Conditional α-diversity for exchangeable Gibbs partitions driven by the stable subordinator

Annalisa Cerquetti

Dept. MeMoTEF - Sapienza University of Rome, Italy

SCo 2011
7th Conference on Statistical Computation and Complex Systems

Padova, September 20th

Outline

- Species sampling and exchangeable Gibbs partitions
  - species sampling sequences and models
  - (ρ_α, γ) Poisson-Kingman partitions
  - (α, θ) 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 $\mathcal{X}$, such that, for H a *non atomic* distribution,

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

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

$$p_{j,n}(\mathbf{n}) = \textit{prob j-th species}, \qquad q_n(\mathbf{n}) = \textit{prob new species}$$

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 $\mathbb{N}$

$$\mathcal{L}(P_1, P_2, \dots) \Leftrightarrow p(n_1, \dots, n_k, \dots)$$

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

$$p_{j,n}(\mathbf{n}) = \frac{p(n_1, \dots, n_j + 1, \dots n_k)}{p(\mathbf{n})} \qquad q_n(\mathbf{n}) = \frac{p(n_1, \dots, n_k, 1)}{p(\mathbf{n})}$$

Ex. For $P \sim Dir(\theta, H)$, ranked (P_i) are *Poisson-Dirichlet* (θ) (Kingman, 1975)

Exchangeable α-Gibbs partitions [Gnedin & Pitman, 2006]

Gnedin and Pitman (2006) describe a convex class of EPPFs in *Gibbs product form of type α* 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 θ of Poisson-Dirichlet (θ) models (from Fisher's model for $\xi \to \infty$, $\alpha \to 0$, $\xi|\alpha| = \theta$)

Poisson-Kingman (ρ_α, γ) models [Pitman, 2003]

For $\alpha \in (0,1)$ Gibbs EPPFs are mixtures of conditional (ρ_α) 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 $T = \sum_i J_i$
- $(P_i) = (J_i/T) \sim$ Poisson-Kingman (ρ_α) on $\mathcal{P}_1^\downarrow$

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

$$PK(\rho_\alpha, \gamma) := \int_0^\infty PK(\rho_\alpha|t)\gamma(dt)$$

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 *tilting* (change of measure), *polynomial* and *exponential* tilting

- *GG*(α, λ) 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 α*-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.

$\rightarrow$ 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 α Gibbs class. (mathematical tractability).

BNP analysis embedded in Pitman's theory [Cerquetti, 2009]

Write a *multi-step prediction rule* for a general EPPF as

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

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

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

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: α 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

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

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

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

In Pitman's language this is the *conditional α 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 $m \to \infty$,

$$\left(\frac{K_m}{m^\alpha} \middle| K_n = k \right) \xrightarrow{a.s.} Z_{n,k}^{\alpha,\theta} \stackrel{d}{=} Y_{(\theta+n)/\alpha} * X$$

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

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

$$\left(\frac{K_m}{m^\alpha} \middle| K_n = k \right) \xrightarrow{a.s.} \tilde{Z}_{n,k}^{\alpha,\theta} \stackrel{d}{=} Y_{(\theta+k\alpha)/\alpha} * W^\alpha$$

for $W \sim Beta(\theta + k\alpha, n - k\alpha)$, but the two results agree.

Asymptotics for general α Gibbs partitions?

- Conditional α 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* α diversity with respect to the change of measure $h(t)$ specified by the mixing $\gamma(t)$.
- A step back to Pitman's *unconditional α diversity* result...

unconditional α-diversity [Pitman, 2003]

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

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

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

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

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 *conditional* limit $S^{\alpha}_{n,k}$ such that, for $PK(\rho_\alpha, f_a)$,

$$\left(\frac{K_m}{m^{\alpha}}\middle| K_n = k\right) \xrightarrow{a.s.} S^{\alpha}_{n,k} \sim g^{\alpha}_{n,k}$$

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

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

for $S^{\alpha,h}_{n,k} \sim \tilde{g}^{\alpha,h}_{n,k}(s) = C^{-1} h(s^{-1/\alpha}) g^{\alpha}_{n,k}(s)$ 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)

$$\left(\frac{K_m}{m^\alpha}\middle| K_n = k\right) \xrightarrow{a.s} S^\alpha_{n,k}$$

for $S^\alpha_{n,k} \stackrel{d}{=} Y_{\alpha,k} * W^\alpha$ 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 $PK(\rho_\alpha, \gamma)$ 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 $PK(\rho_\alpha, \gamma)$ 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

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

in compact form

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

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

$$\mathbb{E}_{n,k}^{\alpha}[h(S^{-1/\alpha})] = V_{n,k,h}\frac{\alpha^{1-k}\Gamma(n)}{\Gamma(k)}. \qquad (3)$$

the number of species represented j times

For $K_{n,j}$ the *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

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

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

$$\left(\frac{K_{m,j}}{m^\alpha}\middle| K_n = k\right) \xrightarrow{d} \frac{\alpha(1-\alpha)_{j-1}}{j!} S_{n,k}^{\alpha,h}$$

which in fact agrees with the result stated *under $PD(\alpha, \theta)$* models in the tomorrow's plenary session paper by Favaro et al.

Selected references

CERQUETTI, A. (2009) A generalized sequential construction of exchangeable Gibbs partitions with application. *Proceedings of S.Co. 2009, Milano, Italy*.

CERQUETTI, A. (2011) On some Bayesian nonparametric estimators for species richness under two-parameter Poisson-Dirichlet priors. *Proceedings of ASMDA - Rome, June, 2011*. (arxiv)

FAVARO, S. LIJOI, AND PRÜNSTER, I. (2011) Asymptotics for a Bayesian nonparametric estimator of species richness. *Bernoulli*, to appear.

FAVARO, S. LIJOI, A., MENA, R. PRÜNSTER, I. (2009) Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *JRSS-B*, 71, 993-1008

GNEDIN, A. AND PITMAN, J. (2006) Exchangeable Gibbs partitions and Stirling triangles. *J. Math. 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) BNP estimators derived from conditional Gibbs structures. *Ann. Appl. Prob.* 18, 1519-1547

PITMAN, J. (2003) Poisson-Kingman partitions. IMS Lecture Notes n. 40.