3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–18

Approximating the clusters' prior distribution in Bayesian nonparametric models

Anonymous Authors
Anonymous Institution

Abstract

Bayesian nonparametric models may induce clustering on the data. Knowledge of the prior distribution induced on the number of clusters is thus key for prior specification and calibration. However, evaluating this prior is infamously difficult even for moderate sample size. We evaluate several statistical approximations to the prior distribution on the number of clusters for Gibbs-type processes, a class including the Pitman–Yor process and the normalized generalized gamma process. We introduce a new approximation based on the predictive distribution of Gibbs-type process, which compares favourably with the existing methods. We thoroughly discuss the limitations of these various approximations by comparing them against an exact implementation of the prior distribution of the number of clusters.

1. Introduction

Bayesian nonparametric (BNP) models may induce clustering on the data, for instance when they are used in mixture models. Knowledge of the prior distribution induced on the number of clusters is thus key for application-driven prior specification. In this note, we focus on the large class of Gibbs-type processes (De Blasi et al., 2015), which contains as special cases the Dirichlet process, the Pitman–Yor process (PY) and the normalized generalized gamma process (NGG), to cite a few. Gibbs-type processes can be defined via Poisson–Kingman partitions, see Pitman (2003). They are parameterized by a discount parameter $\sigma < 1$ and a nonnegative function $\mathfrak{h}$ defined on the positive half-line. We focus on $\sigma \geq 0$ parameter, a restriction which entails infinite-dimensionality of the processes under study, thus ruling out Gibbs-type priors with finitely many species that prevail when $\sigma < 0$ (see Pitman, 2006).

Gibbs-type priors are almost surely discrete, taking the form of a weighted countable sum of Dirac masses. Constructive representations of the weight vector (convenient for sampling) include the stick-breaking representation (see Sethuraman, 1994; Pitman and Yor, 1997; Favaro et al., 2016) and the Ferguson–Klass representation (Ferguson and Klass, 1972, only for some specific cases of Gibbs-type processes). Because of discreteness, modeling observations $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ with a Gibbs-type prior creates ties (clusters) in the observations with positive probability. This clustering procedure inherits a prior distribution from the Gibbs-type prior.

Denote by $K_n \leq n$ the number of clusters implied by the model in data $\boldsymbol{X}_n$. We consider the prior distribution for the number of clusters K_n, supported on the set of integers $\{1, \ldots, n\}$. The probability mass function $(p_{n,k})_{1 \leq k \leq n}$ of this random variable has

© A. Authors.

a fairly simple expression in terms of two triangular arrays of reals, denoted by $\mathcal{V}_{n,k}$ and $\mathcal{C}_{n,k}$, for any positive integers n, k such that $1 \leq k \leq n$ (see De Blasi et al., 2015)

$$p_{n,k} := \mathbb{P}(K_n = k) = \frac{1}{\sigma^k} \mathcal{V}_{n,k} \mathcal{C}_{n,k} \tag{1}$$

The $(\mathcal{V}_{n,k})$ triangular array encodes information about σ and the $\mathfrak{h}$ function, while the $(\mathcal{C}_{n,k})$ triangular array, called generalized factorial coefficients (Charalambides, 2005), only depends on σ. More specifically,

$$\mathcal{V}_{n,k} = \frac{\sigma^k}{\Gamma(n-k\sigma)} \int_{\mathbb{R}^+ \times [0,1]} t^{-k\sigma} p^{n-k\sigma-1} \mathfrak{h}(t) f_\sigma((1-p)t) \mathrm{d}t \mathrm{d}p, \quad \mathcal{C}_{n,k} = \frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} (-i\sigma)_n \tag{2}$$

where Γ is the Gamma function, f_σ is the density of a positive σ-stable random variable, and $(x)_n = x(x+1)\dots(x+n-1)$ denotes the increasing factorial coefficient, for any real number x and integer n. See Appendix A for a more precise description of coefficients $\mathcal{V}_{n,k}$, and $\mathcal{C}_{n,k}$. The special cases of PY and NGG processes are provided in Equations (6) and (7), respectively.

This note addresses the practical evaluation of the prior weights $(p_{n,k})_{1 \leq k \leq n}$, a notoriously difficult task, even for moderate sample sizes of order 10^2 or 10^3. The difficulty stems from the evaluation of the $(\mathcal{V}_{n,k})$ triangular array on the one hand, and of the generalized factorial coefficients $(\mathcal{C}_{n,k})$ on the other. Many of the summands involved in the computation of $\mathcal{C}_{n,k}$ and $\mathcal{V}_{n,k}$ overflow the double precision exponent range, both in the positive and the negative domain, preventing the recourse to the usual log-transformation to address the risk of overflow.

Note that the arduousness of the $(\mathcal{V}_{n,k})$ evaluation actually depends on the specific Gibbs-type class under study. For instance, (2) greatly simplifies in the case of the PY process while it remains unwieldy for the NGG process (see (6) and (7) in Appendix). For this reason, illustrations in this note cover the arduous NGG case, while the more manageable PY case is treated in the Appendix.

Contributions

Sec 2 We propose a new approximation of the prior distribution on K_n based on approximating the predictive distribution of the process. Three other existing approaches are also implemented: the finite-dimensional multinomial processes of Lijoi et al. (2020a,b); and the truncation of two sum representations of the processes, that are stick-breaking (Sethuraman, 1994; Pitman and Yor, 1997; Favaro et al., 2016) and Ferguson–Klass (Ferguson and Klass, 1972). Code is made available in Julia[1] and R.[2]

Sec 3.1 We explore the quality of the proposed approximations by comparing the induced prior probability mass function on K_n against that obtained with an exact implementation based on arbitrary precision arithmetics (Kon Kam King et al.). We find that our new approximation based on the predictive distribution performs favourably when compared to the others, except for a parameter range where the finite-dimensional multinomial process fares better.

Sec 3.2 We demonstrate how knowledge of the prior on K_n can inform the delicate question of specifying hyperparameters on BNP priors, illustrating how the first two prior moments of K_n relate to each other as a function of the hyperparameters.

The functionalities proposed here are highly relevant for probabilistic programming languages supporting BNP modelling, such as Turing (see Turing.jl, Ge et al., 2018), implemented in Julia and maintaining modelling with discrete random probability measures (Trapp et al., 2019), and Edward (Tran et al., 2016), implemented in python. Note that our R package with C++ functions could be easily adapted to python.

1. https://github.com/AABI-Gibbs-type-priors/AABI-Gibbs-type-priors
2. https://github.com/AABI-Gibbs-type-priors/AABI-RGibbs-type-priors

2. Statistical approximations

In this section, we present four approaches, the first of which is novel, leading to approximations of the prior distribution of K_n, for different families of Gibbs-type processes.

2.1. Predictive distribution approximation

The first approximation we consider relies on the second order approximation suggested by Arbel et al. (2017); Arbel and Favaro (2020). This requires the function $\mathfrak{h}$ to be continuously differentiable (with derivative denoted by $\mathfrak{h}'$) in order to define a second order term $\beta_{n,k} := \phi_{\mathfrak{h}}(n/k^{1/\sigma})$, with $\phi_{\mathfrak{h}}(t) := -t\mathfrak{h}'(t)/\mathfrak{h}(t)$, in the approximation. Theorem 1 in Arbel and Favaro (2020) provides a so-called second order approximation for the predictive distribution of the Gibbs-type process in the form of the ratio of successive $\mathcal{V}_{n,k}$ parameters. This, in turn, suggests the following result on the ratio of consecutive prior weights $p_{n+1,k}$:

Proposition 1 *Ratios of consecutive weights satisfy the large n approximation for any $1 \leq k \leq n$:*

$$\frac{p_{n+1,k+1}}{p_{n+1,k}} = \left(k + \frac{\beta_{n,k}}{\sigma}\right) \frac{\mathcal{C}_{n+1,k+1}}{\mathcal{C}_{n+1,k}} + O\left(\frac{1}{n}\right). \tag{3}$$

Based on the large n approximation of Proposition 1, a simple recursive algorithm can be employed in order to compute an approximation to the prior weights $p_{n+1,k}$ without resorting to the cumbersome computation of the $\mathcal{V}_{n,k}$ parameters. In the specific case of the NGG, $\beta_{n,k}$ takes a simple form (see details in Appendix B.1).

2.2. Finite-dimensional approximation

Finite-dimensional alternatives to BNP priors have been developed for cases where the assumption that the number of clusters grows with the sample size does not hold. Their flexibility is such that they could also be considered as approximations to certain nonparametric priors. The recent works by Lijoi et al. (2020a,b) develop finite-dimensional multinomial versions of the PY process and normalized random measures with independent increments (NRMI, Regazzini et al., 2003). The NGG is a special case of an NRMI, and the finite-dimensional NGG counterpart is termed the NGG multinomial process. Being a finite-dimensional process, it requires an integer parameter H which defines the maximum number of clusters that can be obtained. Theorem 4 and Example 4 of Lijoi et al. (2020a) provide the prior distribution of the number of clusters under the NGG multinomial process:

$$\mathbb{P}(K_n = k) = \frac{H!}{(\sigma H)^k (H-k)!} \sum_{\ell=0}^{n-k} \frac{1}{(\sigma H)^\ell} \mathcal{V}_{n,\ell+k} \mathcal{C}_{n,\ell+k} \mathscr{S}_{\ell+k,k}, \tag{4}$$

for any $k \in \{1, \ldots, \min(H, n)\}$, where $\mathscr{S}_{\ell,k}$ is the Stirling number of the second kind. This distribution could alternatively be represented using the probabilities $p_{n,k}$ of the NGG process and in this case, one could use the approximations for NGG to compute the NGG multinomial process. Moreover, it is also possible to compute the expectation through the weights of $p_{n,k}$ (see details in Appendix B.2).

2.3. Stick-breaking truncation

The stick-breaking representation is perhaps the most popular constructive representation of the Dirichlet process (Sethuraman, 1994). More specifically, the stick-breaking describes the distribution of the component weights by recursively breaking a stick of length one (i.e. the total probability of the process). This weight distribution is size-biased, meaning that the weights are stochastically decreasing (decreasing in expectation). It turns out that any Gibbs-type process satisfies such a

stick-breaking representation, see for instance Theorem 14.23 of Ghosal and Van der Vaart (2017) for details. In addition to the Dirichlet process case, Pitman and Yor (1997) present the PY case, while Lau and Cripps (2015); Favaro et al. (2016) present the NGG case. The stick-breaking representation provides a simple way to approximate the process by truncating the representation at some finite level H. Many different approaches have been proposed, with fixed or random truncation levels, yielding varying guarantees on the truncation error, see for instance Ishwaran and James (2001); Muliere and Tardella (1998); Arbel et al. (2019). Here we adopt a simple fixed truncation at level H. We use a stick-breaking representation of NGG process based on fractions of exponentially tilted stable and gamma random variables (see Lemma 3.3 Lau and Cripps, 2015)(see details in Appendix B.3).

2.4. Ferguson–Klass truncation

Lastly, we consider here a sub-class of Gibbs-type processes composed of those processes which are obtained by a normalisation step, called normalized random measures with independent increments (NRMI, Regazzini et al., 2003). An alternative series representation to the stick-breaking in this case has been developed by Ferguson and Klass (1972). It is referred to as the Ferguson–Klass representation. See Barrios et al. (2013) for a review, and Arbel et al. (2020) for an R package devoted to mixture modeling with NRMI. Several truncation strategies have been proposed, see for instance Argiento et al. (2016); Arbel and Prünster (2017). As opposed to the stick-breaking, the Ferguson–Klass representation provides a vector of weights in a non-increasing order. As a result, a similar truncation strategy as that employed for the stick-breaking provides a lower truncation error (almost surely) for a given truncation level H.

3. Comparison of approximations and discussion

3.1. Prior on K_n

We compare all the presented schemes on an approximation of the prior on K_n for a NGG process, for various parameter values (τ, σ, and n the number of observations). Our reference is an exact computation of the prior distribution for the NGG. See Figure 1 for $n = 100$ and Figure 3 in Appendix C for $n = 1000$. For comparison, we specified the same truncation level $H = 250$ for both the truncated methods and for the finite-dimensional process. See results for $H = 1000$ on Figure 4 in Appendix C. The predictive approximation is a good approximation both for small values of τ and large values of σ, but the quality decreases with growing τ, for small values of σ. This is coherent with the results for predictive approximation discussed in Arbel and Favaro (2020). Interestingly, the NGG multinomial process performs well in all situations, outperformed only by the predictive approximation for small values of τ and large values of σ. The stick-breaking truncation performs well for low values of both parameters; but quickly becomes unreliable as they increase. This is expected as the truncation level necessary for a good approximation grows rapidly when the parameters increase. The Ferguson–Klass truncation performs better than the stick-breaking for the same truncation level, because it displays large weights first while there is only a stochastic order for stick-breaking, as discussed previously. Results for $n = 1000$ (Figure 3 in Appendix C) are similar to $n = 100$, and larger truncation level H (Figure 4 in Appendix C) reveal the same trends. Figure 2 shows another comparison for a larger range of parameter values. In line with Figure 1, we use the truncation level $H = 250$ and the sample size $n = 100$. Each point on the graph represents the expectation and standard deviation $(\mathbb{E}[K_n], \mathrm{std}[K_n])$ of the prior distribution induced by a parameter pair (τ, σ), whose values are denoted by colour (see grid legend, panel (d)). All the approximations exhibit a similar pattern, despite huge variations on the range of attainable pairs $(\mathbb{E}[K_n], \mathrm{std}[K_n])$ due to approximation failure for some parameters values, especially for truncated approximations. Increasing the truncation level H would mitigate the disparities

among approximations, but at a higher computational cost. For this large range of parameter values, the predictive approximation seems to perform better than the other approximations except for small values of σ and large values of τ (pink area). The NGG multinomial process performs reasonably well in this area of parameter space, but starts failing for larger values of σ. Both stick-breaking and Ferguson–Klass approximations seem to fail for this low level of truncation, although the Ferguson–Klass approach seems to fare better and could probably give good results with a higher truncation level. All approximations present some degree of challenge for numerical stability, but while the existing solutions for stick-breaking and Ferguson–Klass methods could still be unstable for some parameter values, those challenges seem solved by Kon Kam King et al. for NGG predictive and NGG multinomial.

3.2. Prior calibration

A practical application of the approximations presented here is prior calibration, or in other words choosing prior hyperparameters to reflect pertinent expert information. A natural approach is to choose the moments of the prior distribution to reflect the amount of information known a priori, typically an expected number of clusters and an uncertainty for mixture models. However, the structure of Gibbs-type priors (1) and (2) renders this approach a challenging task, as moments do not have an invertible analytical expression in general that would allow selecting hyperparameters, or even interpreting their qualitative impact on the distribution. The approximations presented in this note allow investigating the impact of the hyperparameters on various aspects of the prior distribution and ultimately choosing them to calibrate the prior. We consider calibrating an NGG process prior or a PY process, both involving two parameters which can be denoted $\tau \in \mathbb{R}^+$ and $\sigma \in [0, 1[$. As a first approximate guide, τ is often presented as influencing the location of the distribution, while σ is often presented as having more bearing on the scale (e.g. Barrios et al., 2013). Figure 2 shows more precisely the impact of the parameters of the NGG process. Dashed (resp. dotted) lines on the graph connect the points where σ (resp. τ) is kept constant. Following dashed lines, we can see that for small fixed σ (dark blue to pink), increasing τ strongly increases the expected number of clusters. For large σ however (green to yellow), increasing τ mostly decreases the prior standard deviation, while for intermediate σ, increasing τ both increases the expectation and decreases the standard deviation. Following dotted lines, we can see that for a fixed τ, increasing σ increases the prior standard deviation up to a point where the standard deviations starts decreasing, in an almost parabolic relationship. The relation appears closer to parabolic for large values of τ, and the maximum of the parabola moves towards smaller expectations with larger values of τ. Another striking feature of Figure 2, particularly relevant for prior calibration, is that some regions of $(\mathbb{E}[K_n], \mathrm{std}[K_n])$ are not attainable for any combination of parameters. This fact must be taken into account when calibrating the prior distribution using its moments. Echoing the observation from the previous section, the figure also suggests that the predictive approximation could be used for prior calibration for large values σ, while NGG multinomial performs better for small values of σ.

3.3. Discussion

We have compared four approximations to the NGG process (and to PY in Appendix) showing that the predictive-based approximation fares best overall, though it fails and should be replaced by the multinomial approximation for small values of σ and large values of τ. To complement these results, note that all three approximations relying on a level H satisfy a weak convergence guarantee when $H \to \infty$. There is no such equivalent for the predictive-based approximation, but the large n result of Proposition 1 holds instead. Further research will address comparing computational efficiency of the methods.

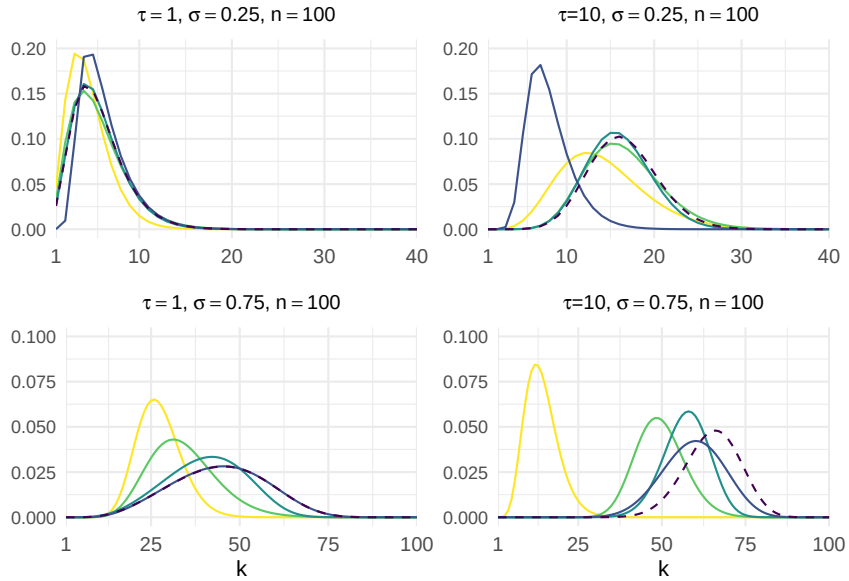
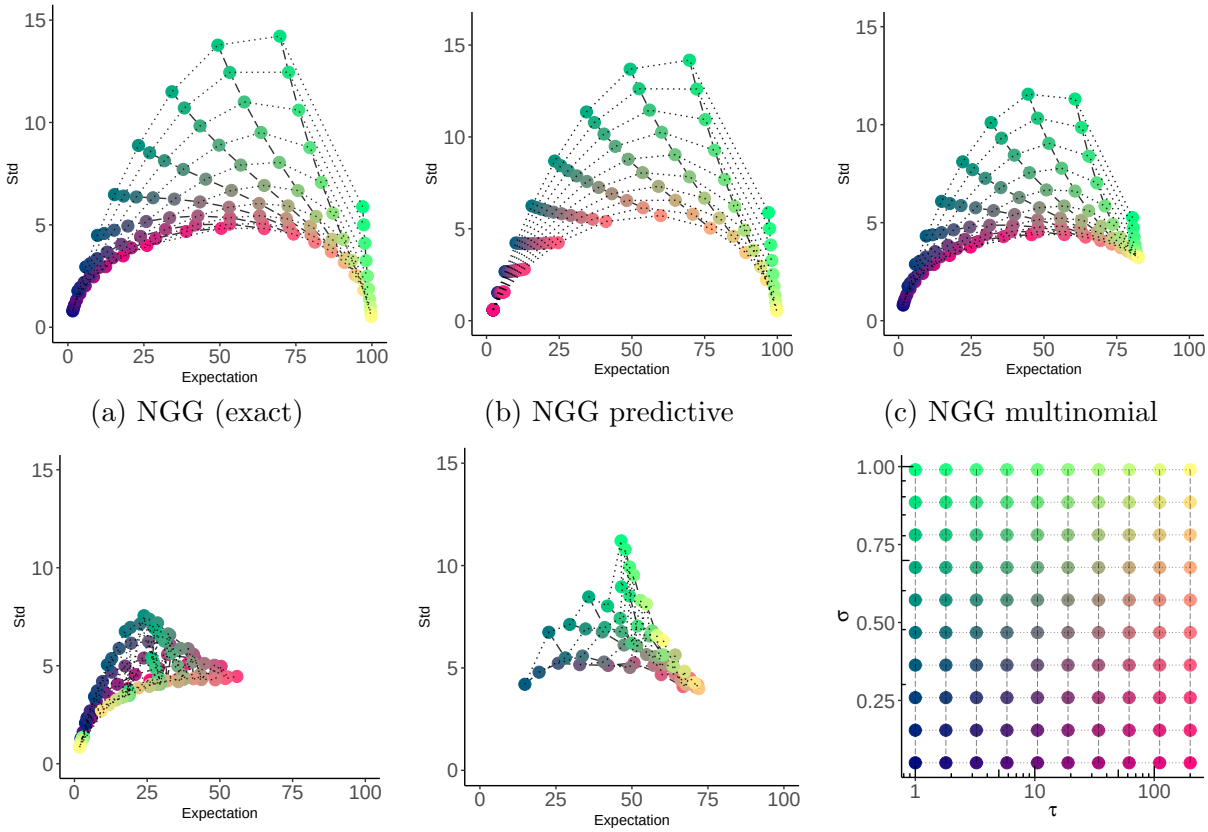


Figure 1: Prior distribution of K_n for exact NGG (dashed line), NGG predictive (dark blue), NGG multinomial (light blue), stick-breaking truncation (yellow), Ferguson–Klass truncation (green), for $n = 100$ and truncation $H = 250$ for $\tau \in \{1, 10\}$ and $\sigma \in \{0.25, 0.75\}$.



(a) NGG (exact) (b) NGG predictive (c) NGG multinomial

(d) Stick-breaking (e) Ferguson–Klass (f) Parameter (τ, σ) legend

Figure 2: Expectation and standard deviation for the distribution of the prior number of clusters, for (a) exact NGG, (b) predictive approximation, (c) NGG multinomial process, (d) stick-breaking truncation, (e) Ferguson–Klass truncation, (f) as a function of parameters (τ, σ) spanning $[1, 200] \times (0, 1)$ (log scale in τ) for $n = 100$ samples and truncation $H = 250$.

References

Julyan Arbel and Stefano Favaro. Approximating predictive probabilities of Gibbs-type priors. *Sankhyā A*, 2020.

Julyan Arbel and Igor Prünster. A moment-matching Ferguson & Klass algorithm. *Statistics and Computing*, 27(1):3–17, 2017.

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

Julyan Arbel, Pierpaolo De Blasi, and Igor Prünster. Stochastic approximations to the Pitman–Yor process. *Bayesian Analysis*, 14(3):753–771, 2019.

Julyan Arbel, Guillaume Kon Kam King, Antonio Lijoi, Luis E. Nieto-Barajas, and Igor Prünster. BNPdensity: Bayesian nonparametric mixture modeling in R. *Technical report*, 2020.

Raffaele Argiento, Ilaria Bianchini, and Alessandra Guglielmi. Posterior sampling from ε-approximation of normalized completely random measure mixtures. *Electronic Journal of Statistics*, 10(2):3516–3547, 2016.

Ernesto Barrios, Antonio Lijoi, Luis E Nieto-Barajas, and Igor Prünster. Modeling with normalized random measure mixture models. *Statistical Science*, 28(3):313–334, 2013.

Richard L Burden and J Douglas Faires. *Numerical Analysis*. Kent Publishing Company, Boston, MA, 1993.

Charalambos A. Charalambides. *Combinatorial methods in discrete distributions*, volume 600. John Wiley & Sons, 2005.

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

S. Favaro, A. Lijoi, C. Nava, B. Nipoti, I. Prünster, and Y. W. Teh. On the stick-breaking representation for homogeneous NRMIs. *Bayesian Analysis*, 11(3):697–724, 09 2016.

Thomas S Ferguson and Michael J Klass. A representation of independent increment processes without gaussian components. *The Annals of Mathematical Statistics*, 43(5):1634–1643, 1972.

Hong Ge, Kai Xu, and Zoubin Ghahramani. Turing: a language for flexible probabilistic inference. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 1682–1690, 2018.

Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

Marius Hofert, Ivan Kojadinovic, Martin Maechler, and Jun Yan. *copula: Multivariate Dependence with Copulas*, 2020. URL https://CRAN.R-project.org/package=copula. R package version 1.0-0.

H. Ishwaran and L.F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001. ISSN 0162-1459.

Guillaume Kon Kam King, François Deslandes, Daria Bystrova, and Julyan Arbel. Computing the prior distribution for the number of clusters induced by Gibbs-type processes. *In preparation*.

John W Lau and Edward Cripps. Stick-breaking representation and computation for normalized generalized gamma processes. *Sankhya A*, 77(2):300–329, 2015.

Antonio Lijoi, Igor Prünster, and Tommaso Rigon. Finite-dimensional discrete random structures and Bayesian clustering. *Preprint*, 2020a.

Antonio Lijoi, Igor Prünster, and Tommaso Rigon. The Pitman–Yor multinomial process for mixture modeling. *Biometrika*, 2020b.

Pietro Muliere and Luca Tardella. Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics*, 26(2):283–297, 1998.

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

Jim Pitman. *Combinatorial stochastic processes*, volume 1875. Springer-Verlag, 2006.

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

Eugenio Regazzini, Antonio Lijoi, and Igor Prünster. Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2):560–585, 2003.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

Dustin Tran, Alp Kucukelbir, Adji B. Dieng, Maja Rudolph, Dawen Liang, and David M. Blei. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*, 2016.

Martin Trapp, Emile Mathieu, Maria Lomeli, and Hong Ge. Turing.jl: Probabilistic programming with discrete random probability measures. *Poster at Bayesian nonparametric conference 12, Oxford University*, 2019.

Contents

1 Introduction **1**

2 Statistical approximations **3**
- 2.1 Predictive distribution approximation . . . 3
- 2.2 Finite-dimensional approximation . . . 3
- 2.3 Stick-breaking truncation . . . 3
- 2.4 Ferguson–Klass truncation . . . 4

3 Comparison of approximations and discussion **4**
- 3.1 Prior on K_n . . . 4
- 3.2 Prior calibration . . . 5
- 3.3 Discussion . . . 5

A Description of specific classes of Gibbs-type processes: PY and NGG **10**
- A.1 $\mathcal{V}_{n,k}$ parameters . . . 10
- A.2 $\mathcal{C}_{n,k}$ parameters . . . 10

B Details on approximations **11**
- B.1 Predictive distribution approximation . . . 11
  - B.1.1 Proof of Proposition 1 . . . 11
  - B.1.2 Algorithm for obtaining prior weights from Proposition 1 . . . 12
- B.2 Finite-dimensional approximation . . . 12
- B.3 Stick-breaking truncation . . . 13
- B.4 Ferguson–Klass truncation . . . 13

C Comparison of approximations for NGG (extra plots) **14**

D Comparison of approximations for PY **16**
- D.1 Finite-dimensional approximation . . . 16
- D.2 Stick-breaking truncation . . . 16
- D.3 Expectation and variance of the PY process . . . 16
- D.4 Comparison of approximations . . . 17

Appendix A. Description of specific classes of Gibbs-type processes: PY and NGG

A.1. $\mathcal{V}_{n,k}$ parameters

The functional parameter $\mathfrak{h}$ defining a Gibbs-type process can be equivalently written in terms of an infinite triangular array of positive parameters $\mathcal{V}_{n,k}$, for any $n \geq 1$ and $1 \leq k \leq n$, as follows:

$$\mathcal{V}_{n,k} = \frac{\sigma^k}{\Gamma(n-k\sigma)} \int_0^{+\infty} \int_0^1 t^{-k\sigma} p^{n-k\sigma-1} \mathfrak{h}(t) f_\sigma((1-p)t) \mathrm{d}t \mathrm{d}p, \tag{5}$$

where Γ is the Gamma function and f_σ is the density of a positive σ-stable random variable.

The PY process, parameterized by $\sigma \in (0,1)$ and $\alpha > -\sigma$, is obtained by choosing

$$\mathfrak{h}(t) = \frac{\sigma\Gamma(\alpha)}{\Gamma(\alpha/\sigma)} t^{-\alpha}, \quad \text{which implies} \quad \mathcal{V}_{n,k} = \frac{\prod_{i=0}^{k-1}(\alpha + i\sigma)}{(\alpha)_n}, \tag{6}$$

where the Pocchammer symbol $(\alpha)_n$ is used to denote the rising factorial $\alpha(\alpha+1)\cdots(\alpha+n-1)$. On the other hand, the NGG is parameterized by $\sigma \in (0,1)$ and $\tau \geq 0$ and is defined by

$$\mathfrak{h}(t) = \mathrm{e}^{\tau^\sigma - \tau t}, \quad \text{such that} \quad \mathcal{V}_{n,k} = \frac{\sigma^{k-1}\mathrm{e}^\tau}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \tau^{i/\sigma} \Gamma\left(k - \frac{i}{\sigma}, \tau\right), \tag{7}$$

where $\Gamma(a,x) = \int_x^\infty s^{a-1}\mathrm{e}^{-s}\mathrm{d}s$ is the incomplete gamma function. Note that the NGG have several equivalent parametrizations. Another parametrization comes from the definition of the NGG through normalization of generalized gamma process. Generalized gamma processes are discrete random measures P of the form

$$P = \sum_{i=1}^{\infty} J_i \delta_{\theta_i}, \tag{8}$$

where weights J_i do not sum to 1 and θ_i are location parameters, sampled iid from a measure P_0, a probability distribution over parameter space Θ (see Barrios et al., 2013). (J_i, θ_i) are the points of Poisson process with mean intensity:

$$\nu(d\nu, d\theta) = \frac{e^{-\kappa\nu}}{\Gamma(1-\sigma)\nu^{1+\sigma}} d\nu \alpha P_0(d\theta), \tag{9}$$

which depends on parameter $\kappa \geq 0$ and $\sigma \in [0,1)$, such that $(\kappa, \sigma) \neq (0,0)$, $\alpha > 0$. The NGG is obtained by a normalization step from P which consists in dividing it by its total mass. It can be denoted $\mathrm{NGG}(\alpha, \kappa, \sigma; P_0)$ (see Barrios et al., 2013). Note the important change of parameter between the (τ, σ) and the (α, κ, σ) parameters:

$$\tau = \alpha\kappa^\sigma/\sigma.$$

A.2. $\mathcal{C}_{n,k}$ parameters

For any Gibbs-type process, the probability mass function $(p_{n,k})_{1\leq k\leq n}$ of the prior distribution on the number of clusters K_n can be described in terms of $\mathcal{V}_{n,k}$ and of another triangular array of reals called generalized factorial coefficients (see Charalambides, 2005) and denoted by $\mathcal{C}_{n,k}$, for any positive integers n, k such that $1 \leq k \leq n$:

$$\mathcal{C}_{n,k} = \frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} (-i\sigma)_n. \tag{10}$$

An interesting property of the generalized factorial coefficients is that they follow a recursive relation:

$$\mathcal{C}_{n+1,k} = (n - \sigma k)\mathcal{C}_{n,k} + \sigma \mathcal{C}_{n,k-1} \tag{11}$$

with the following corner cases: $\mathcal{C}_{0,0} = 1$, $\mathcal{C}_{n,0} = 0$ and for all $k > n$, $\mathcal{C}_{n,k} = 0$.

Appendix B. Details on approximations

B.1. Predictive distribution approximation

B.1.1. Proof of Proposition 1

The second order approximation of Arbel and Favaro (2020) can be written as

$$\frac{\mathcal{V}_{n+1,k+1}}{\mathcal{V}_{n,k}} = \frac{k\sigma + \beta_{n,k}}{n + \beta_{n,k}} + O\left(\frac{1}{n^2}\right) \tag{12}$$

$$\frac{\mathcal{V}_{n+1,k}}{\mathcal{V}_{n,k}} = \frac{1}{n + \beta_{n,k}} + O\left(\frac{1}{n^3}\right), \tag{13}$$

where $\beta_{n,k} = O(1)$. Indeed, inspection of the proof of Theorem 1 in Arbel and Favaro (2020) shows that $o(1/n)$ (resp. $o(1/n^2)$) can be replaced by $O(1/n^2)$ (resp. $O(1/n^3)$). Note that we adopt the form of Equation (1.13) of Arbel and Favaro (2020) which has the advantage of being a proper probability distribution (in contrast to approximations of Equation (1.10) and Equation (1.11) of Arbel and Favaro (2020) which do not lead to a proper probability distribution). A first order approximation is obtained by simply setting $\beta_{n,k} = 0$ in the above equations. But it yields poor approximation quality and shall not be used here. The prior distribution of the number of clusters for $n+1$ samples is defined by weights $p_{n+1,k} = \mathbb{P}(K_{n+1} = k)$, which have the following form

$$p_{n+1,k} = \frac{1}{\sigma^k}\mathcal{V}_{n+1,k}\mathcal{C}_{n+1,k}, \tag{14}$$

where $\sum_{i=1}^{n+1} p_i = 1$. Therefore, the ratio of two consecutive weights is

$$\frac{p_{n+1,k+1}}{p_{n+1,k}} = \frac{\mathcal{V}_{n+1,k+1}\mathcal{C}_{n+1,k+1}}{\sigma\mathcal{V}_{n+1,k}\mathcal{C}_{n+1,k}} \tag{15}$$

This involves a ratio of two consecutive $\mathcal{V}$ parameters which can be obtained as the result of the ratio of Equations (12) and (13):

$$\begin{aligned}\frac{\mathcal{V}_{n+1,k+1}}{\mathcal{V}_{n+1,k}} = \frac{\mathcal{V}_{n+1,k+1}}{\mathcal{V}_{n,k}} \Big/ \frac{\mathcal{V}_{n+1,k}}{\mathcal{V}_{n,k}} &= \frac{\frac{k\sigma+\beta_{n,k}}{n+\beta_{n,k}} + O\left(\frac{1}{n^2}\right)}{\frac{1}{n+\beta_{n,k}} + O\left(\frac{1}{n^3}\right)} \\ &\overset{\text{(a)}}{=} \frac{k\sigma + \beta_{n,k} + O\left(\frac{1}{n}\right)}{1 + O\left(\frac{1}{n^2}\right)} = k\sigma + \beta_{n,k} + O\left(\frac{1}{n}\right),\end{aligned}$$

where step (a) is because $\beta_{n,k} = O(1)$ (Arbel and Favaro, 2020). Hence, incorporating this into (15) yields

$$\frac{p_{n+1,k+1}}{p_{n+1,k}} = \left(k\sigma + \beta_{n,k} + O\left(\frac{1}{n}\right)\right)\frac{\mathcal{C}_{n+1,k+1}}{\sigma\mathcal{C}_{n+1,k}} = \left(k + \frac{\beta_{n,k}}{\sigma}\right)\frac{\mathcal{C}_{n+1,k+1}}{\mathcal{C}_{n+1,k}} + O\left(\frac{1}{n}\right),$$

where we have used the monotony property of the $\mathcal{C}_{n,k}$ sequence provided in Lemma 2 in order to upper bound the ratio $\frac{\mathcal{C}_{n+1,k+1}}{\mathcal{C}_{n+1,k}}$ by one and to ensure that the $O\left(\frac{1}{n}\right)$ term can be placed out of the bracket. This concludes the proof.

Lemma 2 *For any integer n, the sequence $k \mapsto \mathcal{C}_{n,k}$ is non increasing for $k \in \{1, \dots, n\}$.*

Proof The proof is by induction, by noting that the difference $\mathcal{C}_{n+1,k} - \mathcal{C}_{n+1,k+1}$ can be separated into a sum of three non negative terms $(n - \sigma k)(\mathcal{C}_{n,k} - \mathcal{C}_{n,k+1}) + \sigma(\mathcal{C}_{n,k-1} - \mathcal{C}_{n,k}) + \sigma\mathcal{C}_{n,k+1}$. ■

B.1.2. Algorithm for obtaining prior weights from Proposition 1

We describe here a simple algorithm that outputs prior weights $p_{n+1,k}$, $k \in \{1, \dots, n+1\}$, from the predictive approximation ratio of Proposition 1. Note that we consider here a sample size of $n+1$ only for presentation convenience. Let $x_{n,k} := \frac{p_{n+1,k+1}}{p_{n+1,k}}$ as given in Equation 3, for $k \in \{1, \dots, n\}$. Starting from one end or the other (i.e. $p_{n+1,1}$ or $p_{n+1,n+1}$), we can recover all the other weights recursively by using the ratios $x_{n,k}$s:

$$p_{n+1,k+1} = x_k p_{n+1,k} = x_k(x_{k-1}p_{n+1,k-1}) = x_k x_{k-1}(x_{k-2}p_{n+1,k-2}) = \dots = \left(\prod_{i=1}^{k} x_i\right) p_{n+1,1}.$$

Combining this with $\sum_{i=1}^{n+1} p_{n+1,i} = 1$ yields $p_{n+1,1}\left(1 + \sum_{j=1}^{n}\prod_{i=1}^{j} x_i\right) = 1$. Hence all prior weights can be recursively obtained following

$$\begin{cases} p_{n+1,1} = \frac{1}{1+\sum_{j=1}^{n}\prod_{i=1}^{j} x_i} \\ p_{n+1,2} = p_{n+1,1}x_1 \\ \dots \\ p_{n+1,k+1} = p_{n+1,k}x_k \\ \dots \\ p_{n+1,n+1} = p_{n+1,n}x_n. \end{cases}$$

B.2. Finite-dimensional approximation

Distribution of the prior number of clusters for the NGG multinomial process could be computed using the distribution of NGG as follows (Theorem 4 in Lijoi et al., 2020a):

$$\mathbb{P}(K_n = k) = \frac{H!}{(H)^k(H-k)!}\sum_{l=0}^{n-k}\frac{1}{H^l}\mathscr{S}_{l+k,k}, \mathbb{P}(K_{n,\infty} = l+k), \tag{16}$$

for any $k \in \{1, \dots, \min(H, n)\}$, where $\mathbb{P}(K_{n,\infty} = l+k)$ are the probabilities in the NGG process. These probabilities of NGG could be replaced by approximation for faster computation. Moreover, the expected number of clusters induced by NGG multinomial process, could be computed using the approximation for $p_{n,k}$:

$$\mathbb{E}(K_n) = H - H\mathbb{E}\left(\left(1 - \frac{1}{H}\right)^{K_{n,\infty}}\right) = H - H\sum_{l=1}^{n}\left(1 - \frac{1}{H}\right)^l \frac{\mathcal{V}_{n,l}}{\sigma^l}\mathcal{C}_{n,l}. \tag{17}$$

We can use the computed $p_{n,k}$ to compute the expected number of clusters for NGG.

$$\mathbb{E}(K_n) = H - H\sum_{l=1}^{n}\left(1 - \frac{1}{H}\right)^l p_{n,l} \tag{18}$$

B.3. Stick-breaking truncation

We used the algorithm suggested by Lau and Cripps (2015) for stick-breaking representation of the NGG$(\alpha, \kappa, \sigma; P_0)$ (see Appendix A for details on these parameters). The stick-breaking truncation of the NGG can be written as:

$$P_H = \sum_{i=1}^{H} p_i \delta_{Z_i}, \tag{19}$$

where

$$p_i = U_i \prod_{j=1}^{i-1}(1-U_j), \quad i = 2, \ldots, H-1, \quad p_1 = U_1, \tag{20}$$

and $p_H = (1-\sum_{i=j}^{H-1} p_j) = \prod_{j=1}^{H}(1-U_j)$ and H defines truncation level. The sequence $\{U_1, \ldots, U_{H-1}\}$ has conditional density $f_{U_n|U_{n-1},\ldots,U_1}(u_n \mid u_{n-1}, \ldots, u_1)$ equal to

$$\frac{\sigma}{\Gamma(1-\sigma)} u_n^{\sigma}(1-u_n)^{(\theta+n\sigma)-1} \frac{\int_0^\infty e^{-(\kappa s/q_{n-1})/(1-u_n)} s^{-(\alpha+n\sigma)} f_{S_\sigma}(s)\mathrm{d}s}{\int_0^\infty e^{-(\kappa s/q_{n-1})} s^{-(\alpha+(n-1)\sigma)} f_{S_\sigma}(s)\mathrm{d}s}.$$

Here $q_i = \prod_{j=1}^{i}(1-u_j)$ and $q_0 = 1$. For computing the probabilities $p_{n,k}$, $k = 1, \ldots, H$ in (19), we sample U_j, $j = 1, \ldots, H-1$ using the following algorithm proposed in Lau and Cripps (2015) (see Lemma 3.3):

- Sample augmented random variable ξ_n conditional on $\{U_{n-1}, \ldots, U_1\}$, that has density:

$$\frac{e^{-(\kappa/q_{n-1}+\xi_n)^\sigma}(\kappa/q_{n-1}+\xi_n)^{(\sigma-1)}\xi_n^{\alpha+(n-1)\sigma}}{\int_0^\infty e^{-(\kappa/q_{n-1}+\xi_n)^\sigma}(\kappa/q_{n-1}+\xi_n)^{(\sigma-1)}\xi_n^{\alpha+(n-1)\sigma}} \tag{21}$$

- Sample X_n, Z_n conditional on ξ_n: $Z_n \sim \text{gamma}(1-\sigma, \kappa/k_{n-1}+\xi_n)$ and X_n have exponentially tilted stable distribution with parameters σ and $\lambda = \kappa/k_{n-1} + \xi_n$.

For sampling from the exponentially tilted stable distribution we used the R package `copula` (Hofert et al., 2020). To estimate the distribution of the prior number of clusters K_n for the stick-breaking truncation, we averaged over 2×10^2 Monte Carlo simulations then smoothed the result using a Gamma density.

B.4. Ferguson–Klass truncation

We consider discrete random measure of the form (8), with jumps J_i and random locations Z_i (8). Ferguson–Klass representation consists in expressing random jumps J_i in terms of underlying Lévy intensity. This representation produces the jumps in non increasing order $J_1 \geq J_2, \ldots$, which can be obtained by solving numerical equations $\xi_i = N(J_i)$, where $N(v) = \nu([v, \infty)])$ is a decreasing function and $\xi_1, \xi_2 - \xi_1, \ldots$ are jump times of standard Poisson process of unit rate. In case of the generalized gamma process N has the following form:

$$N(v) = \frac{\alpha}{\Gamma(1-\sigma)} \int_\nu^\infty e^{-\kappa u} u^{-(1+\sigma)} \mathrm{d}u. \tag{22}$$

The correspondence between parametrisation in terms of κ and in terms of τ is $\tau = \alpha\kappa^\sigma/\sigma$, c.f. Equation (9). One can truncate the series in Equation (8) at some specified level H. We used the following approach:

- Sample $\xi_i \sim \text{PP}$ for $i = 1, \ldots, H$

- Compute $J_i = N^{-1}(\xi_i)$ for $i = 1, \ldots, H$.

To solve equations $\xi_i = N(J_i)$, we used a combination of quadrature methods to approximate the integral (see, e.g., Burden and Faires, 1993) and a numerical procedure to solve the equation, or a root finding algorithm for numerical inversion and using the implementation of the upper gamma incomplete function from the R package `expint`. Both approaches are bound to be subject to numerical stability over certain parameter ranges. The prior distribution on K_n is obtained by averaging over 2×10^2 Monte Carlo simulations and the density is smoothed using a Gamma density.

Appendix C. Comparison of approximations for NGG (extra plots)

Here we present in Figure 3 the prior distributions for K_n for a large sample size of $n = 1000$. Experiments are also carried out with a large truncation level of $H = 1000$ (for truncated and finite-dimensional approximations), see Figure 4. Comparing to similar Figure 1 and Figure 3, we

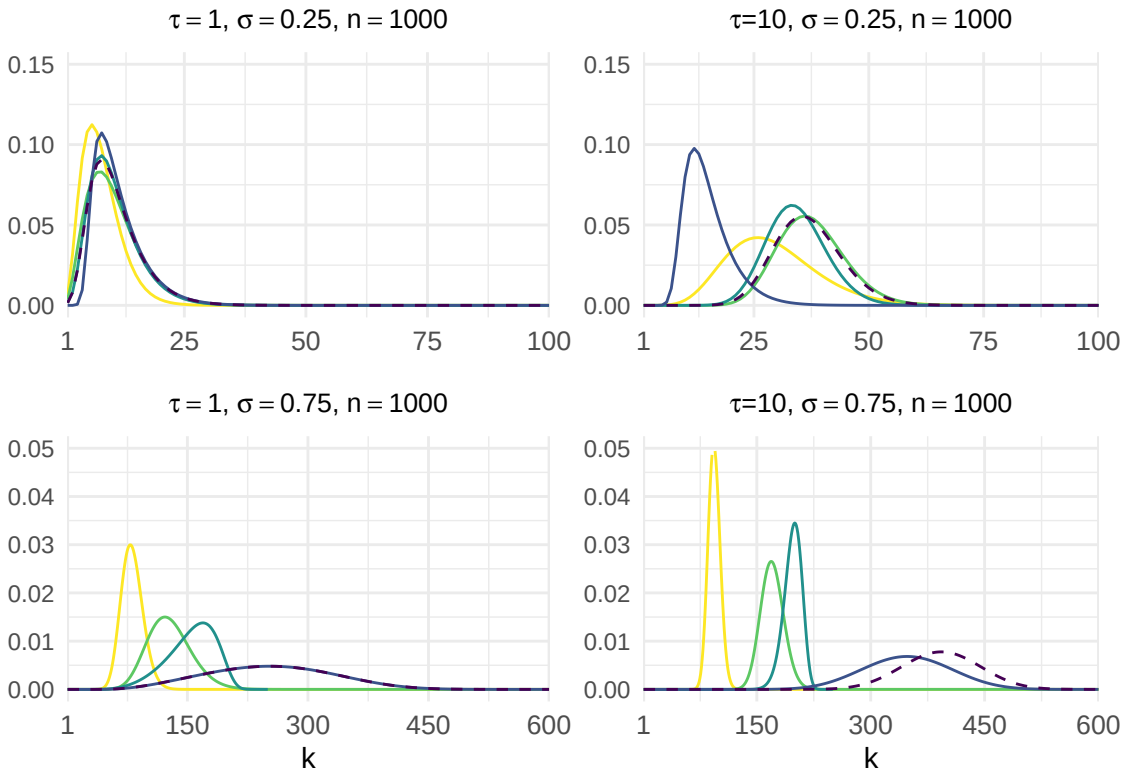


Figure 3: Prior distribution of K_n for exact NGG (dashed line), NGG predictive (dark blue), NGG multinomial (light blue), stick-breaking truncation (yellow), Ferguson–Klass truncation (green), for $n = 1000$ and truncation $H = 250$ for $\tau \in \{1, 10\}$ and $\sigma \in \{0.25, 0.75\}$.

can note that all the distributions that depend on H better approximate the true distribution. The quality of approximation deteriorates with growing sample size.

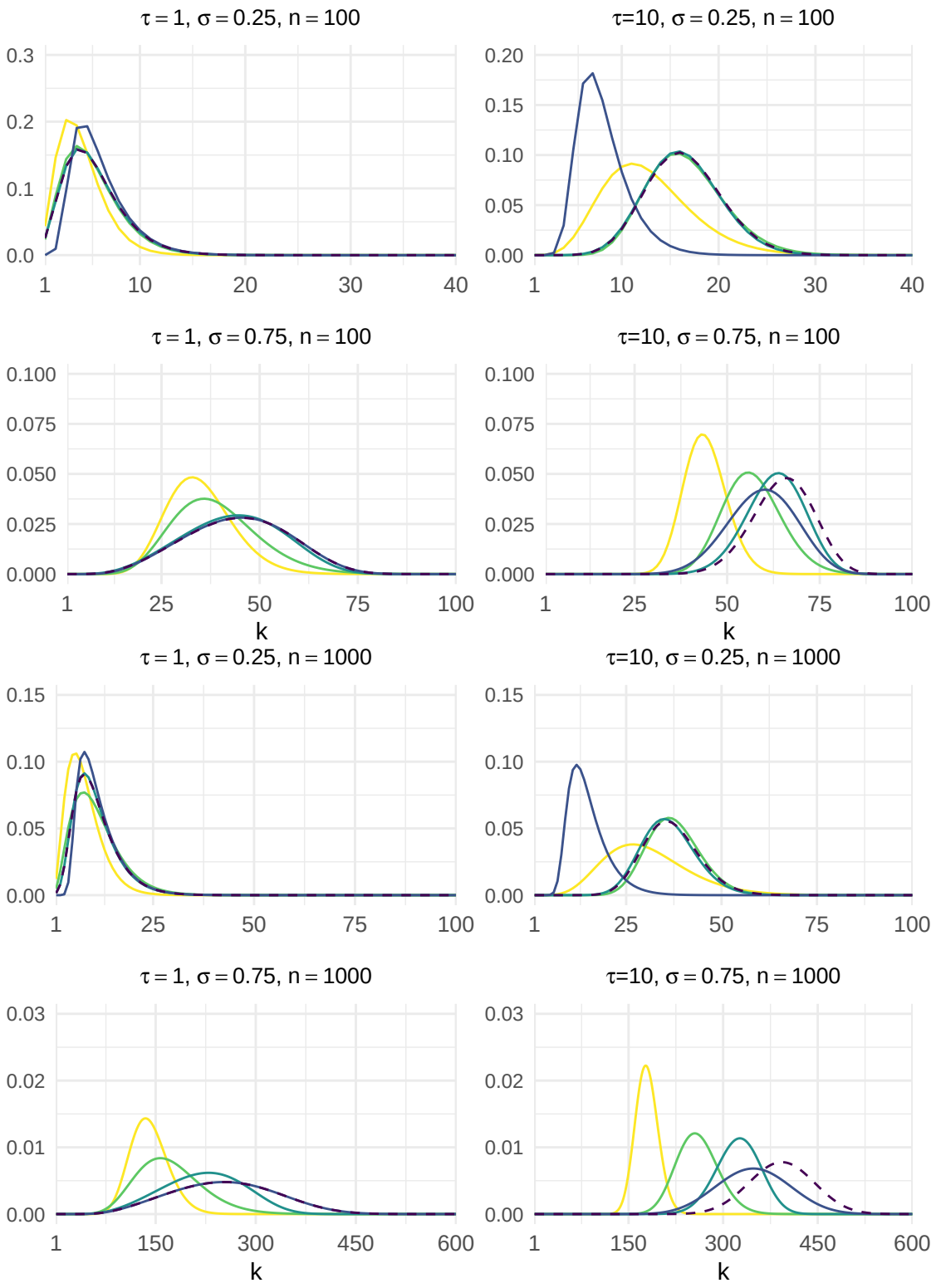


Figure 4: Prior distribution of K_n for exact NGG (dashed line), NGG predictive (dark blue), NGG multinomial (light blue), stick-breaking truncation (yellow), Ferguson–Klass truncation (green), for $n \in \{100, 1000\}$ and truncation $H = 1000$ for $\tau \in \{1, 10\}$ and $\sigma \in \{0.25, 0.75\}$.

Appendix D. Comparison of approximations for PY

D.1. Finite-dimensional approximation

For the PY multinomial process the prior number of clusters is given in Theorem 3 of Lijoi et al. (2020b) and it equals

$$\mathbb{P}(K_n = k) = \frac{H!}{(H-k)!\sigma(\alpha+1)_{n-1}} \sum_{l=k}^{n} \frac{1}{H^l} \frac{\Gamma(\alpha/\sigma + l)}{\Gamma(\alpha/\sigma + 1)} \mathscr{S}_{l,k} \mathcal{C}_{n,l}, \tag{23}$$

for any $k \leq \min\{\text{H,n}\}$, where $\mathscr{S}_{l,k}$ is the Stirling number of second kind.

D.2. Stick-breaking truncation

PY could be defined using stick-breaking construction. Consider two independent families of random variables:

$$V_k \overset{\text{ind}}{\sim} \text{Beta}(1-\sigma, \alpha + k\sigma) \qquad Z_k \overset{\text{iid}}{\sim} P_0 \qquad k = 1, 2, \ldots. \tag{24}$$

Define the random weights as:

$$p_1 := V_1, \qquad p_k := V_k \prod_{j=1}^{k-1} (1 - V_j). \tag{25}$$

Define

$$P = \sum_{k=1}^{\infty} p_k \delta_{Z_k}, \tag{26}$$

then $P \sim \text{PY}(\alpha, \sigma; P_0)$, i.e P is the PY process with concentration parameter α, discount (or diversity) parameter σ (with $0 \leq \sigma \leq 1$) and base measure P_0.

D.3. Expectation and variance of the PY process

In the case of PY process, expectation and variance for the prior number of clusters can be computed directly, without computing the probabilities $p_{n,k}$ for the prior distribution of K_n. We can define (α, σ) based on the values of $\mathbb{E}[K_n]$ and $\text{Var}[K_n]$, defined below. By the Pólya urn representation of PY, we have

$$\mathbb{P}(K_{n+1} = K + 1 \mid K_1, \ldots, K_n = K) = \frac{K\sigma + \alpha}{n + \alpha},$$
$$\mathbb{P}(K_{n+1} = K \mid K_1, \ldots, K_n = K) = \frac{n - K\sigma}{n + \alpha}$$

hence

$$\mathbb{E}[K_{n+1}] = \mathbb{E}[\mathbb{E}[K_{n+1} \mid K_1, \ldots, K_n = K]] = \frac{\alpha}{n+\alpha} + \frac{\sigma + \alpha + n}{n + \alpha} \mathbb{E}[K]. \tag{27}$$

Iterating yields

$$\mathbb{E}[K_n] = \sum_{i=1}^{n} \frac{(\alpha + \sigma)_{i-1}}{(\alpha + 1)_{i-1}}.$$

Using the law of total variance we have,

$$\text{Var}(K_{n+1}) = \mathbb{E}[\text{Var}(K_{n+1} \mid K_1, \ldots, K_n = K)] + \text{Var}(\mathbb{E}[K_{n+1} \mid K_1, \ldots, K_n = K]). \tag{28}$$

Using the fact that the variance of a two-point distribution with outcomes a and b and probabilities p and $1-p$ is $p(1-p)(a-b)^2$, we can write the first term of the right-hand side of (28) as

$$\begin{aligned}\mathbb{E}[\mathrm{Var}(K_{n+1} \mid K_1, \ldots, K_n = K)] &= \mathbb{E}\left[\frac{(K\sigma+\alpha)(n-K\sigma)}{(n+\alpha)^2}\right] \\ &= \frac{\mathbb{E}[K_n]\sigma(n-\alpha)+n\alpha-\sigma^2(\mathbb{E}[K_n]^2+\mathrm{Var}(K_n))}{(n+\alpha)^2}, \end{aligned} \tag{29}$$

while (27) provides for second term of the right-hand side of (28)

$$\mathrm{Var}(\mathbb{E}[K_{n+1} \mid K_1, \ldots, K_n = K]) = \left(\frac{\sigma+n+\alpha}{n+\alpha}\right)^2 \mathrm{Var}(K_n). \tag{30}$$

Then combining (29) and (30), the following recursive formula holds

$$\mathrm{Var}(K_{n+1}) = \mathrm{Var}(K_n)\frac{n+\alpha+2\sigma}{n+\alpha} + \frac{\sigma\mathbb{E}[K_n](n-\alpha\sigma\mathbb{E}[K_n])+n\alpha}{(n+\alpha)^2}.$$

D.4. Comparison of approximations

We compare the prior distribution for K_n under the PY process and its approximations described above. Note that PY is not an NRMI (Regazzini et al., 2003), so as such it cannot be sampled according to Ferguson–Klass algorithm.

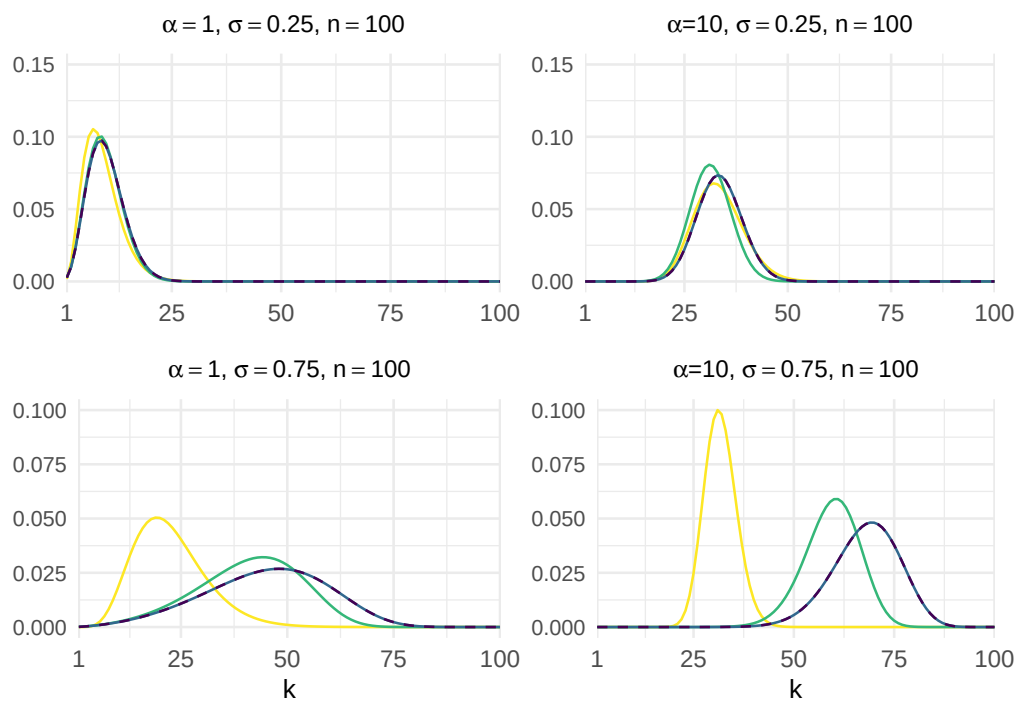
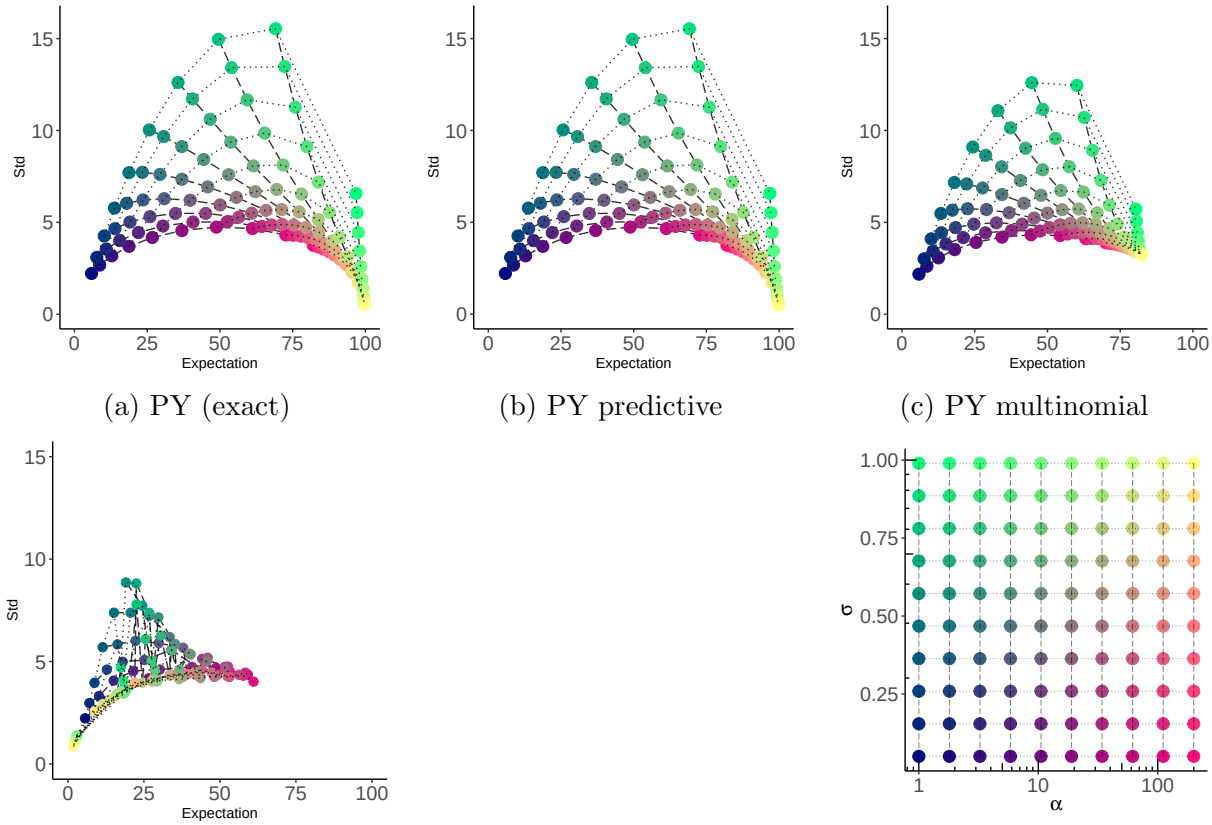


Figure 5: Distribution of the number of clusters for exact PY (dashed line), PY predictive (dark blue), PY multinomial (green), stick-breaking truncation (yellow), for $n = 100$ and truncation $H = 250$ (for truncated and finite dimensional representations) for different values of parameters $\alpha \in \{1, 10\}$ and $\sigma \in \{0.25, 0.75\}$.



(a) PY (exact) (b) PY predictive (c) PY multinomial

(d) PY Stick-breaking (e) No Ferguson–Klass for PY (f) Parameter (α, σ) legend

Figure 6: Expectation and standard deviation for the distribution of the prior number of clusters, for (a) exact PY, (b) predictive approximation, (c) PY multinomial process, (d) stick-breaking truncation, (f) as a function of parameters (σ, α) spanning $(0, 1) \times [1, 200]$ (log scale in α) for $n = 100$ samples and truncation $H = 250$.

Figure 6 shows the relationship between the expected number of clusters and standard deviation for the PY process. This figure is similar to Figure 2 for NGG and for comparison we use the same grid for parameters (α, σ) as the (τ, σ) grid of NGG, and the same values for $n = 100$, $H = 250$.