A generalized sequential construction of exchangeable Gibbs partitions with application

Annalisa Cerquetti

Collegio Nuovo, Pavia, Italy

September 15, 2009
Outline

• Motivation
  • BNP analysis of species sampling problems under Gibbs priors
  • Can be embedded in Pitman’s framework?

• Background
  • Exchangeable Gibbs Partitions (EGPs)
  • Central and non-central Generalized Stirling numbers
  • Chinese Restaurant Process constructions

• Contribution
  • Groups sequential CRP for Gibbs partitions
  • Embedding BNP results in the construction
Bayesian NP analysis of species sampling problems

- Li Joi et al. (2007) *Biometrika*, 94, 769-786

propose BNP analysis under Gibbs priors

- given a sample from $P$ a.s. discrete with atoms $(P_i) \sim PK(\rho_\alpha, \gamma)$
- $k$ distinct species observed with frequencies $(n_1, \ldots, n_k)$

they obtain conditional posterior/predictive results for an additional sample w.r.t.

- the number of new distinct species
- the number of new observations belonging to new species
- the random partition induced by the additional sample
My question was:

Is it possible to embed this analysis in the typical Pitman’s combinatorial framework for exchangeable partitions theory?
Exchangeable random partitions. [Kingman, 1975]

- a sample \((X_1, \ldots, X_n, \ldots)\) from a.s. discrete p.m. \(P\) induces a random partition \(\Pi = \{A_1, \ldots, A_k, \ldots\}\) of \(N\) by

\[ i \approx l \iff X_i = X_l \]

- for each restriction to \([n]\), for \(n_j = |A_j|\),

\[ Pr(\Pi_n = \{A_1, \ldots, A_k\}) = p(n_1, \ldots, n_k) \]

for some symmetric \(p\) called the exchangeable partition probability function (EPPF) of \(\Pi\).
A large class of models for ERP (priors for BNP) is obtained by generalizing infinite sum construction of the Dirichlet process as follows:

- for \( J_1 \geq J_2 \geq \cdots \geq 0 \) ranked points of a Poisson prox on \((0, \infty)\) with general mean intensity \( \rho(x) \), (with some constraints)

- \((P_i) = (J_i/T) \sim \text{Poisson-Kingman} (\rho)\) on \( \mathcal{P}_1 \)

and even a larger class of mixed \( PK(\rho, \gamma) \) models is obtained as

\[
PK(\rho, \gamma) := \int_0^\infty PK(\rho | t) \gamma(dt)
\]

for general mixing density \( \gamma \).
EPPFs in Gibbs product form [Gnedin and Pitman, 2005]

→ an interesting sub-class of $PK(\rho, \gamma)$ for BNP arise for $\rho = \rho_\alpha$ the Levy density of the $\alpha$-stable law for $\alpha \in (0, 1)$

→ $PK(\rho_\alpha, \gamma)$ models are characterized by inducing EPPF in Gibbs form of type $\alpha$ i.e.

$$p(n_1, \ldots, n_k) = V_{n,k} \prod_{j=1}^{k} (1 - \alpha)^{n_j - 1}.$$ 

→ consistency conditions for $\Pi$ to be infinite imply the weights $(V_{n,k})$ must satisfy the backward recursion

$$V_{n,k} = (n - \alpha k) V_{n+1,k} + V_{n+1,k+1}$$
Some known combinatorial formulas

For \( n = 0, 1, 2, \ldots \), and arbitrary real \( x \) and \( h \), \( (x)_{n \uparrow h} \) denotes generalized rising factorial

\[
(x)_{n \uparrow h} := \prod_{i=0}^{n-1} (x + ih) = h^n (x/h)_{n \uparrow}
\]

for which the following multiplicative law holds

\[
(x)_{n+r \uparrow h} = (x)_{n \uparrow h} (x + nh)_{r \uparrow h}
\]  \hspace{1cm} (1)

A binomial formula also holds

\[
(x + y)_{n \uparrow h} = \sum_{k=0}^{n} \binom{n}{k} (x)_{k \uparrow h} (y)_{n-k \uparrow h}
\]
as well as a generalized version of the multinomial theorem,

\[(\sum_{j=1}^{p} z_j)^n \equiv h = \sum_{n_j \geq 0, \sum n_j = n} \frac{n!}{n_1! \cdots n_p!} \prod_{j=1}^{p} (z_j)^{n_j} \equiv h. \quad (2)\]

An application of (1) yields

\[(z_j)^{n_j + m_j - 1} = (z_j)^{m_j - 1} (z_j + m_j - 1)^{n_j} \quad (3)\]

and by (2)

\[\sum_{n_j \geq 0, \sum n_j = n} \frac{n!}{n_1! \cdots n_p!} \prod_{j=1}^{p} (z_j)^{n_j + m_j - 1} = \prod_{j=1}^{p} (z_j)^{m_j - 1} \left( \sum_{j=1}^{p} (z_j + m_j - 1) \right)^{n_j} = \]

\[= \prod_{j=1}^{p} (z_j)^{m_j - 1} \left( m + \sum_{j=1}^{p} z_j - p \right)^{n_j}. \]

which agrees with a Lemma in Lijoi et al. (2008)
Partial Bell polynomials and generalized Stirling numbers

The number of ways to partition \([n]\) into \(k\) blocks and assign each block a combinatorial structure,

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

(4)

Generalized Stirling numbers arise as \(B_{n,k}(w_\bullet)\) for particular \(w_\bullet\),

\[
S_{n,k}^{\eta,\beta} = B_{n,k}((\beta - \eta)_\bullet - 1\downarrow \eta),
\]

where \((x)_{n\downarrow - h} = (x)_{n\uparrow h}\). In general for arbitrary distinct reals \(\eta\) and \(\beta\), are the connection coefficients defined by

\[
(x)_{n\downarrow \eta} = \sum_{k=0}^n S_{n,k}^{\eta,\beta} (x)_{k\downarrow \beta}
\]
Hence for $\eta = -1$, $\beta = -\alpha$, and $\alpha \in (-\infty, 1)$, $S_{n,k}^{-1,-\alpha}$ is defined by

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

and for $w_{n_i} = (1 - \alpha)_{n_i - 1 \uparrow}$ and $\alpha \in [0, 1)$, equation (4) yields

$$
B_{n,k}((1 - \alpha)_{n_i - 1 \uparrow}) = \frac{n!}{k!} \sum_{(n_1, \ldots, n_k)} \prod_{i=1}^{k} \frac{(1 - \alpha)_{n_i - 1 \uparrow}}{n_i!} = S_{n,k}^{-1,-\alpha}.
$$

i.e. the partial Bell we encounter with Gibbs EPPF of type $\alpha$. 
Generalized Stirling vs Generalized Factorial

In Lijoi et al. (2007, 2008) the treatment is in terms of generalized factorial coefficients, \( C_{n,k}^{\alpha} \) defined by

\[
(\alpha y)^{n\uparrow 1} = \sum_{k=0}^{n} C_{n,k}^{\alpha} (y)^{k\uparrow 1}.
\]

From (5), if \( x = y\alpha \) then

\[
(y\alpha)^{n\uparrow 1} = \sum_{k=0}^{n} S_{n,k}^{-1,-\alpha} (y\alpha)^{k\uparrow \alpha} = \sum_{k=0}^{n} S_{n,k}^{-1,-\alpha} \alpha^{k} (y)^{k\uparrow 1},
\]

hence

\[
S_{n,k}^{-1,-\alpha} = C_{n,k}^{\alpha} / \alpha^{k}.
\]
non-central Generalized Stirling numbers [Hsu and Shiue (1998)]

Moreover the following convolution relation holds,

\[ S_{n,k}^{-1,-\alpha,\gamma} = \sum_{s=k}^{n} \binom{n}{s} S_{s,k}^{-1,-\alpha} (-\gamma)^{n-s+1}, \quad (7) \]

hence

\[ C_{n,k}^{\alpha,\gamma} = \alpha^k S_{n,k}^{-1,-\alpha,\gamma} = \sum_{s=k}^{n} \binom{n}{s} C_{s,k}^{\alpha} (-\gamma)^{n-s+1}, \]

and non-central generalized Stirling numbers are the connection coefficients \( S_{n,k}^{-1,-\alpha,\gamma} \) such that

\[ (y\alpha - \gamma)_{n}^{1} = \sum_{k=0}^{n} S_{n,k}^{-1,-\alpha,\gamma} \alpha^k (y)_{k}^{1} = \sum_{k=0}^{n} S_{n,k}^{-1,-\alpha,\gamma} (y\alpha)_{k}^{\alpha}. \quad (8) \]
Sequential CRP construction of exchangeable partitions

Given an infinite EPPF, \( p(n) = p(n_1, \ldots, n_k) \), assume that

- an unlimited number of customers arrives sequentially in a restaurant
- for \( n \geq 1 \), given \((n_1, \ldots, n_k)\), the placement of the first \( n \) customers at \( k \) tables, the \((n + 1)th\) customer is:
  
  → seated at table \( j \), for \( 1 \leq j \leq k_n \), with probability

  \[
p_{j,n} = p_j(n) = \frac{p(n^{j+})}{p(n)}
  \]

  → seated at a new table with probability

  \[
p_{0,n} = p_{0,n}(n) = \frac{p(n^{l+})}{p(n)}
  \]

  for \( l = k_n + 1 \), \( \sum_{j=1}^{k+1} p_{j,n} + p_{0,n} = 1 \), \( p(n^{j+}) = p(n_1, \ldots, n_j + 1, \ldots, n_k) \)
A groups sequential variation of the CRP [Cerquetti, 2008]

Given an infinite EPPF, $p(n) = p(n_1, \ldots, n_k)$, assume that

- an unlimited number of groups of customers arrives sequentially in a restaurant
- for $n \geq 1$, given the placement of the first group of $n$ customers in a $(n_1, \ldots, n_k)$ configuration in $k$ tables, the new group of $m$ customers is:
→ all seated at the old \( k \) tables in configuration \((m_1, \ldots, m_k)\), with probability

\[
p(m|n) = \frac{p(n_1 + m_1, \ldots, n_k + m_k)}{p(n_1, \ldots, n_k)},
\]

(9)

→ all seated at \( k^* \) new tables in configuration \((s_1, \ldots, s_k^*)\) with probability

\[
p(s|n) = \frac{p(n_1, \ldots, n_k, s_1, \ldots, s_k^*)}{p(n_1, \ldots, n_k)},
\]

(10)

→ \( s < m \) are seated at \( k^* \) new tables in configuration \((s_1, \ldots, s_k^*)\) and the remaining \( m - s \) customers at the \( k \) old tables in configuration \((m_1, \ldots, m_k)\) with probability

\[
p(m, s|n) = \frac{p(n_1 + m_1, \ldots, n_k + m_k, s_1, \ldots, s_k^*)}{p(n_1, \ldots, n_k)}.
\]

(11)
Groups sequential rules for EPPFs in Gibbs form

For

\[ p(n_1, \ldots, n_k) = V_{n,k} \prod_{j=1}^{k} (1 - \alpha)_{n_j-1} \]

formulas (9), (10) and (11) specialize as follows,

\[ p(m|n) = \frac{V_{n+m,k}}{V_{n,k}} \prod_{j=1}^{k} (n_j - \alpha)_{m_j} \]  \hspace{1cm} (12)

\[ p(s|n) = \frac{V_{n+m,k+k^*}}{V_{n,k}} \prod_{j=1}^{k^*} (1 - \alpha)_{s_j-1} \]  \hspace{1cm} (13)

\[ p(m, s|n) = \frac{V_{n+m,k+k^*}}{V_{n,k}} \prod_{j=1}^{k} (n_j - \alpha)_{m_j} \prod_{j=1}^{k^*} (1 - \alpha)_{s_j-1}. \]  \hspace{1cm} (14)
Embedding Lijoi et al.’s BNP analysis - some examples

The distribution of the random partition \((s_1, \ldots, s_k^*)\) induced by the additional sample [Lijoi et al. 2008, Prop. 1] arises marginalizing (14) with respect to \((m_1, \ldots, m_k)\):

\[
p(s_1, \ldots, s_k^* | n_1, \ldots, n_k) = \]

\[
\frac{V_{n+m,k+k^*}}{V_{n,k}} \binom{m}{m-s} \sum_{(m_1, \ldots, m_k)} \binom{m-s}{m_1, \ldots, m_k} \prod_{j=1}^{k} (n_j - \alpha)_m \uparrow \prod_{j=1}^{k^*} (1-\alpha)_{s_j-1} \uparrow
\]

\[
\sum_j m_j = m-s, m_j \geq 0
\]

and then exploiting the generalized version of the multinomial theorem

\[
= \frac{V_{n+m,k+k^*}}{V_{n,k}} \binom{m}{m-s} (n-k\alpha)_m \uparrow \prod_{j=1}^{k^*} (1-\alpha)_{s_j-1} \uparrow
\]

(15)
Embedding BNP analysis (continued)

The number of new observations \( s = \sum s_j \) belonging to new species [Lijoi et al. 2008, Eq. 11] arises by (15) by summing over the space of all partitions of \( s \) elements in \( k^* \) blocks, and then marginalizing with respect to \( k^* \), and resorting to the definition of Generalized Stirling numbers,

\[
Pr(S = s | n_1, \ldots, n_k) = \frac{1}{V_{n,k}} \binom{m}{s} (n - k\alpha)^{m-s} \sum_{k^* = 0}^{s} V_{n+m,k+k^*} S^{-1,-\alpha}_{s,k^*}.
\]

(16)
Embedding BNP analysis (continued)

The distribution of the number $k^*$ of new species in the additional sample [Lijoi et al. 2007, Eq. 4] arises marginalizing the joint conditional distribution of $S$ and $K^*$ with respect to $S$, and by the convolution relation defining non-central generalized Stirling numbers,

$$
Pr(K^* = k^* | n_1, \ldots, n_k) = \frac{V_{n+m,k+k^*}}{V_{n,k}} S_{m,k^*}^{-1,-\alpha,-(n-k\alpha)}.
$$

(17)

N.B. Some additional examples and a connection established between the notion of conditional Gibbs structures of Lijoi et al. (2008) and the deletion of classes property of Pitman (2003) may be found on the full paper on the ArXiv repository.
Is it possible to embed Lijoi et al.’s analysis in the typical Pitman’s combinatorial framework of exchangeable partitions theory?

Yes, it is

Is that interesting and worth a submission/a paper?

I was not sure, so I checked Prof. Pitman’s opinion

According to Prof. Jim Pitman (May 2008) it

• provides some simplifications of Lijoi et al.’s work,
• connects it better to the Gneden-Pitman development,
• provides a better platform for future work,

hence it deserves a series of submission attempts. (still ongoing...


