Bayesian nonparametric priors for hidden Markov random fields

Hongliang Lu, Julyan Arbel, Florence Forbes

Noname manuscript No.
(will be inserted by the editor)

Bayesian nonparametric priors for hidden Markov random fields

Hongliang Lü · Julyan Arbel · Florence Forbes

the date of receipt and acceptance should be inserted later

Abstract One of the central issues in statistics and machine learning is how to select an adequate model that can automatically adapt its complexity to the observed data. In the present paper, we focus on the issue of determining the structure of clustered data, both in terms of finding the appropriate number of clusters and of modelling the right dependence structure between the observations. Bayesian nonparametric (BNP) models, which do not impose an upper limit on the number of clusters, are appropriate to avoid the required guess on the number of clusters but have been mainly developed for independent data. In contrast, Markov random fields (MRF) have been extensively used to model dependencies in a tractable manner but usually reduce to finite cluster numbers when clustering tasks are addressed. Our main contribution is to propose a general scheme to design tractable BNP-MRF priors that combine both features: no commitment to an arbitrary number of clusters and a dependence modelling. A key ingredient in this construction is the availability of a stick-breaking representation which has the three-fold advantage to allowing us to extend standard discrete MRFs to infinite state space, to design a tractable estimation algorithm using variational approximation and to derive theoretical properties on the predictive distribution and the number of clusters of the proposed model. This approach is illustrated on a challenging natural image segmentation task for which it shows good performance with respect to the literature.

Keywords Hidden Markov random fields · Bayesian nonparametrics · Variational approximation · Clustering · Image segmentation · Predictive distribution

1 Introduction

Hidden Markov random field (HMRF) models are widely used for clustering data under spatial constraints. Spatial dependencies are encoded by modelling the cluster labels as a discrete state Markov random field (MRF) such as Ising (two clusters or states) or Potts (more than two clusters) model [10, 32]. HMRF can be seen as spatial extensions of independent mixture models. As for standard mixtures, one concern is the automatic selection of the proper number of clusters in the data, or equivalently the number of states in the HMRF. In the independent data case, several criteria exist to select this number automatically based on penalized likelihoods (*e.g.*, AIC, BIC, ICL, etc.) and have been extended in the HMRF framework using variational approximation [17]. They require running several models with different cluster numbers so as to choose the best one, with a potential waste of computational effort as all the other models are usually discarded. Other techniques use a fully Bayesian setting including a prior on the number of components. The most celebrated method in this case is reversible jump Markov chain Monte Carlo [19]. Although simplifications in the inference have been proposed recently in [23], the computational cost of reversible jump techniques remains considerably high.

In the present work, we investigate alternatives based on Bayesian nonparametric (BNP) methods. In particular, Dirichlet process mixture (DPM) models have emerged as promising candidates for clustering applications where the number of clusters is unknown. Nevertheless, applications of DPMs involve observations which are assumed to be independent. For more complex tasks such as unsupervised image segmentation with spatial relationships or dependencies between the observations, DPMs are not satisfactory. Therefore, we propose to

H. Lü · J. Arbel · F. Forbes
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
E-mail: {julyan.arbel, florence.forbes}@inria.fr

introduce MRF dependencies between data points in BNP models, and we term the resulting model BNP-MRF. This requires to extend finite state space MRF models to an infinite number of states. We show that this can be achieved by incorporating a stick-breaking scheme in an MRF formulation more general than the standard Potts model commonly used.

The addition of MRF dependencies between data points in BNP models raises the question of how they impact the natural clustering and rich-get-richer properties of BNP priors? We answer this question by providing theoretical results about two quantities of interest for BNP priors: the predictive distribution, that represents the distribution of one datum conditional on previous observations, and the number of clusters induced by a BNP-MRF prior.

The links to other similar attempts is reviewed in Section 2. The proposed BNP-MRF model is explained in Section 3 and theoretical properties are investigated in Section 4. The model implementation using variational approximation is detailed in Section 5. An illustration of its performance on an image segmentation task is provided in Section 6 and a conclusion ends the paper.

2 Related work

Attempts to build countably infinite state space MRF models using BNP priors have already appeared in the literature. In particular, we can distinguish attempts such as [11, 12, 30] from the work in [2, 26, 38, 31]. The approach in [11, 12, 30] differs in that it is not based on a generalization of the Potts model but on a transformation of an inference algorithm. More specifically in [11, 12], a standard mean field approximation is first considered and then transformed to account for an infinite number of states. In that sense it is closer to an Iterated Conditional Mode (ICM) algorithm [6], but does not provide a spatial generalization of DPMs. Typically, the simple Potts model considered in [11, 12] cannot be extended to an infinite number of states as it will become clear in our Section 3.2. Other attempts include the work in [21], but there the number of states is known to be three and the Dirichlet process (DP) prior is used instead to model intensity distributions non-parametrically. Segmentation with spatially dependent Pitman–Yor processes (PY) has also been considered in [33], but using Gaussian processes.

We build on the approach in [2] which differs from [26, 38, 31] which all use a partition model representation. In particular, [38] generalizes [26] and proposes a more efficient Markov chain Monte Carlo (MCMC) inference by means of the Swendsen–Wang algorithm, while [31] extends this idea to hierarchical DP priors for multiple image segmentation. In contrast to [26, 38, 31], we propose to use a stick-breaking-based scheme for the mixing weights, thus providing a more comprehensive representation than partition models which integrate out the process. In addition, stick-breaking representations lead naturally to variational approximations for performing inference [7]. The advantage is to reduce the computational cost in complex data clustering without suffering from label switching complications. In other words, in our approach the MRF is imposed internally in the BNP mechanics leading to well defined infinite state HMRF models. This construction is valid for any stick-breaking representation. We show how it can be implemented for the DP and PY priors, and provide references for extensions to larger classes of BNP priors.

3 BNP-MRF mixture models

The clustering task is addressed through a missing data model that includes a set $\mathbf{y} = (y_1, \ldots, y_n)$ of observed variables from $\mathbb{R}^d$ and a set $\mathbf{z} = (z_1, \ldots, z_n)$ of missing (also called hidden) variables whose joint distribution $p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\Theta})$ is governed by a set of parameters denoted by $\boldsymbol{\Theta}$ and possibly by additional hyperparameters $\boldsymbol{\phi}$ not specified in the notation. The latter ones are usually fixed and not considered at first. Typically, the z_i's corresponding to group memberships (or labels), take their values in $\{1, \ldots, K\}$ where K is the number of clusters or groups. We shall denote by $\mathcal{Z} = \{1, \ldots, K\}^n$ the set in which $\mathbf{z}$ takes its values and by $\underline{\Theta}$ the parameter space. To account for dependencies between the z_i's, $\mathbf{z}$ can be modeled as a discrete MRF. If in addition, the y_j's are independent conditionally on $\mathbf{z}$, the joint distribution $p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\Theta})$ is referred to as an HMRF model. In this case, the conditional distribution $p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\Theta})$ is also an MRF. For clustering dependent data into K groups, the most commonly used MRF is the so-called Potts model [10, 32].

As already mentioned, our goal is to bypass the issue of selecting the number K of clusters by considering a countably infinite number of them while allowing MRF dependencies between the $\mathbf{y}_i$'s. The construction of the proposed model is explained starting from the link between standard finite mixtures and Dirichlet process mixtures. Basic DP principles and notations are recalled in Section 3.1. The extension of finite state space MRF to a countably infinite number of states is given in Section 3.2 and the resulting BNP-MRF mixture models is summarized in Section 3.3.

3.1 From finite mixtures to DP mixture models

A generative approach to clustering consists of picking one of K clusters from a multinomial distribution with weights parameter $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ and then to generate a data point y from a cluster specific distribution $p(y \mid \theta_k^*)$ with cluster specific parameter θ_k^*. This yields a finite mixture model

$$p(y \mid \boldsymbol{\theta}^*, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \; p(y \mid \theta_k^*) \tag{1}$$

where $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_K^*)$ and $\boldsymbol{\pi}$ are the parameters. For instance, for Gaussian mixtures, $\theta_k^* = (\mu_k, \Sigma_k)$ and $p(y \mid \theta_k^*)$ is a Gaussian distribution with mean μ_k and covariance matrix Σ_k, denoted by $\mathcal{N}(\mu_k, \Sigma_k)$ or $\mathcal{N}(y \mid \mu_k, \Sigma_k)$ when referring to the probability density function (pdf). The observations $(y_1, \dots, y_n)$ are therefore *i.i.d.* and generated from the same mixture (1). It follows that the kth cluster is by definition the set of data points arising from the kth mixture component. This is usually expressed by introducing for each y_j an additional hidden variable Z_j that takes its values in $\{1, \dots, K\}$, so that $p(z_j = k \mid \boldsymbol{\pi}) = \pi_k$. Another way to obtain a sample from a finite mixture model consists of defining a discrete measure $G = \sum_{k=1}^{K} \pi_k \delta_{\theta_k^*}$ and then of considering the following hierarchical representation, for all $j = 1, \dots, n$,

$$\begin{aligned} &\theta_j \mid G \overset{\text{iid}}{\sim} G, \\ &y_j \mid \theta_j \overset{\text{ind}}{\sim} p(\,.\mid \theta_j). \end{aligned}$$

The subset of θ_j's that are equal to θ_k^* corresponds to the y_j's in the kth cluster.

In a Bayesian setting, in addition, a prior distribution is placed on $\boldsymbol{\theta}^*$ and $\boldsymbol{\pi}$. The most common choice for $\boldsymbol{\pi}$ is the Dirichlet distribution $\mathrm{Dir}(\alpha_1, \dots \alpha_K)$ depending on a vector of positive parameters $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$. The choice of the prior on $\boldsymbol{\theta}^*$ (denoted by G_0) is model-specific, usually following a conjugate prior such as a Normal inverse-Wishart distribution for Gaussian mixture models. Other cases are possible and tractable (*e.g.* [13]). It follows the hierarchical representation:

$$\theta_1^*, \dots, \theta_K^* \mid G_0 \sim G_0, \tag{2}$$

$$\boldsymbol{\pi} \mid \boldsymbol{\alpha} \sim \mathrm{Dir}(\alpha_1, \dots \alpha_K), \tag{3}$$

$$G = \sum_{k=1}^{K} \pi_k \delta_{\theta_k^*}, \tag{4}$$

$$\begin{aligned} &\theta_j \mid G \overset{\text{iid}}{\sim} G,\; j = 1, \dots, n, \\ &y_j \mid \theta_j \overset{\text{ind}}{\sim} p(\,.\mid \theta_j)\; j = 1, \dots, n. \end{aligned} \tag{5}$$

To become non-parametric, a first approach is to consider an infinite number of π_k's. Using an infinite number of random variables $\boldsymbol{\tau} = (\tau_1, \tau_2, \dots)$ on $[0, 1]$, we can construct an infinite number of π_k's that sum to one as follows:

$$\pi_1(\boldsymbol{\tau}) = \tau_1 \quad \text{and} \quad \pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l), \quad k = 2, 3, \dots$$

The intuition behind this construction, referred to as *stick-breaking*, is that it consists of recursively breaking a unit-length stick as shown in Fig. 1. It follows an explicit formula for the π_k's. Hence, the τ_k's simulation replaces step (3), and G in (4) can be replaced by

$$G = \sum_{k=1}^{\infty} \pi_k(\boldsymbol{\tau}) \delta_{\theta_k^*}.$$

We can also add after step (5) the fact that $z_j = k$ if $\theta_j = \theta_k^*$ and replace the last step by $y_j | z_j, \boldsymbol{\theta}^* \overset{\text{ind}}{\sim} p(\,.\mid \theta_{z_j}^*)$. Then the distributions of the τ_k's need to be specified. The Dirichlet process [16], denoted by $\mathrm{DP}(G_0, \alpha)$, is characterized by a base distribution G_0 and a positive scaling parameter α. Its stick-breaking representation

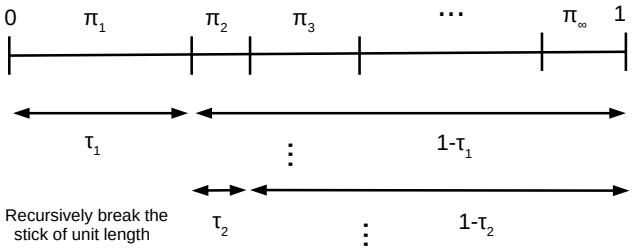

Fig. 1 Illustration of the stick-breaking representation.

corresponds to *i.i.d* τ_k's that follow the same beta $\mathcal{B}(1,\alpha)$ distribution [20]. All together, using the same notation G_0 for the prior of each θ_k^* simulated as *i.i.d.* variables, it comes the following hierarchical representation:

$$\theta_k^* \mid G_0 \overset{\text{iid}}{\sim} G_0, \quad k = 1, 2, \ldots, \tag{6}$$

$$\tau_k \mid \alpha \overset{\text{iid}}{\sim} \mathcal{B}(1,\alpha), \quad k = 1, 2, \ldots,$$

$$\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1}(1-\tau_l), \quad k = 1, 2, \ldots, \tag{7}$$

$$G = \sum_{k=1}^{\infty} \pi_k(\tau)\delta_{\theta_k^*}, \tag{8}$$

$$\theta_j \mid G \overset{\text{iid}}{\sim} G, \text{ and } z_j = k \text{ if } \theta_j = \theta_k^* \tag{9}$$

$$y_j \mid z_j, \boldsymbol{\theta}^* \overset{\text{ind}}{\sim} p(\ . \mid \theta_{z_j}^*). \tag{10}$$

The above hierarchical representation corresponds to a countably infinite mixture model referred to as a Dirichlet process mixture (DPM) model. It is an explicit characterization of the DP (Eq. (6) to (8)) and of the DPM (Eq. (6) to Eq. (10)) using a stick-breaking construction. The stick-breaking representation will be particularly useful in our study for both the definition of our model (Sections 3.2 and 3.3) and its estimation (Section 5).

3.2 Infinite MRF priors

The explicit use of the labels $\mathbf{z} = (z_1, \ldots, z_n)$ in the DPM construction above makes it closer to clustering generative models and opens the way to an HMRF extension. Such a generalization is only possible from Potts models with an external field parameter. In the finite state space case, an MRF model is defined using a dependence structure coded via a graph $\mathcal{G}$ whose nodes correspond to the variables. A K-state Potts model with an external field, defined over $\mathbf{z} = (z_1, \ldots, z_n)$ with for all $j = 1, \ldots, n$, $z_j \in \{1, \ldots, K\}$, corresponds to the following pdf,

$$p(\mathbf{z}; \beta, \boldsymbol{v}) \propto \exp\left(\sum_{j=1}^{n} v_{z_j} + \beta \sum_{i \sim j} \delta_{(z_i = z_j)}\right), \tag{11}$$

where $i \sim j$ means that i and j are neighbors, *i.e.* linked by an edge, in the considered dependence structure described by graph $\mathcal{G}$, $\delta_{(z_i=z_j)}$ is the indicator function which is 1 if $z_i = z_j$ and 0 otherwise, β is a positive scalar interaction parameter and $\boldsymbol{v} = (v_1, \ldots, v_K)$ represents an additional external field parameter where each v_k is a scalar. The distribution (11) is insensitive to an addition of the same constant to all the v_k's. Such non-identifiability can be overcome by an additional constraint on $\boldsymbol{v}$ such as requiring $\sum_{k=1}^{K} \pi_k = 1$ with $v_k = \log \pi_k$. The Potts model in (11) can then be rewritten as

$$p(\mathbf{z}; \beta, \boldsymbol{\pi}) \propto \left(\prod_{j=1}^{n} \pi_{z_j}\right) \exp\left(\beta \sum_{i \sim j} \delta_{(z_i = z_j)}\right). \tag{12}$$

In the finite state space case, we can equivalently use the Gibbs representation,

$$p(\mathbf{z}; \beta, \boldsymbol{\pi}) \propto \mathrm{e}^{V(\mathbf{z}; \beta, \boldsymbol{\pi})}, \tag{13}$$

where $V(\mathbf{z};\beta,\boldsymbol{\pi}) \coloneqq \sum_{j=1}^{n} \log \pi_{z_j} + \beta \sum_{i\sim j} \delta_{(z_i=z_j)}$ is often referred to as the *energy function*. The first sum in V represents the first order potentials while the second sum represents the second order potentials. In the finite state space case, the Hammersley–Clifford theorem [6] applied to the Gibbs representation (13) entails that the distribution in (11) is a Markov random field. What is interesting about formulas (11) and (12) is that they do not involve the number of states K. As long as a stick-breaking construction is available, we can consider a countably infinite number of probabilities π_k that sum to one, *i.e.*, $\sum_{k=1}^{\infty} \pi_k = 1$ and define the same energy function V as before but over an infinite countable set of states. Using the Gibbs representation (13), the Hammersley–Clifford theorem still holds if we can show that $\sum_{\mathbf{z}} \mathrm{e}^{V(\mathbf{z};\boldsymbol{\pi},\beta)} < \infty$, where the sum runs over all n-uples of positive integers $\mathbf{z} \in \{1,2,\ldots\}^n$. Note that this latter condition that is automatically satisfied in the finite state space case (for reasonable potential functions), may not be satisfied in the infinite case. However, the stick-breaking representation of $\boldsymbol{\pi}$ ensures this property since:

$$\sum_{\mathbf{z}} \mathrm{e}^{V(\mathbf{z};\beta,\boldsymbol{\pi})} \overset{(a)}{\leq} \left(\sum_{\mathbf{z}} \prod_{j=1}^{n} \pi_{z_j} \right) \mathrm{e}^{\beta \frac{n(n-1)}{2}}, \\ \overset{(b)}{=} \mathrm{e}^{\beta \frac{n(n-1)}{2}} < \infty$$

where we used for (a) the fact that $n(n-1)/2$ is the maximum number of neighbors among n observations (complete dependence or graph), while (b) comes from $\sum_{\mathbf{z}} \prod_{j=1}^{n} \pi_{z_j} = \left(\sum_{k=1}^{\infty} \pi_k\right)^n = 1$. It follows that $p(\mathbf{z};\beta,\boldsymbol{\pi})$, in the infinite state space case, is still a valid probability distribution and is an MRF by the Hammersley–Clifford theorem. Such a generalization is possible because of the presence of the external field parameters π_k that satisfy $\sum_{k=1}^{\infty} \pi_k = 1$ as ensured by the stick-breaking construction. A standard Potts model with equal or no external field parameters cannot be as simply extended to an infinite countable state space because in the K-state case this Potts model is equivalent to $\pi_k = 1/K$ for all k which possesses a degenerate limit when K tends to infinity.

3.3 BNP-MRF mixture models

The stick-breaking representation amounts to identifying a set of random variables $\boldsymbol{\tau} = (\tau_k)_{k=1}^{\infty}$ with each $\tau_k \in [0,1]$ and so that the weights π_k are defined by (7). Then the Potts model construction (12) is valid for any set of parameters $\boldsymbol{\tau} = (\tau_k)_{k=1}^{\infty}$ with each $\tau_k \in [0,1]$. Bayesian non-parametric priors specify a prior distribution on τ_k's. For instance, as already mentioned for the DP stick-breaking, all τ_k's are independent and identically distributed according to a $\mathcal{B}(1,\alpha)$ distribution. For the Pitman–Yor (PY) process [28], the τ_k's are independent but not identically distributed with

$$\tau_k \mid \alpha, \sigma \overset{\text{ind}}{\sim} \mathcal{B}(1-\sigma, \alpha+k\sigma) \quad \text{for } k = 1,2,\ldots, \tag{14}$$

where $\sigma \in (0,1)$ is a discount parameter and α a concentration parameter $\alpha > -\sigma$. The PY is a two-parameter generalisation of the DP which allows to control the tail behavior when modeling data with either exponential or power-law tails [20,28]. When $\sigma = 0$, the PY reduces to a DP. More general stick-breaking representations are possible (*e.g.*, for Gibbs-type priors [14,18] or homogeneous normalized random measures with independent increments (NRMIs) [15]) but the Pitman–Yor case provides a clear interpretation in terms of number of clusters. The rich-gets-richer property of the DP is preserved meaning that there are a small number of large clusters, but there is also a large number of small clusters with parameter σ decreasing the probability that observations join small clusters. The PY yields a power-law behavior which can make it more suitable for a number of applications. In other words, the number of clusters grows as $\mathcal{O}(n^{\sigma})$ for the PY while it grows more slowly at $\mathcal{O}(\log n)$ for the DP.

The extension we propose is therefore to augment the original HMRF formulation with additional variables $(\tau_k)_{k=1}^{\infty}$. We refer to it as the BNP-MRF mixture model. It corresponds to the following hierarchical construction

written here in the PY case:

$$\theta_k^* \mid G_0 \overset{\text{iid}}{\sim} G_0,\ k = 1, 2, \ldots, \tag{15}$$

$$\tau_k \mid \alpha, \sigma \overset{\text{ind}}{\sim} \mathcal{B}(1-\sigma, \alpha + k\sigma),\ k = 1, 2, \ldots, \tag{16}$$

$$\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l), \tag{17}$$

$$p(\mathbf{z} \mid \boldsymbol{\tau}; \beta) \propto \left(\prod_{j=1}^{n} \pi_{z_j}(\boldsymbol{\tau}) \right) \exp\left(\beta \sum_{i \sim j} \delta_{(z_i = z_j)} \right), \tag{18}$$

$$y_j \mid z_j, \boldsymbol{\theta}^* \overset{\text{ind}}{\sim} p(y_j \mid \theta^*_{z_j})\ . \tag{19}$$

The prior on τ_k's from (16) can be adapted to more general classes of BNP priors, see for example Theorem 14.23 of [18] for Gibbs-type priors, and [15] for NRMIs. Importantly, in the BNP-MRF model above, the θ_j's and z_j's are not *i.i.d* conditionally on G anymore. The joint distribution (18) on $\mathbf{z} = (z_1, \ldots, z_n)$ induces a joint distribution on $(\theta_1, \ldots, \theta_n)$ using that $\theta_j = \theta^*_{z_j}$. If we still denote for simplicity by G this joint distribution, we can define it in a similar manner as in the *i.i.d.* case, using its conditional specifications,

$$\theta_j | \theta_{\mathcal{N}_j}; G \sim \sum_{k=1}^{\infty} p(z_j = k | z_{\mathcal{N}_j}, \boldsymbol{\tau}; \beta)\ \delta_{\theta^*_k},$$

where $\mathcal{N}_j$ denotes the neighbors of j in the graph dependence structure $\mathcal{G}$ and the $p(z_j = k | z_{\mathcal{N}_j}, \boldsymbol{\tau}; \beta)$'s are the conditional specifications of (18).

In Section 5.2, we detail the case when cluster specific distributions are Gaussian, with $\theta^*_k = (\mu_k, \Sigma_k)$ and $p(y_j \mid \theta^*_k) = \mathcal{N}(y_j \mid \mu_k, \Sigma_k)$.

4 Predictive distribution and number of clusters for a BNP-MRF prior

In this section, we provide theoretical results about two quantities of interest for Bayesian nonparametric priors: the predictive distribution, that represents the distribution of one datum conditional on previous observations, and the number of clusters induced by a BNP-MRF prior. We consider data of varying sample size, and denote by $\mathcal{G}_n$ the subgraph of $\mathcal{G}$ induced by node $\{1, \ldots, n\}$.

We focus on the large class of Gibbs-type priors [14], of which the DP and PY are special cases. Consider n observations $(\theta_1, \ldots, \theta_n)$ sampled from a BNP-MRF prior Eq (15)-(18) but using a Gibbs-type prior instead of PY prior (16). We are interested in the predictive distribution of observation θ_{n+1} conditional on $(\theta_1, \ldots, \theta_n)$, but unconditional on G. With a BNP-MRF prior, this predictive distribution depends on the structure of the graph $\mathcal{G}$, more specifically on the neighbors of θ_{n+1}. Denote by K_n the number of clusters in $(\theta_1, \ldots, \theta_n)$, by $(\theta^\star_1, \ldots, \theta^\star_{K_n})$ their K_n different values[1] and by $(n_1, \ldots, n_{K_n})$ their size. We first consider the Gibbs-type prior case without the addition of a Markov component. The predictive distribution [18] is given by,

$$p(\theta_{n+1} \mid \theta_1, \ldots, \theta_n) = \frac{V_{n+1,K_n+1}}{V_{n,K_n}} G_0 + \frac{V_{n+1,K_n}}{V_{n,K_n}} \sum_{\ell=1}^{K_n} (n_\ell - \sigma) \delta_{\theta^\star_\ell} \tag{20}$$

where the triangular array of nonnegative parameters $V_{n,k}$, $1 \leq k \leq n$, satisfy the backward recurrence relation

$$V_{n,k} = (n - \sigma k) V_{n+1,k} + V_{n+1,k+1}, \tag{21}$$

with $V_{1,1} = 1$. This predictive can be specialized to the PY case with

$$V_{n,k} = \frac{\sigma^k (1 + \frac{\alpha}{\sigma})_{(k-1)}}{(1+\alpha)_{(n-1)}},$$

where $(a)_{(x)} \coloneqq \Gamma(a+x)/\Gamma(a)$ denotes the rising factorial. It follows

$$p(\theta_{n+1} \mid \theta_1, \ldots, \theta_n) = \frac{\alpha + \sigma K_n}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{\ell=1}^{K_n} (n_\ell - \sigma) \delta_{\theta^\star_\ell}, \tag{22}$$

[1] Note that the notation introduced for the different $\theta^\star_j$ differs from that devoted to the stick-breaking variables, θ^*_j.

while the case of the DP is obtained by setting $\sigma = 0$ above.

For the sake of simplicity, we propose to use the labels notation $z_{1:n} = (z_1, \ldots, z_n)$ defined so that $z_j = \ell$ when $\theta_j = \theta^\star_\ell$, we denote by $\{z_1, \ldots, z_n\}$ the set of label values which includes only K_n different labels. In the Gibbs-type prior case, it is clear from (20) that

$$p(z_{n+1} \mid z_{1:n}) = \frac{V_{n+1,K_n+1}}{V_{n,K_n}} \quad \text{if } z_{n+1} \notin \{z_1, \ldots, z_n\}, \tag{23}$$
$$p(z_{n+1} = \ell \mid z_{1:n}) = \frac{V_{n+1,K_n}}{V_{n,K_n}}(n_\ell - \sigma) \quad \text{if } \ell \in \{z_1, \ldots, z_n\}\ .$$

The next proposition indicates how the predictive is impacted by the addition of a Markov dependence. The neighbors of θ_{n+1} in $\mathcal{G}_n$ is denoted by $\mathcal{N}_{n+1}$ and $\tilde{n}_\ell$ is the number of neighbors of θ_{n+1} which belong to cluster ℓ, hence satisfying $\tilde{n}_\ell \leq n_\ell$. Also $z_{\mathcal{N}_{n+1}} = \{z_i, i \in \mathcal{N}_{n+1}\}$ denotes the labels in the neighborhood. The proof of the proposition is given in Appendix A.1.

Proposition 1 (Predictive distribution of a Gibbs-MRF prior) *The predictive distribution for a Gibbs-MRF prior is given by*

$$p(\theta_{n+1} \mid \theta_1, \ldots, \theta_n) = \frac{V_{n+1,K_n+1}}{V_{n,K_n} + V_{n+1,K_n}\boldsymbol{\eta}_{n+1}} G_0 + \frac{V_{n+1,K_n}}{V_{n,K_n} + V_{n+1,K_n}\boldsymbol{\eta}_{n+1}} \sum_{\ell=1}^{K_n} \boldsymbol{\lambda}_{n+1,\ell}\ \delta_{\theta^\star_\ell} \tag{24}$$

where

$$\boldsymbol{\eta}_{n+1} = \boldsymbol{\eta}_{n+1}(\sigma, \beta) = \sum_{\ell \in z_{\mathcal{N}_{n+1}}} (n_\ell - \sigma)(\mathrm{e}^{\beta\tilde{n}_\ell} - 1),$$

$$\boldsymbol{\lambda}_{n+1,\ell} = \boldsymbol{\lambda}_{n+1,\ell}(\sigma, \beta) = (n_\ell - \sigma)\ \mathrm{e}^{\beta\tilde{n}_\ell \delta_{\mathcal{N}_{n+1}}(\ell)}.$$

and $\delta_{\mathcal{N}_{n+1}}(\ell)$ is 1 when ℓ is a label present in the neighborhood of θ_{n+1} and 0 otherwise.

Remark 1 When $\beta = 0$, $\boldsymbol{\eta}_{n+1}(\sigma, 0) = 0$ and $\boldsymbol{\lambda}_{n+1,\ell}(\sigma, 0) = n_\ell - \sigma$ so that the Gibbs-type prior predictive (20) is recovered. In contrast, for $\beta > 0$, the above predictive specialized to the PY-MRF case is,

$$p(\theta_{n+1} \mid \theta_1, \ldots, \theta_n) = \frac{\alpha + \sigma K_n}{\alpha + n + \boldsymbol{\eta}_{n+1}} G_0 + \frac{1}{\alpha + n + \boldsymbol{\eta}_{n+1}} \sum_{\ell=1}^{K_n} \boldsymbol{\lambda}_{n+1,\ell}\ \delta_{\theta^\star_\ell}, \tag{25}$$

while the case of the DP-MRF is obtained by setting $\sigma = 0$. Comparing the probability of a new draw for a Gibbs-type prior, $\frac{V_{n+1,K_n+1}}{V_{n,K_n}}$, with that of a new draw for a Gibbs-MRF prior, $\frac{V_{n+1,K_n+1}}{V_{n,K_n}+V_{n+1,K_n+1}\boldsymbol{\eta}_{n+1}}$, we see that the MRF has the effect of reducing this probability. In the PY case, this increase corresponds to increasing the sample size from n to $n+\boldsymbol{\eta}_{n+1}$ when $\beta > 0$, where $\boldsymbol{\eta}_{n+1}$ can be quite a large number. More specifically for a label ℓ in the neighborhood of z_{n+1}, the weight of each previous observations with label ℓ (in the neighborhood or not) is multiplied by a factor $(\mathrm{e}^{\beta\tilde{n}_\ell} - 1)$. The effect is then all the more important as β is large and as n_ℓ is large.

The predictive distribution (24) provides in turn the following lower bounds on the prior expectation of the number of clusters. The proof of Proposition 2 is given in Appendix A.2.

Proposition 2 (Lower bound for expected number of clusters) *Assume that the graph $\mathcal{G}$ has maximal degree D. Then the expected prior number of clusters for a BNP-MRF distribution has the following lower bound*

$$\mathbb{E}[K_n] \gtrsim \frac{\alpha}{\mathrm{e}^{D\beta}} \log n \tag{26}$$

for the Dirichlet process and

$$\mathbb{E}[K_n] \gtrsim c n^{\sigma \mathrm{e}^{-D\beta}}, \tag{27}$$

for the Pitman–Yor process, with some positive constant c, and where $a_n \gtrsim b_n$ stands for $\limsup a_n/b_n \geq 1$.

Remark 2 We do not have a proof for the general case of Gibbs-type priors, but we conjecture that the same power-law lower bound (27) as for PY holds.

Note that the MRF component of a BNP prior can only *reduce* the prior expected number of clusters. For instance, for the DP with a simple graph where the first two nodes are connected, we have

$$\mathbb{E}[K_2] = 1 + \frac{\alpha}{\alpha + \mathrm{e}^{\beta}} \leq 1 + \frac{\alpha}{\alpha + 1} = \mathbb{E}[K_2; \beta = 0]$$

where the last two terms above correspond the expectation of K_2 for a DP, *i.e.* when $\beta = 0$. Thus natural upper bounds that complement the lower bounds of Proposition 2 are given by

$$\mathbb{E}[K_n] \lesssim \alpha \log n$$

for the Dirichlet process and

$$\mathbb{E}[K_n] \lesssim \frac{\Gamma(\alpha + 1)}{\sigma \Gamma(\alpha + \sigma)} n^{\sigma} \tag{28}$$

for the Pitman–Yor process (see [27]).

5 Inference using Variational approximation

Sampling based inference (MCMC) for a similar BNP-MRF model has been proposed in [26,38] for the case of a DP prior. As an alternative, we propose a variational approximation that is facilitated by the stick-breaking representation. For that purpose, we shall briefly recall the variational principle.

5.1 Variational Bayesian Expectation Maximization

The clustering task consists primarily of estimating the unknown labels $\mathbf{z} = (z_1, \ldots, z_n)$ from observed $\mathbf{y} = (y_1, \ldots, y_n)$ assuming a joint distribution $p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\Theta}; \boldsymbol{\phi})$ governed by a set of parameters denoted by $\boldsymbol{\Theta}$ and often by additional hyperparameters $\boldsymbol{\phi}$. However to perform good label estimation, the parameters $\boldsymbol{\Theta}$ values (and hyperparameters $\boldsymbol{\phi}$) have to be available. A natural approach for parameter estimation is based on maximum likelihood, where $\boldsymbol{\Theta}$ is estimated by $\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta} \in \underline{\Theta}} p(\mathbf{y} \mid \boldsymbol{\Theta})$. Then an estimate of $\mathbf{z}$ can be obtained by maximizing $p(\mathbf{z} \mid \mathbf{y}, \hat{\boldsymbol{\Theta}})$. However, $p(\mathbf{y} \mid \boldsymbol{\Theta})$ is a marginal distribution over the unknown $\mathbf{z}$ variables, so that direct maximum likelihood is intractable in general. The Expectation-Maximization (EM) algorithm [22] is a general iterative technique for maximum likelihood estimation in the presence of unobserved latent variables or missing data. An EM iteration consists of two steps usually referred to as the E-step in which the expectation of the so-called complete log-likelihood is computed and the M-step in which this expectation is maximized over $\boldsymbol{\Theta}$. An equivalent way to define EM is the following. As discussed in [25], EM can be viewed as an alternating maximization procedure of a function $\mathcal{F}_0$ defined, for any probability distribution q_Z on $\mathcal{Z}$ by

$$\begin{aligned} \mathcal{F}_0(q_Z, \boldsymbol{\Theta}, \boldsymbol{\phi}) &= \sum_{\mathbf{z} \in \mathcal{Z}} q_Z(\mathbf{z}) \ \log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\Theta}; \boldsymbol{\phi}) + I[q_Z] \\ &= \mathbb{E}_{q_Z}\left[\log \frac{p(\mathbf{y}, \mathbf{Z} \mid \boldsymbol{\Theta}; \boldsymbol{\phi})}{q_Z(\mathbf{z})}\right] \end{aligned} \tag{29}$$

where $I[q_Z] = -\mathbb{E}_{q_Z}[\log q_Z(\mathbf{Z})]$ is the entropy of q_Z ($\mathbb{E}_q$ denotes the expectation with regard to q). The function $\mathcal{F}_0$ depends on observations $\mathbf{y}$ which are fixed throughout, hence are omitted from the notation.

Instead of considering only point estimation of $\boldsymbol{\Theta}$, a fully Bayesian approach can be carried out, for instance when prior knowledge on the parameters $\boldsymbol{\Theta}$ is available. In this case, we have to compute

$$p(\mathbf{z} \mid \mathbf{y}) = \int_{\Theta} p(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\Theta}) \ p(\boldsymbol{\Theta} \mid \mathbf{y}) \mathrm{d}\boldsymbol{\Theta} \tag{30}$$

Integrating out $\boldsymbol{\Theta}$ in this way requires the computation of $p(\boldsymbol{\Theta} \mid \mathbf{y})$ which is not usually available in closed-form. As an alternative to costly simulation-based methods (MCMC), an EM-like procedure using variational approximation can provide approximations of the marginal posterior distributions $p(\boldsymbol{\Theta} \mid \mathbf{y})$ and $p(\mathbf{z} \mid \mathbf{y})$. This approach is referred to as VBEM for Variational Bayesian EM [5]. Let q_Z and q_Θ denote respectively

distributions over $\mathbf{Z}$ and $\boldsymbol{\Theta}$ that will serve as approximations to the true posteriors. Similarly to standard EM, VBEM is maximizing the following *free energy* function defined for any q_Z and q_Θ distributions

$$\mathcal{F}(q_Z, q_\Theta, \boldsymbol{\phi}) = \mathbb{E}_{q_Z q_\Theta}\left[\log \frac{p(\mathbf{y}, \mathbf{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi})}{q_Z(\mathbf{z}) q_\Theta(\boldsymbol{\Theta})}\right]$$

alternatively over q_Z, q_Θ and $\boldsymbol{\phi}$. Adding a prior on $\boldsymbol{\Theta}$ is formally the same as adding $\boldsymbol{\Theta}$ to the missing variables, while the hyperparameters $\boldsymbol{\phi}$ play the role of the parameters of interest in maximum likelihood estimation.

The alternate maximization of $\mathcal{F}$ yields the VBEM algorithm that decomposes into three steps. It is easy to show, using the Kullback–Leibler (KL) divergence properties, that the maximization over q_Z and q_Θ leads to the following E-steps (see Appendix A of [9]). At the rth iteration, using current values $\boldsymbol{\phi}^{(r-1)}$ and $q_\Theta^{(r-1)}$, we get the following updating,

$$\begin{aligned}
&\textbf{VB-E-Z:} \quad q_Z^{(r)}(\mathbf{z}) \propto \exp \mathbb{E}_{q_\Theta^{(r-1)}}[\log p(\mathbf{y}, \mathbf{z}, \boldsymbol{\Theta}; \boldsymbol{\phi}^{(r-1)})],\\
&\textbf{VB-E-}\boldsymbol{\Theta}\textbf{:} \quad q_\Theta^{(r)}(\boldsymbol{\Theta}) \propto \exp \mathbb{E}_{q_Z^{(r)}}[\log p(\mathbf{y}, \mathbf{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi}^{(r-1)})],\\
&\textbf{VB-M-}\boldsymbol{\phi}\textbf{:} \quad \boldsymbol{\phi}^{(r)} = \arg\max_{\boldsymbol{\phi}} \mathbb{E}_{q_Z^{(r)} q_\Theta^{(r)}}[\log p(\mathbf{y}, \mathbf{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi})].
\end{aligned}$$

Also, it is worth noticing that if $\mathbf{Y}$ and $\mathbf{Z}$ are independent of $\boldsymbol{\phi}$ conditionally on $\boldsymbol{\Theta}$, as this is often the case when $\boldsymbol{\phi}$ gathers the parameters that describe the prior on $\boldsymbol{\Theta}$, then the VB-M-step simplifies into

$$\boldsymbol{\phi}^{(r)} = \arg\max_{\boldsymbol{\phi}} \mathbb{E}_{q_\Theta^{(r)}}[\log p(\boldsymbol{\Theta}; \boldsymbol{\phi})] = \arg\min_{\boldsymbol{\phi}} \mathrm{KL}(q_\Theta^{(r)} \| p(\boldsymbol{\Theta}; \boldsymbol{\phi})). \tag{31}$$

Then $\boldsymbol{\phi}^{(r)}$ is the value that minimizes the KL distance between the prior $p(\boldsymbol{\Theta}; \boldsymbol{\phi})$ and the variational posterior $q_\Theta^{(r)}(\boldsymbol{\Theta})$. In the conjugate exponential family case, it is known that both distributions belong to the same family [5]. If this family is identifiable it follows that $\boldsymbol{\phi}^{(r)} = \hat{\boldsymbol{\phi}}^{(r)}$ where $\hat{\boldsymbol{\phi}}^{(r)}$ are the variational parameters defining $q_\Theta^{(r)}(\boldsymbol{\Theta})$. A more detailed example is given in Section 5.2.

In practice, we can decide which parameters are treated as genuine parameters $\boldsymbol{\Theta}$ or as hyperparameters $\boldsymbol{\phi}$, depending on whether some prior knowledge is available only for a subset of the parameters or whether the model has hyperparameters $\boldsymbol{\phi}$ for which no prior information is available. Also for complex models, q_Θ and q_Z may need to be further restricted to simpler forms, such as factorized forms, in order to ensure tractable VB-E-steps. This is illustrated in the next section for the PY-MRF inference.

5.2 VBEM for a PY-MRF mixture model with Gaussian components

The VBEM steps are described for a PY-MRF mixture model as defined in Eq. (16) to (19), with Gaussian distributed observations $\mathbf{y}$. As hyperparameters α and σ may have a significant effect on the growth of the number of clusters with data sample size, it is possible to specify priors on them. For the DP case obtained with $\sigma = 0$, it is suggested in [7] to use a gamma prior over α with two hyperparameters s_1 and s_2, *i.e.* $\alpha \sim \mathcal{G}(s_1, s_2)$ where s_1 and s_2 can be estimated or fixed. A natural question that arises is then whether one can also find a tractable prior for the discount parameter σ. We propose to use the following prior that accounts for the constraints $\sigma \in (0, 1)$ and $\alpha > -\sigma$,

$$p(\alpha, \sigma; s_1, s_2, a) = p(\alpha \mid \sigma; s_1, s_2)\; p(\sigma; a) \tag{32}$$

where $p(\alpha \mid \sigma; s_1, s_2)$ is a shifted gamma distribution $\mathcal{SG}(s_1, s_2, \sigma)$ and $p(\sigma; a)$ is a distribution depending on some parameter a not specified for the moment but that can typically be taken as the uniform distribution on the interval $(0, 1)$. Such a shifted gamma distribution is the distribution of a variable $U - \sigma$ where σ is considered as fixed and U follows a gamma distribution $\mathcal{G}(s_1, s_2)$. The pdf of this shifted gamma distribution is obtained from the standard gamma distribution as $p(\alpha \mid \sigma; s_1, s_2) = \mathcal{G}(\alpha + \sigma; s_1, s_2)$. It follows that the joint distribution of the observed data $\mathbf{y}$ and all latent variables becomes

$$p(\mathbf{y}, \mathbf{z}, \boldsymbol{\Theta}; \boldsymbol{\phi}) = p(\alpha, \sigma; s_1, s_2, a) \prod_{j=1}^{n} p(y_j | z_j, \boldsymbol{\theta}^*) p(\mathbf{z} | \boldsymbol{\tau}; \beta) \prod_{k=1}^{\infty} p(\tau_k | \alpha, \sigma) \prod_{k=1}^{\infty} p(\theta_k^*; \rho_k),$$

where the notation $\prod_{k=1}^{\infty}$ is a distributional notation, and in addition to the terms already defined in (16) and (18), we specify the likelihood term (19) as a Gaussian distribution $p(y_j | \theta_{z_j}^*) = \mathcal{N}(y_j | \mu_{z_j}, \Sigma_{z_j})$ and the G_0

prior on cluster specific parameters $\theta_k^* = (\mu_k, \Sigma_k)$ as a Normal-inverse-Wishart distribution parameterized by $\rho_k = (m_k, \lambda_k, \Psi_k, \nu_k)$ with a pdf

$$p(\theta_k^*; \rho_k) = \mathcal{NIW}(\mu_k, \Sigma_k; \rho_k) = \mathcal{N}(\mu_k; m_k, \lambda_k^{-1}\Sigma_k)\, \mathcal{IW}(\Sigma_k; \Psi_k, \nu_k).$$

In the above notation, we consider as hyperparameters the set $\boldsymbol{\phi} = (s_1, s_2, a, \beta, (\rho_k)_{k=1}^{\infty})$ while $\boldsymbol{\Theta} = (\boldsymbol{\tau}, \alpha, \sigma, \boldsymbol{\theta}^*)$.

In most variational approximations, the posteriors are approximated in a factorized form (mean-field approximation). In particular, the intractable MRF posterior on $\mathbf{z}$ is approximated as $q_{\mathbf{z}}(\mathbf{z})$ that factorizes so as to handle intractability due to spatial dependencies, namely

$$q_{\mathbf{z}}(\mathbf{z}) = \prod_{j=1}^{n} q_{z_j}(z_j).$$

Then, the infinite state space for each z_i is dealt with by choosing a truncation of the state space to a maximum label K [7]. In practice, this consists of assuming that the variational distributions q_{z_j}, for $j = 1, \ldots, n$, satisfy $q_{z_j}(k) = 0$ for $k > K$ and that the variational distribution on $\boldsymbol{\tau}$ also factorizes as $q_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \prod_{k=1}^{K-1} q_{\tau_k}(\tau_k)$, with the additional condition that $\tau_K = 1$. Thus, the truncated variational posterior of parameters $\boldsymbol{\Theta}$ is given by

$$q_{\Theta}(\boldsymbol{\Theta}) = q_{\alpha,\sigma}(\alpha, \sigma) \prod_{k=1}^{K-1} q_{\tau_k}(\tau_k) \prod_{k=1}^{K} q_{\theta_k^*}(\theta_k^*). \tag{33}$$

These forms of $q_{\mathbf{z}}$ and $q_{\boldsymbol{\Theta}}$ lead to four VB-E steps and three VB-M steps summarized below with details in the Appendix. Set the initial value of $\boldsymbol{\phi}$ to $\boldsymbol{\phi}^{(0)}$. Then, repeat iteratively the following steps. The iteration index is omitted in the update formulas for simplicity.

VB-E-$\boldsymbol{\tau}$ step

The VB-E-$\boldsymbol{\tau}$ step corresponds to a variational approximation in the exponential family case and results in a posterior from the same family as the prior. It comes for $k = 1, \ldots, K$,

$$q_{\tau_k}(\tau_k) = \mathcal{B}(\tau_k; \hat{\gamma}_{k,1}, \hat{\gamma}_{k,2}) \tag{34}$$

with

$$\hat{\gamma}_{k,1} = 1 - \mathbb{E}_{q_\sigma}[\sigma] + \bar{n}_k, \quad \hat{\gamma}_{k,2} = \mathbb{E}_{q_\alpha}[\alpha] + k\mathbb{E}_{q_\sigma}[\sigma] + \sum_{\ell=k+1}^{K} \bar{n}_\ell, \tag{35}$$

where

$$\text{for} \quad k = 1, \ldots, K, \quad \bar{n}_k = \sum_{j=1}^{n} q_{z_j}(k) \tag{36}$$

corresponds to the weight of cluster k.

VB-E-(α, σ) step

The (α, σ) variational posterior is more complex but has a simple gamma form in the DP ($\sigma = 0$) case. More specifically, we need to compute

$$\hat{s}_1 = s_1 + K - 1, \ \text{ and } \ \hat{s}_2 = s_2 - \sum_{k=1}^{K-1} \psi(\hat{\gamma}_{k,2}) - \psi(\hat{\gamma}_{k,1} + \hat{\gamma}_{k,2}) \tag{37}$$

where $\psi(\cdot)$ is the digamma function defined by $\psi(z) = \frac{\mathrm{d}}{\mathrm{d}z} \log \Gamma(z) = \frac{\Gamma'(z)}{\Gamma(z)}$. When $\sigma = 0$ then q_α is a gamma distribution $\mathcal{G}(\hat{s}_1, \hat{s}_2)$ and $\mathbb{E}_{q_\alpha}[\alpha] = \dfrac{\hat{s}_1}{\hat{s}_2}$. Otherwise (PY case), $q_{\alpha,\sigma}$ is only identified up to a normalizing constant but the required $\mathbb{E}_{q_\alpha}[\alpha]$ and $\mathbb{E}_{q_\sigma}[\sigma]$ can be computed by importance sampling (see Appendix A.4 for details).

VB-E-Z step

Due to the mean field approximation and the truncation, this step consists in computing (see details in Appendix A.5), for all $j = 1, \ldots, n$ and for $k \leq K$,

$$q_{z_j}(k) = \frac{\tilde{q}_j(k)}{\sum_{\ell=1}^{K} \tilde{q}_j(\ell)}, \tag{38}$$

where $\log \tilde{q}_j(k)$ is defined by

$$\begin{aligned} &-\frac{1}{2}\left\{\log\left|\frac{\hat{\Psi}_k}{2}\right| - \sum_{i=1}^{d}\psi\left(\frac{\hat{\nu}_k + (1-i)}{2}\right) + \hat{\nu}_k(y_j - \hat{m}_k)^T\hat{\Psi}_k^{-1}(y_j - \hat{m}_k) + \frac{d}{\hat{\lambda}_k}\right\} + \\ &\psi(\hat{\gamma}_{k,1}) - \psi(\hat{\gamma}_{k,1} + \hat{\gamma}_{k,2}) + \sum_{l=1}^{k-1}\psi(\hat{\gamma}_{l,2}) - \psi(\hat{\gamma}_{l,1} + \hat{\gamma}_{l,2}) + \beta\sum_{i\in\mathcal{N}_j} q_{z_i}(k), \end{aligned} \tag{39}$$

where in the last sum, $\mathcal{N}_j$ represents the neighbours of j. In the above formula, symbols $(\hat{m}_k, \hat{\lambda}_k, \hat{\boldsymbol{\Psi}}_k, \hat{\nu}_k)$ are the variational hyperparameters for $q_{\boldsymbol{\theta}_k^*}$ more specifically defined in the following step and d is the dimension of the data. The advantage of Eq. (38) is that it provides assignment probabilities $q_{z_i}(k)$ and does not require intermediate commitments to hard assignments of the z_j's. The hard assignments can be postponed to the end if desired to get a segmentation through the following maximum a posteriori (MAP) estimation:

$$\hat{z}_j = \underset{k\in\{1,\ldots,K\}}{\arg\max}\ q_{z_j}(k). \tag{40}$$

VB-E-$\boldsymbol{\theta}^*$ step

This step is divided into K parts where the computation is similar to that in standard Bayesian finite mixtures with a choice of conjugate prior, here for Gaussian distributions. Hence, for each $k \leq K$, the variational posterior is a Normal-inverse-Wishart distribution defined as

$$q_{\theta_k^*}(\mu_k, \Sigma_k) = \mathcal{NIW}(\mu_k, \Sigma_k; \hat{m}_k, \hat{\lambda}_k, \hat{\Psi}_k, \hat{\nu}_k), \tag{41}$$

where the hyperparameters are updated as follows (see for instance [24])

$$\begin{aligned} \hat{\lambda}_k &= \lambda_k + \bar{n}_k, \quad \hat{\nu}_k = \nu_k + \bar{n}_k, \\ \hat{\Psi}_k &= \Psi_k + S_k + \frac{\lambda_k\bar{n}_k}{\lambda_k + \bar{n}_k}(m_k - \bar{\mu}_k)(m_k - \bar{\mu}_k)^T, \\ \hat{m}_k &= \frac{\lambda_k m_k + \bar{n}_k\bar{\mu}_k}{\lambda_k + \bar{n}_k} = \frac{\lambda_k m_k + \bar{n}_k\bar{\mu}_k}{\hat{\lambda}_k}, \end{aligned} \tag{42}$$

with $\bar{n}_k$ defined in (36) and

$$\begin{aligned} \bar{\mu}_k &= \frac{1}{\bar{n}_k}\sum_{j=1}^{n} q_{z_j}(k) y_j, \\ S_k &= \sum_{j=1}^{n} q_{z_j}(k)(y_j - \bar{\mu}_k)(y_j - \bar{\mu}_k)^T. \end{aligned} \tag{43}$$

VB-M steps

The maximization step consists of updating the hyperparameters $\boldsymbol{\phi} = (\beta, s_1, s_2, a, \boldsymbol{\rho})$, where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)$, by maximizing the free energy, if they are not set heuristically:

$$\boldsymbol{\phi}^{(r)} = \underset{\boldsymbol{\phi}}{\arg\max}\ \mathbb{E}_{q_Z^{(r)} q_\tau^{(r)} q_{\alpha,\sigma}^{(r)} q_{\theta^*}^{(r)}}[\log p(\boldsymbol{y}, \mathbf{Z}, \boldsymbol{\tau}, \alpha, \sigma, \boldsymbol{\theta}^*; \boldsymbol{\phi})]\ . \tag{44}$$

The VB-M-step can therefore be divided into 3 independent sub-steps as listed below. From the conditional independence of $(s_1, s_2, a, \boldsymbol{\rho})$ and $(\boldsymbol{Y}, \mathbf{Z})$ given $(\boldsymbol{\tau}, \alpha, \sigma, \boldsymbol{\theta}^*)$, the VB-M-step writes as in (31) so that the solutions for the VB-M-(s_1, s_2) (in the DP case) and VB-M-$\boldsymbol{\rho}$ steps are straightforward. Only the β step and the M-(s_1, s_2, a) step (in the PY case) are more involved.

VB-M-β: The maximization of (44) with respect to β leads to

$$\beta^{(r)} = \arg\max_{\beta} \mathbb{E}_{q_Z^{(r)} q_\tau^{(r)}} [\log p(\mathbf{Z}|\boldsymbol{\tau}; \beta)] \ . \tag{45}$$

This step does not admit a closed-form solution but can be solved numerically. More details are given in Appendix A.6.

VB-M-(s_1, s_2, a): This step is straightforward in the DP case ($\sigma = 0$). It can be expressed easily using the fact that both the prior and the variational posterior are Gamma distributions, and using the cross-entropy properties,

$$(s_1, s_2)^{(r)} = \arg\max_{(s_1,s_2)} \mathbb{E}_{q_\alpha^{(r)}} [\log p(\alpha; s_1, s_2)] \ = \ (\hat{s}_1^{(r)}, \hat{s}_2^{(r)}) \tag{46}$$

where $(\hat{s}_1^{(r)}, \hat{s}_2^{(r)})$ is given in (37). In the more general PY case, we can solve this step numerically using also importance sampling. See Appendix A.7.

VB-M-$\boldsymbol{\rho}$: This step divides into K sub-steps that involve again cross-entropies,

$$\rho_k^{(r)} = \arg\max_{\rho} \mathbb{E}_{q_{\theta_k^*}^{(r)}} [\log p(\theta_k^*; \rho_k)] \ = \ \hat{\rho}_k^{(r)} \tag{47}$$

where $\hat{\rho}_k^{(r)} = (\hat{\lambda}_k^{(r)}, \hat{\nu}_k^{(r)}, \hat{\Psi}_k^{(r)}, \hat{m}_k^{(r)})$ is given in Eq. (42).

6 Application to image segmentation

To validate the proposed approach, we consider its application to unsupervised image segmentation as a spatial clustering task. Image segmentation consists of partitioning a digital image into distinct regions that contain pixels with similar properties. Extensive research work has been done in this field using various clustering techniques. In practice, to be meaningful for image analysis and interpretation, the segmented regions should closely relate to depicted objects or features of interest. A number of tasks in image analysis often depends on the reliability of preliminary segments, but an accurate partitioning of an image is still quite challenging.

6.1 Feature extraction for image segmentation

The color and texture features in a natural image are often very complex. For our experiments, we mainly focus on two special types of features based on the HSV (Hue, Saturation, Value) color space and the maximum response (MR) filter bank. The HSV color space is often used in natural image analysis because it corresponds better to how people experience color than the RGB color space does. Regarding the texture information, we shall consider the MR8 filter bank [36], which consists of 38 filters but only 8 filter responses. More precisely, the MR8 filter bank contains filters at multiple orientations but their outputs are compressed by recording only the maximum filter response across all orientations. This achieves rotation invariance. Furthermore, the images are presegmented into superpixels that group pixels similar in color and other low-level properties [1]. In this respect, superpixels are regarded as more natural entities that allow reducing the number of observations drastically for running clustering algorithms. In all our experiments, each image is presegmented into approximately $1\,000$ superpixels using the SLIC algorithm proposed in [1]. Finally, we compute the feature vectors at superpixel level, *i.e.*, the average of features on the centroid of each superpixel. The entire segmentation procedure is summarized in Algorithm 1.

Algorithm 1: Summary of the image segmentation procedure.

Input: An input color image.

Output: A segmented image.

Procedure:

- Use the SLIC algorithm to form an over-segmentation [1].
- Compute HSV color and texture features at superpixel level [36].
- Partition the aggregated features by BNP-MRF.

Return: Multiple segmented regions.

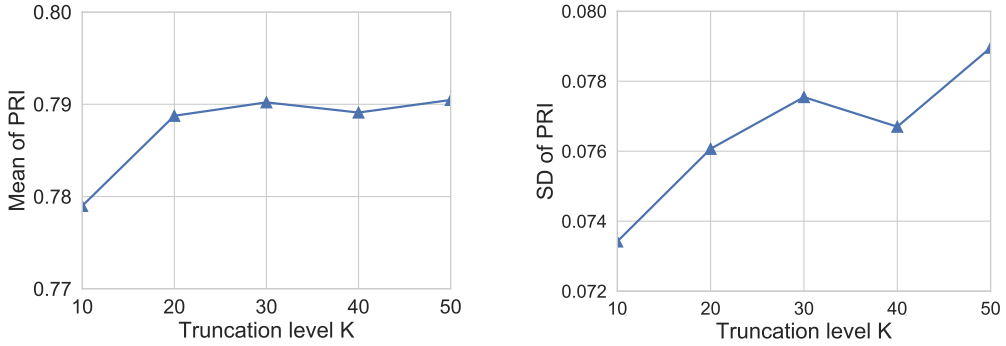

Fig. 2 PY-MRF mixture model: Mean and standard deviation of the PRI score over the considered subset of the BSDS500 data set as a function of the truncation level K.

	Proposed model	Results given in [11]			
PRI (%)	**PY-MRF**	DPM	iHMRF	MRF-PYP	Graph Cuts
Mean	**79.05**	74.15	75.50	76.49	76.10
Median	**80.62**	75.49	76.89	78.08	77.59
St. Dev.	**7.9**	8.4	8.2	7.9	8.3

Table 1 Performance comparison: Summary statistics of the PRI score over the 154 images from BSDS500 studied by [12] for our PY-MRF mixture model and the approaches tested in [12, 11].

6.2 Berkeley Segmentation Data Set

To quantify the performance of our segmentation algorithm, numerical experiments were conducted on a subset of images selected from the Berkeley Segmentation Data Set 500 (BSDS500) already studied by [3, 11], which provides multiple human annotated segments as many ground truths for each image. The considered subset consists of 154 images as listed in Tables 1 and 2 in [11].

In the literature, a standard measure for comparing a test segmentation to another is the rand index (RI) [29]. The RI is one when two segmentations are exactly the same. However, when having for one image a set of ground truths which do not completely agree, the probabilistic rand index (PRI) [35] is preferable. Given a set of ground truths $\mathcal{S} = (S_1, \ldots, S_T)$, the PRI is defined as follows:

$$\mathrm{PRI}(S_{\mathrm{test}}, \mathcal{S}) = \frac{2}{n(n-1)} \sum_{i<j} [c_{ij} p_{ij} + (1 - c_{ij})(1 - p_{ij})] \tag{48}$$

where $c_{ij} = 1$ if pixels i and j belong to the same segment in S_{test} and $c_{ij} = 0$ otherwise, n is the number of image pixels and p_{ij} is the probability of two pixels i and j having the same label, *i.e.*, the fraction of all available ground truths in $\mathcal{S}$ where pixels i and j belong to the same segment. In fact, it can be shown that Eq. (48) is simply the mean of the RI computed between each pair (S_{test}, S_k), namely $\frac{1}{T} \sum_{k=1}^{T} \mathrm{RI}(S_{\mathrm{test}}, S_k)$. By definition, the PRI always takes values in $[0, 1]$, where 0 means that S_{test} and $(S_1, \ldots, S_T)$ have no similarities and 1 means all segments are identical. The larger the PRI, the better. In practice, PRI values are often reported as percentages in $[0, 100]$.

Our approach has been tested on the considered subset of the BSDS500 and the summary statistics of the PRI score are shown in Figure 2 as a function of the truncation level K for the PY-MRF case. Similar results were observed for the DP. It appears that for $K \geq 30$, the global performance does not change much and is satisfying with respect to existing results in the literature. We compared our best results with those reported in [11]. Table 1 shows that our approach outperforms the existing results. The improvement in PRI may appear overall small but it can be assessed by visualizing original images and their segmentations. We show in Figure 3 segmentation results for four images. The main differences between the non spatial PY and PY-MRF mixture models can be visualized for the first image in the ground and water which are segmented in the latter case into a smaller number of regions whose shapes are in addition smoother. This is typical of more spatial interaction in the clustering process. Similarly, the same phenomenon is also visible in the peak part of the second image, in the sky and grass parts of the third image and in the plant parts of the fourth image.

We also examined, for the PY-MRF mixture model with $K = 50$, the values of the expected α, σ and β for each of the 154 segmented images presented in Figure 4 as scatter plots (one point per image). Recall from Section 5.2 that α and σ are elements of the parameters $\boldsymbol{\Theta}$ while β is considered as a hyperparameter from

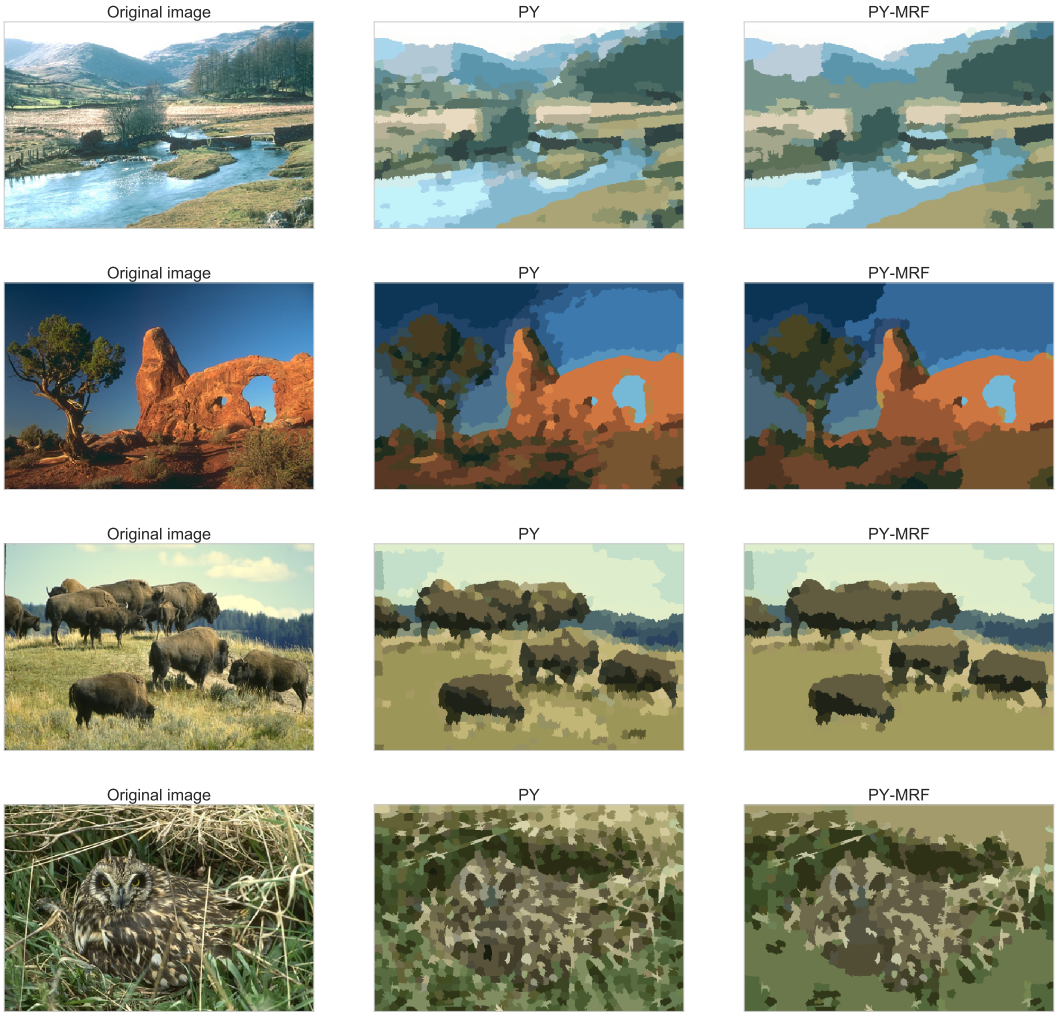

Fig. 3 Segmentation results for four images from the BSDS500 data set. From left to right, columns show respectively, the original images, the segmentation results with the PY and PY-MRF mixture models.

ϕ. Figure 4 shows also the correlations (across the 154 images) between the expected values of α, σ and β. It appears that the estimated σ values are most of the time smaller than 0.5 and sometimes closer to 0 with some anti-correlation with respect to α values. In contrast, β values appear quite independent from α or σ.

In terms of pure PRI performance, the BSDS500 data set is not an easy example because the ground truth segmentations are labeled manually by humans and are sometimes quite subjective and inconsistent across users. However, this example allows comparison of methods and visualization. Two interesting findings are that the choice of K does not seem to be too sensitive as soon as K is large enough, and there seems to be some correlation between α and σ while β is rather independent of the latest. Further analysis would be needed to confirm these properties but in practice, they could be used to guide the segmentations into more or less spatially smooth versions without risking to eliminate too small segments.

7 Conclusion and perspectives

In this paper, we proposed a general scheme to build BNP priors that can model dependencies through the addition of a Markov random field term. In contrast to other existing attempts that reduce to spatially constrained standard BNP priors such as [11, 12], our proposal leads to proper spatial priors. Our construction is based on the stick-breaking representation and was illustrated starting from the Dirichlet and Pitman–Yor processes, although this approach could be extended to other forms of BNP priors admitting a stick-breaking representation such as Gibbs-type priors. The stick-breaking representation was further exploited to derive clustering properties of the

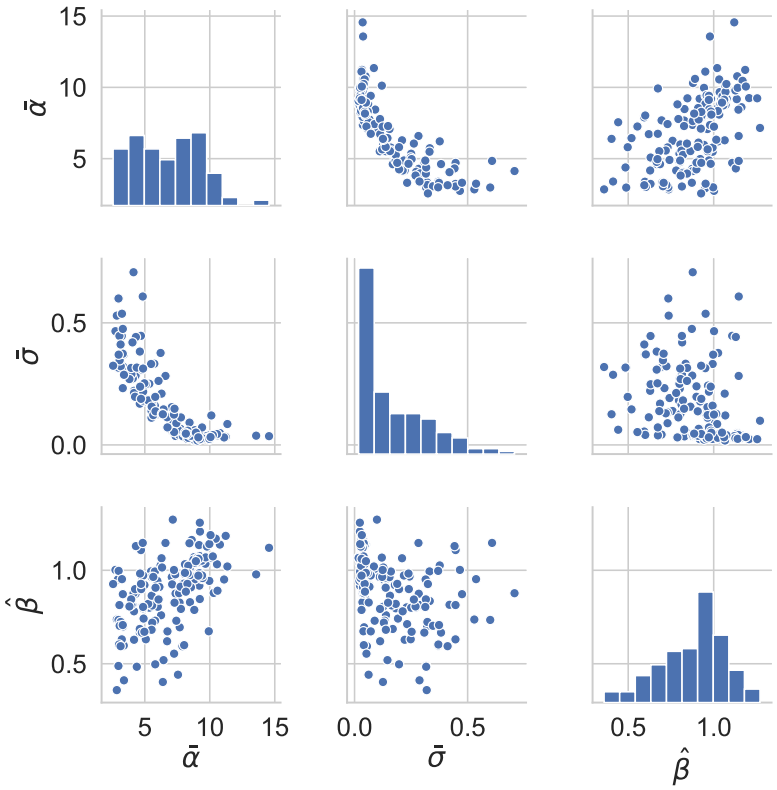

Fig. 4 Estimated parameter values $(\bar{\alpha}, \bar{\sigma})$ obtained from **VB-E** steps and $\hat{\beta}$ obtained from a **VB-M** step using the PY-MRF model with truncation level $K = 50$, on the 154 images from the Berkeley benchmark.

model and to provide a variational inference algorithm. In addition to the usual BNP parameters, an estimation of the Markov interaction parameter β was proposed. The variational approximation chosen was based on a standard truncation but it would be interesting to investigate other approximations, *e.g.* [37]. Also the variational algorithm is greatly simplified for standard stick-breaking representations (*e.g.* DP and PY) with independent weight variables. Nevertheless, it would be interesting to investigate more general stick-breaking representations possibly using some MCMC counterpart for estimation.

The approach was illustrated on a challenging unsupervised image segmentation task with good results with respect to the literature, but the proposed scheme is quite flexible and can be used in more general settings including community detection or disease mapping in epidemiology.

A Appendix

Proofs of propositions 1 and 2 are given in the two first sections. Details on the VBEM steps, to complete the previous developments when necessary, follow.

A.1 Proof of Proposition 1

Integrating out $\boldsymbol{\pi}$ from distribution (12) and noticing that the Markov term does not depend on $\boldsymbol{\pi}$, leads to

$$p(z_{1:n+1}; \beta) \propto p(z_{1:n+1}; \beta = 0)\ \mathrm{e}^{\beta H(z_{1:n+1})}$$

where $H(z_{1:n+1}) = \sum_{i \sim j} \delta_{z_i = z_j}$ is counting the number of homogeneous edges. It follows that

$$p(z_{n+1} | z_{1:n}; \beta) \propto p(z_{n+1} | z_{1:n}; \beta = 0)\ \mathrm{e}^{\beta H(z_{1:n+1})}\ .$$

When $\beta = 0$, the model (12) reduces to standard Gibbs-type priors for which the quantities above have been given in (23). Using (23) and enumerating how z_{n+1} can affect the number of homogeneous edges, it comes that:

for $z_{n+1} \notin \{z_1, \dots, z_n\}$,

$$p(z_{n+1}|z_{1:n};\beta) \propto \frac{V_{n+1,K_n+1}}{V_{n,K_n}}\ \mathrm{e}^{\beta H(z_{1:n})}$$

and for $\ell \in \{z_1 \dots z_n\}$,

$$p(z_{n+1} = \ell \mid z_{1:n};\beta) \propto \frac{V_{n+1,K_n}}{V_{n,K_n}}(n_\ell - \sigma)\ \mathrm{e}^{\beta\tilde{n}_\ell \delta_{\mathcal{N}_{n+1}}(\ell)}\ \mathrm{e}^{\beta H(z_{1:n})}\ .$$

The result follows by normalizing the quantities above, using (21) and noticing that

$$\sum_{\ell \in \{z_1 \dots z_n\}} (n_\ell - \sigma)\mathrm{e}^{\beta\tilde{n}_\ell \delta_{\mathcal{N}_{n+1}}(\ell)} = \boldsymbol{\eta}_{n+1} + n - \sigma K_n\ .$$

■

A.2 Proof of Proposition 2

Consider the DP first. Let $D_n = \delta_{\theta_n \text{ is new}|\theta_{1:n-1}}$ be the Bernoulli random variable equal to one when θ_n is a fresh draw, which for the DP-MRF happens with probability

$$\frac{\alpha}{\alpha + n - 1 + \boldsymbol{\eta}_n(0,\beta)}.$$

Then $K_n = \sum_{i=1}^n D_i$, so we have for the DP-MRF

$$\mathbb{E}[K_n] = \sum_{i=0}^{n-1} \mathbb{E}\left[\frac{\alpha}{\alpha + i + \boldsymbol{\eta}_{i+1}(0,\beta)}\right]. \tag{49}$$

By definition of the maximal degree D of the graph, the multiplicity $\tilde{n}_\ell$ is at most D for any ℓ, hence we have the following upper bound

$$\sum_{\ell \in z_{\mathcal{N}_{i+1}}} n_\ell(\mathrm{e}^{\beta\tilde{n}_\ell} - 1) \leq \sum_{\ell \in z_{\mathcal{N}_{i+1}}} n_\ell(\mathrm{e}^{D\beta} - 1) \leq i(\mathrm{e}^{D\beta} - 1). \tag{50}$$

Plugging (50) into (49) provides the desired inequality

$$\begin{aligned}\mathbb{E}[K_n] &\geq \sum_{i=0}^{n-1} \frac{\alpha}{\alpha + i\mathrm{e}^{D\beta}} \\ &= \frac{\alpha}{\mathrm{e}^{D\beta}} \sum_{i=0}^{n-1} \frac{1}{i + \alpha\mathrm{e}^{-D\beta}} \gtrsim \frac{\alpha}{\mathrm{e}^{D\beta}} \log n.\end{aligned}$$

Turning to the Pitman–Yor process, we follow the proof technique of [34]. The prior expectation of the number of clusters satisfies the following recursion

$$\mathbb{E}[K_{n+1}] = \mathbb{E}[K_n] + \mathbb{E}\left[\frac{\alpha + \sigma K_n}{\alpha + n + \boldsymbol{\eta}_{n+1}}\right].$$

Assuming here that $\alpha \geq 0$ and using the lower bound (50) yields

$$\mathbb{E}[K_{n+1}] \geq \mathbb{E}[K_n]\left(1 + \frac{\sigma}{\alpha + n\mathrm{e}^{D\beta}}\right).$$

By induction, and using $K_1 = 1$,

$$\begin{aligned}\mathbb{E}[K_n] &\geq \prod_{i=1}^{n-1}\left(1 + \frac{\sigma}{\alpha + i\mathrm{e}^{D\beta}}\right) = \exp\left(\sum_{i=1}^{n-1} \log\left(1 + \frac{\sigma}{\alpha + i\mathrm{e}^{D\beta}}\right)\right) \\ &\simeq \exp\left(\sum_{i=1}^{n-1} \frac{\sigma}{\alpha + i\mathrm{e}^{D\beta}}\right) \simeq \exp\left(\frac{\sigma}{\mathrm{e}^{D\beta}} \log n\right) = n^{\sigma\mathrm{e}^{-D\beta}},\end{aligned}$$

where $a_n \simeq b_n$ means a_n/b_n converges to some positive constant when $n \to \infty$. The desired inequality for PY is obtained by writing this constant equal to c. ■

A.3 **VB-E-τ step**

$$\begin{aligned} q_{\tau_k}(\tau_k) &\propto \exp\left(\mathbb{E}_{q_{\alpha,\sigma}}[\log p(\tau_k \mid \alpha,\sigma)] + \sum_{j=1}^{n} \mathbb{E}_{q_{z_j} q_{\tau\setminus\{k\}}}[\log \pi_{z_j}(\boldsymbol{\tau})]\right) \\ &\propto \exp\Big(-\mathbb{E}_{q_{\alpha,\sigma}}[\sigma]\log\tau_k + (\mathbb{E}_{q_{\alpha,\sigma}}[\alpha] + k\mathbb{E}_{q_{\alpha,\sigma}}[\sigma] - 1)\log(1-\tau_k) \\ &+ \sum_{j=1}^{n}\sum_{l=k+1}^{K} q_{z_j}(l)\log(1-\tau_k) + \sum_{j=1}^{n} q_{z_j}(k)\log(\tau_k)\Big). \end{aligned} \tag{51}$$

Considering the terms involving τ_k, we recognize the beta distribution $\mathcal{B}(\tau_k; \hat{\gamma}_{k,1}, \hat{\gamma}_{k,2})$ specified in (34). It follows the expressions of the following quantities,

$$\begin{aligned} \mathbb{E}_{q_{\tau_k}}[\log(\tau_k)] &= \psi(\hat{\gamma}_{k,1}) - \psi(\hat{\gamma}_{k,1} + \hat{\gamma}_{k,2}) \\ \mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)] &= \psi(\hat{\gamma}_{k,2}) - \psi(\hat{\gamma}_{k,1} + \hat{\gamma}_{k,2}) \end{aligned} \tag{52}$$

where $\psi(\cdot)$ is the digamma function defined in Section 5.2.

A.4 **VB-E-(α, σ) step**

In the PY case, $q_{\alpha,\sigma}(\alpha,\sigma)$ is proportional to

$$\begin{aligned} \tilde{q}_{\alpha,\sigma}(\alpha,\sigma) &\propto p(\alpha,\sigma; s_1, s_2, a)\ \exp\left(\sum_{k=1}^{K-1} \mathbb{E}_{q_{\tau_k}}[\log p(\tau_k \mid \alpha,\sigma)]\right) \\ &\propto p(\alpha,\sigma; s_1, s_2, a)\ \prod_{k=1}^{K-1} B(1-\sigma, \alpha+k\sigma)^{-1} \times \\ &\exp\Big(-\sigma(\sum_{k=1}^{K-1}\mathbb{E}_{q_{\tau_k}}[\log\tau_k] - \sum_{k=1}^{K-1} k\ \mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]) + (\alpha-1)\sum_{k=1}^{K-1}\mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\Big), \end{aligned}$$

where $B(a,b) \coloneqq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ represents the beta function. The last term simplifies into

$$\begin{aligned} \prod_{k=1}^{K-1} B(1-\sigma, \alpha+k\sigma)^{-1} &= \frac{\Gamma(\alpha)}{\Gamma(1-\sigma)^{K-1}\,\Gamma(\alpha+(K-1)\sigma)} \\ &\times \prod_{k=1}^{K-1}(\alpha+(k-1)\sigma). \end{aligned} \tag{53}$$

The difficulty is that except in the DP-MRF case, the normalizing constant for $\tilde{q}_{\alpha,\sigma}(\alpha,\sigma)$ is not tractable. Nevertheless, to carry out the VBEM algorithm, we do not need the full $q_{\alpha,\sigma}$ distribution but only the means $\mathbb{E}_{q_\sigma}[\sigma]$ and $\mathbb{E}_{q_\alpha}[\alpha]$. One solution is therefore to use importance sampling or MCMC to compute these expectations via Monte Carlo sums. Using the prior on (α,σ) given in (32), it comes that

$$\begin{aligned} \tilde{q}_{\alpha,\sigma}(\alpha,\sigma) &= \mathcal{G}(\alpha+\sigma; \hat{s}_1, \hat{s}_2)\ \mathrm{e}^{-\sigma\xi} \\ &\quad p(\sigma; a)\ \frac{\Gamma(\alpha)}{\Gamma(1-\sigma)^{K-1}\,\Gamma(\alpha+(K-1)\sigma)} \prod_{k=1}^{K-1}\left(\frac{\alpha+(k-1)\sigma}{\alpha+\sigma}\right) \end{aligned} \tag{54}$$

where $\mathcal{G}(\alpha+\sigma; \hat{s}_1, \hat{s}_2)$ is the pdf of a σ-shifted gamma distribution with $\hat{s}_1, \hat{s}_2$ given in (37). The parameter ξ is defined as

$$\xi = \sum_{k=1}^{K-1}\mathbb{E}_{q_{\tau_k}}[\log\tau_k] - \sum_{k=1}^{K-1}(k-1)\ \mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)], \tag{55}$$

which can be computed using (52). We propose then to use as importance distribution $\nu(\alpha,\sigma) = \mathcal{G}(\alpha+\sigma; \hat{s}_1, \hat{s}_2)\ p(\sigma; a)$ with $p(\sigma; a)$ the uniform distribution on [0,1], $\mathcal{U}_{[0,1]}(\sigma)$. It comes an expression for the importance weights,

$$\begin{aligned} W(\alpha,\sigma) &= \frac{\tilde{q}_{\alpha,\sigma}(\alpha,\sigma)}{\nu(\alpha,\sigma)} \\ &= \mathrm{e}^{-\sigma\xi}\ \frac{\Gamma(\alpha)}{\Gamma(1-\sigma)^{K-1}\,\Gamma(\alpha+(K-1)\sigma)} \prod_{k=1}^{K-1}\left(\frac{\alpha+(k-1)\sigma}{\alpha+\sigma}\right). \end{aligned} \tag{56}$$

The importance sampling scheme then consists of:

- For $i = 1$ to M, simulate first independently σ_i from $\mathcal{U}_{[0,1]}(\sigma)$ and then simulate conditionnaly α_i using the σ_i-shifted gamma $\mathcal{SG}(\sigma_i, \hat{s}_1, \hat{s}_2)$. This later simulation is easily obtained by simulating a standard $\mathcal{G}(\hat{s}_1, \hat{s}_2)$ and then substracting σ_i to the result.
- Compute the importance weights $w_i = W(\alpha_i, \sigma_i)$.
- Approximate the means $\mathbb{E}_{q_\sigma}[\sigma] \approx \dfrac{\sum_{i=1}^M w_i \sigma_i}{\sum_{i=1}^M w_i}$ and $\mathbb{E}_{q_\alpha}[\alpha] \approx \dfrac{\sum_{i=1}^M w_i \alpha_i}{\sum_{i=1}^M w_i}$

Note that this complication is due to the PY. In the DP-MRF case, the E-α step is considerably simpler as it reduces to computing the approximate posterior expectation of α, namely $\mathbb{E}_{q_\alpha}[\alpha] = \hat{s}_1/\hat{s}_2$.

A.5 **VB-E-Z step**

The VB-E-**Z** is divided into n steps. Since we assume $q_{z_j}(z_j) = 0$ for $z_j > K$, it is only necessary to compute the distributions for $z_j \leq K$, namely

$$q_{z_j}(z_j) \propto \exp\left(\mathbb{E}_{q_{\theta^*_{z_j}}}[\log p(y_j \mid \theta^*_{z_j})] + \mathbb{E}_{q_{\boldsymbol{\tau}}}[\log \pi_{z_j}(\boldsymbol{\tau})] + \beta \sum_{i \in \mathcal{N}_j} q_{z_i}(z_j)\right), \tag{57}$$

where for $z_j = k$,

$$\mathbb{E}_{q_{\boldsymbol{\tau}}}[\log \pi_k(\boldsymbol{\tau})] = \mathbb{E}_{q_{\tau_k}}[\log \tau_k] + \sum_{l=1}^{k-1} \mathbb{E}_{q_{\tau_l}}[\log(1-\tau_l)]. \tag{58}$$

The term $\mathbb{E}_{q_{\theta^*_{z_j}}}[\log p(y_j|\theta^*_{z_j})]$ is computed using the fact that $q_{\theta^*_k}$ is a Normal-inverse-Wishart distribution as described in Eq. (41) of the next VB-E-$\boldsymbol{\theta}^*$ step, namely

$$\begin{aligned} q_{\theta^*_k}(\mu_k, \Sigma_k) &= \mathcal{NIW}(\mu_k, \Sigma_k; \hat{m}_k, \hat{\lambda}_k, \hat{\Psi}_k, \hat{\nu}_k) \\ &= \mathcal{N}\left(\mu_k; \hat{m}_k, \frac{\Sigma_k}{\hat{\lambda}_k}\right) \mathcal{IW}(\Sigma_k; \hat{\Psi}_k, \hat{\nu}_k). \end{aligned} \tag{59}$$

It comes out that (d is the dimension of y_j)

$$\begin{aligned} \mathbb{E}_{q_{\theta^*_k}}[\log p(y_j \mid \theta^*_k)] = &-\frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{q_{\theta^*_k}}[\log|\Sigma_k|] \\ &- \frac{1}{2}\mathbb{E}_{q_{\theta^*_k}}[(y_j - \mu_k)^T \Sigma_k^{-1}(y_j - \mu_k)]. \end{aligned} \tag{60}$$

Using the decomposition (cf. Eq. (59)) and the fact that Σ_k^{-1} follows a Wishart distribution, it comes out that

$$\begin{aligned} \mathbb{E}_{q_{\theta^*_k}}[\log|\Sigma_k|] &= \mathbb{E}_{q_{\Sigma_k}}[\log|\Sigma_k|] = -\mathbb{E}_{q_{\Sigma_k}}[\log|\Sigma_k^{-1}|] \\ &= -\sum_{i=1}^{d} \psi\left(\frac{\hat{\nu}_k + (1-i)}{2}\right) + \log\left|\frac{\hat{\Psi}_k}{2}\right|, \end{aligned} \tag{61}$$

Then, one has

$$\begin{aligned} &\mathbb{E}_{q_{\theta^*_k}}[(y_j - \mu_k)^T \Sigma_k^{-1}(y_j - \mu_k)] \\ &= \mathbb{E}_{q_{\Sigma_k}}[(y_j - \hat{m}_k)^T \Sigma_k^{-1}(y_j - \hat{m}_k) + \mathrm{Tr}(\Sigma_k^{-1}\Sigma_k/\hat{\lambda}_k)] \\ &= \hat{\nu}_k (y_j - \hat{m}_k)^T \hat{\Psi}_k^{-1}(y_j - \hat{m}_k) + \frac{d}{\hat{\lambda}_k}. \end{aligned} \tag{62}$$

Plugging in all of the above expressions back into Eq. (57) yields, for $z_j = k$ and $k = 1, \ldots, K$, $q_{z_j}(k) \propto \tilde{q}_j(k)$ with $\tilde{q}_j(k)$ given in (39).

A.6 **VB-M-β step**

This step does not admit a closed-form expression but can be solved numerically. The maximization in β admits a unique solution. Indeed, it is equivalent to solve the following equation:

$$\begin{aligned} \hat{\beta} &= \arg\max_\beta \mathbb{E}_{q_{\mathbf{z}} q_{\boldsymbol{\tau}}}[\log p(\mathbf{z} \mid \boldsymbol{\tau}; \beta)] \\ &= \arg\max_\beta \mathbb{E}_{q_{\mathbf{z}} q_{\boldsymbol{\tau}}}[V(\mathbf{z}; \boldsymbol{\tau}, \beta)] - \mathbb{E}_{q_{\boldsymbol{\tau}}}[\log C(\boldsymbol{\tau}, \beta)]. \end{aligned} \tag{63}$$

where C denotes the normalizing constant that depends on $\boldsymbol{\tau}$ and β. Denoting the gradient vector and Hessian matrix respectively by ∇_β and ∇^2_β, it comes that

$$\begin{aligned}\nabla_\beta \mathbb{E}_{q_{\mathbf{z}} q_{\boldsymbol{\tau}}}[\log p(\mathbf{z} \mid \boldsymbol{\tau}; \beta)] &= \mathbb{E}_{q_{\mathbf{z}} q_{\boldsymbol{\tau}}}[\nabla_\beta V(\mathbf{z} \mid \boldsymbol{\tau}; \beta)] - \mathbb{E}_{p(\mathbf{z}|\boldsymbol{\tau};\beta) q_{\boldsymbol{\tau}}}[\nabla_\beta V(\mathbf{z} \mid \boldsymbol{\tau}; \beta)], \\ \nabla^2_\beta \mathbb{E}_{q_{\mathbf{z}} q_{\boldsymbol{\tau}}}[\log p(\mathbf{z} \mid \boldsymbol{\tau}; \beta)] &= \mathbb{E}_{q_{\mathbf{z}} q_{\boldsymbol{\tau}}}[\nabla^2_\beta V(\mathbf{z} \mid \boldsymbol{\tau}; \beta)] - \mathbb{E}_{p(\mathbf{z}|\boldsymbol{\tau};\beta) q_{\boldsymbol{\tau}}}[\nabla^2_\beta V(\mathbf{z} \mid \boldsymbol{\tau}; \beta)] \\ &\quad - \mathbb{E}_{q_{\boldsymbol{\tau}}}[\mathrm{Var}_{p(\mathbf{z}|\boldsymbol{\tau};\beta)}[\nabla_\beta V(\mathbf{z} \mid \boldsymbol{\tau}; \beta)]].\end{aligned} \tag{64}$$

It follows that whenever $V(\mathbf{z} \mid \boldsymbol{\tau}; \beta)$ is linear in β, $\nabla^2_\beta V(\mathbf{z} \mid \boldsymbol{\tau}; \beta)$ is zero, the Hessian matrix is negative semidefinite and the function to optimize is concave.

Unfortunately, due to the intractable normalizing constant C, it is necessary to evaluate the terms involving $p(\mathbf{z} \mid \boldsymbol{\tau}; \beta)$ in an approximate manner. A natural approach is to use a mean-field-like approximation that consists of replacing all neighbors in the interaction term by non-random values. Thus, the Potts model is approximated by

$$p(\mathbf{z} \mid \boldsymbol{\tau}; \beta) \simeq \prod_{j=1}^{n} p^{\mathrm{MF}}_{z_j}(z_j \mid \boldsymbol{\tau}; \beta), \tag{65}$$

with $p^{\mathrm{MF}}_{z_j}(z_j \mid \boldsymbol{\tau}; \beta)$ defined as

$$p^{\mathrm{MF}}_{z_j}(z_j = k \mid \boldsymbol{\tau}; \beta) \propto \exp\left(\log \pi_k(\boldsymbol{\tau}) + \beta \sum_{i \in \mathcal{N}_j} q_{z_i}(k)\right). \tag{66}$$

This approximation induced by the posterior variational approximation has been proposed in [8,17] and also exploited in [9]. It thus follows that

$$\begin{aligned}\mathbb{E}_{q_{\mathbf{z}} q_{\boldsymbol{\tau}}}[\nabla_\beta V(\mathbf{z} \mid \boldsymbol{\tau}; \beta)] &= \sum_{k=1}^{K} \sum_{i \in \mathcal{N}_j} q_{z_j}(k) q_{z_i}(k) \\ \mathbb{E}_{p(\mathbf{z}|\boldsymbol{\tau};\beta) q_{\boldsymbol{\tau}}}[\nabla_\beta V(\mathbf{z} \mid \boldsymbol{\tau}; \beta)] &\simeq \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\sum_{k=1}^{K} \sum_{i \in \mathcal{N}_j} p^{\mathrm{MF}}_{z_j}(k \mid \boldsymbol{\tau}; \beta) p^{\mathrm{MF}}_{z_i}(k \mid \boldsymbol{\tau}; \beta)\right].\end{aligned} \tag{67}$$

The additional difficulty is that we have to compute the expectation with respect to each q_{τ_k}. This can be done using simulations. To avoid the Monte Carlo sum, we can use instead of (66) another approximation where the random $\boldsymbol{\tau}$ is replaced by a set of fixed values $\widetilde{\boldsymbol{\tau}}$ given by

$$\widetilde{\tau}_k = \mathbb{E}_{q_{\tau_k}}[\tau_k] = \frac{\hat{\gamma}_{k,1}}{\hat{\gamma}_{k,1} + \hat{\gamma}_{k,2}}. \tag{68}$$

Thus, Eq. (66) turns into

$$p^{\mathrm{MF}}_{z_j}(z_j = k; \beta) \propto \exp\left(\log \pi_k(\widetilde{\boldsymbol{\tau}}) + \beta \sum_{i \in \mathcal{N}_j} q_{z_i}(k)\right), \tag{69}$$

where $\log \pi_k(\widetilde{\boldsymbol{\tau}}) = \log \widetilde{\tau}_k + \sum_{l=1}^{k-1} \log(1 - \widetilde{\tau}_l)$. Similarly, it follows that

$$\mathbb{E}_{p(\mathbf{z}|\boldsymbol{\tau};\beta) q_{\boldsymbol{\tau}}}[\nabla_\beta V(\mathbf{z} \mid \boldsymbol{\tau}; \beta)] \simeq \sum_{k=1}^{K} \sum_{i \in \mathcal{N}_j} p^{\mathrm{MF}}_{z_j}(k; \beta) p^{\mathrm{MF}}_{z_i}(k; \beta). \tag{70}$$

By setting them equal to each other and solving this equation for β over an interval, say $[0, 10]$, one obtains an updated value $\hat{\beta}$.

A.7 VB-M-(s_1, s_2, a) step

For the sake of simplicity, we use for σ a uniform prior so that parameter a does not have to be taken into account. With the previous choice of priors on α and σ for PY-MRF mixture models, this VB-M-step becomes

$$(s_1, s_2)^{(r)} = \underset{(s_1,s_2)}{\arg\max}\, \mathbb{E}_{q^{(r)}_{\alpha,\sigma}}[\log p(\alpha \mid \sigma; s_1, s_2)] \tag{71}$$

But the issue is now that the precise form of $q^{(r)}_{\alpha,\sigma}$ is not known. We can use again importance sampling for the optimization.

More specifically, these steps do not admit an explicit closed-form expression but can be solved numerically using gradient ascent schemes. Indeed, for s_1, s_2, it is equivalent to solve

$$\begin{aligned}\nabla_{s_1} \mathbb{E}_{q_\alpha}[\log p(\alpha \mid \sigma; s_1, s_2)] &= \mathbb{E}_{q_\alpha}[\nabla_{s_1} \log p(\alpha \mid \sigma; s_1, s_2)] \\ &= \log s_2 + \mathbb{E}_{q_{\alpha,\sigma}}[\log(\alpha + \sigma)] - \Psi(s_1) \\ &= 0 \\ \nabla_{s_2} \mathbb{E}_{q_\alpha}[\log p(\alpha \mid \sigma; s_1, s_2)] &= \mathbb{E}_{q_\alpha}[\nabla_{s_2} \log p(\alpha \mid \sigma; s_1, s_2)] \\ &= \frac{s_1}{s_2} - \mathbb{E}_{q_\alpha}[\alpha] - \mathbb{E}_{q_\sigma}[\sigma] \\ &= 0.\end{aligned} \tag{72}$$

As before, when $\sigma = 0$, this step simplifies into $s_1^{(r)} = \hat{s}_1^{(r)}$ and $s_2^{(r)} = \hat{s}_2^{(r)}$, namely to the DP-MRF case.

A.8 Initialization of the VBEM algorithm

An important question which is not often addressed in details is how to initialize the VBEM algorithm. In contrast to the standard EM that we can equivalently start with an initial E-step or an initial M-step, VBEM requires several steps to be initialized depending on the complexity of the model.

In the present work, we propose the following procedure for initializing the VBEM algorithm. First, we set values for s_1 and s_2, which are taken in our experiments to $s_1 = 1$ and $s_2 = 200/K$. From this, we can initialize the VB-E-(α, σ) step by setting $\mathbb{E}[\alpha] = s_1/s_2$ and $\mathbb{E}[\sigma] = 0$. It is then required to set values for the other hyperparameters. The interaction parameter β can be initialized to $\beta = 0$ assuming no initial spatial interaction and for the ρ_k's defining the Normal-inverse-Wishart priors, we suggest to use for all k, $m_k = \mathbf{0}$, $\Psi_k = \mathbf{1} \cdot 10^3$, $\nu_k = d$ and $\lambda_k = 1$ in order to start with a large variance for the μ_k's.

In addition, we need to initialize the cluster assignments which correspond to the VB-E-$\mathbf{Z}$ step. Several approaches are possible depending on the available information and model. A common way often used in image segmentation is to initialize the $q_Z(z_j)$'s randomly or using an initial segmentation coming either from preliminary information or from a simpler non spatial clustering procedure, *e.g. k-means*. In our experiments, an initialization into K clusters obtained with *k-means++* [4] was used. *k-means++* is basically identical to the *k-means* algorithm, except for the selection of initial centers. More concretely, *k-means++* starts with allocating one cluster center at random and then searches for the next ones which will be selected with a probability proportional to the distance to the closest center already chosen. The essential idea is to make all centers that would be selected as far away as possible from each other. It should be noted that *k-means++* also uses random initialization as a starting point, so it can give different results on different runs. To overcome the issue of poor initialization, we propose to run *k-means++* several times and use the labels that yield the best compactness (the sum of squared distances from each point to their corresponding center) to initialize the $q_Z(z_j)$'s and thus update the ρ_k's. From the initialization of $q_Z(\mathbf{z})$ and $\boldsymbol{\rho}$, the VB-E-$\boldsymbol{\tau}$ and VB-E-$\boldsymbol{\theta}^*$ steps can then be derived. This simple scheme has the advantage of accelerating the convergence of the VBEM algorithm.

Acknowledgements This article was developed in the framework of the Grenoble Alpes Data Institute, supported by the French National Research Agency under the "Investissements d'avenir" program (ANR-15-IDEX-02).

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2282 (2012)
2. Albughdadi, M., Chaâri, L., Tourneret, J., Forbes, F., Ciuciu, P.: A Bayesian non-parametric hidden Markov random model for hemodynamic brain parcellation. Signal Processing **135**, 132–146 (2017)
3. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 898–916 (2011)
4. Arthur, D., Vassilvitskii, S.: K-means++: The advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07, pp. 1027–1035. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2007). URL `http://dl.acm.org/citation.cfm?id=1283383.1283494`
5. Beal, M., Ghahramani, Z.: The variational Bayesian EM Algorithm for incomplete data: with application to scoring graphical model structures. In: J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, M. West (eds.) Bayesian Statistics, pp. 453–464. Oxford University Press (2003)
6. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society. Series B (Methodological) pp. 192–236 (1974)
7. Blei, D.M., Jordan, M.I.: Variational inference for Dirichlet process mixtures. Bayesian Anal. **1**(1), 121–143 (2006)
8. Celeux, G., Forbes, F., Peyrard, N.: EM procedures using mean field-like approximations for Markov model-based image segmentation. Pattern Recognition **36**, 131–144 (2003)
9. Chaari, L., Vincent, T., Forbes, F., Dojat, M., Ciuciu, P.: Fast joint detection-estimation of evoked brain activity in event-related fmri using a variational approach. IEEE Trans. Med. Imag. **32**(5), 821–837 (2013)
10. Chandler, D.: Introduction to modern statistical mechanics. Oxford University Press, New York, Oxford (1987). URL `http://opac.inria.fr/record=b1081336`
11. Chatzis, S.P.: A Markov random field-regulated Pitman-Yor process prior for spatially constrained data clustering. Pattern Recognition **46**(6), 1595–1603 (2013)
12. Chatzis, S.P., Tsechpenakis, G.: The infinite hidden Markov random field model. IEEE Trans. Neural Networks **21**(6), 1004–1014 (2010)
13. Corduneanu, A., Bishop, C.M.: Variational Bayesian Model Selection for Mixture Distributions. In: Proceedings Eighth International Conference on Artificial Intelligence and Statistics, pp. 27–34. Morgan Kaufmann (2001)
14. De Blasi, P., Favaro, S., Lijoi, A., Mena, R.H., Prünster, I., Ruggiero, M.: Are Gibbs-type priors the most natural generalization of the Dirichlet process? IEEE Trans. Pattern Anal. Mach. Intell. **37**(2), 212–229 (2015)
15. Favaro, S., Lijoi, A., Nava, C., Nipoti, B., Prünster, I., Teh, Y.W.: On the Stick-Breaking Representation for Homogeneous NRMIs. Bayesian Anal. **11**(3), 697–724 (2016)
16. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. The Annals of Statistics pp. 209–230 (1973)
17. Forbes, F., Peyrard, N.: Hidden Markov Random Field model selection criteria based on mean field-like approximations. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1089–1101 (2003)
18. Ghosal, S., Van der Vaart, A.: Fundamentals of nonparametric Bayesian inference, vol. 44. Cambridge University Press (2017)
19. Green, P.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82**, 711–732 (1995)
20. Ishwaran, H., James, L.F.: Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association **96**(453), 161–173 (2001)
21. Johnson, T.D., Liu, Z., Bartsch, A.J., Nichols, T.E.: A Bayesian non-parametric Potts model with application to pre-surgical fMRI data. Statistical Methods in Medical Research **22**(4), 364–381 (2013)
22. McLachlan, G., Krishnan, T.: The EM Algorithm and Extensions. Wiley (1996)

23. Miller, J.W., Harrison, M.T.: Mixture models with a prior on the number of components. Journal of the American Statistical Association **113**, 340–356 (2018)
24. Murphy, K.P.: Conjugate bayesian analysis of the gaussian distribution. def **1**($2\sigma 2$), 16 (2007)
25. Neal, R.M., Hinton, G.E.: A view of the EM algorithm that justifies incremental, sparse and other variants. In: Jordan (ed.) Lear. in Graph. Mod., pp. 355–368 (1998)
26. Orbanz, P., Buhmann, J.M.: Nonparametric Bayesian image segmentation. International Journal of Computer Vision **77**(1-3), 25–45 (2008)
27. Pitman, J.: Combinatorial stochastic processes, *Lecture Notes in Mathematics*, vol. 1875. Springer-Verlag, Berlin (2006). Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002
28. Pitman, J., Yor, M.: The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator. The Annals of Probability **25**(2), 855–900 (1997)
29. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association **66**(336), 846–850 (1971)
30. da Silva, A.R.F.: A Dirichlet process mixture model for brain MRI tissue classification. Medical Image Analysis **11**(2), 169–182 (2007)
31. Sodjo, J., Giremus, A., Dobigeon, N., Giovannelli, J.F.: A generalized Swendsen-Wang algorithm for Bayesian nonparametric joint segmentation of multiple images. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1882–1886. IEEE, La Nouvelle Orléans, LA, United States (2017)
32. Stoehr, J.: A review on statistical inference methods for discrete Markov random fields. arXiv e-prints arXiv:1704.03331 (2017)
33. Sudderth, E.B., Jordan, M.I.: Shared segmentation of natural scenes using dependent Pitman-Yor processes. In: Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008, pp. 1585–1592 (2008)
34. Teh, Y.W.: A Bayesian interpretation of interpolated Kneser-Ney. Technical report (2006)
35. Unnikrishnan, R., Pantofaru, C., Hebert, M.: A measure for objective evaluation of image segmentation algorithms. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops, pp. 34–34 (2005)
36. Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. International Journal of Computer Vision **62**(1), 61–81 (2005)
37. Wang, C., Blei, D.M.: Truncation-free stochastic variational inference for Bayesian nonparametric models. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12, pp. 413–421 (2012)
38. Xu, D., Caron, F., Doucet, A.: Bayesian nonparametric image segmentation using a generalized Swendsen-Wang algorithm. ArXiv e-prints (2016)