On the role of normalized inverse-Gaussian priors in continuous-time models

Matteo Ruggiero

Abstract This short paper discusses the role that normalized inverse-Gaussian priors assume in certain continuous-time models. These describe the time evolution of infinitely-many frequencies, together with a measure of their heterogeneity, based on an infinite normalized inverse-Gaussian sample, whose individuals are subject to random genetic drift and mutation in a randomly evolving environment.

Key words: Alpha-diversity, generalized gamma processes, Gibbs partitions, mutation, normalized inverse-Gaussian priors.

1 Normalized inverse-Gaussian priors

Normalized inverse-Gaussian priors are special cases of normalized generalized gamma processes. These are discrete random probability measures with representation

$$\mu = \sum_{i=1}^{\infty} P_i \delta_{X_i} \tag{1}$$

where the locations $\{X_i\}_{i\geq 1}$ are iid samples from a non atomic probability measure P_0 defined on some Polish space $\mathbb{X}$, the weights $\{P_i, i \in \mathbb{N}\}$ are obtained by means of the normalization

$$P_i = J_i \Big/ \sum_{k=1}^{\infty} J_k, \tag{2}$$

and $\{J_i, i \in \mathbb{N}\}$ are the points of a generalized gamma process ([1]). This is obtained from a Poisson random process on $[0,\infty)$ with mean intensity

$$\lambda(\mathrm{d}s) = \frac{1}{\Gamma(1-\alpha)} \exp(-\tau s) s^{-(1+\alpha)} \,\mathrm{d}s, \qquad s \geq 0,$$

with $0 < \alpha < 1$ and $\tau \geq 0$, so that if $N(A)$ is the number of J_i's which fall in $A \in \mathscr{B}([0,\infty))$, then $N(A)$ is Poisson distributed with mean $\lambda(A)$. [5] showed that a generalized gamma random measure defined via (1)-(2), where $\beta = a\tau^{\alpha}/\alpha$ with $a > 0$ and $\tau > 0$, induces a random partition of Gibbs-type ([3]). This can also be

Matteo Ruggiero
University of Torino, Department of Statistics and Applied Mathematics, C.so Unione Sovietica 218/bis, 10134, Torino, Italy. e-mail: matteo.ruggiero@unito.it

generated by means of the urn scheme

$$\mathbb{P}\{X_{n+1} \in \cdot | X_1, \ldots, X_n\} = g_0(n, K_n) P_0(\cdot) + g_1(n, K_n) \sum_{j=1}^{K_n} (n_j - \alpha)\, \delta_{X_j^*}(\cdot) \tag{3}$$

where $X_1^*, \ldots, X_{K_n}^*$ are the K_n distinct values observed in $X_1, \ldots, X_n$ with absolute frequencies $n_1, \ldots, n_{K_n}$, with coefficients $g_0(n, K_n)$ and $g_1(n, K_n)$ given by

$$\begin{aligned} g_0(n,k) &= \frac{\alpha \sum_{i=0}^{n} \binom{n}{i} (-1)^i \beta^{i/\alpha} \Gamma(k+1-i/\alpha;\beta)}{n \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \beta^{i/\alpha} \Gamma(k-i/\alpha;\beta)} \\ g_1(n,k) &= \frac{\sum_{i=0}^{n} \binom{n}{i} (-1)^i \beta^{i/\alpha} \Gamma(k-i/\alpha;\beta)}{n \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \beta^{i/\alpha} \Gamma(k-i/\alpha;\beta)} \end{aligned} \tag{4}$$

where $\Gamma(c;x)$ denotes the upper incomplete gamma function

$$\Gamma(c;x) = \int_x^\infty s^{c-1} \exp(-s) \mathrm{d}s. \tag{5}$$

Special cases of a generalized gamma process with parameters (β, α) are the Dirichlet process ([2]), obtained by letting $\tau = 1$ and $\alpha \to 0$, the normalized stable process ([4]), obtained by setting $\beta = 0$, and the normalized inverse-Gaussian process, obtained by setting $\alpha = 1/2$.

In the next section, which reviews the results contained in [7], we construct and analyze a continuous-time model which is closely related with the class of normalized inverse-Gaussian priors.

2 Results

It is well known that a first order approximation of the weights $g_0(n,k)$ and $g_1(n,k)$ in (4) is $g_0(n,k) \approx 1/n$ and $g_1(n,k) \approx \alpha k/n$. The next proposition determines a second order approximation of the weights, which is crucial for the derivation of the subsequent results.

Proposition 1. *Let $g_0(n,k)$ and $g_1(n,k)$ be as above. When $\alpha = 1/2$ we have*

$$g_0(n,k) = \alpha k/n + \beta/(s_n n) + o(n^{-1}), \qquad g_1(n,k) = 1/n - \beta/(s_n n^2) + o(n^{-2})$$

where $s_n = k/n^\alpha$ and $\beta = a\tau^\alpha/\alpha$.

Consider now a population evolving in time, such that marginally at each time point the population is a normalized inverse-Gaussian sample of size n, i.e. generated

from (3) with weights as in Proposition 1. Every transition consists in substituting a uniformly selected coordinate with a sample from (3), which leaves the marginals unchanged due to the exchangeability of the sample. The next results describes the dynamic species heterogeneity in the population as the sample size increase, where "$\Rightarrow$" denotes convergence in distribution and $C_B(A)$ the space of continuous functions from A to B.

Theorem 1. *Let $\{K_n(m), m \in \mathbb{N}_0\}$ denote the Markov chain which tracks the number of distinct types in the above described dynamic sample, and define $\{\tilde{K}_n(t), t \geq 0\}$ to be such that $\tilde{K}_n(t) = K_n(\lfloor n^{3/2}t \rfloor)/n^{\alpha}$. Let also $\{S_t, t \geq 0\}$ be a diffusion process driven by the stochastic differential equation*

$$\mathrm{d}S_t = (\beta/S_t)\mathrm{d}t + \sqrt{S_t}\,\mathrm{d}B_t, \qquad S_t \geq 0, \tag{6}$$

where B_t is a standard Brownian motion. If $\tilde{K}_n(0) \Rightarrow S_0$, then

$$\{\tilde{K}_n(t), t \geq 0\} \Rightarrow \{S_t, t \geq 0\} \qquad \text{in } C_{[0,\infty)}([0,\infty)) \text{ as } n \to \infty.$$

Furthermore, the points 0 and ∞ are respectively an entrance and a natural boundary for S_t.

Hence the heterogeneity in the sample, once normalized and time-rescaled, is described for large n by a diffusion process which has strictly positive sample paths. Furthermore, (6) can be seen as a particular instance of a continuous-time analog of the notion of α-diversity, introduced by [6] for Poisson-Kingman models (see [7] for details).

The following result joins the above heterogeneity dynamics, with those of the random frequencies. Consider the closure of the infinite ordered simplex

$$\overline{\nabla}_\infty = \Big\{ z = (z_1, z_2, \ldots) : z_1 \geq z_2 \geq \ldots \geq 0,\ \textstyle\sum_{i=1}^{\infty} z_i \leq 1 \Big\},$$

and define the second order differential operator

$$\mathscr{A} = \frac{\beta}{s}\frac{\partial}{\partial s} + \frac{1}{2}s\frac{\partial^2}{\partial s^2} + \frac{1}{2}\sum_{i,j=1}^{\infty} z_i(\delta_{ij} - z_j)\frac{\partial^2}{\partial z_i \partial z_j} - \frac{1}{2}\sum_{i=1}^{\infty}\left(\frac{\beta}{s}z_i + \alpha\right)\frac{\partial}{\partial z_i}. \tag{7}$$

The domain $\mathscr{D}(\mathscr{A})$ of the operator (7) is taken to be the sub-algebra of $C_0([0,\infty) \times \overline{\nabla}_\infty)$ generated by $f = f_0 \times f_1$, with $f_0 \in \mathscr{D}(\mathscr{A}_0)$, $f_1 \in \mathscr{D}(\mathscr{A}_1)$, and where

$$\begin{aligned}
\mathscr{D}(\mathscr{A}_0) &= \{ f \in C_0([0,\infty)) \cap C^2((0,\infty)) : \mathscr{A}_0 f \in C_0([0,\infty)) \}, \\
\mathscr{D}(\mathscr{A}_1) &= \Big\{ \text{sub-algebra of } C(\overline{\nabla}_\infty) \text{ generated by } 1, \textstyle\sum_{i=1}^{\infty} z_i^2, \sum_{i=1}^{\infty} z_i^3, \ldots \Big\},
\end{aligned}$$

with $\mathscr{A}_0$ given by the first two terms in (7). The following theorem states that (7) characterizes a Feller diffusion which almost surely has paths in $C_{[0,\infty)\times\overline{\nabla}_\infty}([0,\infty))$.

Theorem 2. *Let $\mathscr{A}$ be (7) with domain $\mathscr{D}(\mathscr{A})$. The closure in $C_0([0,\infty) \times \overline{\nabla}_\infty)$ of $\mathscr{A}$ generates a Feller semigroup $\{\mathscr{T}(t)\}$ on $C_0([0,\infty) \times \overline{\nabla}_\infty)$. For every $\nu \in \mathscr{P}([0,\infty) \times$*

$\overline{\nabla}_\infty$), *there exists a strong Markov process* $Z(\cdot)$ *corresponding to* $\{\mathscr{T}(t)\}$ *with initial distribution* ν *and sample paths in* $C_{[0,\infty)\times\overline{\nabla}_\infty}([0,\infty))$ *with probability one.*

The last two terms of (7) describe the time evolution of the frequencies of infinitely-many types, where $z_i(\delta_{ij}-z_j)$ are the covariance terms and $-[(\beta/s)z_i+\alpha]$ defines the structure of the drift terms, driven by mutation forces. The positive coefficient $\theta_t=\beta/S_t$ varies in time, and is driven by the diffusion (6).

The connection between (7) and normalized inverse-Gaussian priors is given by the following result.

Theorem 3. *Let* $X^{(n)}(\cdot)$ *be the* $\mathbb{X}^n$*-valued process described at the beginning of the section, with transition occurring at exponential times with mean one, let* $w(X^{(n)}(t))$ *denote the vector of decreasingly ordered frequencies of the distinct types in* $X^{(n)}$ *at time* t*, whose total amount is* $K_n(t)$*, and let* $Z(\cdot)$ *be as in Theorem 2. If the initial distributions converge, then*

$$\left[K_n(n^{3/2}t)/n^\alpha, w(X^{(n)}(n^2t/2))\right] \Rightarrow Z(t)$$

in $C_{[0,\infty)\times\overline{\nabla}_\infty}([0,\infty))$.

The previous theorem states that the diffusion process characterized in Theorem 2 can be constructed as the limit in distribution of the sequence $(K_n(t), w(X^{(n)}(t)))$, once appropriately transformed and rescaled, whose components represent the heterogeneity and the frequencies of the different types in a sample of size n from a normalized inverse-Gaussian prior.

References

[1] Brix, A.: Generalized gamma measures and shot-noise Cox processes. Adv. Appl. Probab. **31**, 929–953 (1999)

[2] Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Ann. Statist. **1**, 209–230 (1973)

[3] Gnedin, A. and Pitman, J.: Exchangeable Gibbs partitions and Stirling triangles. J. Math. Sciences **138**, 5674–5685 (2006)

[4] Kingman, J.F.C.: Random discrete distributions. J. Roy. Statist. Soc. Ser. B **37**, 1–22 (1975)

[5] Lijoi, A., Mena, R.H. and Prünster, I.: Controlling the reinforcement in Bayesian non-parametric mixture models. J. Roy. Statist. Soc. Ser. B **69**, 715–740 (2007)

[6] Pitman, J.: Poisson-Kingman partitions. In Goldstein D.R. (ed.), Lecture Notes, Monograph Series. IMS, Hayward (2003)

[7] Ruggiero, M., Walker, S.G. and Favaro, S.: Alpha-diversity processes and normalized inverse-Gaussian diffusions. Ann. Appl. Probab., in press (2012)