Conditional $\alpha$-diversity for exchangeable Gibbs partitions driven by the stable subordinator

Annalisa Cerquetti

Dept. MeMoTEF - Sapienza University of Rome, Italy

7th Conference on Statistical Computation and Complex Systems

Padova, September 20th
Outline

- Species sampling and exchangeable Gibbs partitions
  - species sampling sequences and models
  - \((\rho_\alpha, \gamma)\) Poisson-Kingman partitions
  - \((\alpha, \theta)\) Poisson-Dirichlet and generalized Gamma partitions
- BNP approach to species sampling problems
  - finite sample posterior species richness under \(PK(\rho_\alpha, \gamma)\) priors
  - asymptotics under \(PD(\alpha, \theta)\) and GG priors
- Contribution
  - asymptotics for species richness under \(PK(\rho_\alpha, \gamma)\) models
  - a collateral result
Species sampling sequences and models [Pitman, 1996]

If \((X_n)\) is an infinite exchangeable sequence of labels/species with values in \(X\), such that, for \(H\) a non atomic distribution,

\[
\mathbb{P}(X_{n+1} \in \cdot | X_1, \ldots, X_n) = \sum_{j=1}^{k} p_{j,n}(n) \delta_{X_j^*}(\cdot) + q_n(n) H(\cdot)
\]

for \(n = (n_1, \ldots, n_k)\) the partition of \([n]\) induced by \((X_1^*, \ldots, X_k^*)\), the distinct values in \((X_1, \ldots, X_n)\), and

\[
p_{j,n}(n) = \text{prob } j\text{-th species}, \quad q_n(n) = \text{prob new species}
\]

then there exists an infinite sequence of unknown species proportions \((P_n)\) whose law is in one-to-one correspondence with...
a consistent symmetric law \( p \) on partitions of \( \mathbb{N} \)

\[
\mathcal{L}(P_1, P_2, \ldots) \iff p(n_1, \ldots, n_k, \ldots)
\]

called the exchangeable partition probability function (EPPF), such that the directing (de Finetti) measure of \((X_n)\) is the law of the a.s. discrete random \( P \) representable as

\[
P(\cdot) = \sum_{i=1}^{\infty} P_i \delta_{\hat{X}_i}(\cdot)
\]

for \( \hat{X}_i \) iid \( \sim H \), independent of the \((P_i)\), and

\[
p_{j,n}(n) = \frac{p(n_1, \ldots, n_j + 1, \ldots n_k)}{p(n)} \quad q_n(n) = \frac{p(n_1, \ldots, n_k, 1)}{p(n)}
\]

Ex. For \( P \sim \text{Dir}(\theta, H) \), ranked \((P_i)\) are Poisson-Dirichlet (\( \theta \)) (Kingman, 1975)
Exchangeable $\alpha$-Gibbs partitions [Gnedin & Pitman, 2006]

Gnedin and Pitman (2006) describe a convex class of EPPFs in *Gibbs product form of type $\alpha$* i.e.

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

as mixtures of extreme partitions, for $\alpha \in (-\infty, 1)$ and weights satisfying $V_{n,k} = (n - k\alpha)V_{n+1,k} + V_{n+1,k+1}$

In terms of distributions on the ranked atoms ($P_i$) of $P$ those correspond

- for $\alpha \in (-\infty, 0)$, to mixtures over $\xi = 1, 2, 3, \ldots$, of Poisson-Dirichlet ($\alpha, \xi|\alpha|$) (Fisher, 1943) models

- for $\alpha = 0$, to mixtures over $\theta$ of Poisson-Dirichlet ($\theta$) models (from Fisher’s model for $\xi \to \infty$, $\alpha \to 0$, $\xi|\alpha| = \theta$)
Poisson-Kingman \((\rho_\alpha, \gamma)\) models [Pitman, 2003]

For \(\alpha \in (0, 1)\) Gibbs EPPFs are mixtures of conditional \((\rho_\alpha)\) Poisson-Kingman models, i.e.

- for \(J_1 \geq J_2 \geq \cdots \geq 0\) the ranked points of a Poisson prox on \((0, \infty)\) with mean intensity \(\rho_\alpha(x) = \alpha x^{-\alpha-1}[\Gamma(1 - \alpha)]^{-1}\) the Lévy density of the stable subordinator, and \(T = \sum_i J_i\)

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

then for \(\text{PK}(\rho_\alpha|t)\) the law of \((P_i)|(T = t)\)

\[
\text{PK}(\rho_\alpha, \gamma) := \int_0^\infty \text{PK}(\rho_\alpha|t) \gamma(dt)
\]

for \(\gamma(t) = h(t)f_\alpha(t)\) a general mixing density.
Three notable classes in the $PK(\rho_\alpha, \gamma)$ class

$\rightarrow$ for $\alpha \in (0, 1)$, $\theta > -\alpha$, and $\gamma_{\alpha, \theta}(t) = \frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)} t^{-\theta} f_\alpha(t)$

$PK(\rho_\alpha, \gamma_{\alpha, \theta}) = PD(\alpha, \theta)$, two-parameter PD models [Pitman & Yor, 1997]

$\rightarrow$ for $\alpha \in (0, 1)$, $\psi_{\alpha}(t) = (2\lambda)^\alpha$, and $\gamma_{\alpha, \lambda}(t) = \exp\{\psi_{\alpha}(\lambda) - \lambda t\} f_\alpha(t)$

$PK(\rho_\alpha, \gamma_{\alpha, \lambda}) = GG(\alpha, \lambda)$, generalized Gamma models [Pitman, 2003]

$\rightarrow$ for $\alpha = 1/2$, $\gamma_{1/2, \lambda}(t) = \exp\{\psi_{1/2}(\lambda) - \lambda t\} f_{1/2}(t)$

$PK(\rho_{1/2}, \gamma_{1/2, \lambda}) = IG(1/2, \lambda)$, inverse Gaussian models [Pitman, 2003]

All are operations of \textit{tilting} (change of measure), \textit{polynomial} and \textit{exponential} tilting
• GG($\alpha$, $\lambda$) and $IG(1/2, \lambda)$ models have been exploited in BNP to build alternatives *priors* to the Dirichlet process and in BNP *hierarchical mixtures modeling* [Lijoi et al. 2005, 2007a].

• recently a *BNP approach to species sampling problems* under $\alpha$-*Gibbs priors* has been devised in Lijoi et al. (2007b, 2008), further results are in Favaro et al. (2009, 2011).

• Here I just give a quick overview to locate my results.

→ Don’t forget BNP for SSP will be the topic of tomorrow’s plenary lecture, h 11:30, I. Prünster.
BNP approach to SSP [Lijoi et al., 2007, 2008]

In a population of different species, both the number and the kind of the different species is unknown. After \( n \) observations

- \( (x_1, \ldots, x_n) \), the vector of the species labels observed
- \( k_n \) distinct species observed with frequencies \( (n_1, \ldots, n_{k_n}) \),

Interest is on conditional posterior/predictive results for an additional \( m \)-sample \((X_{n+1}, \ldots, X_{n+m})\) w.r.t.

- the number \( K_m \) of new species observed (species richness)
- the asymptotic behaviour of posterior species richness.

Choosing a BNP prior corresponds to choose an EPPF for \((n_1, \ldots, n_k)\). A convenient choice is a \( p \) in the \( \alpha \) Gibbs class. (mathematical tractability).
Write a multi-step prediction rule for a general EPPF as

$$p_{s,m}(n) = \frac{p(s_1, \ldots, s_k^*, n_1 + m_1, \ldots, n_k + m_k)}{p(n_1, \ldots, n_k)}$$

for $(s_1, \ldots, s_k^*)$ the allocation of $s \leq m$ observations in new species, and $(m_1, \ldots, m_k)$ the allocation of $m - s$ in old species, and specialize for Gibbs partitions of type $\alpha \in (-\infty, 1)$ as

$$p_{s,m}(n) = \frac{V_{n+m,k+k^*}}{V_{n,k}} \prod_{j=1}^k (n_j - \alpha)^{m_j} \prod_{j=1}^{k^*} (1 - \alpha)^{s_j - 1}$$

then by marginalization the conditional EPPF corresponds to

$$p(s_1, \ldots, s_k^* | n_1, \ldots, n_k) = \frac{V_{n+m,k+k^*}}{V_{n,k}} \binom{m}{s} (n - k\alpha)^{m-s} \prod_{j=1}^{k^*} (1 - \alpha)^{s-j-1}.$$
Posterior species richness: $\alpha$ Gibbs EPPF [Lijoi et al., 2007]

Some combinatorial calculus plus the convolution definition of non-central generalized Stirling numbers $S_{m,k^*}^{-1,-\alpha,-(n-k\alpha)}$, yield

$$
\mathbb{P}_\alpha(K_m = k^* | K_n = k) = \frac{V_{n+m,k+k^*}}{V_{n,k}} S_{m,k^*}^{-1,-\alpha,-(n-k\alpha)},
$$

which agrees with the result firstly obtained in Lijoi et al. (2007).

By the need to obtain HPD intervals for point estimates of $K_m$, and involved computational burden, interest arises (Favaro et al. 2009) in the asymptotic behaviour, for $m \to \infty$, of

$$
\left( \frac{K_m}{m^{\alpha}} | K_n = k \right).
$$

In Pitman’s language this is the conditional $\alpha$ diversity of a $PK(\rho_\alpha, \gamma)$ model.
Asymptotics for conditional species richness: \( PD(\alpha, \theta) \)

By adopting the same technique in Pitman’s proof of the unconditional result, Favaro et al. (2009) show that a.s., for \( m \to \infty \),

\[
\left( \frac{K_m}{m^\alpha} \middle| K_n = k \right) \overset{a.s.}{\to} Z_{n,k}^{\alpha,\theta} \overset{d}{=} Y_{(\theta+n)/\alpha} \ast X
\]

for \( X \sim \text{Beta}((\theta/\alpha + k, n/\alpha - k)) \), \( Y_\beta \sim f_{Y_\beta} = \frac{\Gamma(\beta \alpha+1)}{\Gamma(\beta+1)\alpha} y^{\beta-1/\alpha-1} f_\alpha(y^{-1/\alpha}) \)

A different argument (Cerquetti, 2011), exploiting some known facts about \( PD(\alpha, \theta) \) models, yields a different scale mixture representation

\[
\left( \frac{K_m}{m^\alpha} \middle| K_n = k \right) \overset{a.s.}{\to} \tilde{Z}_{n,k}^{\alpha,\theta} \overset{d}{=} Y_{(\theta+k\alpha)/\alpha} \ast W^\alpha
\]

for \( W \sim \text{Beta}(\theta + k\alpha, n - k\alpha) \), but the two results agree.
Asymptotics for general $\alpha$ Gibbs partitions?

- Conditional $\alpha$ diversity under N-GG priors (PK models obtained by exponential tilting of the stable density) have been derived in Favaro et al. (2011) by means of the same technique adopted in Favaro et al. (2009).

- In the same paper the possibility to obtain a general result for the entire $PK(\rho_\alpha, \gamma)$ class is conjectured based on a similar behaviour of unconditional and conditional $\alpha$ diversity with respect to the change of measure $h(t)$ specified by the mixing $\gamma(t)$.

- A step back to Pitman’s unconditional $\alpha$ diversity result...
unconditional $\alpha$-diversity [Pitman, 2003]

For $(P_i) \sim PK(\rho_\alpha, f_\alpha) = PK(\rho_\alpha)$ then

$$\frac{K_n}{n^\alpha} \overset{a.s.}{\to} S = T^{-\alpha}$$

for $T \sim f_\alpha(\cdot)$. For a general mixed $PK(\rho_\alpha, \gamma)$ model where, (without loss of generality) $\gamma_{\alpha,h}(t) = h(t)f_\alpha(t)$ on $(0, \infty)$ then

$$\frac{K_n}{n^\alpha} \overset{a.s.}{\to} S_h = T_h^{-\alpha}$$

for $T_h \sim \gamma_{\alpha,h}(t) = h(t)f_\alpha(t)$.

So, on the unconditional limit, the same change of measure $h(t)$ applies which identifies the specific partition model.
This implies that if we are able to find the \textit{conditional} limit \( S_{n,k}^\alpha \) such that, for \( PK(\rho_\alpha, f_\alpha) \),

\[
\left( \frac{K_m}{m^\alpha} \middle| K_n = k \right) \xrightarrow{a.s.} S_{n,k}^\alpha \sim g_{n,k}^\alpha
\]

then we can apply the \textit{same change of measure} to the conditional limit distribution and state that, for \( PK(\rho_\alpha, h \ast f_\alpha) \),

\[
\left( \frac{K_m}{m^\alpha} \middle| K_n = k \right) \xrightarrow{a.s.} S_{n,k}^{\alpha,h}
\]

for \( S_{n,k}^{\alpha,h} \sim \tilde{g}_{n,k}^{\alpha,h}(s) = C^{-1} h(s^{-1/\alpha})g_{n,k}^\alpha(s) \) and \( C \) a normalizing constant.
Notice that

\[ PK(\rho_\alpha, f_\alpha) = PD(\alpha, 0) \]

then, by the result in Lijoi et al. (2009) (using Cerquetti’s scale mixture)

\[ \left( \frac{K_m}{m^{\alpha}} \, | \, K_n = k \right) \overset{a.s.}{\rightarrow} S_{n,k}^{\alpha} \]

for \( S_{n,k}^{\alpha} \overset{d}{=} Y_{\alpha,k} \ast W^{\alpha} \) where \( Y_{\alpha,k} \) has density

\[ g_{\alpha,k\alpha}(y) = \frac{\Gamma(k\alpha + 1)}{\Gamma(k + 1)} y^k g_{\alpha}(y) \]

for \( g_{\alpha}(y) = \alpha^{-1} y^{-1-1/\alpha} f_\alpha(y^{-1/\alpha}) \), and \( W \sim \beta(k\alpha, n - k\alpha) \).

But it seems the result for \( PD(\alpha, \theta) \) model is not necessary to obtain the general result. We can resort to Bayes’ rule...
and write the law of $S_{\alpha, \gamma} | K_n = k$ for a general $PK(\rho, \gamma)$ model/prior as

$$f_{S_{\alpha, \gamma}}(s|k_n) = f_{S_{\alpha, \gamma}}(s|n_1, \ldots, n_k) = \frac{p_\alpha(n_1, \ldots, n_k | s^{-1/\alpha}) \gamma(s^{-1/\alpha})}{\int_0^\infty p_\alpha(n_1, \ldots, n_k | s^{-1/\alpha}) \gamma(s^{-1/\alpha}) ds}$$

for $\gamma(s^{-1/\alpha}) = h(s^{-1/\alpha}) f_\alpha(s^{-1/\alpha}) \alpha^{-1} s^{-1/\alpha - 1}$.

By Pitman (2003) the general conditional EPPF for a $PK(\rho, \gamma)$ model is given by

$$p_\alpha(n_1, \ldots, n_k | s^{-1/\alpha}) =$$

$$= \frac{\alpha^k s^k}{\Gamma(n - k\alpha)} [f_\alpha(s^{-1/\alpha})]^{-1} \int_0^1 p^{n-1-k\alpha} f_\alpha((1 - p)s^{-1/\alpha}) dp \prod_{j=1}^k (1 - \alpha) n_{j-1},$$

which yields
Conditional $\alpha$ diversity for $PK(\rho_\alpha, \gamma)$

A collateral result

\[ f_{S_{\alpha,h}}(s|K_n = k) = \frac{h(s^{-1/\alpha})s^{k-1/\alpha-1} \int_0^1 p^{n-1-k\alpha} f_\alpha((1 - p)s^{-1/\alpha}) dp}{\int_0^\infty h(s^{-1/\alpha})s^{k-1/\alpha-1} \left[ \int_0^1 p^{n-1-k\alpha} f_\alpha((1 - p)s^{-1/\alpha}) dp \right] ds}, \]

in compact form

\[ f_{n,k}^{h,\alpha}(s) = \frac{h(s^{-1/\alpha})\tilde{g}_{n,k}^\alpha(s)}{\mathbb{E}_{n,k}^{\alpha}[h(S^{-1/\alpha})]} \] \hspace{1cm} (2)

for

\[ \tilde{g}_{n,k}^\alpha(s) = \frac{\Gamma(n)}{\Gamma(n - k\alpha)\Gamma(k)} s^{k-1/\alpha-1} \int_0^1 p^{n-1-k\alpha} f_\alpha((1 - p)s^{-1/\alpha}) dp \]

which is in fact the density of the scale mixture $Y_{\alpha,k} \times [W]^\alpha$.

The normalizing constant may be obtained through the known result

\[ \mathbb{E}_{n,k}^{\alpha}[h(S^{-1/\alpha})] = V_{n,k,h} \frac{\alpha^{1-k} \Gamma(n)}{\Gamma(k)}. \] \hspace{1cm} (3)
the number of species represented \( j \) times

For \( K_{n,j} \) the \textit{number of species represented \( j \) times}, \( \sum_j K_{n,j} = K_n \), from Pitman (2006) \( S_\alpha = T^{-\alpha} \sim \gamma(t) \) is even the unconditional limit in distribution for

\[
\frac{K_{n,j}}{n^\alpha\alpha(1-\alpha)_{j-1}}\frac{j!}{\alpha(1-\alpha)_{j-1}}.
\]

It follows that the general result for the conditional \( \alpha \) diversity may provide the following additional result for general \( PK(\alpha, h \ast f_\alpha) \) models

\[
\left\{ \frac{K_{m,j}}{m^\alpha}\middle| K_n = k \right\} \overset{d}{\rightarrow} \frac{\alpha(1-\alpha)_{j-1}}{j!} S_{\alpha,h}^{\alpha,h}
\]

which in fact agrees with the result stated \textit{under PD}(\( \alpha, \theta \)) models in the tomorrow’s plenary session paper by Favaro et al.
Conditional $\alpha$-diversity for $PK(\rho_\alpha, \gamma)$

A collateral result

Selected references


