Yet another application of marginals of multivariate Gibbs distributions*

ANNALISA CERQUETTI†
Department of Methods and Models for Economics, Territory and Finance
Sapienza University of Rome, Italy

20 Dicembre 2013

Abstract

We give yet another example of the usefulness of working with marginals of multivariate Gibbs distributions (Cerquetti, 2013) in deriving Bayesian non-parametric estimators under Gibbs priors in species sampling problems. Here in particular we substantially reduce length and complexity of the proofs in Bacallado et al. (2013, Th. 1, and Th. 2) for looking backward probabilities under incomplete information.

1 Introduction

In the typical setting of Bayesian nonparametric inference for species sampling problems under Gibbs priors (Lijoi et al. 2007, 2008, Favaro et al. 2012, 2013) Bacallado et al. (2013) obtain conditional falling factorial moments, given an initial sample of observations, for the number of old species re-observed and for the number of old species re-observed a certain number of times in an additional sample, under complete and incomplete information. As the Authors explain, here incomplete information stands for the possibility to know the number \( j \) of different species observed in the initial \( n \)-sample, but not the specific multiplicities \( (n_1, \ldots, n_j) \).

Recently it has been shown that relying on marginals of multivariate Gibbs distributions, posterior predictive inference in species sampling problems can be considerably simplified (Cerquetti, 2013, 2013b). Here we give another example of the usefulness of this new technique deriving the main results in Bacallado et al. (2013) quickly and easily. For the sake of brevity we just recall here the essential preliminaries and notation. We refer the interested reader to the cited literature for a comprehensive introduction to BNP in species sampling.

---

*AMS (2000) subject classification. Primary: 60G58. Secondary: 60G09.
†Corresponding author, SAPIENZA University of Rome, Via del Castro Laurenziano, 9, 00161 Rome, Italy. E-mail: annalisa.cerquetti@gmail.com
2 Some preliminaries on notation

Gibbs priors are a large class of laws for infinite discrete distributions \((P_i)_{i \geq 1}\) introduced by Gnedin and Pitman (2006) as generalizations of the two-parameter Poisson-Dirichlet priors (Pitman and Yor, 1997). The Gibbs class is conveniently identified by the corresponding family of exchangeable partition probability functions (EPPFs) it induces by random sampling. Those are distributions on the space of consistent sequences of random partitions \(\{A_1, \ldots, A_j\}\) of the finite sets \([n] := \{1, \ldots, n\}\) characterized by the Gibbs product form

\[
p_{\alpha,V}(n_1, \ldots, n_j) = V_{n,j} \prod_{i=1}^{j} (1 - \alpha)_{n_i - 1},
\]

where \(p\) is a symmetric function of compositions \((n_1, \ldots, n_j)\) of \(n\) corresponding to the sizes of the \(j\) different blocks in order of appearance, \(V = (V_{n,j})\) are weights that identify the specific Gibbs model and \(\alpha \in (-\infty, 1)\). Notice that here, with respect to Bacallado et al. (2013) we use \(\alpha\) in place of \(\sigma\) as it is standard in exchangeable random partitions literature and the lighter notation \((x)_{s} = (x)(x+1)\cdots(x+s-1)\) for rising factorials. By Theorem 12 in Gnedin and Pitman (2006) each exchangeable Gibbs partition arises as a probability mixture of extreme partitions, which differ for \(\alpha < 0, \alpha = 0\) and \(\alpha \in (0, 1)\). The joint law of the corresponding random vector \((N_1, \ldots, N_{K_n}, K_n)\) of the blocks’ multiplicities in exchangeable random order and the number of blocks, first introduced in Pitman (2006, eq. (2.7)), has been termed multivariate Gibbs distribution in Cerquetti (2013) and is given by

\[
\mathbb{P}_{\alpha,V}(N_1 = n_1, \ldots, N_{K_n} = n_j, K_n = j) = \frac{n!}{\prod_i n_i!} V_{n,j} \prod_{i=1}^{j} (1 - \alpha)_{n_i - 1}.
\]

The law of \(K_n\) follows by marginalizing (1) over the space of compositions of \(n\) with \(j\) blocks

\[
\mathbb{P}_{\alpha,V}(K_n = j) = V_{n,j} S_{n,j}^{-1,-\alpha}
\]

where \(S_{n,j}^{-1,-\alpha}\) are generalized Stirling numbers as arising from Bell polynomials

\[
B_{n,j}(w_\bullet) = \frac{n!}{j!} \sum_{(n_1, \ldots, n_j)} \prod_{i=1}^{j} \frac{w_{n_i}}{n_i!}
\]

for \(w_\bullet = (1 - \alpha)_{\bullet-1}\).

Notice that in Bacallado et al. (2013) equation (2) is expressed in terms of generalized factorial coefficients \(C(n, j; \alpha)\) as

\[
\mathbb{P}_{\alpha,V}(K_n = j) = V_{n,j} \frac{C(n, j; \alpha)}{\alpha^j}
\]
where $C(n, j; \alpha)$ are the connection coefficients

$$\alpha x_n = \sum_{j=0}^{n} C(n, j; \alpha)(x)_j,$$

and are not defined for $\alpha = 0$. Since the general Gibbs class is defined for $\alpha \in (-\infty, 1)$ here, as in Cerquetti (2013, 2013b) we use generalized Stirling numbers $S_{n,j}^{-1,-\alpha}$ defined as connection coefficients

$$S_{n,j}^{-1,- \alpha} = \sum_{j=0}^{n} S_{n,j}^{-1,- \alpha}(x)_j \uparrow \alpha$$

where $(x)_j \uparrow \alpha = x(x+\alpha)(x+2\alpha)\cdots(x+(j-1)\alpha)$ are generalized rising factorials. For $\alpha \neq 0$ then $S_{n,j}^{-1,- \alpha} = \frac{C(n, j; \alpha)}{\alpha}$. For $\alpha = 0$ then $(x)_j \uparrow 0 = x^j$ and those numbers correspond to signless Stirling numbers of the first kind $S_{n,j}^{-1,0}$.

3 Looking backward via marginals of multivariate Gibbs

In species sampling problems given an initial sample of size $n$ with $j$ different species observed with multiplicities $(n_1, \ldots, n_j)$, interest usually lies on inferring the behaviour of an additional sample of size $m$ with respect to the number $K_m$ of new species, the total number $L_m$ of observations belonging to new species, the multiplicities $M_1, \ldots, M_{K_m}$ of the new species observed, and the number $S_1, \ldots, S_j$ of new observations belonging to old species. (See Cerquetti, 2013, Lijoi et al., 2007, 2008, Favaro et al., 2012, 203). The main results in Bacallado et al. (2013) are in Theorem 1. and Theorem 2. which provide respectively:

- conditional falling factorial moments for $R_{m}^{(j,n,n)}$ and $R_{m}^{(j,m)}$, the total number of old species re-observed in the additional sample under complete and incomplete information (Theorem 1, eq. (3.4) and (3.5)).

- conditional falling factorial moments for $R_{l,m}^{(j,n,n)}$ and $R_{l,m}^{(j,n)}$ the number of old species re-observed $l$ times in the additional sample under complete and incomplete information (Theorem 2, eq. (3.12) and (3.13)).

To obtain results (3.4) and (3.12), as the same Authors state in the proofs, it is enough to specialize for $l = 0$ the result in Theorem 1 in Favaro et al. (2013) and to resort to $R_{m}^{(j,n,n)} = j - R_{0,m}^{(j,n,n)}$. See also Cerquetti (2013, eq. 26) for a result in terms of generalized Stirling numbers.

To obtain results (3.5) and (3.13) the Authors adopt a complicated procedure resulting in two different proofs of about six pages each. Here we present a far more easy route relying on marginals of multivariate Gibbs distributions as introduced in Cerquetti (2013). Previous examples of the usefulness of this new technique are in Cerquetti (2013b).
First notice that since the total number of old species is observed, then even under incomplete information
\[ R^{(j,n)}_m = j - R^{(j,n)}_{0,m}. \]
Therefore, by properties of rising factorials and the definition of non central Lah numbers, which correspond to generalized Stirling numbers for \( \alpha = -1 \),
\[ S_{r,v}^{-1,1} = \frac{r!}{v!} \binom{r-j-1}{r-v} \]
then
\[ [(R^{(j,n)}_m)_r] = [(j - R^{(j,n)}_{0,m})_r] = (-1)^r \sum_{v=0}^r S_{r,v}^{-1,1-j}(R^{(j,n)}_{0,m})_v. \]
This implies that it is enough to give a proof for (3.13), since (3.5) follows easily.

**Remark 1.** Eq. (3.13) in Bacallado et al. (2013) is written in terms of generalized factorial coefficients \( C(n, j; \alpha) \) (central and non central) namely
\[ \mathbb{E}[(R^{(n,j)}_{l,m})_r] = \frac{r!}{C(n, j; \alpha)} \frac{m!}{(l!)(m-rl)!} [\alpha(1-\alpha)l-1]^r \]
\[ \times \sum_{s=r}^{n-j} \binom{n}{s} C(s, r; \alpha-s) \sum_{k=0}^{m-rl} \frac{V_{n+m,j+k} C(m-rl, k, \alpha, -(n-s-(j-r)\alpha)}{\alpha^k} \]
Since generalized factorial coefficients are not defined for \( \alpha = 0 \) the previous expression is not directly applicable for posterior inference under priors belonging to the Gibbs class for \( \alpha = 0 \). To recover the value of the estimator for \( \alpha = 0 \) it is necessary to resort to the limit value for \( \alpha \to 0 \) exploiting the known relationship
\[ \lim_{\alpha \to 0} \frac{C(n, j; \alpha)}{\alpha^l} = S_{n,j}^{-1,0}, \]
where \( S_{n,j}^{-1,0} \) are signless Stirling numbers of the first kind. The very same problem applies to equations (3.4), (3.5) and (3.12). To avoid this kind of drawbacks we advocate here the use of generalized Stirling numbers in Bayesian nonparametrics under Gibbs priors, as it is standard in the exchangeable Gibbs partitions literature (cf. e.g. Gnedin and Pitman, 2006; Pitman, 2006). We stress that, despite still not completely explored, the class of Gibbs priors with \( \alpha = 0 \) is theoretically infinite. The corresponding Gibbs weights \( V_{n,k} \) can be obtained mixing over \( \theta \) the weights of Dirichlet priors \( V_{n,k}^\theta = \frac{\theta^k}{\theta_1 \cdots \theta_n} \) with a general density \( \gamma(\theta) \) on \((0, \infty)\).
The following Proposition gives an easy and short proof for $\mathbb{E}[(R_{i,m}^{j,n})_{r,i}]$ as obtained in eq. (3.13) in Bacallado et al. (2013).

**Proposition 1.** Let $n$ and $j$ be respectively the size and the number of different species observed in a sample from an unknown population of infinite species. Let $m$ be the size of an additional sample, then the $r$-th falling factorial moment of the number $R_{i,m}$ of old species reobserved $l$ times, given $n$ and $j$ under $(\alpha, V)$-Gibbs priors, for $\alpha \in (-\infty, 1)$, is given by

$$
\mathbb{E}^\alpha [(R_{i,m}^{j,n})_{r,i}] = \frac{m!}{(l!)^r (m-r)!} \left(1 - \alpha\right)^l \sum_{s=r}^{n-r+j} \frac{n-s}{s} \frac{S_{s,r}^{1,\alpha} S_{s-n-r}^{1,\alpha}}{S_{n,j}^{1,\alpha}}
$$

(9)

\[ \times \sum_{k=0}^{m-r} \frac{V_{n+m,j+k}}{V_{n,j}} S_{m-r-l,k}^{1,\alpha} - \alpha - (n-s-j) \alpha. \]

**Proof.** Let $\mathbf{N} = (N_1, \ldots, N_j, K_n = j)$ then by definition of incomplete information

$$
\mathbb{E}^\alpha [(R_{i,m}^{j,n})_{r,i}] = \mathbb{E}_{\mathbf{N}}^\alpha[a_{K_n=j}] \left[ \mathbb{E}^\alpha [(R_{i,m}^{j,n})_{r,i}] \right]
$$

where, by (3.12) in Bacallado et al. (2013) expressed in terms of generalized Stirling numbers

$$
\mathbb{E}^\alpha [(R_{i,m}^{j,n})_{r,i}] = \frac{m!}{(l!)^r (m-r)!} \left(1 - \alpha\right)^l \sum_{(c_1, \ldots, c_r) \in C_{j,r}} \prod_{i=1}^r (N_{c_i} - \alpha)_i
$$

\[ \times \sum_{k=0}^{m-r} \frac{V_{n+m,j+k}}{V_{n,j}} S_{m-r-l,k}^{1,\alpha} - \alpha - (n-S_j-\alpha). \]

Now for $T_r = \sum_{i=1}^r N_{c_i}$ taking values in $[r, n - (j - r)]$, we can write in short form

$$
\mathbb{E}^\alpha [(R_{i,m}^{j,n})_{r,i}] = \frac{m!}{(l!)^r (m-r)!} \sum_{(c_1, \ldots, c_r) \in C_{j,r}} \prod_{i=1}^r (N_{c_i} - \alpha)_i g_{m,r,l,j}^\alpha (T_r)
$$

where $C_{j,r}$ is the space of $r$-combinations of $[j]$, hence

$$
\mathbb{E}_{\mathbf{N}}^\alpha[a_{K_n=j}] \left[ \mathbb{E}^\alpha [(R_{i,m}^{j,n})_{r,i}] \right] = \mathbb{E}_{N_{c_1}, \ldots, N_{c_r}}^\alpha[(R_{i,m}^{j,n})_{r,i}]
$$

and by exchangeability

$$
= \binom{j}{r} \frac{m!}{(l!)^r (m-r)!} \mathbb{E}_{N_{c_1}, \ldots, N_{c_r}}^\alpha \left\{ \prod_{i=1}^r (N_i - \alpha)_i \right\} g_{m,r,l,j}^\alpha (T_r). \tag{10}
$$

By (7) in Cerquetti (2013) and (2) for $r \leq j$

$$
\mathbb{P}_{\alpha}(N_1 = n_1, \ldots, N_r = n_r | K_n = j) = \frac{n!}{\prod_{i=1}^r n_i! (n - \sum_i n_i)!} \prod_{i=1}^r (1 - \alpha)_i \frac{S_{n-\sum_i n_i-j}^{1-\alpha}}{S_{n,j}^{1-\alpha}}.
$$
Now notice that \((n_i - \alpha)l(1 - \alpha)_{n_i - 1} = (1 - \alpha)_l(1 - \alpha + l)_{n_i - 1}\), hence (11) corresponds to

\[
\begin{align*}
\binom{j}{r} \frac{r!m!}{(l!)^r(m - rl)!} [(1 - \alpha)_l]^r & \mathbb{E}^\alpha_l \left[ \sum_{\sum_{i=1}^r n_i = j} [g_{m,r,l,j}(T_r)]^{S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j-r} \frac{S_{n_j}^{-1,\alpha}}{S_{n_j}^{S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j-r}}} \right] \\
& = \binom{j}{r} \frac{r!m!}{(l!)^r(m - rl)!} [(1 - \alpha)_l]^r \left[ \sum_{\sum_{i=1}^r N_i = j} [g_{m,r,l,j}(T_r)]^{S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j-r} \frac{S_{n_j}^{-1,\alpha}}{S_{n_j}^{S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j-r}}} \right]
\end{align*}
\]

and

\[
\begin{align*}
\binom{j}{r} \frac{r!m!}{(l!)^r(m - rl)!} [(1 - \alpha)_l]^r & \mathbb{E}^\alpha_l \left[ \sum_{\sum_{i=1}^r n_i = j} [g_{m,r,l,j}(T_r)]^{S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j-r} \frac{S_{n_j}^{-1,\alpha}}{S_{n_j}^{S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j-r}}} \right] \\
& = \binom{j}{r} \frac{r!m!}{(l!)^r(m - rl)!} [(1 - \alpha)_l]^r \left[ \sum_{\sum_{i=1}^r N_i = j} [g_{m,r,l,j}(T_r)]^{S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j-r} \frac{S_{n_j}^{-1,\alpha}}{S_{n_j}^{S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j-r}}} \right]
\end{align*}
\]

By equation (9) in Cerquetti (2013) and the definition of generalized Stirling numbers in terms of Bell polynomials (3), multiplying and dividing by \(s!\) and \(r!\)

\[
\mathbb{P}_{\alpha-l}(T_r = s|K_n = j) = \sum_{\sum_{i=1}^r r_i = s} \mathbb{P}_{\alpha-l}(N_1, \ldots, N_r|K_n = j) = \sum_{\sum_{i=1}^r r_i = s} \frac{n! (j - r)! \prod_{i=1}^r (1 - \alpha + l)n_i^{-1,\alpha+l} S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j-r} \frac{S_{n_j}^{-1,\alpha}}{S_{n_j}^{S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j-r}}} \right] \\
\binom{j}{r} \frac{r!m!}{(l!)^r(m - rl)!} [(1 - \alpha)_l]^r \left[ \sum_{\sum_{i=1}^r N_i = j} [g_{m,r,l,j}(T_r)]^{S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j-r} \frac{S_{n_j}^{-1,\alpha}}{S_{n_j}^{S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j-r}}} \right]
\end{align*}
\]

Therefore, writing explicitly (11) yields

\[
\begin{align*}
\binom{j}{r} \frac{r!m!}{(l!)^r(m - rl)!} [(1 - \alpha)_l]^r & \sum_{\sum_{i=1}^r n_i = j} \binom{n}{s} S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j-r} \frac{S_{n_j}^{-1,\alpha}}{S_{n_j}^{S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j-r}}} [g_{m,r,l,j}(s)].
\end{align*}
\]

**Corollary 1.** For \(l = 0\) (3) yields

\[
(R_{0,m}^{(j,n)})_{r_i} = r! \sum_{s=r}^{n-j-r} \binom{n}{s} S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j} \frac{S_{n_j}^{-1,\alpha}}{S_{n_j}^{S_{n_j}^{-1,\alpha-l} \sum_{\sum_{i=1}^r n_i = j}}} [g_{m,r,l,j}(s)].
\]

Applying (7)

\[
(R_{m}^{(j,n)})_{r_i} = (j - R_{0,m}^{(j,n)})_{r_i} = (-1)^r \sum_{v=0}^r S_{r,v}^{-1,j} [(R_{0,m}^{(j,n)})_{v_i}]
\]

and by (6) and recalling that \((x)_s = (-1)^s(-x)_{s-1}\) equation (3.5) in Theorem 1. in Bacallado et al. (2013) rewritten in terms of generalized Stirling numbers is recovered by elementary combinatorics.

**Acknowledgements**

The author wishes to thank Lorenzo Trippa for kindly explaining the functioning of formulas (3.4), (3.5), (3.12) and (3.13) in Bacallado et al. (2013) under \(\alpha = 0\).
References

Bacallado, S., Favaro S. and Trippa, L. (2013) Looking-backward probabilities for Gibbs type exchangeable random partitions. Bernoulli (to appear).

Cerquetti, A. (2013) Marginals of multivariate Gibbs distributions with applications in Bayesian species sampling Elect. J. Stat.,7, 697–716.

Cerquetti, A. (2013b) A note on a Bayesian nonparametric estimator of the discovery probability. [arXiv:1304.1030] [math.ST]

Favaro, S., Lijoi, A. and Prünster, I. (2012) A new estimator of the discovery probability. Biometrics, 68, 1188-1196.

Favaro, S., Lijoi, A. and Prünster, I. (2013) Conditional formulae for Gibbs-type exchangeable random partitions. Ann. Appl. Probab. 23, 5, 1721–1754.

Gnedin, A. and Pitman, J. (2006) Exchangeable Gibbs partitions and Stirling triangles. Journal of Mathematical Sciences, 138, 3, 5674–5685.

Lijoi, A., Mena, R.H. and Prünster, I. (2007) Bayesian nonparametric estimation of the probability of discovering new species. Biometrika, 94, 769–786.

Lijoi, A., Prünster, I. and Walker, S.G. (2008) Bayesian nonparametric estimator derived from conditional Gibbs structures. Annals of Applied Probability, 18, 1519–1547.

Pitman, J. (2006) Combinatorial Stochastic Processes. Ecole d’Eté de Probabilité de Saint-Flour XXXII - 2002. Lecture Notes in Mathematics N. 1875, Springer.

Pitman, J. and Yor, M. (1997) The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann. Probab., 25, 855–900.