Majorization-Minimization mixture model determination in image segmentation

Giorgos Sfikas[1]*, Christophoros Nikou[2], Nikolaos Galatsanos[3] and Christian Heinrich[1]

[1]LSIIT (UMR CNRS-ULP 7005), University of Strasbourg, France
[2]Department of Computer Science, University of Ioannina, Greece
[3]Department of Electrical and Computer Engineering, University of Patras, Greece

Abstract

A new Bayesian model for image segmentation based on a Gaussian mixture model is proposed. The model structure allows the automatic determination of the number of segments while ensuring spatial smoothness of the final output. This is achieved by defining two separate mixture weight sets: the first set of weights is spatially variant and incorporates an MRF edge-preserving smoothing prior; the second set of weights is governed by a Dirichlet prior in order to prune unnecessary mixture components. The model is trained using variational inference and the Majorization-Minimization (MM) algorithm, resulting in closed-form parameter updates. The algorithm was successfully evaluated in terms of various segmentation indices using the Berkeley image data base.

1. Introduction

Image segmentation is the process of grouping image pixels according to the coherence of certain attributes such as intensity, spatial relation, texture. As such, it has been popularly addressed as a special type of data clustering problem by various techniques [16].

Choosing the appropriate number of clusters for a given data set is an important issue, on which several approaches have been proposed. The most straightforward model selection approach is fitting a number of models with varying number of components, and evaluating the solutions using a suitable criterion. Such penalty terms, inspired by coding theory and minimum description length, try to avoid data overfitting by penalizing solutions with high number of components. Examples include Akaike's information criterion, the Bayesian information criterion and the Minimum message length criterion [1].

In the more specific context of assuming the data being generated by a mixture model, methodologies include notably the Bayesian approaches [5, 21] where the number of kernels and the model parameters are estimated simultaneously. In this family of methods, the model is initialized on a large number of components, and progressively removes those components that reside in the same region of the data space. On the contrary, in [4], the model starts with a low number of components and more kernels are progressively added by splitting existing kernels when necessary.

It is straightforward to adapt these Bayesian models into image segmentation, simply by assuming our image feature vectors to be the data to be clustered. However, the important feature of an image's spatial structure would not be accounted for. Natural images have a spatial smoothness property which is neglected by standard mixture model approaches. Approaches like the method proposed in [12] are based on MRF priors [9] to account for spatial characteristics. However, they assume an *a priori* known number of segments.

In this paper, we present a Bayesian model for image segmentation that enables the estimation of the number of segments during the training process while accounting for image spatial smoothness. We assume that the distribution of the hidden class labels is controlled by two distinct sets of probability weight vectors, tagged correspondingly as *local* and *global* weights.

The *local* weights are varying with each pixel. Local differences in these weights follow a Student's-*t* distribution. The Student's-*t* distribution decomposes on two levels: the lower level is a Gaussian pdf with precision (inverse variance) that is spatially variant, while the higher level is a Gamma pdf that generates the aforementioned precision values. This precision variability of the Gaussians allows the model to incorporate elegantly the image edge structure along with imposing smoothness constraints.

The *global* weights control the number of image segments that are active in the model by imposing a Dirichlet prior on them. In this way, more probable solutions, which otherwise exhibit high model complexity by comprising many kernels, are penalized as low probability states. This

*Giorgos Sfikas was supported by a grant from *Région Alsace* (France)

allows the model to estimate the number of classes in the segmentation process, by starting from an initial high number of classes estimate and pruning mixing kernels gradually during the model training process.

The variational inference framework [1, 7] is used to train the model. Variational inference involves iteratively optimizing a lower bound of the model evidence with regard to the posterior distribution of the hidden variables, and the model parameters. The mean field approximation is employed on the posterior distribution of the hidden variables, so as to render its estimation tractable. Let us note that the proposed model is different with respect to standard Dirichlet priors imposed on the mixing proportions of a mixture [1]. In our model, the hidden variables depend on two priors (*local* and *global* weights) and model inference is not trivial with standard inference techniques. Therefore, we optimize the variational lower bound by making use of the Majorization-Minimization (MM) methodology [10].

Thus, unlike state-of-the-art methods in image segmentation like normalized cuts [20] and standard or spatially varying mixtures [19], the proposed model can produce an estimate of the number of image classes while at the same time ensuring a smooth segmentation result. Methods supporting automatic determination of the number of classes typically depend on a scalar parameter, or a small set of parameters, that more or less directly control the number of classes / fit likelihood trade-off (e.g. the bandwidth in mean shift [3]). Such parameters are meant to be beforehand empirically adjusted. Concerning the rest of the parameters, affecting the quality of the segmentation itself, the proposed model determines them automatically. We consider this issue as an advantage in comparison with other methods; graph-cut based methods fall under this latter category [19], as well as recently proposed extensions that can handle number of components determination [8].

Producing a smooth segmentation result while determining the number of classes has also been addressed in the Dirichlet process prior models proposed in [6, 15], among others. However such approaches rely on sampling techniques which are notoriously computationally expensive, in contrast to the Majorization-Minimization iterative scheme we propose in this work.

2. Model description

Let $\mathbf{X} = \{x^n\}_{n=1}^N$ be the observed set of the image intensities. Consider also that there exist *at most* K classes in our segmentation. Each datum x^n is governed by different statistics, according to which class it belongs to. Let us assume a hidden variable set $\mathbf{Z} = \{z_j^n\}_{n=1..N,\, j=1..K}$, grouped as N one-zero $K \times 1$ vectors that control pixel class membership.

It is a popular choice in computer vision to choose the data to be Gaussian and *i.i.d* distributed, assuming knowledge of class memberships $\mathbf{Z}$:

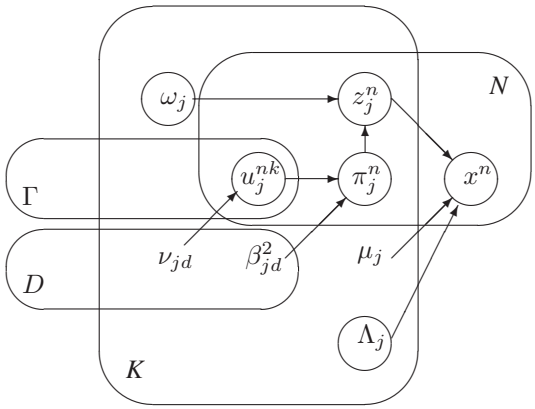

Figure 1. Graphical representation of the proposed model. Superscript $n \in [1, N]$ denotes pixel index, subscript $j \in [1, K]$ denotes kernel (segment) index, $d \in [1, D]$ describes the neighborhood direction type and $k \in [1, \Gamma]$ denotes neighbor index.

$$x^n | z_j^n = 1 \;\sim\; \mathcal{N}(\mu_j, \Lambda_j) \tag{1}$$

with $\mathcal{N}$ representing a Gaussian distribution with μ_j and Λ_j being the mean vector and the precision (inverse covariance) matrix respectively.

The distribution choice of $\mathbf{Z}$ plays a drastic role on the model behavior. Let us recall that under a multinomial and *i.i.d* assumption on the z^n, the model is essentially a Gaussian mixture [1] governed by a set of K weights. In [19] this idea is extended by using spatially varying weights along with a smoothness prior on them.

In this work we make a design choice meant to lie between the spatially and non-spatially varying hypotheses for the mixing proportions. Let z^n, $\forall n \in [1, N]$ be independently distributed with

$$p(z^n | \mathbf{\Omega}, \mathbf{\Pi}) = \frac{\prod_{j=1}^K (\pi_j^n \omega_j)^{z_j^n}}{\sum_{j=1}^K \pi_j^n \omega_j}. \tag{2}$$

In the equation above we have introduced $\mathbf{\Pi} = \{\pi_j^n\}_{n=1..N,\, j=1..K}$ and $\mathbf{\Omega} = \{\omega_j\}_{j=1..K}$ weight variable sets, which are constrained by $\sum_{j=1}^K \pi_j^n = 1, \forall n \in [1, N]$ and $\sum_{j=1}^K \omega_j = 1$. In view that the denominator in (2) is the probability distribution normalization constant, eq. (2) acts closely like a multinomial distribution with weights given by the set $[\pi_1^n \omega_1, \pi_2^n \omega_2, ..., \pi_K^n \omega_K]$. Thus, for each pixel $n \in [1, N]$, a class membership is determined by two sets of weight vectors. At first, it depends on the set $\mathbf{\Pi} = [\pi_1^n, \pi_2^n, ..., \pi_K^n]$ whose components are vectors, now called *local weights*. The local weights are spatially varying as they depend on the position (indexed by n). Secondly, it depends on the set $\mathbf{\Omega} = [\omega_1, \omega_2, ..., \omega_K]$ whose components are scalars and are now called *global weights*.

For further insight, note that, if in eq. (2), for a given pixel indexed by n, we treat its local weights π_j^n as parameters with probabilities equal to the uninformative $1/K$, the distribution of the hidden variables z^n is a multinomial distribution. This is also true for the global weights ω_j. Thus, in eq.(2), for a given a pixel, both the local and global weights should have high values for a certain class in order to be dominant.

2.1. Local weights

Considering the set of *local weights* $\mathbf{\Pi}$ as random variables and assuming a proper prior, we can incorporate the spatial smoothness trait by forcing neighbouring vectors to be more likely to share the same class label. We assume a Markov random field on $\mathbf{\Pi}$, which equivalently means that $\mathbf{\Pi}$ is governed by a Gibbs distribution [9], generally expressed by:

$$p(\mathbf{\Pi}) \propto \prod_C e^{-\Psi_c(\mathbf{\Pi})}, \tag{3}$$

where Ψ_c is a function on clique c, called *clique potential* function in the literature, and the product is over all minimal cliques of the Markov random field.

An appropriate clique distribution choice would be to assume that the differences of *local weights* $\mathbf{\Pi}$ follow a Student's-*t* distribution with its peak set at zero. This setting, proposed previously in [19], also provides our model with the properties of an edge-preserving line-process [9]. The probability law for local differences is thus expressed by

$$\pi_j^n - \pi_j^k \sim \mathcal{S}t(0, \beta_{jd}^2, \nu_{jd}), \quad \forall n, j, d, \forall k \in \gamma_d(n). \tag{4}$$

The parameters β_{jd} control how tightly smoothed we need the vectors of segment j to be. In eq. (4), D stands for the number of a pixel's neighbourhood adjacency types and $\gamma_d(n)$ is the set of neighbours of pixel indexed by n, with respect to the d^{th} adjacency type, where $d \in [1, D]$. In our model, we assume 4 neighbours for each pixel (*first-order* neighbourhood), and partition the corresponding adjacency types into horizontal and vertical, thus, setting $D = 2$. This variability of parameter aims to capture the intuitive property that smoothness statistics may vary along clusters and spatial directions.

It can be observed that the assumption in (4) is equivalent to

$$\pi_j^n - \pi_j^k \sim \mathcal{N}(0, \beta_{jd}^2 / u_j^{nk}),$$

$$u_j^{nk} \sim \mathcal{G}(\nu_{jd}/2, \nu_{jd}/2), \quad \forall n, j, d, \quad \forall k \in \gamma_d(n),$$

where $\mathcal{N}$ and $\mathcal{G}$ represent a Gaussian and a Gamma distribution respectively. This breaking-down of the Student's-*t* distribution allows clearer insight on how our implicit edge-preserving line-process works. Since u_j^{nk} depends on datum indexed by n, each weight difference in the MRF can be described by a different instance of a Gaussian distribution. Therefore, as $u_j^{nk} \to +\infty$ the distribution tightens around zero, and forces neighboring *local weights* to be smooth. On the other hand, $u_j^{nk} \to 0$ signifies the existence of an edge and consequently no smoothing.

2.2. Global weights

The *global weights* $\mathbf{\Omega}$ are introduced in the model in order to cover our model's second important property which is the automatic estimation of the number image segments. The idea is that starting from a predefined maximum segments figure K, during the training process some segments "fade out" eventually to zero-weights [5]. While it is possible that for a certain class j all local weights π_j^n, $\forall n \in [1, N]$ may attain negligible values, in practice this is difficult; this is due to the fact that updating each π_j^n $\forall n \in [1, N]$ individually must account for MRF local dependencies, which will not lead to an update far from each site's neighbours.

Thus, assigning for each class a single *global weight* scalar allows us to conveniently treat each class as a distinct entity during model training. Therefore, $\mathbf{\Omega}$ is considered to be a random vector governed by a Dirichlet distribution:

$$p(\mathbf{\Omega}; \alpha_0) \propto \prod_{j=1}^{K} \omega_j^{\alpha_0 - 1}. \tag{5}$$

By these means, we can penalize solutions with numerous non-zero components. As hyperparameter $\alpha_0 \to -\infty$, solutions with less segments are encouraged, and as $\alpha_0 \to +\infty$ all K initial segments tend to be preserved. While $\alpha_0 < 0$ may enforce the prior in (5) to be improper, in practice negative α_0 values are applicable since it is not necessary to compute the normalizing constant of eq. (5) during inference as it will be explained in the next section.

Finally we impose a Wishart prior on precision matrices $\Lambda_j, \forall j \in [1, K]$:

$$p(\Lambda_j; W_0, \eta_0) \propto |\Lambda_j|^{(\eta_0 - \Delta - 1)/2} e^{-\frac{1}{2} Tr(W_0^{-1} \Lambda_j)}, \tag{6}$$

where the matrix W_0 and the scalar η_0 are such that $\mathcal{E}\{\Lambda_j\} = W_0 \eta_0$. Parameter Δ stands for the number of variates of the feature vectors x^n. Imposing this prior on precision matrices avoids degenerate cases, for instance, when the corresponding covariance matrix Λ_j^{-1} has zero eigenvalues or equivalently $|\Lambda_j| \to +\infty$ [1]. For an overview of the proposed model, see fig.(1).

3. Model inference

To perform inference and consequently segmentation, the model likelihood with respect to model parameters has to be optimized:

$$\underset{\mu, \Pi, \Omega, \beta}{\operatorname{argmax}} \ln p(X, \Pi, \Omega; \mu, \beta, \nu).$$

Due to the functional form of the involved distributions the above optimization problem is practically intractable. Therefore, we resort to variational inference [1]. This involves calculating approximations of the posterior distributions $q(\cdot)$ of the hidden variables Z, U, Λ, then using them to find parameter estimates that maximize a lower bound of the model likelihood.

3.1. Variational inference

Adapting the standard variational methodology [1] to our problem, the lower bound to be optimized is

$$\mathcal{L}(q,\Pi,\Omega,\mu,\beta,\nu) \triangleq$$

$$\sum_{Z}\int_{U,\Lambda} q(Z,U,\Lambda)\ln\frac{p(X,\Pi,Z,U,\Omega;\mu,\beta,\nu)}{q(Z,U,\Lambda)}dUd\Lambda$$

$$=<\ln p(X,\Pi,Z,U,\Omega;\mu,\beta,\nu)>_{Z,U,\Lambda} -$$

$$<\ln q(Z,U,\Lambda)>_{Z,U,\Lambda}$$

$$=<\ln p(X|Z,\Lambda;\mu)>_{Z,\Lambda} + <\ln p(\Lambda)>_{\Lambda} +$$

$$<\ln p(Z|\Pi,\Omega)>_{Z} + <\ln p(\Pi|U;\beta)>_{U} + \ln p(\Omega) +$$

$$<\ln p(U;\nu)>_{U} - <\ln q(Z,U,\Lambda)>_{Z,U,\Lambda}\ . \qquad (7)$$

Model evidence is decomposed to the lower bound $\mathcal{L}$ and the Kullback-Leibler distance between the approximation of the posterior and the posterior itself:

$$\ln p(X,\Pi,\Omega;\mu,\beta,v) = \mathcal{L}(q,\Pi,\Omega,\mu,\beta,\nu) + KL(q||p).$$

To proceed with the computation of the optimal distribution q on $\mathcal{L}$, we recur to the *mean field approximation* which stems from statistical physics [1]:

$$q(Z,U,\Lambda) = q(Z)q(U)q(\Lambda). \qquad (8)$$

Note that in the proposed model, we only need to assume $q(Z,U,\Lambda) = q(Z,U)q(\Lambda)$, as $q(Z,U) = q(Z)q(U)$ is induced from the model structure (fig. 1). Thence we can obtain update equations for the expected values of hidden variables Z,U,Λ:

$$<z_j^n>^{(t)} = \frac{\pi_j^{n(t)}\omega_j^{(t)}\mathcal{N}(x^n;\mu_j^{(t)}|\ <\Lambda_j>^{(t)})}{\sum_{l=1}^{K}\pi_l^{n(t)}\omega_j^{(t)}\mathcal{N}(x^n;\mu_l^{(t)}|\ <\Lambda_l>^{(t)})}, \qquad (9)$$

$$<u_j^{nk}>^{(t)} = \zeta_j^{nk(t)}/\theta_j^{nk(t)},$$

$$<\ln u_j^{nk}>^{(t)} = \psi(\zeta_j^{nk(t)}) - \ln\theta_j^{nk(t)},$$

where $\psi(\cdot)$ stands for the digamma function, and parameters ζ, θ being:

$$\zeta_j^{nk(t)} = \frac{1}{2}\left(\nu_{jd}^{(t)} + 1\right),$$

$$\theta_j^{nk(t)} = \frac{1}{2}\left(\nu_{jd}^{(t)} + \frac{(\pi_j^{n(t)} - \pi_j^{k(t)})^2}{\beta_{jd}^{2(t)}}\right).$$

The required moment for variables Λ_j are given by

$$<\Lambda_j>^{(t)} = W_j^{(t)}\eta_j^{(t)} \qquad \eta_j^{(t)} = \eta_0 + \sum_{n=1}^{N} <z_j^n>^{(t)}$$

$$W_j^{-1(t)} = W_0^{-1} + \sum_{n=1}^{N} <z_j^n>^{(t)}\ (x^n - \mu_j^{(t)})(x^n - \mu_j^{(t)})^T$$

3.2. Majorization-Minimization

Estimation of the deterministic parameters μ, Π, Ω, β, ν is achieved by maximization of (7) with respect to them. However, optimizing (7) with respect to Ω is difficult due to the normalizing factor in (2). We can work around this obstacle and find a closed form update for Ω as well, by making use of the Majorization-Minimization (MM) methodology [10]. MM in its philosophy is quite close to variational inference and the EM algorithm [1], in the sense that the problem of minimizing a given objective function is transformed to successive minimizations of surrogate functions, i.e. majorizers of the original objective function that can be minimized in closed-form.

For the term of (7) involving $\ln p(Z|\Pi,\Omega)$, we note the following inequality:

$$<\ln p(Z|\Pi,\Omega)>_{Z} =$$

$$-\sum_{n=1}^{N}\ln\sum_{j=1}^{K}\pi_j^n\omega_j + \sum_{j=1}^{K}\ln\omega_j\sum_{n=1}^{N}\pi_j^n\ <z_j^n>\ \geqslant$$

$$-\sum_{n=1}^{N}\ln y^n - \sum_{n=1}^{N}\frac{1}{y^n}\left(\sum_{j=1}^{K}\pi_j^n\omega_j - y^n\right) + \sum_{j=1}^{K}\ln\omega_j\sum_{n=1}^{N}\pi_j^n\ <z_j^n>$$

$$\triangleq \varphi(Z,\Pi,\Omega,y) \qquad (10)$$

where we have introduced $\mathbf{y} = \{y^1, y^2, ..., y^N\}$ as a new set of auxiliary real parameters. In eq. (10), we made use of the linear *minorization*

$$f(x) \geqslant f(y) + \frac{df(y)}{dy}(x - y),$$

which holds for any convex function f. Here $f(x) = -\ln x$ and $x = \sum_{j=1}^{K}\pi_j^n\omega_j$.

Consequently, we define our *minorant* as

$$\mathcal{L}^{MM}(q,\mu,\beta,\nu,\Pi,\Omega,y) \triangleq \qquad (11)$$

$$\mathcal{L}(q,\mu,\beta,\nu,\Pi,\Omega) - <\ln p(Z|\Pi,\Omega)>_{Z} + \varphi(Z,\Pi,\Omega,y),$$

which according to (10) is easily confirmed to be a lower bound of (7). Therefore, we have an MM approach in a *Minorization-Maximization* sense.

Optimization of (11) leads to parameter value updates for μ, y, β, ν, $\mathbf{\Pi}$, $\mathbf{\Omega}$. For the first three parameter sets, the updates are

$$\mu_j^{(t+1)} = \frac{\sum_{n=1}^N <z_j^n>^{(t)} x^n}{\sum_{n=1}^N <z_j^n>^{(t)}}, \quad y^{n(t+1)} = \sum_{j=1}^K \omega_j \pi_j^n,$$

$$\beta_{jd}^{2(t+1)} = \frac{\sum_{n=1}^N \sum_{k\in\gamma_d(n)} <u_j^{nk}>^{(t)} (\pi_j^{n(t)} - \pi_j^{k(t)})^2}{\sum_{n=1}^N |\gamma_d(n)|}.$$

Setting the derivative of the lower bound (11) with respect to the degrees of freedom of the Student's -*t* distributions equal to zero we obtain $\nu_{jd}^{(t+1)}$ as the solutions of the equation:

$$\ln(\nu_{jd}^{(t+1)}/2) - \psi(\nu_{jd}^{(t+1)}/2) +$$
$$+\left[\frac{\sum_{n=1}^N \sum_{k\in\gamma_d(n)} (<\ln u_j^{nk}>^{(t)} - <u_j^{nk}>^{(t)})}{\sum_{n=1}^N |\gamma_d(n)|}\right] + 1 = 0,$$

The solution for parameter ν is obtained using the bisection method [17].

In order to estimate $\mathbf{\Pi}$ and $\mathbf{\Omega}$ we have the difficulty that the optimization is under positivity and sum-to-unity constraints as they have to be probability vectors defined by (2). The *local weights* π_j^n are computed as the roots of a quadratic equation:

$$a_j^n \left(\pi_j^{n(t+1)}\right)^2 + b_j^n \left(\pi_j^{n(t+1)}\right) + c_j^{n(t+1)} = 0 \qquad (12)$$

with coefficients:

$$a_j^n = -\sum_{d=1}^D \left\{ \beta_{jd}^{-2(t)} \sum_{k\in\gamma_d(n)} <u_j^{nk}>^{(t)} \right\},$$

$$b_j^n = \sum_{d=1}^D \left\{ \beta_{jd}^{-2(t)} \sum_{k\in\gamma_d(n)} <u_j^{nk}>^{(t)} \pi_j^{k(t)} \right\} - \frac{\omega_j}{2y^n},$$

$$c_j^n = \frac{1}{2} <z_j^n>^{(t)}.$$

The solutions of (12) for a given pixel, indexed by n, will not in general satisfy the constraints $\pi_j^n \geqslant 0$, $\sum_{j=1}^K \pi_j^n = 1$. In order to get proper mixing weight vectors we perform a projection step onto the constraints subspace using the quadratic programming algorithm described in [19].

Motivated by the form of the objective function to be optimized, we follow a different strategy for the estimation of the *global weights*. At first, the unconstrained optimizers are computed:

$$\tilde{\omega}_j = \frac{\sum_{n=1}^N <z_j^n> + \alpha_0 - 1}{\sum_{n=1}^N \pi_j^n / y^n} \qquad (13)$$

If $\tilde{\omega}_j < 0$, we fix the corresponding constrained solution to $\omega_j = 0$, so that they comply with the positivity constraint $\omega_j \geqslant 0$.

We carry on to the second step with the remaining $J \leqslant K$ non-zero components after relabelling them as $\{\omega_1, ..., \omega_J\}$ and the $K - J$ zero components as $\{\omega_{J+1}, ..., \omega_K\}$. Solving the corresponding equation subject to the constraint $\sum_{j=1}^J \omega_j = 1$, we obtain

$$\omega_j^{(t+1)} = \frac{\sum_{n=1}^N <z_j^n>^{(t)} + \alpha_0 - 1}{\lambda + \sum_{n=1}^N \pi_j^{n(t)} / y^{n(t)}} \qquad (14)$$

where λ is the Lagrange multiplier. Substituting ω_j, $j = 1, ..., J$ from (14) to the sum-to-unity constraint yields

$$\sum_{j=1}^J \frac{\sum_{n=1}^N <z_j^n>^{(t)} + \alpha_0 - 1}{\lambda + \sum_{n=1}^N \pi_j^{n(t)} / y^{n(t)}} - 1 = 0. \qquad (15)$$

Note that the left-hand side in (15) is continuous and monotonically decreasing function of:

$$\lambda \in \left[\max_j \left\{ -\sum_{n=1}^N \pi_j^{n(t)} / y^{n(t)} \right\}, \ +\infty \right)$$

Also, as $\lambda \to \max_j \left\{ -\sum_{n=1}^N \pi_j^{n(t)} / y^{n(t)} \right\}$, the left hand of (15) goes to $+\infty$ and as $\lambda \to +\infty$ the left hand side of (15) goes to -1. Thus, we can determine the solution for λ using the bisection method. Substituting it into (14) yields the updates for the constrained global weights.

Summing up, the updates presented here for the posterior distribution in section 3.1 and the updates for the model deterministic parameters in 3.2 constitute an iterative model training scheme. During the training process, some of the K ω_j global weight coefficients may gradually go down to zero. In view of update (9), no pixels will any longer be assigned to the corresponding class, and effectively these classes are pruned from the model.

The iterations terminate with lower bound convergence. Since bound convergence is guaranteed for both *MM* [10] and variational inference [1], the proposed MM-derived lower bound $\mathcal{L}^{MM}$ will also converge in a finite number of iterations.

4. Numerical Experiments

At first, we have applied the proposed model to the segmentation of a piecewise constant image slightly corrupted by white Gaussian noise at SNR of 20 dB(*Mondrian* [15], fig. 2). As this is a relatively easy segmentation problem, we use this example to show our algorithm results over varying Dirichlet hyperparameter values α_0. For convenience, we express α_0 in (5) as a function of the image size N and the maximum number of classes K, with

$\alpha_0 = -\epsilon^{-1}NK^{-1}$. In view of the global weights updates (14), the value of α_0 is compared to the sum of the expected values z_j^n which is of the order of NK^{-1}. Thus, values of ϵ close to one will encourage pruning of kernels. On the other hand, as ϵ approaches $N^{-1}K$ kernel pruning is progressively less encouraged. For the *Mondrian* image (fig. 2) we have set $\epsilon = 1, 5, 50, 1000$, starting from an initial number of kernels $K = 12$. Low values for parameter ϵ lead to underestimation of the true number of segments. High values of ϵ yield over-fitting problems.

Furthermore, to test the dependency of the estimated number of kernels on the initial number of segments K and the hyperparameter ϵ (and consequently the Dirichlet parameter α_0), we have run tests with varying parameter values on the *Mondrian* (fig. 2) and *Church* (fig. 3). The results presented in figure 4 show that for low values of ϵ, that is, penalizing configurations with high number of kernels, the final number of segments are almost invariant with regard to its initial value K, as it would be desired.

Figure 3. Natural image segmentation using Lab features for $\epsilon = 5$, $K = 7$ initial number of segments. The algorithm converged to four segments.

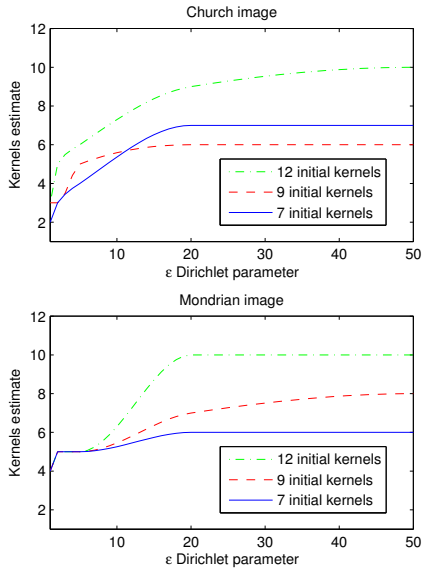

Figure 4. Estimate of the number of kernels for varying values of initial segments K and Dirichlet hyperparameter ϵ. Curves are interpolants for values at points $\epsilon = 1, 2, 5, 20, 50$.

In order to evaluate the combination of both the smoothing MRF, and the automatic determination of the number of components, we have compared the proposed algorithm to two other models. The first model comprises an MRF prior without any automatic component number selection. In that case, the *global weights* $\mathbf{\Omega}$ are inactive. The second model consists of estimating the number of segments but incorporates no smoothing prior. Therefore, in this case, the *local weights* $\mathbf{\Pi}$ are inactive. Models close to these two may be found in [19] and [5] respectively.

The experiment was run on a test 3-class piecewise constant image degraded by white additive Gaussian (SNR of 18 dB) as presented in figure 5. The initial number of segments was set to $K = 7$ and the Dirichlet parameter $\epsilon = 5$. The model with no kernel number selection fails as it identifies erroneously the noise as separate classes. Both models with kernel number selection successfully prune the extra kernels to the correct number of three. However, the proposed model succeeds also to deal with the noise due to its smoothing property.

Finally, we have tested our algorithm on the Berkeley natural image database [13]. We have used a superpixels initialization [14] as described in [22]. We start by oversegmenting the images, all at full resolution of 480×320, to typically around 200 superpixels each. Then, we associate to each superpixel the medoid of the color feature vectors of the pixels that belong to the superpixel in question. These superpixels medoids play the role of the X observed set for our algorithm and represent the whole set of pixels belonging to the corresponding superpixel. A region adjecency graph is also computed to keep track of the superpixel neighborhoods for the MRF imposed on the *local* weights. To this end, we consider for the MRF prior, those medoids that represent spatially adjacent superpixels.

We have quantitatively evaluated the segmentations using four performance measures [22]: the Rand index (RI), the variation of information (VI), the global consistency error (GCE) and the Boundary displacement error (BDE). The RI measures the consistency between human segmentations and the computed segmentation map. VI measures the amount of information one segmentation conveys about the other. GCE measures the degree of refinement between two segmentations. Finally, BDE measures the average chamfer distance between the boundaries of two segmentation maps. The mean values of the results over the 300 images of the data base are summarized in table 1. As it can be observed, all of the indices have values comparable to the ones obtained by state of the art techniques (for example [22]). It should be noted that the BDE depends highly on the image size. In figure 6, we present some representative results using an initial number of kernels $K = 15$ and $\epsilon = \{3, 10\}$. As a final remark concerning the parameter ϵ (or equivalently α_0), let us stress that (a) its choice amounts

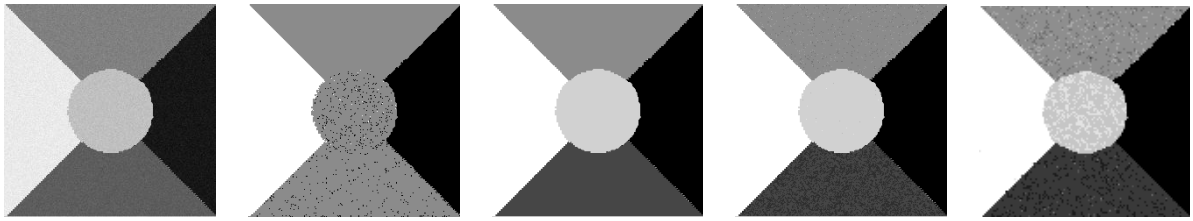

Figure 2. Segmentation results for the *Mondrian* image, over various Dirichlet hyperparameter values. From left to right: the original image after degradation by additive white Gaussian noise (SNR of 20 dB), segmentations using $\epsilon = 1$, $\epsilon = 5$, $\epsilon = 20$, $\epsilon = 1000$. For too low ϵ, the number of segments is undervalued. For too high ϵ, extra segments are formed erroneously out of image noise.

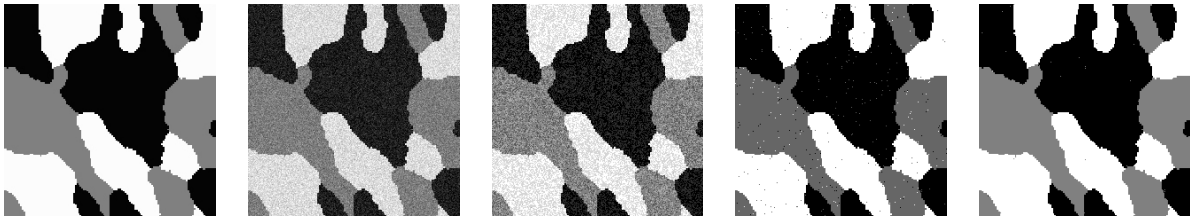

Figure 5. Segmentation results for a 3-class piecewise constant image. From left to right: original image, image degraded by additive Gaussian noise (SNR of 18 dB), segmentation using the proposed model without global weights (Rand index = 82.14%), segmentation using the proposed model with $\epsilon = 5$ and without local weights (Rand index = 99.2%), segmentation using the proposed model with $\epsilon = 5$ (Rand index = 99.86%).

Table 1. *Segmentation evaluation of the algorithm on the 300 images of the Berkeley image data base. The mean values of the corresponding indices are presented (see text for abbreviations). The model was applied to the original images of size* 480×320 *pixels and it was initialized to* $K = 15$ *starting number of classes. Results are computed over two different segmentation scales,* $\epsilon = 3$ *(lower number of classes) and* $\epsilon = 10$ *(higher number of classes).*

$Index$	RI	BDE	GCE	VI
$\epsilon = 3$	0.71	15.6	0.27	2.27
$\epsilon = 10$	0.72	14.4	0.31	2.52

to how coarse or fine we want the segmentation to be, (b) user-defined parameters with a similar role, affecting the resulting number of classes are used in other number-of-class-determining segmentation algorithms to our knowledge (e.g. the Dirichlet hyperparameter in [15] or the bandwidth in [3]). Nonetheless, in the experiments on natural images described here values of $\epsilon \simeq 5$ seem to have given the best results.

Algorithm runtime is in the order of a few minutes. On a 2 Ghz workstation each algorithm iteration took around 20 seconds for a 480×320 color image (MATLAB code), converging at $20 - 30$ iterations.

5. Conclusion

In this paper, we proposed a segmentation algorithm based on a Bayesian model. The main novelty of this work is the use of a smoothness MRF prior along with automatic selection of the number of segmentation classes. Updates for the model training are obtained in an efficient manner by variational inference and the Majorization-Minimization (MM) methodology. Recently proposed algorithms that combine a smoothness prior with automatic selection of the number of kernels have to resort to computationally expensive Monte Carlo sampling instead. As future work, novel MRF energy minimization techniques such as proposal-based fusion [11] could be integrated to our model, as MRF optimization is a critical step. The evaluation of the model to image data bases using more sophisticated features for natural images such as the MRF texture features [18] and the Blobworld features [2] is also envisaged.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 2169, 2170, 2171, 2172, 2173

[2] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: image segmentation using Expectation-Maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, 2002. 2175

[3] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002. 2170, 2175

[4] C. Constantinopoulos and A. Likas. Unsupervised learning of Gaussian mixtures based on variational component splitting. *IEEE Transactions on Neural Networks*, 18(3):745–755, 2007. 2169

[5] A. Corduneanu and C. M. Bishop. Variational Bayesian model selection for mixture distributions. *Artificial Intelligence and Statistics*, pages 27–34, 2001. 2169, 2171, 2174

Figure 6. Segmentation results on images from the Berkeley database, using a superpixels initialization. All images were initialized at $K = 15$ classes. From left to right, each column shows the original image, the segmentation for $\epsilon = 3$ and the segmentation for $\epsilon = 10$.

[6] A. R. F. DaSilva. A Dirichlet process mixture model for brain MRI tissue classification. *Medical Image Analysis*, 11:169–182, 2007. 2170

[7] M. I. J. E. B. Sudderth. Shared segmentation of natural scenes using dependent Pitman-Yor processes. *Proceedings of Neural Information Processing Systems (NIPS)*, 21, 2008. 2170

[8] W. Feng, J. Jia, and Z.-Q. Liu. Self-validated labeling of markov random fields for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(10):1871–1887, 2010. 2170

[9] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):721–741, 1984. 2169, 2171

[10] K. Lange. *Optimization*. Springer, 2004. 2170, 2172, 2173

[11] V. Lempitsky, C. Rother, S. Roth, and A. Blake. Fusion moves for markov random field optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1392–1405, 2010. 2175

[12] J. Marroquin, E. Arce, and S. Botello. Hidden Markov measure field models for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11):1380–1387, 2003. 2169

[13] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the 8th International Conference on Computer Vision (ICCV '01)*, volume 2, pages 416–423, July 2001. 2174

[14] G. Mori. Guiding model search using segmentation. In *ICCV'05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, volume 2, pages 1417–1423, 2005. 2174

[15] P. Orbanz and J. M. Buhmann. Nonparametric Bayesian image segmentation. *International Journal of Computer Vision*, 77:25–45, 2007. 2170, 2173, 2175

[16] N. Pal and S. Pal. A review of image segmentation techniques. *Pattern Recognition*, 26:1277–1294, 1993. 2169

[17] D. Peel and G. J. McLachlan. Robust mixture modeling using the *t*-distribution. *Statistics and Computing*, 10:339–348, 2000. 2173

[18] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV'03: Proceedings of the IEEE International Conference on Computer Vision*, volume 1, pages 10–17, 2003. 2175

[19] G. Sfikas, C. Nikou, N. Galatsanos, and C. Heinrich. Spatially varying mixtures incorporating line processes forimage segmentation. *Journal of Mathematical Imaging and Vision*, 36:91–110, 2010. 2170, 2171, 2173, 2174

[20] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. 2170

[21] N. Ueda and Z. Ghahramani. Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, 15(10):1223–1241, 2002. 2169

[22] A. Yang, J. Wright, S. Sastry, and Y. Ma. Unsupervised segmentation of natural images via lossy data compression. *Computer Vision and Image Understanding*, 110(2):212–225, 2007. 2174