 1 are divided by $\binom{n+m}{q}$, then for each q_1, $\binom{n}{q_1}\binom{m}{q-q_1}/\binom{n+m}{q}$ is interpreted as the probability that exactly q_1 objects are defective in a sample of q distinctive objects drawn from an urn with $n+m$ objects in which n are defective, i.e. there are $\binom{n+m}{q}$ possible samples (without replacement); there are $\binom{n}{q_1}$ ways to obtain q_1 defective objects and there are $\binom{m}{q-q_1}$ ways to fill out the rest of the sample with non-defective objects.

The Vandermonde identity can be generalized to non-integer arguments. In this case, it is known as the Chu–Vandermonde's identity and takes on the form

$$(a_1+a_2)_q = \sum_{q_1=0}^{q} \binom{q}{q_1} (a_1)_{q_1} (a_2)_{(q-q_1)} \tag{2}$$

for any complex-valued a_1 and a_2 with $(a)_n$ being the Pochhammer symbol for the ascending (or rising) factorial of a of order n, i.e. $(a)_n := a(a+1)\cdots(a+n-1) = \prod_{i=0}^{n-1}(a+i)$ (see Comtet 1974 and references therein).

In this paper we introduce a new generalization of the Chu–Vandermonde identity. In particular, the multivariate version of this new generalization of the Chu–Vandermonde identity is then derived by inductive reasoning in terms of the fourth Lauricella function. The motivation for studying such a generalization of the Chu–Vandermonde identity stems from applications to species diversity estimation and, in particular, to prediction problems in Genomics. In fact, by adopting a Bayesian nonparametric approach for predicting the number of new genes to be discovered in sequencing a cDNA library, the determination of suitable estimators crucially relies on obtaining closed form solutions for multivariate convolutions generalizing the one of Chu–Vandermonde; see Lijoi et al. (2008) and reference therein. The proposed results and its application in Bayesian nonparametrics highlights once again the interplay between Bayesian nonparametrics on one side and the theory of Lauricella functions on the other. Further examples of this close connection can be found in Regazzini (1998), Lijoi and Regazzini (2004) and James (2005) where functionals of the Dirichlet process are considered. It is worth noting that there is growing literature concerning Bayesian nonparametric approaches to species sampling and related prediction and estimation problems. See, for instance, Müller and Quintana (2004), Navarrete et al. (2008), Petrone et al. (2009), Quintana (2006) and Hjort et al. (2010) for a recent review of the discipline.

2 Generalized Chu–Vandermonde Identity

The topic of multiple hypergeometric functions was first approached, in a systematic way, by Lauricella (1893) at the end of the 19th century and further investigated by Appell and Kampé de Fériet (1926). See the comprehensive and stimulating monograph by Exton (1976). The original paper by Lauricella (1893) proceeded to define and study four n-dimensional functions which bear his name. In particular, here we focus on the fourth Lauricella function, which, for any $n \in \mathbb{N}$, is characterized by the following Laplace-type integral representation

$$F_D^{(n)}(a, b_1, \ldots, b_n; c; x_1, \ldots, x_n) = \frac{1}{\Gamma(b_1)\cdots\Gamma(b_n)} \times \int_{(\mathbb{R}^+)^n} e^{-\sum_{i=1}^n t_i} \prod_{i=1}^n t_i^{b_i - 1} {}_1F_1\left(a; c; \sum_{i=1}^n x_i t_i\right) dt_1 \cdots dt_n. \tag{3}$$

for any $a, c \in \mathbb{R}$ and any $b_1, \ldots, b_n \in \mathbb{R}^+$, with Γ being the Gamma function and ${}_1F_1$ being the confluent hypergeometric function of the first kind.

If $n = 2$, the fourth Lauricella function reduces to the Appell hypergeometric function F_1, whereas, if $n = 1$, it becomes the Gauss hypergeometric function ${}_2F_1$ which has been the starting point in the definition of the $F_D^{(n)}$.

The following proposition provides an extension of the Chu–Vandermonde identity, which, to the authors' knowledge and although simple to derive, is not present in the literature.

Proposition 1 *For any $q \geq 1$, $w_1, w_2 \in \mathbb{R}^+$ and $a_1, a_2 > 0$*

$$\sum_{q_1=0}^{q} \binom{q}{q_1} w_1^{q_1} w_2^{q-q_1} (a_1)_{q_1} (a_2)_{(q-q_1)} = w_2^q (a)_q \, {}_2F_1\left(-q, a_1; a; \frac{w_2 - w_1}{w_2}\right) \tag{4}$$

where $a := a_1 + a_2$.

Proof Several proofs can be given by using different known characterizations of the Gauss hypergeometric function ${}_2F_1$. Here, a straightforward proof is given by the direct application of two known representation for the Gauss hypergeometric function ${}_2F_1$:

i) for any $a, b \in \mathbb{R}$ and $n \in \mathbb{N}$

$$ {}_2F_1(a, b; b - n; z) = (1 - z)^{-a-n} \sum_{k=0}^{n} \frac{(-n)_k (b - a - n)_k z^k}{(b - n)_k k!} \tag{5}$$

and

ii) for any $a, b \in \mathbb{R}$ and $n \in \mathbb{N}$

$$ {}_2F_1(a, b; b-n; z) = \frac{(-1)^n (a)_n}{(1-b)_n} (1-z)^{-a-n} {}_2F_1(-n, b-a-n; 1-a-n; 1-z). \tag{6} $$

Now set $n := q$, $b := 1 - a_2$, $a := -a_2 - q + 1 - a_1$, $k := q_1$ and $z := w_1/w_2$ in Eqs. 5 and 6. Then, by using the representation (Eq. 5) we obtain the relation

$$ \sum_{q_1=0}^{q} \binom{q}{q_1} w_1^{q_1} w_2^{q-q_1} (a_1)_{q_1} (a_2)_{(q-q_1)} = {}_2F_1\left(a_1, -q; -a_2 - q + 1; \frac{w_1}{w_2}\right) w_2^q (a_2)_q. $$

and by resorting to Eq. 6 we have

$$ {}_2F_1\left(a_1, -q; 1 - a_2 - q; \frac{w_1}{w_2}\right) = \frac{(a_1+a_2)_q}{(a_2)_n} {}_2F_1\left(-q, a_1; a_1 + a_2; \frac{w_2 - w_1}{w_2}\right). $$

which implies Eq. 4. □

Note that the Chu–Vandermonde identity (Eq. 2) is immediately recovered from Eq. 4 by setting $w_1 = w_2 = 1$. The following proposition, obtained by inductive reasoning from Eq. 4, provides the multivariate extension of the identity given in Proposition 1 and represents the main result of the paper. In fact, as concisely illustrated in Section 3, it represents a crucial tool for determining computable expressions for the estimators of interest and may turn out to be useful also in different applied contexts.

Proposition 2 *For any $q \geq 1$, $j \geq 1$ let $\mathcal{D}_{j,q} := \{(q_1, \ldots, q_j) \in \{1, \ldots, q\}^j : \sum_{i=1}^{j} q_i = q\}$ and let $w_1, \ldots, w_j \in \mathbb{R}^+$ and $a_1, \ldots, a_j > 0$. Then*

$$ \sum_{(q_1,\ldots,q_j)\in\mathcal{D}_{j,q}} \binom{q}{q_1, \ldots, q_j} \prod_{i=1}^{j} w_i^{q_i} (a_i)_{q_i} $$
$$ = w_j^q (a)_q F_D^{(j-1)}\left(-q, a_1, \ldots, a_{j-1}, a; \frac{w_j - w_1}{w_j}, \ldots, \frac{w_j - w_{j-1}}{w_j}\right) \tag{7} $$

where $a := \sum_{i=1}^{j} a_i$.

Proof Using Eq. 4, the proof follows by inductive reasoning. Suppose the identity holds true for $j-1$, i.e.,

$$ \sum_{(q_1,\ldots,q_{j-1})\in\mathcal{D}_{j-1,q}} \binom{q}{q_1, \ldots, q_{j-1}} \prod_{i=1}^{j-1} w_i^{q_i} (a_i)_{q_i} $$
$$ = \sum_{(q_1,\ldots,q_{j-1})\in\mathcal{D}_{j,q}} \frac{q!}{q_1! \cdots q_{j-1}!} w_{j-1}^{q_{j-1}} (a_{j-1})_{q_{j-1}} \prod_{i=1}^{j-2} w_i^{q_i} (a_i)_{q_i} $$
$$ = w_{j-1}^q (a - a_j)_q F_D^{(j-2)}\left(-q, a_1, \ldots, a_{j-2}, a - a_j; \frac{w_{j-1} - w_1}{w_{j-1}}, \ldots, \frac{w_{j-1} - w_{j-2}}{w_{j-1}}\right) $$

and we show it holds for j as well. Observe that

$$\sum_{(q_1,\dots,q_j)\in\mathcal{D}_{j,q}} \frac{q!}{q_1!\cdots q_j!}\prod_{i=1}^{j} w_i^{q_i}(a_i)_{q_i} = \sum_{q_j=0}^{q} \frac{q!}{q_j!(q-q_j)!} w_j^{q_j}(a_j)_{q_j} \times \sum_{(q_1,\dots,q_{j-1})\in\mathcal{D}_{j-1,q-q_j}} \frac{(q-q_j)!}{q_1!\cdots q_{j-1}!}\prod_{i=1}^{j-1} w_i^{q_i}(a_i)_{q_i}.$$

For any $n\in\mathbb{N}$ let $\Delta^{(n)} := \{(u_1,\dots,u_n) : u_i \ge 0, i=1,\dots,n, \sum_{i=1}^{n} u_i \le 1\}$ be the n–dimensional simplex; then, we can write

$$\begin{aligned}
&\sum_{(q_1,\dots,q_j)\in\mathcal{D}_{j,q}} \frac{q!}{q_1!\cdots q_j!}\prod_{i=1}^{j} w_i^{q_i}(a_i)_{q_i}\\
&= \sum_{q_j=0}^{q} \frac{q!}{q_j!(q-q_j)!} w_j^{q_j}(a_{j-1})_{q_j} w_{j-1}^{q-q_j}(a-a_j)_{(q-q_j)}\\
&\quad\times F_D^{(j-2)}\left(-q+q_j, a_1,\dots,a_{j-2}, a-a_j; \frac{w_{j-1}-w_1}{w_{j-1}},\dots,\frac{w_{j-1}-w_{j-2}}{w_{j-1}}\right)\\
&= \frac{\Gamma(a-a_j)}{\Gamma(a_1)\cdots\Gamma(a_{j-1})}\int_{\Delta^{(j-2)}} \prod_{i=1}^{j-2} z_i^{a_i-1}\left(1-\sum_{i=1}^{j-2} z_i\right)^{a_{j-1}-1}\\
&\quad\times \sum_{q_j=0}^{q} \frac{q!}{q_j!(q-q_j)!} w_j^{q_j}(a_j)_{q_j} w_{j-1}^{q-q_j}(a-a_j)_{(q-q_j)}\left(1-\sum_{i=1}^{j-2} z_i\frac{w_{j-1}-w_i}{w_{j-1}}\right)^{q-q_j} dz_1\cdots dz_{j-2}\\
&= \frac{\Gamma(a-a_j)}{\Gamma(a_1)\cdots\Gamma(a_{j-1})}\int_{\Delta^{(j-2)}} \prod_{i=1}^{j-2} z_i^{a_i-1}\left(1-\sum_{i=1}^{j-2} z_i\right)^{a_{j-1}-1} \sum_{q_j=0}^{q} \frac{q!}{q_j!(q-q_j)!} w_j^{q}(a_j)_{q_j}\\
&\quad\times \left(\frac{1}{w_j}\right)^{q-q_j}\left[w_j-\sum_{i=1}^{j-2} z_i(w_j-w_i)-\left(1-\sum_{i=1}^{j-2} z_i\right)(w_j-w_{j-1})\right]^{q-q_j}\\
&\quad\times\ (a-a_j)_{(q-q_j)} dz_1\cdots dz_{j-2}\\
&= \frac{\Gamma(a-a_j)}{\Gamma(a_1)\cdots\Gamma(a_{j-1})}\int_{\Delta^{(j-2)}} \prod_{i=1}^{j-2} z_i^{a_i-1}\left(1-\sum_{i=1}^{j-2} z_i\right)^{a_{j-1}-1}\\
&\quad\times \left(1-\sum_{i=1}^{j-2} z_i\frac{w_j-w_i}{w_j}-\left(1-\sum_{i=1}^{j-2} z_i\right)\frac{w_j-w_{j-1}}{w_j}\right)^{q}\\
&\quad\times \sum_{q_j=0}^{q} \frac{q!}{q_j!(q-q_j)!} w_j^{q-q_j}(a_j)_{q_j}\left(\frac{-w_j}{\sum_{i=1}^{j-2} z_i\frac{w_j-w_i}{w_j}+\left(1-\sum_{i=1}^{j-2} z_i\right)\frac{w_j-w_{j-1}}{w_j}-1}\right)^{q_j}\\
&\quad\times (a-a_j)_{(q-q_j)} dz_1\cdots dz_{j-2}.
\end{aligned}$$

By applying Eq. 4, from the last equation we obtain

$$\frac{\Gamma(a-a_j)}{\Gamma(a_1)\cdots\Gamma(a_{j-1})} w_j^q (a)_q \int_{\Delta^{(j-2)}} \prod_{i=1}^{j-2} z_i^{a_i-1} \left(1-\sum_{i=1}^{j-2} z_i\right)^{a_{j-1}-1}$$

$$\times \left(1-\sum_{i=1}^{j-2} z_i \frac{w_j-w_i}{w_j} - \left(1-\sum_{i=1}^{j-2} z_i\right)\frac{w_j-w_{j-1}}{w_j}\right)^q$$

$$\times {}_2F_1\left(-q, a_j; a; \frac{\sum_{i=1}^{j-2} z_i \frac{w_j-w_i}{w_j} + \left(1-\sum_{i=1}^{j-2} z_i\right)\frac{w_j-w_{j-1}}{w_j}}{\sum_{i=1}^{j-2} z_i \frac{w_j-w_i}{w_j} + \left(1-\sum_{i=1}^{j-2} z_i\right)\frac{w_j-w_{j-1}}{w_j} - 1}\right) dz_1\cdots dz_{j-2}$$

or, equivalently,

$$\frac{\Gamma(a-a_j)}{\Gamma(a_1)\cdots\Gamma(a_{j-1})} w_j^q (a)_q \int_{\Delta^{(j-2)}} \prod_{i=1}^{j-2} z_i^{a_i-1} \left(1-\sum_{i=1}^{j-2} z_i\right)^{a_{j-1}-1}$$

$$\times {}_2F_1\left(-q, a-a_j; a; \sum_{i=1}^{j-2} z_i \frac{w_j-w_i}{w_j} + \left(1-\sum_{i=1}^{j-2} z_i\right)\frac{w_j-w_{j-1}}{w_j}\right) dz_1\cdots dz_{j-2}.$$

Since $a-a_j>0$ and

$$1 > \max\left\{0, \Re\left(\sum_{i=1}^{j-2} z_i \frac{w_j-w_i}{w_j} + (w_j-w_{j-1})\left(1-\sum_{i=1}^{j-2} z_i\right)\right)\right\}$$

then we can apply Eq. 7.621.4 in Gradshteyn and Ryzhik (2000) in order to obtain the expression

$$\frac{1}{\Gamma(a_1)\cdots\Gamma(a_{j-1})} w_j^q (a)_q \int_0^{+\infty} \mathrm{e}^{-z_{j-1}} z_{j-1}^{a-a_j-1} \int_{\Delta^{(j-2)}} \prod_{i=1}^{j-2} z_i^{a_i-1} \left(1-\sum_{i=1}^{j-2} z_i\right)^{a_{j-1}-1}$$

$$\times {}_1F_1\left(-q; a; z_{j-1}\left(\sum_{i=1}^{j-2} z_i \frac{w_j-w_i}{w_j} + \left(1-\sum_{i=1}^{j-2} z_i\right)\frac{w_j-w_{j-1}}{w_j}\right)\right) dz_1\cdots dz_{j-2} dz_{j-1}$$

Finally, using the change of variable $y_i = z_i z_{j-1}$ for $i=1,\ldots,j-2$ and $y_{j-1}=z_{j-1}$ we obtain the expression

$$\frac{1}{\Gamma(a_1)\cdots\Gamma\left(a_{j-1}\right)} w_j^q (a)_q \int_0^{+\infty} \mathrm{e}^{-y_{j-1}} \int_{B(y_j)} \prod_{i=1}^{j-2} y_i^{a_i-1} \left(y_{j-1}-\sum_{i=1}^{j-2} y_i\right)^{a_{j-1}-1}$$

$$\times {}_1F_1\left(-q; a; \sum_{i=1}^{j-2} y_i \frac{w_j-w_i}{w_j} + \left(y_{j-1}-\sum_{i=1}^{j-2} y_i\right)\frac{w_j-w_{j-1}}{w_j}\right) dy_1\cdots dy_{j-1}$$

where

$$B\left(y_j\right) = \left\{\left(y_1,\ldots,y_{j-1}\right) : y_i \geq 0, \sum_{i=1}^{j-1} y_i \leq y_j\right\}$$

and using the change of variable $u_i = y_i$ per $i = 1, \dots, j-2$ e $u_{j-1} = y_{j-1} - \sum_{i=1}^{j-2} y_i$ we have

$$\frac{w_j^q (a)_q}{\Gamma(a_1)\cdots\Gamma\left(a_{j-1}\right)} \int_{(\mathbb{R}^+)^{j-1}} \mathrm{e}^{-\sum_{i=1}^{j-1} u_i} \prod_{i=1}^{j-1} u_i^{a_i-1}\, {}_1F_1\left(-q; a; \sum_{i=1}^{j-1} u_i \frac{w_j - w_i}{w_j}\right) du_1 \cdots du_{j-1}$$

and the proof is completed by applying the identity (Eq. 3). □

In the following corollary identity (Eq. 7) in Proposition 2 is specialized to the setup arising in the derivation of the estimators.

Corollary 1 *For any $q \geq 1$, $j \geq 1$ let $w_1, \dots, w_j \in \mathbb{R}^+$, $a_1, \dots, a_j > 0$ and $p_1, \dots, p_j \in \mathbb{N}$. Then*

$$\sum_{(q_1,\dots,q_j)\in\mathcal{D}_{j,q}} \binom{q}{q_1, \dots, q_j} \prod_{i=1}^{j} w_i^{q_i} (a_i)_{(q_i+p_i)}$$

$$= w_j^q (p+a)_q \prod_{i=1}^{j} (a_i)_{p_i} F_D^{(j-1)}\left(-q, a_1, \dots, a_{j-1}, p+a; \frac{w_j - w_1}{w_j}, \dots, \frac{w_j - w_{j-1}}{w_j}\right) \quad (8)$$

where $a := \sum_{i=1}^{j} a_i$ and $p := \sum_{i=1}^{j} p_i$.

3 Application to Species Diversity Estimation

We first introduce the framework and then highlight the usefulness of the multivariate generalized Chu–Vandermonde identity derived in Section 2. Let $(X_n)_{n\geq1}$ be a sequence of exchangeable random variables defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with values in a complete and separable metric space $\mathbb{X}$ equipped with the corresponding Borel σ–field $\mathscr{X}$. Then, by de Finetti's representation theorem, there exists a random probability measure $\tilde{P}$ such that given $\tilde{P}$, a sample $X_1, \dots, X_n$ from the exchangeable sequence is independent and identically distributed with distribution $\tilde{P}$. That is, for every $n \geq 1$ and any $A_1, \dots, A_n \in \mathscr{X}$

$$\mathbb{P}\left(X_1 \in A_1, \dots, X_n \in A_n | \tilde{P}\right) = \prod_{i=1}^{n} \tilde{P}(A_i).$$

By assuming the random probability measure $\tilde{P}$ to be almost surely discrete, ties will appear in the sample with positive probability, namely $(X_1, \dots, X_n)$ will contain $K_n \leq n$ distinct observations $X_1^*, \dots, X_{K_n}^*$ with frequencies $\mathbf{N}_n := (N_1, \dots, N_{K_n})$ such that $\sum_{j=1}^{k} N_j = n$.

The joint distribution of K_n and $\mathbf{N}_n$ provides the partition distribution of the exchangeable sample $X_1, \dots, X_n$ and plays an important role in a variety of research areas such as population genetics, machine learning, Bayesian nonparametrics, combinatorics, excursion theory and statistical physics. See Pitman (2006) for an exhaustive and stimulating account. In particular, recent applications of exchangeable partition distributions concern species sampling problems, which gained a renewed interest due to their importance in Genomics where the population is typically a cDNA library and the species are unique genes which are progressively sequenced; see Lijoi et al. (2007a, b, 2008) and references therein. Specifically, given an exchangeable sample $(X_1, \ldots, X_n)$ from some almost surely discrete random probability measure $\tilde{P}$ consisting of a collection of $K_n = j$ distinct species with labels $(X_1^*, \ldots, X_j^*)$ and frequencies $(n_1, \ldots, n_j)$, the main interest relies in estimating the number of distinct species to be observed in a hypothetical additional sample of size m.

Formally, let $X_1, \ldots, X_n$ be the so–called "basic sample" of size n containing K_n distinct observations with frequencies $\mathbf{N}_n$ and corresponding to the typically available information. Denote by $K_m^{(n)} = K_{m+n} - K_n$ the number of new partition sets $C_1, \ldots, C_{K_m^{(n)}}$ generated by the additional sample $X_{n+1}, \ldots, X_{n+m}$. Furthermore, if $C := \cup_{i=1}^{K_m^{(n)}} C_i$ whenever $K_m^{(n)} \geq 1$ and $C \equiv \emptyset$ if $K_m^{(n)} = 0$, we set $L_m^{(n)} := \text{card}(\{X_{n+1}, \ldots, X_{n+m}\} \cap C)$ as the number of observations belonging to the new clusters C_i. It is clear that $L_m^{(n)} \in \{0, 1, \ldots, m\}$ and that $m - L_m^{(n)}$ observations belong to the sets defining the partition of the original n observations. According to this, if $\mathbf{S}_{L_m^{(n)}} := (S_{1, L_m^{(n)}}, \ldots, S_{K_m^{(n)}, L_m^{(n)}})$, then the distribution of $\mathbf{S}_{L_m^{(n)}}$ conditional on $L_m^n = s$, is supported by all vectors $(s_1, \ldots, s_{K_m^{(n)}})$ of positive integers such that $\sum_{i=1}^{K_m^{(n)}} s_i = s$. The remaining $m - L_m^{(n)}$ observations are allocated to the "old" K_n clusters with vector of nonnegative frequencies $\mathbf{R}_{m-L_m^{(n)}} := (R_{1,m-L_m^{(n)}}, \ldots, R_{K_n,m-L_m^{(n)}})$ such that $\sum_{i=1}^{K_n} R_{i,m-L_m^{(n)}} = m - L_m^{(n)}$. Based on this setup of random variables, the issue we address consists in evaluating, conditionally on the partition induced by the basic sample of size n, the probability of sampling in m further draws a certain number of new partition groups (species), i.e. ,

$$\mathbb{P}\left(K_m^{(n)} = k | X_1, \ldots, X_n\right) = \sum_{\mathscr{P}_{m,k+j}} \frac{\mathbb{P}\left(L_m^{(n)}=s, K_n=j, \mathbf{N}_n=\left(n_1,\ldots,n_{K_n}\right), K_m^{(n)}=k, \mathbf{S}_{L_m^{(n)}}=\left(s_1,\ldots,s_{K_m^{(n)}}\right), \mathbf{R}_{m-L_m^{(n)}}=\left(r_1,\ldots,r_{K_n}\right)\right)}{\mathbb{P}\left(K_n = j, \mathbf{N}_n = \left(n_1, \ldots, n_{K_n}\right)\right)} \tag{9}$$

where $\mathscr{P}_{m,j+k}$ denotes the set of all allocations of m observations into $q \leq m$ classes, with $q \in \{k, \ldots, k+j\}$; in other terms k observations are new species and $q - k \leq j$ coincide with some of the j already observed distinct species in $X_1, \ldots, X_n$. In particular, expression (Eq. 9) can be written as

$$\begin{aligned}\mathbb{P}\left(K_m^{(n)} = k | X_1, \ldots, X_n\right) \propto \sum_{s=k}^{m} \binom{m}{s} \sum_{(r_1,\ldots,r_j)\in\mathcal{D}_{j,n}} \binom{m-s}{r_1, \ldots, r_j} \frac{1}{k!} \sum_{(s_1,\ldots,s_k)\in\mathcal{D}_{k,s}^*} \binom{s}{s_1, \ldots, s_k} \\ \times \mathbb{P}\left(L_m^{(n)} = s, K_n = j, \mathbf{N}_n = \left(n_1, \ldots, n_{K_n}\right), K_m^{(n)} = k,\right. \\ \left.\mathbf{S}_{L_m^{(n)}} = \left(s_1, \ldots, s_{K_m^{(n)}}\right), \mathbf{R}_{m-L_m^{(n)}} = \left(r_1, \ldots, r_{K_n}\right)\right)\end{aligned}$$

with

$$\mathcal{D}_{k,s}^* := \left\{(s_1, \ldots, s_k) : s_i \geq 1 \text{ for } i = 1, \ldots, k, \sum_{i=1}^{k} s_i = s\right\}.$$

At this point the usefulness of Corollary 1 becomes evident. Consider a species sampling problem characterized by a joint distribution $\mathbb{P}(L_m^{(n)} = s, K_n = j, \mathbf{N}_n = (n_1, \ldots, n_{K_n}), K_m^{(n)} = k, \mathbf{S}_{L_m^{(n)}} = (s_1, \ldots, s_{K_m^{(n)}}), \mathbf{R}_{m-L_m^{(n)}} = (r_1, \ldots, r_{K_n}))$ assuming the following quite general form, which includes all explicitly known instances,

$$\mathbb{P}\left(L_m^{(n)} = s, K_n = j, \mathbf{N}_n = \left(n_1, \ldots, n_{K_n}\right), K_m^{(n)} = k, \mathbf{S}_{L_m^{(n)}} = \left(s_1, \ldots, s_{K_m^{(n)}}\right),\right.$$
$$\left.\mathbf{R}_{m-L_m^{(n)}} = \left(r_1, \ldots, r_{K_n}\right)\right) = g\left(n, m, j, k\right) \prod_{i=1}^{j} w_i^{r_i} (a_i)_{(n_i+r_i)} \prod_{i=1}^{k} f_i(m, k, s_i)$$

for some positive functions $g(\cdot)$ and $f_i(\cdot)$ for $i = 1, \ldots, k$ and for some $w_1, \ldots, w_j \in \mathbb{R}^+$ and $a_1, \ldots, a_j \in \mathbb{R}^+$. Then the identity (Eq. 8) provided Corollary 1 can be usefully applied in order to obtain closed form solutions for the multivariate convolutions generalizing the one of Chu–Vandermonde, i.e.

$$\begin{aligned}
&\mathbb{P}\left(K_m^{(n)} = k | X_1, \ldots, X_n\right) \\
&\propto \sum_{s=k}^{m} \binom{m}{s} \sum_{(r_1,\ldots,r_j)\in\mathcal{D}_{j,n}} \binom{m-s}{r_1, \ldots, r_j} \frac{1}{k!} \sum_{(s_1,\ldots,s_k)\in\mathcal{D}^*_{k,s}} \binom{s}{s_1, \ldots, s_k} \\
&\quad \times g(n, m, j, k) f\left(m, k, (s_1, \ldots, s_k)\right) \prod_{i=1}^{j} w_i^{r_i} (a_i)_{(n_i+r_i)} \qquad (10)\\
&= g(n, m, j, k) \sum_{s=k}^{m} \binom{m}{s} w_j^{m-s} (n+a)_{(m-s)} \prod_{i=1}^{j} (a_i)_{(n_i)} \\
&\quad \times F_D^{(j-1)}\left(-m+s, a_1, \ldots, a_{j-1}, n+a; \frac{w_j - w_1}{w_j}, \ldots, \frac{w_j - w_{j-1}}{w_j}\right) \\
&\quad \times \frac{1}{k!} \sum_{(s_1,\ldots,s_k)\in\mathcal{D}^*_{k,s}} \binom{s}{s_1, \ldots, s_k} \prod_{i=1}^{k} f_i(m, k, s_i) \qquad (11)
\end{aligned}$$

With reference to the sum over the set of partitions $\mathcal{D}^*_{k,s}$ it has to be evaluated according to the analytic form of the functions $f_i(m, k, s_i)$ for $i = 1, \ldots, k$. In particular, if $f_i(m, k, s_i) = f(m, k, s_i)$ for $i = 1, \ldots, k$, for some positive function $f(\cdot)$, then it is well–known that

$$\frac{1}{k!} \sum_{(s_1,\ldots,s_k)\in\mathcal{D}^*_{k,s}} \binom{s}{s_1, \ldots, s_k} \prod_{i=1}^{k} f(m, k, s_i) = B_{s,k}(v_\bullet)$$

where $B_{s,k}(v_\bullet)$ is the (s, k)–partial Bell polynomial with weight sequence $v_\bullet := \{v_i, i \geq 1\}$ such that $v_i := h(m, k, i)$ for $i \geq 1$; see Comtet (1974). For some examples, where Eq. 11 can be evaluated explicitly leading to a readily applicable estimator of $K_m^{(n)} | K_n = j$ we refer to Lijoi et al. (2007a, 2008) and Favaro et al. (2010).

References

Appell P, Kampé de Fériet J (1926) Functions hypergéométriques et hypersphériques: polynomes d'hermite. Gauthier-Villars, Paris

Comtet L (1974) Advanced combinatorics. Reidel, Boston

Exton H (1976) Multiple hypergeometric functions and application. Ellis Horwood, Chichester

Favaro S, Prünster I, Walker SG (2010) On a class of random probability measures with general predictive structure. Scand J Statist. doi:10.1111/j.1467-9469.2010.00702.x

Gradshteyn LS, Ryzhik LM (2000) Table of integrals, series, and products. Academic, New York

Hjort NL, Holmes CC, Müller P, Walker SG (eds) (2010) Bayesian nonparametrics. Cambridge University Press, Cambridge

James LF (2005) Functionals of Dirichlet processes, the Cifarelli–Regazzini identity and beta-gamma processes. Ann Stat 33:647–660

Lauricella G (1893) Sulle funzioni ipergeometriche a più variabili. Rend Circ Mat Palermo 7:111–158

Lijoi A, Mena RH, Prünster I (2007a) Bayesian nonparametric estimation of the probability of discovering a new species Biometrika 94:769–786

Lijoi A, Mena RH, Prünster I (2007b) A Bayesian nonparametric method for prediction in EST analysis. BMC Bioinformatics 8:339

Lijoi A, Prünster I, Walker SG (2008) Bayesian nonparametric estimators derived from conditional Gibbs structures. Ann Appl Probab 18:1519–1547

Lijoi A, Regazzini E (2004) Means of a Dirichlet process and multiple hypergeometric functions. Ann Probab 32:1469–1495

Müller P, Quintana FA (2004) Nonparametric Bayesian data analysis. Stat Sci 19:95–110

Navarrete C, Quintana FA, Müller P (2008) Some issues on nonparametric Bayesian modeling using species sampling models. Stat Model 8:3–21

Petrone S, Guindani M, Gelfand AE (2009) Hybrid Dirichlet mixture models for functional data. J R Stat Soc Ser B 71:755–782

Pitman J (2006) Combinatorial stochastic processes. Ecole d'Eté de Probabilités de Saint-Flour XXXII. In: Lecture notes in mathematics, vol 1875. Springer, New York

Quintana FA (2006) A predictive view of Bayesian clustering. J Stat Plan Inference 136:2407–2429

Regazzini E (1998) An example of the interplay between statistics and special functions. In: Tricomi's ideas and contemporary applied mathematics. Accademina Nazionale dei Lincei, Rome, pp 303–320