A Spatially-Constrained Normalized Gamma Process for Data Clustering

Sotirios P. Chatzis, Dimitrios Korkinof, and Yiannis Demiris

Abstract

In this work, we propose a novel nonparametric Bayesian method for clustering of data with spatial interdependencies. Specifically, we devise a novel normalized Gamma process, regulated by a simplified (pointwise) Markov random field (Gibbsian) distribution with a countably infinite number of states. As a result of its construction, the proposed model allows for introducing spatial dependencies in the clustering mechanics of the normalized Gamma process, thus yielding a novel nonparametric Bayesian method for spatial data clustering. We derive an efficient truncated variational Bayesian algorithm for model inference. We examine the efficacy of our approach by considering an image segmentation application using a real-world dataset. We show that our approach outperforms related methods from the field of Bayesian nonparametrics, including the infinite hidden Markov random field model, and the Dirichlet process prior.

Keywords: Clustering; Markov random field; normalized Gamma process.

1. Introduction

Nonparametric Bayesian modeling techniques, especially Dirichlet process mixture (DPM) models, have become very popular in statistics over the last few years, for performing nonparametric density estimation (Walker et al., 1999; Neal, 2000; Muller and Quintana, 2004). This theory is based on the observation that an infinite number of component distributions in an ordinary finite mixture model (clustering model) tends on the limit to a Dirichlet process (DP) prior (Neal, 2000; Antoniak, 1974). Indeed, although theoretically a DPM model has an infinite number of parameters, it turns out that inference for the model is possible, since only the parameters of a finite number of the mixture components need to be represented explicitly. Eventually, the nonparametric Bayesian inference scheme induced by a DPM model yields a posterior distribution on the proper number of model component densities (inferred clusters) (Blei and Jordan, 2004), rather than selecting a fixed number of mixture components. Hence, the obtained nonparametric Bayesian formulation eliminates the need of doing inference (or making arbitrary choices) on the number of mixture components (clusters) necessary to represent the modeled data.

Markov random fields (MRFs) (Orbanz and Buhmann, 2008) are a classical methodology for modeling spatially-interdependent data. In essence, MRFs

impose a Gibbsian distribution over the allocation of the modeled data into states (clusters), which enforces the belief that spatially adjacent data are more likely to cluster together. As the Gibbsian prior imposed by MRFs entails complex calculations that make it intractable in real-world problems dealing with large datasets, efficient approximations of the full MRF distribution are usually employed. For example, a pointwise simplification of the MRF prior based on the *mean-field principle* from statistical mechanics (Zhang, 1993) was employed in Celeux et al. (2003). Recently, MRFs have also been used in the context of Bayesian nonparametrics yielding the infinite hidden Markov random field (iHMRF) model (Chatzis and Tsechpenakis, 2009, 2010). This model obtains a joint MRF-Dirichlet process prior for spatially-constrained data clustering. As such, it introduces a nonparametric Bayesian approach to hidden MRF models, that is a novel formulation for such models that entails a countably infinite number of constituent states.

Inspired by these advances, in this paper we come up with a different approach towards clustering data with spatial interdependencies. We propose a spatially-adaptive random measure, coined the Markov random field normalized Gamma process (MRF-NGP). Our model is based on the introduction of a normalized Gamma process (NGP) controlled by an additionally postulated pointwise Markov random field imposed over the data allocation into model states, obtained by application of the mean-field principle (Chatzis and Tsechpenakis, 2009). As a result of its construction, the proposed prior discounts or increases the probability of cluster allocation for each observed data point depending on the allocation of the rest of the data points in its neighborhood, where the neighborhoods are defined as sets of spatially interdependent data points in the modeled datasets. We provide an efficient truncated algorithm for model inference based on the variational Bayesian paradigm. We empirically study the performance of the MRF-NGP prior in an image segmentation application, using a publicly available benchmark dataset, and compare it to the iHMRF model and the Dirichlet process prior.

The remainder of this paper is organized as follows: In Section 2, we provide a brief presentation of the theoretical background of the proposed method. Initially, we review the Dirichlet process and its function as a prior in nonparametric Bayesian models; subsequently, we briefly describe the theory of Markov random fields, and their pointwise approximations obtained on the basis of the mean-field principle. In Section 3, the proposed nonparametric prior for clustering data with spatial dependencies is introduced, and an efficient variational Bayesian algorithm for model inference is derived. In Section 4, the experimental evaluation of the proposed algorithm is conducted, considering an unsupervised image segmentation application using benchmark data. In the final section, our results are summarized and discussed.

2. Theoretical Background

2.1. The Dirichlet Process

Dirichlet process (DP) models were first introduced in Ferguson (1973). A DP is characterized by a base distribution G_0 and a positive scalar α, usually referred to as the innovation parameter, and is denoted as $\mathrm{DP}(\alpha, G_0)$. Essentially, a DP is a distribution placed over a distribution. Let us suppose we randomly draw a sample distribution G from a DP, and, subsequently, we independently draw M random variables $\{\Theta_m^*\}_{m=1}^M$ from G:

$$G|\alpha, G_0 \sim \mathrm{DP}(\alpha, G_0) \tag{1}$$

$$\Theta_m^*|G \sim G, \quad m = 1, \dots M \tag{2}$$

Integrating out G, the joint distribution of the variables $\{\Theta_m^*\}_{m=1}^M$ can be shown to exhibit a clustering effect. Specifically, given the first $M-1$ samples of G, $\{\Theta_m^*\}_{m=1}^{M-1}$, it can be shown that a new sample Θ_M^* is either (a) drawn from the base distribution G_0 with probability $\frac{\alpha}{\alpha+M-1}$, or (b) is selected from the existing draws, according to a multinomial allocation, with probabilities proportional to the number of the previous draws with the same allocation (Blackwell and MacQueen, 1973). Let $\{\Theta_c\}_{c=1}^C$ be the set of distinct values taken by the variables $\{\Theta_m^*\}_{m=1}^{M-1}$. Denoting as f_c^{M-1} the number of variables in $\{\Theta_m^*\}_{m=1}^{M-1}$ that equal to Θ_c, the distribution of Θ_M^* given $\{\Theta_m^*\}_{m=1}^{M-1}$ can be shown to be of the form (Blackwell and MacQueen, 1973)

$$\begin{aligned} p(\Theta_M^*|\{\Theta_m^*\}_{m=1}^{M-1}, \alpha, G_0) = &\frac{\alpha}{\alpha + M - 1} G_0 \\ &+ \sum_{c=1}^{C} \frac{f_c^{M-1}}{\alpha + M - 1} \delta_{\Theta_c} \end{aligned} \tag{3}$$

where δ_{Θ_c} denotes the distribution concentrated at a single point Θ_c. These results illustrate two key properties of the DP scheme. First, the innovation parameter α plays a key-role in determining the number of distinct parameter values. A larger α induces a higher tendency of drawing new parameters from the base distribution G_0; indeed, as $\alpha \to \infty$ we get $G \to G_0$. On the contrary, as $\alpha \to 0$ all $\{\Theta_m^*\}_{m=1}^M$ tend to cluster to a single random variable. Second, the more often a parameter is shared, the more likely it will be shared in the future.

A characterization of the (unconditional) distribution of the random variable G drawn from a Dirichlet process $\mathrm{DP}(G_0, \alpha)$ is provided by the stick-breaking construction of Sethuraman (1994). Consider two infinite collections of independent random variables $\boldsymbol{v} = (v_c)_{c=1}^\infty$, $\{\Theta_c\}_{c=1}^\infty$, where the v_c are drawn from the Beta distribution $\mathrm{Beta}(1, \alpha)$, and the Θ_c are independently drawn from the base distribution G_0. The stick-breaking representation of G is then given by (Sethuraman, 1994)

$$G = \sum_{c=1}^{\infty} \varpi_c(\boldsymbol{v}) \delta_{\Theta_c} \tag{4}$$

where

$$\varpi_c(\boldsymbol{v}) = v_c \prod_{j=1}^{c-1}(1 - v_j) \quad \in [0, 1] \tag{5}$$

and

$$\sum_{c=1}^{\infty} \varpi_c(\boldsymbol{v}) = 1 \tag{6}$$

The stick-breaking representation of the DP makes clear that the random variable G drawn from a DP is discrete. It shows explicitly that the support of G consists of a countably infinite sum of atoms located at Θ_c, drawn independently from G_0. It is also apparent that the innovation parameter α controls the mean value of the stick variables, v_c, as a hyperparameter of their prior distribution; hence, it regulates the effective number of the distinct values of the drawn atoms (Sethuraman, 1994).

2.2. Markov random fields

We consider an alphabet $Q = \{1, ..., K\}$. Let S be a finite index set, $S = \{1, ..., N\}$; we shall refer to this set, S, as the set of sites or locations. Let us consider for every site $j \in S$ a finite space $\mathcal{Z}_j$ of states z_j, such as $\mathcal{Z}_j = \{z_j : z_j \in Q\}$. The product space $\mathcal{Z} = \prod_{j=1}^{N} \mathcal{Z}_j$ will be denoted as the space of the configurations of the state values of the considered sites set, $\boldsymbol{z} = (z_j)_{j \in S}$. A strictly positive probability distribution, $p(\boldsymbol{z})$, $\boldsymbol{z} \in \mathcal{Z}$, on the product space $\mathcal{Z}$ is called a random field (Maroquin et al., 1987).

Let ∂ denote a neighborhood system on S, i.e. a collection $\partial = \{\partial_j : j \in S\}$ of sets, such as $j \notin \partial_j$ and $l \in \partial_j$ if and only if $j \in \partial_l$ $\forall l, j \in S$. Then, the previously considered random field, $p(\boldsymbol{z})$, is a Markov random field with respect to the introduced neighborhood system ∂ if (Geman and Geman, 1984)

$$p(z_j | \boldsymbol{z}_{S-\{j\}}) = p(z_j | \boldsymbol{z}_{\partial_j}) \; \forall j \in S \tag{7}$$

The distribution $p(\boldsymbol{z})$ of a Markov random field can be shown to be of a Gibbsian form (Clifford, 1990):

$$p(\boldsymbol{z}) \triangleq \frac{1}{W(\gamma)} \exp\left(- \sum_{c \in \mathcal{C}} V_c(\boldsymbol{z}|\gamma) \right) \tag{8}$$

where γ is the inverse temperature of the model, $W(\gamma)$ is the (normalizing) partition function of the model, $V_c(\boldsymbol{z}|\gamma)$ are the clique potentials of the model, and $\mathcal{C}$ is the set of the cliques included in the model neighborhood system.

A significant problem of MRF models concerns computational tractability, as the normalizing term $W(\gamma)$ is hard to compute in applications dealing with large datasets. Usually, these computations are conducted by means of Bayesian sampling, e.g. using Markov chain Monte Carlo methods Chalmond (1989). Nevertheless, such methods still require a large amount of computation. An alternative to these approaches is the mean-field approximation (Zhang, 1993;

Chatzis and Varvarigou, 2008). It is based on the idea of neglecting the fluctuations of the sites interacting with a considered site, so that the resulting system behaves as one composed of independent variables for which computation becomes tractable. That is, given an estimate $\hat{\boldsymbol{z}}$ of the unknown site labels vector $\boldsymbol{z}$, obtained by means of a stochastic restoration criterion, such as the iterative conditional modes (ICM) or the marginal posterior modes (MPM) algorithm (see, e.g., Geman and Geman (1984); Chatzis and Varvarigou (2008)), we make the hypothesis (Qian and Titterington, 1991)

$$p(\boldsymbol{z}) = \prod_{j=1}^{N} p(z_j | \hat{\boldsymbol{z}}_{\partial_j}; \gamma) \tag{9}$$

where

$$p(z_j = i | \hat{\boldsymbol{z}}_{\partial_j}; \gamma) = \frac{\exp(-\sum_{c \ni j} V_c(\tilde{\boldsymbol{z}}_{ij} | \gamma))}{\sum_{h=1}^{K} \exp(-\sum_{c \ni j} V_c(\tilde{\boldsymbol{z}}_{hj} | \gamma))} \tag{10}$$

$\tilde{\boldsymbol{z}}_{ij} \triangleq (z_j = i, \hat{\boldsymbol{z}}_{\partial_j})$, $\hat{\boldsymbol{z}}_{\partial_j}$ is the estimate of the jth site neighborhood, and the indexes c refer to the cliques that contain the jth site.

3. Proposed Approach

3.1. Model Formulation

Let us consider a set of observations $Y = \{\boldsymbol{y}_n\}_{n=1}^{N}$, $\boldsymbol{y}_n \in \mathcal{Y}$, measured over a set of sites $\mathcal{S} = \{1, ..., S\}$ on which a neighborhood system ∂ is defined. Let us denote as $X = \{x_n\}_{n=1}^{N}$, $x_n \in \mathcal{S}$, the sites where the observed data points $\{\boldsymbol{y}_n\}_{n=1}^{N}$ were measured. We aim to obtain a clustering algorithm which takes into account the prior information regarding the adjacencies of the observed data in the neighborhood system ∂, promoting clustering of data measured in positions adjacent in the neighborhood system ∂, and discouraging clustering of data points relatively near in the feature space $\mathcal{Y}$ but measured in remote locations in ∂. For this purpose, we seek to provide an MRF-driven nonparametric prior for clustering the observed data Y.

Let us introduce the latent variables $\{z_n\}_{n=1}^{N}$ denoting the model state (cluster) where an observed data point $\boldsymbol{y}_n$ measured at the location x_n is assigned by our model. Motivated by merits and the theory of the DP discussed in the previous section, to derive the sought model, we make the key-assumption, based on the mean-field-based approximation of the MRF distribution, that for any given site x_n, we have available an estimate $\hat{\boldsymbol{z}}_{\partial_n}$ of the value $\boldsymbol{z}_{\partial_n} \triangleq (z_m)_{m \in \partial_n}$ of the latent cluster assignment variables of the observations measured at sites in the neighborhood of site x_n. Apparently, this assumption entails a priori application of a methodology for obtaining an initial estimate of the latent variables $\{z_n\}_{n=1}^{N}$ for the modeled data, as discussed in Section 2.2. Additionally, it requires updating of these estimates on each iteration of the model inference algorithm, as we shall discuss in Section 3.2. Further, we consider the following

predictor (location)-dependent random measure

$$G(x) = \sum_{i=1}^{\infty} \varpi_i(x)\delta_{\Theta_i} \tag{11}$$

where

$$\varpi_i(x) = \frac{\Lambda_i(x)}{\sum_{j=1}^{\infty} \Lambda_j(x)} \tag{12}$$

the random variables Λ_i follow a Gamma distribution as

$$\Lambda_i | x_n \sim \mathcal{G}(\alpha k_i(x_n; \hat{\boldsymbol{z}}_{\partial_n}), 1) \tag{13}$$

α is the innovation parameter of the process, $k_i\left(x_n; \hat{\boldsymbol{z}}_{\partial_n}\right)$ is the probability of the nth site being assigned to the ith cluster as computed by the employed pointwise MRF distribution

$$\begin{aligned} k_i\left(x_n; \hat{\boldsymbol{z}}_{\partial_n}\right) \triangleq & p\left(z_n = i \middle| \hat{\boldsymbol{z}}_{\partial_n}; \gamma\right) \\ = & \frac{\exp(-\sum_{c \ni x_n} V_c(\tilde{\boldsymbol{z}}_{ni}|\gamma))}{\sum_{h=1}^{\infty} \exp(-\sum_{c \ni x_n} V_c(\tilde{\boldsymbol{z}}_{nh}|\gamma))} \end{aligned} \tag{14}$$

$\tilde{\boldsymbol{z}}_{ni} \triangleq (z_n = i, \hat{\boldsymbol{z}}_{\partial_n})$, $\hat{\boldsymbol{z}}_{\partial_n}$ is the current estimate of the nth site neighborhood, $V_c(\cdot)$ are the employed clique potential functions, and the indexes c refer to the cliques that include the nth site, x_n. The utility of the pointwise MRF distribution $k_i\left(x_n; \hat{\boldsymbol{z}}_{\partial_n}\right)$ in our model consists in reducing the probability (discounting) of clusters that seem rather unlikely from the viewpoint of the postulated neighborhood system. We dub this random probability measure $G(x)$ the MRF-NGP process. A proof that the normalizing constant in the denominator of (12) is finite almost surely is provided in the Appendix.

3.2. Variational Bayesian Inference

Inference for nonparametric models can be conducted under a Bayesian setting, typically by means of variational Bayes (e.g., Blei and Jordan (2006)), or Monte Carlo techniques (e.g., Qi et al. (2007)). Here, we prefer a variational Bayesian approach, due to its considerably better scalability in terms of computational costs, which becomes of importance when dealing with large datasets. Let us a consider a set of observations $Y = \{\boldsymbol{y}_n\}_{n=1}^N$ with corresponding locations $X = \{x_n\}_{n=1}^N$. We postulate for our observed data a likelihood function of the form

$$p(\boldsymbol{y}_n | z_n = i) = p(\boldsymbol{y}_n | \boldsymbol{\theta}_i) \tag{15}$$

while for the latent assignment variables z_n we consider

$$p(z_n = i | x_n) = \varpi_i(x_n) \tag{16}$$

where the $\varpi_i(x)$ are given by (12), with the prior over the $\Lambda_i(x)$ given by (13). Regarding the likelihood parameters $\boldsymbol{\theta}_i$, we impose a suitable conjugate exponential prior over them; for instance, in case of a Gaussian likelihood function

$$p(\boldsymbol{y}_n | \boldsymbol{\theta}_i) = \mathcal{N}(\boldsymbol{y}_n | \boldsymbol{\mu}_i, \boldsymbol{R}_i) \tag{17}$$

we impose a Normal-Wishart prior over the likelihood parameters $\boldsymbol{\theta}_i = \{\boldsymbol{\mu}_i, \boldsymbol{R}_i\}$, i.e.

$$p(\boldsymbol{\mu}_i, \boldsymbol{R}_i) = \mathcal{NW}(\boldsymbol{\mu}_i, \boldsymbol{R}_i | \lambda_i, \boldsymbol{m}_i, \omega_i, \boldsymbol{\Omega}_i) \tag{18}$$

Regarding the MRF temperature parameter γ, and the innovation parameter α, we choose to optimize them as model hyperparameters, as part of the variational inference procedure discussed next.

Our variational Bayesian inference formalism consists in derivation of a family of variational posterior distributions $q(.)$ which approximate the true posterior distribution over the infinite sets $\{z_n\}_{n=1}^{N}$, $\{\Lambda_i(x_n)\}_{i,n=1}^{\infty,N}$, and $\{\boldsymbol{\theta}_i\}_{i=1}^{\infty}$. Apparently, under this infinite dimensional setting, Bayesian inference is not tractable. For this reason, we employ a common strategy in the literature of Bayesian nonparametrics: we fix a value K and we let the variational posterior over the $\Lambda_k(x)$ have the property $q(\Lambda_{k>K}(x) = 0) = 1,\ \forall x \in \mathcal{S}$ (Blei and Jordan, 2006). In other words, we set $\varpi_k(x)$ equal to zero for $k > K$, $\forall x \in \mathcal{S}$. Note that, under this setting, the treated model involves a full MRF-NGP prior; truncation is not imposed on the MRF-NGP prior itself, but only on the variational distribution to allow for a tractable inference procedure. Hence, the truncation level K is a variational parameter which can be freely set, and not part of the prior model specification.

Let $W = \{\{z_n\}_{n=1}^{N}, \{(\Lambda_k(x_n))_{k=1}^{K}\}_{n=1}^{N}, \{\boldsymbol{\theta}_k\}_{k=1}^{K}\}$ be the set of the parameters of our truncated model over which a prior distribution has been imposed, and Ξ be the set of the hyperparameters of the model, comprising the γ, the innovation parameter α, and the hyperparameters of the priors over the likelihood parameters $\boldsymbol{\theta}_k$ of the model. Variational Bayesian inference consists in derivation of an approximate posterior $q(W)$ by maximization (in an iterative fashion) of the variational free energy

$$\mathcal{L}(q) = \int \mathrm{d}W q(W) \log \frac{p(X, Y, W|\Xi)}{q(W)} \tag{19}$$

which provides a lower bound to the computationally intractable log marginal likelihood (log evidence), $\log p(X, Y)$, of the model (Jordan et al., 1998).

Having considered a conjugate exponential prior configuration, the variational posterior $q(W)$ is expected to take the same functional form as the prior, $p(W)$ (Bishop, 2006). Thus, the variational free energy of our model reads

(ignoring constant terms)

$$\begin{aligned}\mathcal{L}(q) = & \\ & \sum_{k=1}^{K-1}\sum_{n=1}^{N}\int \mathrm{d}\Lambda_k(x_n)q(\Lambda_k(x_n))\log\frac{p(\Lambda_k(x_n))}{q(\Lambda_k(x_n))} \\ & +\sum_{k=1}^{K}\int \mathrm{d}\boldsymbol{\theta}_k q(\boldsymbol{\theta}_k)\log\frac{p(\boldsymbol{\theta}_k)}{q(\boldsymbol{\theta}_k)} + \sum_{k=1}^{K}\sum_{n=1}^{N} q(z_n=k) \\ & \times\left\{\int \mathrm{d}\boldsymbol{\Lambda}(x_n)q(\boldsymbol{\Lambda}(x_n))\log p(z_n=k|x_n)\right. \\ & \left. -\log q(z_n=k) + \int \mathrm{d}\boldsymbol{\theta}_k q(\boldsymbol{\theta}_k)\log p(\boldsymbol{y}_n|\boldsymbol{\theta}_k)\right\}\end{aligned} \tag{20}$$

where $\boldsymbol{\Lambda}(x) = (\Lambda_k(x))_{k=1}^{K}$. Based on Jensen's inequality, the term $\int \mathrm{d}\boldsymbol{\Lambda}(x_n)q(\boldsymbol{\Lambda}(x_n))\log p(z_n = i|x_n)$ in (20) yields

$$\begin{aligned}&\int \mathrm{d}\boldsymbol{\Lambda}(x_n)q(\boldsymbol{\Lambda}(x_n))\log p(z_n=i|x_n) \\ &= \int \mathrm{d}\boldsymbol{\Lambda}(x_n)q(\boldsymbol{\Lambda}(x_n))\log\frac{\Lambda_i(x_n)}{\sum_{j=1}^{K}\Lambda_j(x_n)} \\ &= \int \mathrm{d}\boldsymbol{\Lambda}(x_n)q(\boldsymbol{\Lambda}(x_n))\left[\log\Lambda_i(x_n) - \log\sum_{j=1}^{K}\Lambda_j(x_n)\right] \\ &\geq \int \mathrm{d}\boldsymbol{\Lambda}(x_n)q(\boldsymbol{\Lambda}(x_n))\log\Lambda_i(x_n) \\ &\quad -\log\sum_{j=1}^{K}\int \mathrm{d}\boldsymbol{\Lambda}(x_n)q(\boldsymbol{\Lambda}(x_n))\Lambda_j(x_n)\end{aligned} \tag{21}$$

This latter result shall be exploited to obtain the variational posteriors of our model in the analysis that follows.

3.3. Variational Posteriors

Derivation of the variational posterior distribution $q(W)$ involves maximization of the variational free energy $\mathcal{L}(q)$ over each one of the factors of $q(W)$ in turn, holding the others fixed, in an iterative manner (Chandler, 1987). By construction, this iterative, consecutive updating of the variational posterior distribution is guaranteed to monotonically and maximally increase the free energy $\mathcal{L}(q)$, which functions as the convergence criterion of the derived inference algorithm for our model (Chatzis et al., 2008).

The derived algorithm is in essence an expectation-maximization-like algorithm. Each iteration comprises an E-step, on which the variational posteriors over the model latent variables are computed, and an M-step, on which the variational posteriors over the model parameters are updated. Let us denote as $\langle . \rangle$ the posterior expectation of a quantity.

3.3.1. M-step

This step comprises the updates of the Gamma-distributed variables $\Lambda_i(x_n)$

$$q(\Lambda_i(x_n)) = \mathcal{G}(\Lambda_i(x_n)|\beta_{ni}, \xi_{ni}) \tag{22}$$

where

$$\beta_{ni} = \alpha k_i(x_n; \hat{\boldsymbol{z}}_{\partial_n}) + q(z_n = i) \tag{23}$$

$$\xi_{ni} = 1 + \frac{1}{\sum_{j=1}^{K} \langle \Lambda_j(x_n) \rangle} \tag{24}$$

and

$$\langle \Lambda_j(x_n) \rangle = \frac{\beta_{nj}}{\xi_{nj}} \tag{25}$$

as well as of the parameters $\boldsymbol{\theta}_i$, for which we obtain the general solution

$$\log q(\boldsymbol{\theta}_i) \propto \log p(\boldsymbol{\theta}_i) + \sum_{n=1}^{N} q(z_n = i) \log p(\boldsymbol{y}_n|\boldsymbol{\theta}_i) \tag{26}$$

which is similar to the corresponding solution for models imposing simple DP priors over their cluster assignment distributions.

3.3.2. E-step

This step comprises the updates of the posteriors $q(z_n = i)$:

$$q(z_n = i) \propto \exp\left(\langle \log \Lambda_i(x_n) \rangle\right) \exp\left(\varphi_{ni}\right) \tag{27}$$

where

$$\langle \log \Lambda_i(x_n) \rangle = \psi(\beta_{ni}) - \log \xi_{ni} \tag{28}$$

and

$$\varphi_{nj} = \langle \log p(\boldsymbol{y}_n|\boldsymbol{\theta}_j) \rangle \tag{29}$$

It also consists in updating the estimates of the assignment variables $\hat{\boldsymbol{z}} = (\hat{z}_n)_{n=1}^{N}$ which are used in computing the pointwise MRF priors employed in our model to regulate cluster discounting. For this purpose, we simply set

$$\hat{z}_n = \text{argmax}_{i=1}^{K} q(z_n = i) \tag{30}$$

Finally, regarding the model hyperparameters Ξ, their values can be either heuristically selected or estimated by means of type-II maximum likelihood. Indeed, in this work, we obtain estimates of the hyperparameter γ by maximization of the lower bound $\mathcal{L}(q)$, and we heuristically select the values of the rest of the model hyperparameters.

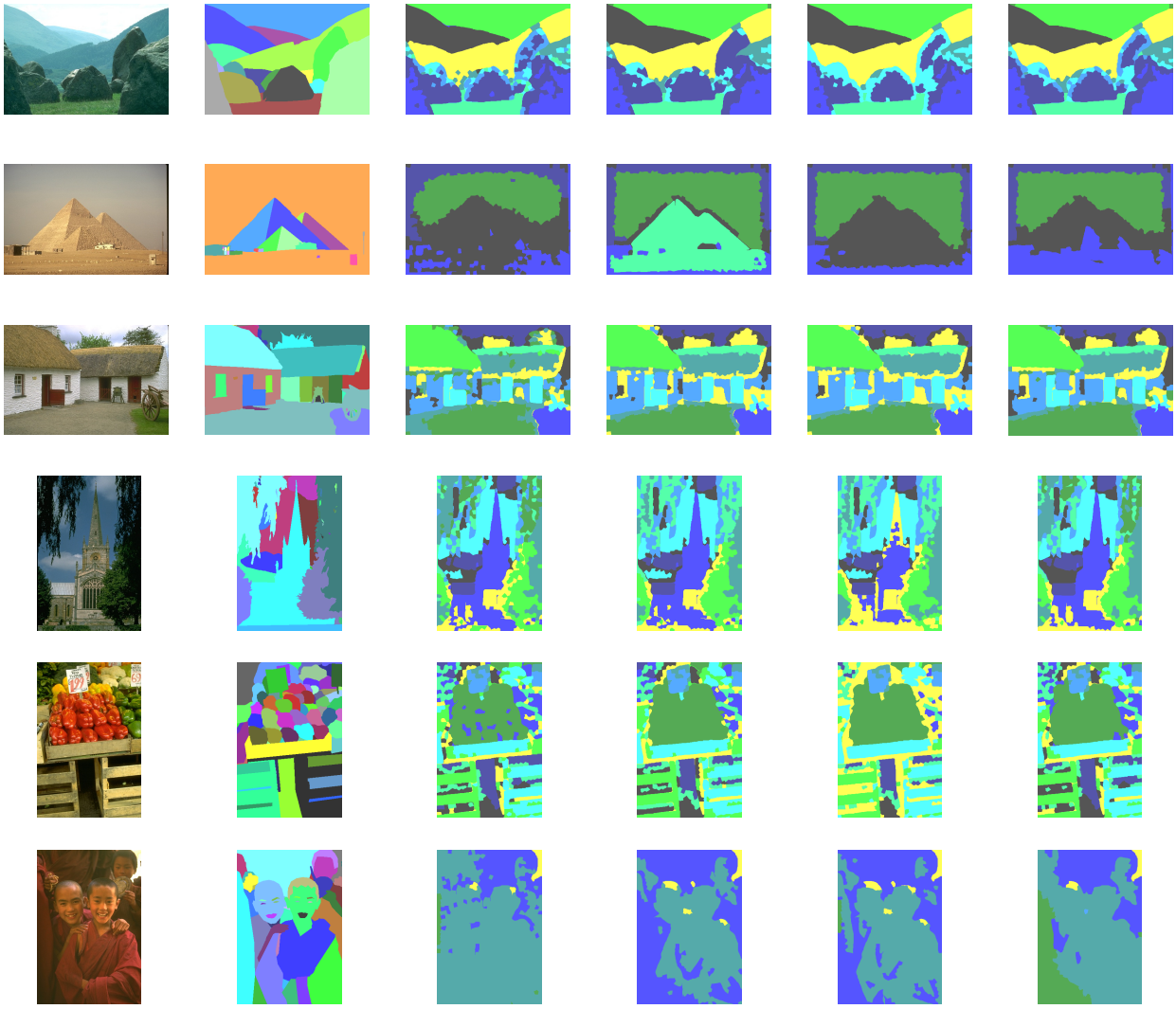

Figure 1: Few selected 321x481 images from the Berkeley image segmentation dataset. **Left-to-right**: a) Original image, b) One human groundtruth, c) K-means initialization, d) iHMRF, e) MRF-NGP. **Top-to-bottom**: a) #241004, b) #161062, c) #385028, d) #301007, e)#25098, f)#246053.

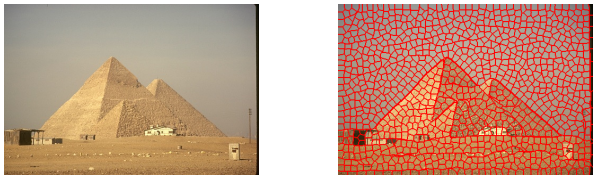

Figure 2: Example of superpixel segmentation (Mori, 2005).

Table 1: Obtained PRI results for the considered subset of the Berkeley benchmark.

Image #	DPM	iHMRF	MRF-NGP
159029	0.7688	0.7727	0.7842
20008	0.8376	0.8514	0.8478
100075	0.7851	0.7795	0.7861
301007	0.8438	0.8432	0.8460
122048	0.7396	0.7520	0.7421
145053	0.6189	0.6315	0.7304
236017	0.5997	0.6035	0.6346
170054	0.6841	0.7453	0.7628
385028	0.8520	0.8393	0.8544
67079	0.7344	0.7347	0.7599
209070	0.6335	0.7006	0.7380
27059	0.8359	0.8470	0.8669
176019	0.6930	0.7470	0.7343
246053	0.5896	0.6006	0.6526
239096	0.7570	0.7957	0.7871
323016	0.8283	0.8436	0.8479
231015	0.8019	0.8185	0.8138
25098	0.8270	0.8231	0.8394
8143	0.6294	0.6605	0.7011
35010	0.7701	0.7854	0.8051
15004	0.7561	0.7865	0.7994
100080	0.8067	0.8093	0.8036
161062	0.6610	0.6280	0.6742
159045	0.6984	0.7315	0.7482
170057	0.6948	0.7243	0.7498
89072	0.7377	0.7590	0.7841
175032	0.5594	0.6783	0.6686
86016	0.7379	0.7664	0.7573
103070	0.7164	0.7307	0.7530
241004	0.8643	0.8642	0.8738

Table 2: Mean and median of the PRI metric across the considered subset of the Berkeley benchmark.

PRI(%)	DPM	iHMRF	MRF-NGP
Mean	73.54	75.51	77.15
Median	73.88	76.27	77.34

4. Experimental Evaluation

Here, we investigate the efficacy of our approach considering an unsupervised image segmentation application. Specifically, we consider segmentation of real-world images, using a subset of the Berkeley image segmentation benchmark (Martin et al., 2001). The Berkeley image segmentation benchmark comprises a set of 300, 321x481 real-world color images along with their segmentation maps provided by different individuals. Given the multiple groundtruths available for each image within the used dataset, to obtain an objective performance evaluation of the proposed algorithm, we employ the probabilistic rand index (PR index or PRI) (Unnikrishnan et al., 2005). The PR index counts the fraction of pairs of pixels whose labelings are consistent between a computed segmentation and the given groundtruth, averaging across multiple groundtruth segmentations to account for scale variation in human perception. Denoting as $G = \{G_1, G_2, ...G_M\}$ a set of groundtruth images, and as G_{eval} a segmentation map under evaluation, it holds

$$PR(G_{eval}, G) = \frac{2}{s(s-1)} \sum_i \sum_{j>i} [c_{ij} p_{ij} + (1 - c_{ij})(1 - p_{ij})] \tag{31}$$

where $c_{ij} = 1$ if pixels i and j belong to the same segment in G_{eval}, $c_{ij} = 0$ otherwise, s is the number of image pixels, and p_{ij} is the groundtruth probability that pixels i and j belong to the same segment, that is the fraction of the available groundtruths where pixels i and j belong to the same segment. It has been shown (Unnikrishnan and Hebert, 2005) that the PR index possesses the desirable property of being robust to segmentation maps resulting from groundtruth segment splitting or merging. It takes values between 0 and 1, with the values close to 0 indicating a bad segmentation result, and the values close to 1 indicating a good result.

In all our experiments, we use Gaussian likelihoods and choose to impose a Normal-Wishart prior over the likelihood parameters. We compare the performance of our approach to iHMRF and DPM, both using the same likelihood function and prior over the likelihood function parameters as in the case of our model. All the evaluated algorithms are initialized by means of the k-means algorithm. Regarding the potential functions of the imposed pointwise MRFs for both the evaluated iHMRF and MRF-NGP models, we opt for a simple Potts model with a second order (8-neighbors) neighborhood system, yielding

$$p(z_n = c | \hat{\boldsymbol{z}}_{\partial_n}; \gamma) = \frac{\exp(\gamma \sum_{l \in \partial_n} \delta(c - \hat{z}_l))}{\sum_{h=1}^{K} \exp(\gamma \sum_{l \in \partial_n} \delta(h - \hat{z}_l))} \tag{32}$$

for the pointwise MRF priors, where K is the truncation threshold, and $\delta(.)$ stands for the Kronecker's delta function, given by

$$\delta(x_j - x_l) = \begin{cases} 1, \text{ if } x_j = x_l \\ 0, \text{ otherwise} \end{cases}$$

Feature extraction is effected as follows: First, each image is segmented into approximately $N = 1000$ superpixels using the method proposed in Mori (2005); an example superpixel segmentation is shown in Fig. 2. We then compute feature vectors at superpixel level, comprising RGB and HSV color information along with the values of the Maximum Response (MR) filter banks (Varma and Zisserman, 2002). The truncation level of the variational Bayesian algorithm for all the treated models is set to $K = 10$.

In an attempt to account for the effect of poor model initialization, which may lead model training to yield bad local optima as model estimators, we execute our experiments multiple times for each image, with different initializations each time, common for all the evaluated algorithms. The visual segmentation result is presented for 6 selected images in Fig. 1, along with the original image, one human groundtruth, and the initialization. The mean PRI results (over the executed repetitions) for the whole considered dataset are presented in Table 1. Total results across all images are presented in Table 2. Based on the obtained PRI metric results, we can conclude that the MRF-NGP performs considerably better than all the considered rival methods. Note also that small differences in the values of the PRI metric correspond to significant differences in the quality of the obtained segmentation results (Nikou et al., 2007). The illustrated segmentation results vouch for this assertion.

5. Conclusions

In this paper, we proposed a method for nonparametric clustering of data with general spatial interdependencies. Our method, coined the MRF-NGP, consists in postulating a normalized Gamma process, the cluster prior probabilities of which are discounted by means of a simplified pointwise Markov random field imposed over data point allocation into clusters. As a result of this construction, the MRF-NGP imposes the belief that spatially proximate data are more likely to cluster together. To examine the efficacy of our approach, we evaluated it in unsupervised image segmentation tasks using a real-life benchmark dataset, namely the Berkeley image segmentation benchmark (Martin et al., 2001). We showed that it yields a considerable improvement in the obtained performance of the clustering algorithm compared to both the DPM, and the recently proposed iHMRF model.

The source codes allowing for the replication of the here presented results shall be made available through the website of the authors: `http://www.iis.ee.ic.ac.uk/~sotirios`.

Acknowledgment

This work has been funded by the EU FP7 ALIZ-E project (contract #248116).

Appendix

Here, we prove the almost sure finiteness of the normalizing factor $\sum_{j=1}^{\infty} \Lambda_j(x_n)$ in (12). Let

$$S_T \triangleq \sum_{j=1}^{T} \Lambda_j(x_n) \tag{33}$$

It follows that $S_1 \leq S_2 \leq \cdots \leq S_T \leq \cdots \leq S$, where

$$S \triangleq \lim_{T\to\infty} S_T \tag{34}$$

since the random variables $\Lambda_j(x_n)$ are non-negative, as they follow a Beta distribution.

Then, to prove that S is finite almost surely, we only need to prove that $\mathbb{E}[S]$ is finite. From the monotone convergence theorem, we yield

$$\mathbb{E}[S] = \lim_{T\to\infty} \mathbb{E}[S_T] = \lim_{T\to\infty} \sum_{j=1}^{T} \mathbb{E}[\Lambda_j(x_n)] = \alpha \tag{35}$$

since $\lim_{T\to\infty} \sum_{j=1}^{T} k_j(x_n; \hat{\boldsymbol{z}}_{\partial_n}) = 1$, as the $k_j(x_n; \hat{\boldsymbol{z}}_{\partial_n})$ comprise prior MRF-derived probabilities of the observation at the nth site being assigned to any of the postulated model states. Hence, we have proven that S is finite almost surely.

References

Antoniak, C., 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics 2 (6), 1152–1174.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Springer, New York.

Blackwell, D., MacQueen, J., 1973. Ferguson distributions via Pólya urn schemes. The Annals of Statistics 1 (2), 353–355.

Blei, D., Jordan, M., July 2004. Variational methods for the Dirichlet process. In: 21st Int. Conf. Machine Learning. New York, NY, USA, pp. 12–19.

Blei, D. M., Jordan, M. I., 2006. Variational inference for Dirichlet process mixtures. Bayesian Analysis 1 (1), 121–144.

Celeux, G., Forbes, F., Peyrard, N., 2003. EM procedures using mean field-like approximations for Markov model-based image segmentation. Pattern Recognition 36 (1), 131–144.

Chalmond, B., 1989. An iterative Gibbsian technique for reconstruction of m-ary images. Pattern Recognition 22 (6), 747–761.

Chandler, D., 1987. Introduction to Modern Statistical Mechanics. Oxford University Press, New York.

Chatzis, S., Kosmopoulos, D., Varvarigou, T., March 2008. Signal modeling and classification using a robust latent space model based on t distributions. IEEE Trans. Signal Processing 56 (3), 949–963.

Chatzis, S. P., Tsechpenakis, G., September 2009. The infinite hidden Markov random field model. In: Proc. 12th International IEEE Conference on Computer Vision (ICCV). Kyoto, Japan, pp. 654–661.

Chatzis, S. P., Tsechpenakis, G., 2010. The infinite hidden Markov random field model. IEEE Transactions on Neural Networks 21 (6), 1004–1014.

Chatzis, S. P., Varvarigou, T. A., October 2008. A fuzzy clustering approach toward hidden Markov random field models for enhanced spatially constrained image segmentation. IEEE Transactions on Fuzzy Systems 16 (5), 1351–1361.

Clifford, P., 1990. Markov random fields in statistics. In: Grimmett, G., Welsh, D. (Eds.), Disorder in physical systems. A volume in honour of John M. Hammersley on the occasion of his 70th birthday. Oxford Science Publication, Clarendon Press, Oxford.

Ferguson, T., 1973. A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1, 209–230.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–741.

Jordan, M., Ghahramani, Z., Jaakkola, T., Saul, L., 1998. An introduction to variational methods for graphical models. In: Jordan, M. (Ed.), Learning in Graphical Models. Kluwer, Dordrecht, pp. 105–162.

Maroquin, J., Mitte, S., Poggio, T., 1987. Probabilistic solution of ill-posed problems in computational vision. Journal of The American Statistical Association 82, 76–89.

Martin, D., Fowlkes, C., Tal, D., Malik, J., July 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. 8th Int'l Conf. Computer Vision. Vancouver, Canada, pp. 416–423.

Mori, G., 2005. Guiding model search using segmentation. In: Proc. 10th IEEE Int. Conf. on Computer Vision (ICCV).

Muller, P., Quintana, F., 2004. Nonparametric Bayesian data analysis. Statist. Sci. 19 (1), 95–110.

Neal, R., 2000. Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Statist. 9, 249–265.

Nikou, C., Galatsanos, N., Likas, A., 2007. A class-adaptive spatially variant mixture model for image segmentation. IEEE Transactions on Image Processing 16 (4), 1121–1130.

Orbanz, P., Buhmann, J., 2008. Nonparametric Bayes image segmentation. International Journal of Computer Vision 77, 25–45.

Qi, Y., Paisley, J. W., Carin, L., 2007. Music analysis using hidden Markov mixture models. IEEE Transactions on Signal Processing 55 (11), 5209–5224.

Qian, W., Titterington, D., 1991. Estimation of parameters in hidden Markov models. Philosophical Transactions of the Royal Society of London A 337, 407–428.

Sethuraman, J., 1994. A constructive definition of the Dirichlet prior. Statistica Sinica 2, 639–650.

Unnikrishnan, R., Hebert, M., January 2005. Measures of similarity. In: Proc. 7th IEEE Workshop on Applications of Computer Vision. Breckenridge, Colorado, pp. 394–400.

Unnikrishnan, R., Pantofaru, C., Hebert, M., June 2005. A measure for objective evaluation of image segmentation algorithms. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. San Diego, CA, USA, pp. 34–41.

Varma, M., Zisserman, A., 2002. Classifying images of materials: Achieving view-point and illumination independence. In: Proc. 7th IEEE European Conf. on Computer Vision (ECCV).

Walker, S., Damien, P., Laud, P., Smith, A., 1999. Bayesian nonparametric inference for random distributions and related functions. J. Roy. Statist. Soc. B 61 (3), 485–527.

Zhang, J., 1993. The mean field theory in EM procedures for Markov random fields. IEEE Transactions on Image Processing 2 (1), 27–40.