Investigating predictive probabilities of Gibbs-type priors

✉ julyan.arbel@inria.fr   🖱 www.julyanarbel.com

Inria Grenoble - Rhône-Alpes & Laboratoire Jean Kuntzmann

S. Favaro (University of Turin & Collegio Carlo Alberto)

Mathematical Methods of Modern Statistics, Luminy, July 11, 2017

Investigating predictive probabilities of Gibbs-type priors

✉ julyan.arbel@inria.fr 🖱 www.julyanarbel.com

University of Texas at Austin

S. Favaro (University of Turin & Collegio Carlo Alberto)

Mathematical Methods of Modern Statistics, Luminy, July 11, 2017

Bayesian nonparametric applications

- random sparse graphs [Caron and Fox, 2017]

- nonparametric mixture modeling [De Blasi et al., 2015]
- feature allocation [Broderick et al., 2013]
- ranked data [Caron et al., 2014]
- textual data [Chen et al., 2013]
- species sampling [Arbel et al., 2017]
- diversity indices estimation [Cerquetti, 2014]

Bayesian nonparametric applications

- random sparse graphs [Caron and Fox, 2017]

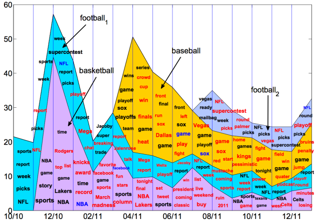
- nonparametric mixture modeling [De Blasi et al., 2015]
- feature allocation [Broderick et al., 2013]
- ranked data [Caron et al., 2014]
- textual data [Chen et al., 2013]
- species sampling [Arbel et al., 2017]
- diversity indices estimation [Cerquetti, 2014]

Bayesian nonparametric applications

- random sparse graphs [Caron and Fox, 2017]

- nonparametric mixture modeling [De Blasi et al., 2015]
- feature allocation [Broderick et al., 2013]
- ranked data [Caron et al., 2014]

- textual data [Chen et al., 2013]



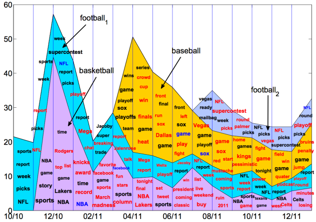
- species sampling [Arbel et al., 2017]
- diversity indices estimation [Cerquetti, 2014]

Bayesian nonparametric applications

- random sparse graphs [Caron and Fox, 2017]

- nonparametric mixture modeling [De Blasi et al., 2015]
- feature allocation [Broderick et al., 2013]
- ranked data [Caron et al., 2014]

- textual data [Chen et al., 2013]



- species sampling [Arbel et al., 2017]
- diversity indices estimation [Cerquetti, 2014]

Why Gibbs-type priors instead of Dirichlet process?

$$X_1, \ldots X_n | P \overset{\text{iid}}{\sim} P, \text{ and } P \sim \mathcal{Q},$$
$$k_n \leq n,\ X_1^*, \ldots, X_{k_n}^*,\ n_1, \ldots, n_{k_n}$$

Gibbs-type: most natural generalization of the Dirichlet process [De Blasi et al., 2015]

- More flexible clustering control
- Power-law vs log rate

Dirichlet process by Ferguson [1973]: $P \sim DP(\theta, G_0)$

$$\mathbb{P}[X_{n+1} \in \cdot \,|\, \boldsymbol{X}^n] = \frac{\theta}{\theta + n} G_0(.) + \frac{1}{\theta + n} \sum_{j=1}^{k_n} n_j \delta_{X_j^*}(.)$$

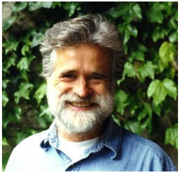
Log rate for number of clusters $k_n \asymp \theta \log n$

Why Gibbs-type priors instead of Dirichlet process?

$$X_1, \ldots X_n | P \overset{\text{iid}}{\sim} P, \text{ and } P \sim \mathcal{Q},$$
$$k_n \leq n,\ X_1^*, \ldots, X_{k_n}^*,\ n_1, \ldots, n_{k_n}$$

Gibbs-type: most natural generalization of the Dirichlet process [De Blasi et al., 2015]

- More flexible clustering control
- Power-law vs log rate



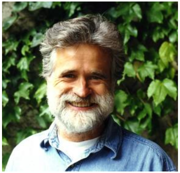

Pitman–Yor process by Pitman and Yor [1997]: $P \sim PY(\alpha, \theta, G_0)$, $\alpha \in (0, 1)$

$$\mathbb{P}[X_{n+1} \in \cdot \,|\, \boldsymbol{X}^n] = \frac{\theta + \alpha k_n}{\theta + n} G_0(.) + \frac{1}{\theta + n} \sum_{j=1}^{k_n} (n_j - \alpha) \delta_{X_j^*}(.)$$

Power law rate for number of clusters $k_n \asymp S n^{\alpha}$

Why Gibbs-type priors instead of Dirichlet process?

$$X_1, \dots X_n | P \stackrel{\text{iid}}{\sim} P, \text{ and } P \sim \mathcal{Q},$$
$$k_n \leq n,\ X_1^*, \dots, X_{k_n}^*,\ n_1, \dots, n_{k_n}$$



Gibbs-type: most natural generalization of the Dirichlet process [De Blasi et al., 2015]

- More flexible clustering control
- Power-law vs log rate

Gibbs-type processes by Pitman [2003]; Gnedin and Pitman [2006]:
$P \sim Gibbs(\alpha, (V_{n,k})_{n,k}, G_0)$, $\alpha < 1$

$$\mathbb{P}[X_{n+1} \in \cdot \,|\, \boldsymbol{X}^n] = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} G_0(.) + \frac{V_{n+1,k_n}}{V_{n,k_n}} \sum_{j=1}^{k_n} (n_j - \alpha) \delta_{X_j^*}(.)$$

$$k_n \asymp \begin{cases} K \text{ random variable a.s. finite if } \alpha < 0 \\ \theta \log n \text{ if } \alpha = 0 \\ S n^{\alpha} \text{ if } \alpha \in (0,1), (S \text{ random variable}). \end{cases}$$

Highly flexible, but barely tractable

V_{n,k_n} parameterized by a positive function h such that $h(t)f_\alpha(t)$ is a density where f_α is the density of a positive α-stable r.v

$$V_{n,k_n} = \frac{\alpha^{k_n}}{\Gamma(n - \alpha k_n)} \int_0^{+\infty} \int_0^1 h(t) t^{-\alpha k_n} p^{n-1-\alpha k_n} f_\alpha((1-p)t) \mathrm{d}p \mathrm{d}t$$

Gibbs-type processes include

- Dirichlet process as a limiting case when $\alpha \to 0$
- Normalized α-stable process: $h(t) = 1$
- Pitman–Yor process: $h(t) = t^{-\theta} \frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)}$
- Normalized generalized gamma process: $h(t) = \exp(\beta^\alpha - \beta t)$

$$V_{n,k_n} = \frac{\alpha^{k_n} e^{\tau}}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-\tau^{1/\alpha})^i \Gamma\left(k_n - \frac{i}{\alpha}, \tau\right)$$

- Outside of these: no explicit expression for weights

Highly flexible, but barely tractable

V_{n,k_n} parameterized by a positive function h such that $h(t)f_\alpha(t)$ is a density where f_α is the density of a positive α-stable r.v

$$V_{n,k_n} = \frac{\alpha^{k_n}}{\Gamma(n-\alpha k_n)} \int_0^{+\infty} \int_0^1 h(t) t^{-\alpha k_n} p^{n-1-\alpha k_n} f_\alpha((1-p)t) \mathrm{d}p \mathrm{d}t$$

Gibbs-type processes include

- Dirichlet process as a limiting case when $\alpha \to 0$
- Normalized α-stable process: $h(t) = 1$
- Pitman–Yor process: $h(t) = t^{-\theta} \frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)}$
- Normalized generalized gamma process: $h(t) = \exp(\beta^\alpha - \beta t)$

$$V_{n,k_n} = \frac{\alpha^{k_n} e^\tau}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-\tau^{1/\alpha})^i \Gamma\left(k_n - \frac{i}{\alpha}, \tau\right)$$

- Outside of these: no explicit expression for weights

Highly flexible, but barely tractable

V_{n,k_n} parameterized by a positive function h such that $h(t)f_\alpha(t)$ is a density where f_α is the density of a positive α-stable r.v

$$V_{n,k_n} = \frac{\alpha^{k_n}}{\Gamma(n-\alpha k_n)} \int_0^{+\infty} \int_0^1 h(t) t^{-\alpha k_n} p^{n-1-\alpha k_n} f_\alpha((1-p)t) \mathrm{d}p \mathrm{d}t$$

Gibbs-type processes include

- Dirichlet process as a limiting case when $\alpha \to 0$
- Normalized α-stable process: $h(t) = 1$
- Pitman–Yor process: $h(t) = t^{-\theta} \frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)}$
- Normalized generalized gamma process: $h(t) = \exp(\beta^\alpha - \beta t)$

$$V_{n,k_n} = \frac{\alpha^{k_n} e^\tau}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-\tau^{1/\alpha})^i \Gamma\left(k_n - \frac{i}{\alpha}, \tau\right)$$

- Outside of these: no explicit expression for weights

Highly flexible, but barely tractable

V_{n,k_n} parameterized by a positive function h such that $h(t)f_\alpha(t)$ is a density where f_α is the density of a positive α-stable r.v

$$V_{n,k_n} = \frac{\alpha^{k_n}}{\Gamma(n-\alpha k_n)} \int_0^{+\infty} \int_0^1 h(t) t^{-\alpha k_n} p^{n-1-\alpha k_n} f_\alpha((1-p)t) \mathrm{d}p \mathrm{d}t$$

Gibbs-type processes include

- Dirichlet process as a limiting case when $\alpha \to 0$
- Normalized α-stable process: $h(t) = 1$
- Pitman–Yor process: $h(t) = t^{-\theta} \frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)}$
- Normalized generalized gamma process: $h(t) = \exp(\beta^\alpha - \beta t)$

$$V_{n,k_n} = \frac{\alpha^{k_n} e^\tau}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-\tau^{1/\alpha})^i \Gamma\left(k_n - \frac{i}{\alpha}, \tau\right)$$

- Outside of these: no explicit expression for weights

Highly flexible, but barely tractable

V_{n,k_n} parameterized by a positive function h such that $h(t)f_\alpha(t)$ is a density where f_α is the density of a positive α-stable r.v

$$V_{n,k_n} = \frac{\alpha^{k_n}}{\Gamma(n-\alpha k_n)} \int_0^{+\infty} \int_0^1 h(t) t^{-\alpha k_n} p^{n-1-\alpha k_n} f_\alpha((1-p)t) \mathrm{d}p \mathrm{d}t$$

Gibbs-type processes include

- Dirichlet process as a limiting case when $\alpha \to 0$
- Normalized α-stable process: $h(t) = 1$
- Pitman–Yor process: $h(t) = t^{-\theta} \frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)}$
- Normalized generalized gamma process: $h(t) = \exp(\beta^\alpha - \beta t)$

$$V_{n,k_n} = \frac{\alpha^{k_n} \mathrm{e}^\tau}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-\tau^{1/\alpha})^i \Gamma\left(k_n - \frac{i}{\alpha}, \tau\right)$$

- Outside of these: no explicit expression for weights

Highly flexible, but barely tractable

V_{n,k_n} parameterized by a positive function h such that $h(t)f_\alpha(t)$ is a density where f_α is the density of a positive α-stable r.v

$$V_{n,k_n} = \frac{\alpha^{k_n}}{\Gamma(n-\alpha k_n)} \int_0^{+\infty} \int_0^1 h(t) t^{-\alpha k_n} p^{n-1-\alpha k_n} f_\alpha((1-p)t) \mathrm{d}p \mathrm{d}t$$

Gibbs-type processes include

- Dirichlet process as a limiting case when $\alpha \to 0$
- Normalized α-stable process: $h(t) = 1$
- Pitman–Yor process: $h(t) = t^{-\theta} \frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)}$
- Normalized generalized gamma process: $h(t) = \exp(\beta^\alpha - \beta t)$

$$V_{n,k_n} = \frac{\alpha^{k_n} \mathrm{e}^\tau}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-\tau^{1/\alpha})^i \Gamma\left(k_n - \frac{i}{\alpha}, \tau\right)$$

- Outside of these: no explicit expression for weights

Efficient approximation of predictive weights

Discovery probability: exact

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{\iint \dots}{\iint \dots}$$

Predictive distribution:

$$\mathbb{P}[X_{n+1} \in \cdot \mid \boldsymbol{X}^n] = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} G_0(.) + \frac{V_{n+1,k_n}}{V_{n,k_n}} \sum_{j=1}^{k_n} (n_j - \alpha) \delta_{X_j^*}(.)$$

Efficient approximation of predictive weights

Discovery probability: first order approximation

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{\alpha k_n}{n} + o\left(\frac{k_n}{n}\right)$$

Predictive distribution:

$$\mathbb{P}[X_{n+1} \in \cdot \mid \boldsymbol{X}^n] \sim \frac{\alpha k_n}{n} G_0(.) + \frac{1}{n}\sum_{j=1}^{k_n}(n_j - \alpha)\delta_{X_j^*}(.)$$

At 1^{st} order: any Gibbs-type boils down to normalized α-stable

Efficient approximation of predictive weights

Discovery probability: second order approximation

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{\alpha k_n}{n} + \frac{\theta_n}{n} + o\left(\frac{1}{n}\right)$$

Predictive distribution:

$$\mathbb{P}[X_{n+1} \in \cdot \mid \boldsymbol{X}^n] \approx \frac{\theta_n + \alpha k_n}{\theta_n + n} G_0(.) + \frac{1}{\theta_n + n} \sum_{j=1}^{k_n} (n_j - \alpha)\delta_{X_j^*}(.)$$

At 2^{nd} order: any Gibbs-type boils down to Pitman–Yor with random precision parameter $\theta_n = \varphi_h\left(nk_n^{-1/\alpha}\right)$ with φ_h being defined as $\varphi_h(t) = -th'(t)/h(t)$

Sketch of proof by Laplace Approximation

1. Rewrite discovery probability as expectation

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \mathbb{E}_n[1-P]$$

according to density $f_n(p,t) \propto t^{-\alpha k_n} p^{n-1-k_n\alpha} h(t) f_\alpha((1-p)t)$,

2. Using power law behavior $k_n \asymp Sn^\alpha$, differentiate density f_n leads to characterize its mode (t_n, p_n) by system $\begin{cases} t_n \asymp nk_n^{-1/\alpha} \\ 1-p_n \asymp \frac{\varphi_h(t_n)+\alpha k_n}{\varphi_h(t_n)+n} \end{cases}$

3. By Laplace approximation

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} \asymp \frac{\varphi_h(nk_n^{-1/\alpha}) + \alpha k_n}{\varphi_h(nk_n^{-1/\alpha}) + n}$$

Sketch of proof by Laplace Approximation

1. Rewrite discovery probability as expectation

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \mathbb{E}_n[1-P]$$

according to density $f_n(p,t) \propto t^{-\alpha k_n} p^{n-1-k_n\alpha} h(t) f_\alpha((1-p)t)$,

2. Using power law behavior $k_n \asymp Sn^\alpha$, differentiate density f_n leads to characterize its mode (t_n, p_n) by system $\begin{cases} t_n \asymp nk_n^{-1/\alpha} \\ 1-p_n \asymp \frac{\varphi_h(t_n)+\alpha k_n}{\varphi_h(t_n)+n} \end{cases}$

3. By Laplace approximation

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} \asymp \frac{\varphi_h(nk_n^{-1/\alpha}) + \alpha k_n}{\varphi_h(nk_n^{-1/\alpha}) + n}$$

Sketch of proof by Laplace Approximation

1. Rewrite discovery probability as expectation

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \mathbb{E}_n[1-P]$$

   according to density $f_n(p,t) \propto t^{-\alpha k_n} p^{n-1-k_n\alpha} h(t) f_\alpha((1-p)t)$,

2. Using power law behavior $k_n \asymp Sn^\alpha$, differentiate density f_n leads to characterize its mode (t_n, p_n) by system $\begin{cases} t_n \asymp n k_n^{-1/\alpha} \\ 1-p_n \asymp \frac{\varphi_h(t_n)+\alpha k_n}{\varphi_h(t_n)+n} \end{cases}$

3. By Laplace approximation

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} \asymp \frac{\varphi_h(nk_n^{-1/\alpha}) + \alpha k_n}{\varphi_h(nk_n^{-1/\alpha}) + n}$$

Sketch of proof by Laplace Approximation

1. Rewrite discovery probability as expectation

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \mathbb{E}_n[1-P]$$

according to density $f_n(p,t) \propto t^{-\alpha k_n} p^{n-1-k_n\alpha} h(t) f_\alpha((1-p)t)$,

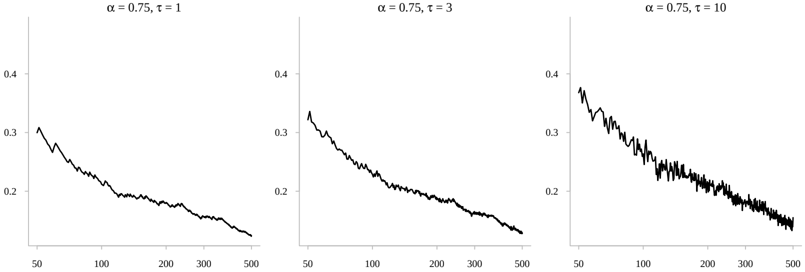
2. Using power law behavior $k_n \asymp Sn^\alpha$, differentiate density f_n leads to characterize its mode (t_n, p_n) by system $\begin{cases} t_n \asymp nk_n^{-1/\alpha} \\ 1-p_n \asymp \frac{\varphi_h(t_n)+\alpha k_n}{\varphi_h(t_n)+n} \end{cases}$

3. By Laplace approximation

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} \asymp \frac{\varphi_h(nk_n^{-1/\alpha}) + \alpha k_n}{\varphi_h(nk_n^{-1/\alpha}) + n}$$

Numerical illustration

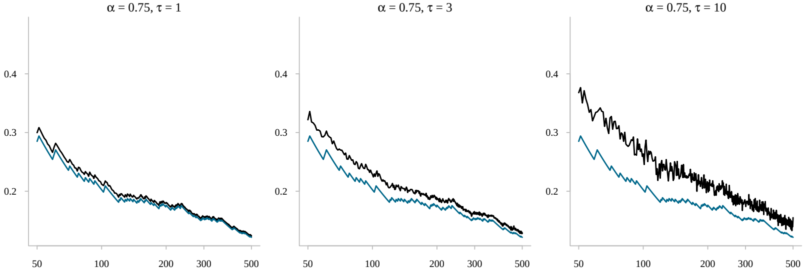
Discovery probabilities plots: data generated from Zipf law parameter 1.5, sample size n from 50 to 500 $\longrightarrow$ defines some k_n sequence
Normalized generalized gamma with $\alpha = 0.75$, $\tau \in \{1, 3, 10\}$.



Exact $\dfrac{V_{n+1,k_n+1}}{V_{n,k_n}}$

Numerical illustration

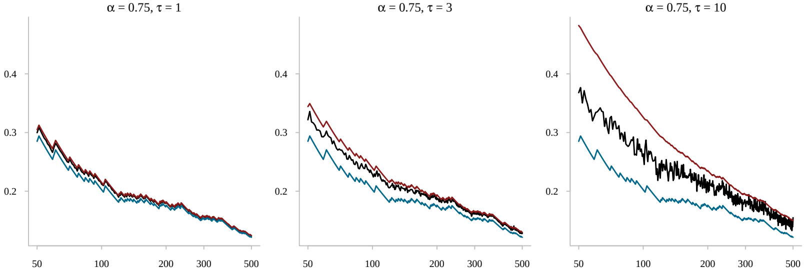
Discovery probabilities plots: data generated from Zipf law parameter 1.5, sample size n from 50 to 500 $\longrightarrow$ defines some k_n sequence
Normalized generalized gamma with $\alpha = 0.75$, $\tau \in \{1, 3, 10\}$.



Exact $\dfrac{V_{n+1,k_n+1}}{V_{n,k_n}}$, 1^{st} order $\dfrac{\alpha k_n}{n}$

Numerical illustration

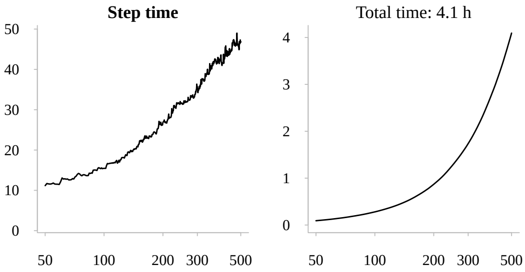
Discovery probabilities plots: data generated from Zipf law parameter 1.5, sample size n from 50 to 500 $\longrightarrow$ defines some k_n sequence
Normalized generalized gamma with $\alpha = 0.75$, $\tau \in \{1, 3, 10\}$.



$$\text{Exact } \frac{V_{n+1,k_n+1}}{V_{n,k_n}}, \; 1^{st} \text{ order } \frac{\alpha k_n}{n}, \; 2^{nd} \text{ order } \frac{\theta_n + \alpha k_n}{\theta_n + n}$$

Numerical illustration

Discovery probabilities plots: data generated from Zipf law parameter 1.5, sample size n from 50 to 500 $\longrightarrow$ defines some k_n sequence
Normalized generalized gamma with $\alpha = 0.75$, $\tau \in \{1, 3, 10\}$.



Monte Carlo computing time: exponential

Conclusion

- In this talk we have presented a large *n* fast and reliable approximation to Gibbs-type predictive weights
- Better understanding of the parameterization
- Make easier to use

Links

- paper on arXiv later this week
- R code available at http://www.julyanarbel.com/software
- Some random thoughts on https://statisfaction.wordpress.com/

Conclusion

- In this talk we have presented a large n fast and reliable approximation to Gibbs-type predictive weights
- Better understanding of the parameterization
- Make easier to use

Links

- paper on arXiv later this week
- R code available at http://www.julyanarbel.com/software
- Some random thoughts on https://statisfaction.wordpress.com/

References I

Julyan Arbel and Stefano Favaro (2017). Investigating predictive probabilities of Gibbs-type priors. **soon on arXiv**.

Julyan Arbel, Stefano Favaro, Bernardo Nipoti, and Yee Whye Teh (2017). Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. **Statistica Sinica**, 27:839–858.

Tamara Broderick, Jim Pitman, and Michael I Jordan (2013). Feature allocations, probability functions, and paintboxes. **arXiv preprint arXiv:1301.6647**.

Francois Caron and Emily B Fox (2017). Sparse graphs using exchangeable random measures. **Journal of the Royal Statistical Society: Series B (Statistical Methodology) with discussion**.

Francois Caron, Yee Whye Teh, and Thomas Brendan Murphy (2014). Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes. **The Annals of Applied Statistics**, 8(2):1145–1181.

Annalisa Cerquetti (2014). Bayesian nonparametric estimation of Patil-Taillie-Tsallis diversity under Gnedin-Pitman priors. **arXiv preprint arXiv:1404.3441**.

Changyou Chen, Vinayak Rao, Wray Buntine, and Yee Whye Teh (2013). Dependent Normalized Random Measures. **Int. Conf. Mach. Learn.**

Pierpaolo De Blasi, Stefano Favaro, Antonio Lijoi, Ramsés H Mena, Igor Prünster, and Matteo Ruggiero (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, 37(2):212–229.

T.S. Ferguson (1973). A Bayesian analysis of some nonparametric problems. **The Annals of Statistics**, 1(2):209–230.

Alexander Gnedin and Jim Pitman (2006). Exchangeable Gibbs partitions and Stirling triangles. **Journal of Mathematical Sciences**, 138(3):5674–5685.

Jim Pitman (2003). Poisson-Kingman partitions. **Lecture Notes-Monograph Series**, pages 1–34.

Jim Pitman and Marc Yor (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. **The Annals of Probability**, 25(2):855–900.