

Image Segmentation with Hierarchical Potts-Markov Priors

Satvik Saha*

Department of Statistics, Columbia University

Abstract

Image segmentation generally describes the process of separating an image into different groups or regions of pixels, with pixels in a group sharing similar characteristics. A crude method of image segmentation involves clustering pixels by their color, with no regard to their spatial arrangement. In this work, we introduce spatial structure to a mixture model for segmenting images, by imposing suitable priors on group labels. In particular, we focus on hierarchical Potts-Markov-like priors which encourage smoothing based on pixel neighborhoods and group-specific temperatures. We use MCMC-based inference procedures to apply this method on the Berkeley Segmentation Dataset, and find that it generally outperforms the corresponding spatially-naïve mixtures.

1 Introduction

Image segmentation, the process of dividing an image into different parts (segments), is a problem in computer vision that has applications in many areas, ranging from biomedical imaging (detecting diseased tissue) to autonomous vehicles (detecting objects on the road). Ideally, each segment represents a recognizable object in a scene, i.e. is a region of similar pixels. This lets us frame segmentation as a clustering problem, for which mixture models are well suited.

Interpreting an image as a bag of pixel colors $\mathbf{x} = \{x_\ell\}_{\ell \in \mathcal{L}}$ suggests a crude segmentation method: pick K colors, then declare a pixel to be part of the k -th object if its color is most similar to the k -th color. While such a procedure works reasonably well, it often leads to noisy segments with fuzzy boundaries; this kind of algorithm discards a lot of potentially useful information, in particular the spatial configuration of the pixels in the image. For instance, supposing that our objects of interest occupy continuous regions in space, we may believe that neighboring pixels will more often than not belong to the same segment. In this project, we build a simple, spatially-aware probabilistic model that exploits this heuristic and produces cleaner segmentations than a naïve mixture.

*ss7228@columbia.edu

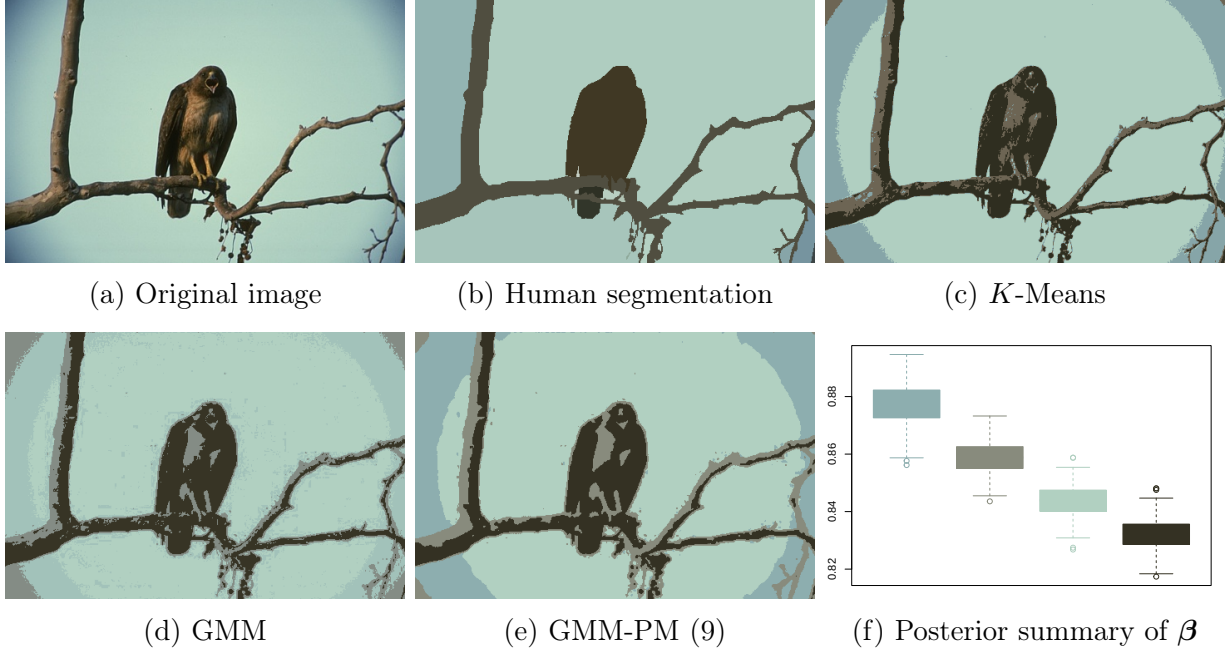


Figure 1: Segmentations of an image (1a, 42049 in the Berkeley Segmentation Dataset) by humans (1b), K -Means (1c), a simple Gaussian Mixture (GMM, 1d), and a Gaussian Mixture with a hierarchical Potts-Markov prior (GMM-PM, 1e). The last three models use $K = 4$ segments. The posterior distribution of β in GMM-PM (1f) gives us information about the relative smoothness of different segments.

1.1 Image Segmentation with Mixture Models

A general mixture model for $\mathbf{x} = \{x_\ell\}_{\ell \in \mathcal{L}}$ looks like

$$\begin{aligned}
 \beta &\sim p(\beta), \\
 \theta &\sim p(\theta), \\
 \mathbf{z} \mid \beta &\sim p(\mathbf{z} \mid \beta), \\
 x_\ell \mid \theta, z_\ell = k &\sim p(x_\ell \mid \theta, z_\ell = k) =: f(x_\ell; \theta_k), \quad \ell \in \mathcal{L}.
 \end{aligned} \tag{1}$$

Each data item x_ℓ is associated with a latent label $z_\ell \in \{1, \dots, K\}$ which determines which component it belongs to. The distribution $f(\cdot; \theta_k)$ of items within the k -th component is controlled by the parameter θ .

A simple mixture model may choose *a priori* independent labels $z_\ell \mid \beta$, i.e. allow

$$p(\mathbf{z} \mid \beta) = \prod_{\ell \in \mathcal{L}} p(z_\ell \mid \beta) \tag{2}$$

to factorize. This makes the pairs $\{(z_\ell, x_\ell)\}_{\ell \in \mathcal{L}}$ exchangeable: the location $\ell \in \mathcal{L}$ of a data point x_ℓ plays no role in the model. In the context of image segmentation, with $x_\ell \in \mathbb{R}^d$ denoting the color of a pixel, its location ℓ within the image is ignored.

Instead, we will choose our prior $p(\mathbf{z} \mid \beta)$ in a way that respects the spatial structure of $\mathcal{L}$. Specifically, we would like to use the following heuristic — *neighboring pixels often belong to*

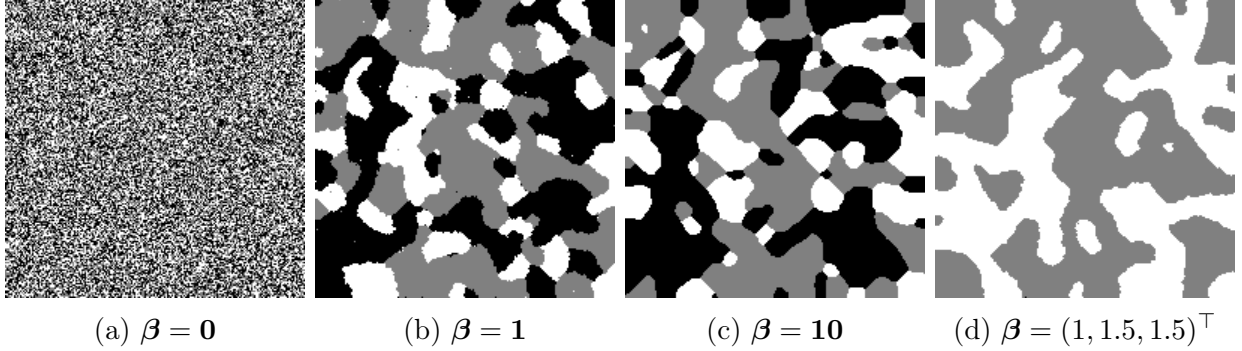


Figure 2: Draws from $\text{Potts}(\beta)$, where each pixel is connected to its (at most) 8 immediate neighbors. Larger values of β_k promote more cohesive components with smoother boundaries: components in 2b have noticeably more ‘islands’ of foreign labels compared to 2c. Note that imbalanced β , as in 2d, leads to some components effectively vanishing.

the same segment — motivated by the following observation. A segment within an image is rarely uniform in color; small fluctuations and scattered islands of pixels with very different colors may appear within the larger region. A simple mixture model using (2) would see these islands as part of a completely different segment (or as individual segments of their own). For example, looking closely at the segmentation produced by such a model in Figure 1d reveals a visually noisy blue sky. The kind of spatially aware model we explore here chooses to smooth out these irregularities and group them as part of the encompassing segment. This ought to result in a cleaner, less noisy segmented image; for instance, note the more uniform sky in Figure 1e. Additionally, such a model can also give us the opportunity to learn about the smoothness of each segment.

Note that this is very much a modeling choice/goal, which is not always desirable: in applications like medical imaging, such small irregularities are often the target and should *not* be glossed over! The aim here is to identify high level structure in an image where all objects are of similar importance, rather than picking out fine detail.

1.2 The Potts-Markov Prior

Let $\mathcal{G}(\mathcal{L}, \mathcal{E})$ be a graph on the set of locations $\mathcal{L}$, and notate $\ell \sim \ell'$ when $\ell, \ell' \in \mathcal{L}$ are neighbors in $\mathcal{G}$. We denote $\mathbf{z} \sim \text{Potts}(\beta)$ for $\beta \succeq \mathbf{0}$, when the distribution of $\mathbf{z}$ has density

$$p(\mathbf{z} \mid \beta) = \frac{1}{Z(\beta)} \exp \left(\sum_{\ell \sim \ell'} \sum_{k=1}^K \beta_k \mathbf{1}(z_{\ell'} = z_{\ell} = k) \right). \quad (3)$$

This produces a smoothing effect on the class labels $\mathbf{z}$ with respect to the spatial structure $\mathcal{G}$. The ‘reward’ for agreement between neighboring nodes assigned class k is controlled by a class-specific temperature β_k : this allows different components to have smoother boundaries than others. See Figure 2 for illustrative samples from $\text{Potts}(\beta)$.

Observe that (3) is an exponential family in β , with sufficient statistic $\mathbf{T}(\mathbf{z})$ described

by

$$T_k(\mathbf{z}) := \sum_{\ell \sim \ell'} \mathbf{1}(z_\ell = z_{\ell'} = k), \quad 1 \leq k \leq K. \quad (4)$$

Each T_k counts the number of edges in $\mathcal{G}$ connecting items assigned the same class k . With this, we have $p(\mathbf{z} \mid \boldsymbol{\beta}) = \exp(\boldsymbol{\beta}^\top \mathbf{T}(\mathbf{z})) / Z(\boldsymbol{\beta})$. Further keeping track of the statistics

$$U_{k\ell}(\mathbf{z}) := \sum_{\ell' \sim \ell} \mathbf{1}(z_{\ell'} = k), \quad 1 \leq k \leq K, \quad \ell \in \mathcal{L}, \quad (5)$$

which count the number of neighbors of z_ℓ assigned the class k , the full conditionals

$$p(z_\ell = k \mid \mathbf{z}_{-\ell}, \boldsymbol{\beta}) \propto \exp(\beta_k U_{k\ell}(\mathbf{z})) \quad (6)$$

suggest a simple Gibbs sampling algorithm for Potts($\boldsymbol{\beta}$).

The partition function $Z(\boldsymbol{\beta}) = \sum_{\mathbf{z} \in \{1, \dots, K\}^{|\mathcal{L}|}} \exp(\boldsymbol{\beta}^\top \mathbf{T}(\mathbf{z}))$ is an intractable normalizing constant. This makes *both* the joint density of the mixture model (1), as well as the full-conditional

$$p(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\theta}, \mathbf{x}) \propto \frac{p(\boldsymbol{\beta})}{Z(\boldsymbol{\beta})} \exp(\boldsymbol{\beta}^\top \mathbf{T}(\mathbf{z})) \quad (7)$$

intractable, which presents problems for standard posterior inference algorithms such as Gibbs sampling and mean field variational inference. For the former choice, sampling from (7) using ordinary Metropolis-Hastings requires computing $Z(\boldsymbol{\beta}^*) / Z(\boldsymbol{\beta})$ as part of the acceptance ratio. This issue of a *doubly intractable likelihood* can be sidestepped using an *auxiliary variable* technique.

1.3 The Exchange Algorithm

The following auxiliary variable Metropolis-Hastings scheme [Murray et al., 2006] is used to produce samples from (7) without directly dealing with the partition function $Z(\cdot)$.

1. Propose $\boldsymbol{\beta}^* \sim q(\boldsymbol{\beta}^*; \boldsymbol{\beta})$.
2. Draw auxiliary $\mathbf{z}^* \sim p(\mathbf{z}^* \mid \boldsymbol{\beta}^*)$.
3. Accept $\boldsymbol{\beta}^*$ with probability

$$\frac{q(\boldsymbol{\beta}; \boldsymbol{\beta}^*) p(\boldsymbol{\beta}^*)}{q(\boldsymbol{\beta}^*; \boldsymbol{\beta}) p(\boldsymbol{\beta})} \cdot \frac{p(\mathbf{z} \mid \boldsymbol{\beta}^*)}{p(\mathbf{z} \mid \boldsymbol{\beta})} \cdot \frac{p(\mathbf{z}^* \mid \boldsymbol{\beta})}{p(\mathbf{z}^* \mid \boldsymbol{\beta}^*)}, \quad (8)$$

and repeat from step 1.

Note that the problematic terms $Z(\boldsymbol{\beta})$, $Z(\boldsymbol{\beta}^*)$ neatly cancel out in the ratio (8). This is a simpler version of the algorithm in Møller et al. [2006].

Running this algorithm is fairly expensive in practice for image data, especially since step 2 involves sampling from Potts($\boldsymbol{\beta}^*$), with the benefit of targeting the correct stationary distribution. Mean field variational methods are much faster, at the cost of using inexact approximations for $Z(\cdot)$. See Section A, step (c) for details on implementation.

2 Gaussian Mixture Model with Potts-Markov Prior

Let $x_\ell \in \mathbb{R}^d$ be the color of the pixel at location $\ell \in \mathcal{L}$. We use a $K \in \mathbb{N}$ component mixture model of the form (1), with a Potts-Markov prior (3) on the latent class labels $\mathbf{z} = \{z_\ell\}_{\ell \in \mathcal{L}}$, multivariate Gaussians as component distributions with means and covariances $\{(\mu_k, \Sigma_k)\}_{k=1}^K \subseteq \mathbb{R}^d \times \mathbb{R}^{d \times d}$, and conjugate Normal-Inverse-Wishart priors on the latter component parameters. We further impose weak Gamma priors on the smoothness parameters β , which yields the following hierarchical model.

$$\begin{aligned} \beta_k &\stackrel{iid}{\sim} \text{Gamma}(a_\beta, b_\beta), & 1 \leq k \leq K, \\ \mathbf{z} \mid \beta &\sim \text{Potts}(\beta), \\ \Sigma_k &\stackrel{iid}{\sim} \text{Inverse-Wishart}(\Sigma_0, \nu_0), & 1 \leq k \leq K, \\ \mu_k \mid \Sigma_k &\sim \text{Normal}(m_0, \kappa_0^{-1} \Sigma_k), & 1 \leq k \leq K, \\ x_\ell \mid \mu, \Sigma, z_\ell = k &\sim \text{Normal}(\mu_k, \Sigma_k), & \ell \in \mathcal{L}. \end{aligned} \tag{9}$$

We will abbreviate (9) as a GMM-PM model. Its joint density factors as

$$\prod_{k=1}^K p(\beta_k) \cdot p(\mathbf{z} \mid \beta) \cdot \prod_{k=1}^K p(\Sigma_k) p(\mu_k \mid \Sigma_k) \cdot \prod_{\ell \in \mathcal{L}} p(x_\ell \mid \mu_k, \Sigma_k, z_\ell = k). \tag{10}$$

Here, $\mathcal{L} = \{(i, j) : 1 \leq i \leq m, 1 \leq j \leq n\}$ is a collection of tuples representing positions on an $m \times n$ image grid, and we choose the spatial structure $\mathcal{G}$ such that each pixel is connected to its (at most) 8 immediate neighbors on the grid. Formally, the neighborhood of ℓ in $\mathcal{G}$ looks like $\mathcal{N}_\ell = (\ell + \Delta) \cap \mathcal{L}$ for the set of translations $\Delta = \{-1, 0, 1\}^2 \setminus \{(0, 0)\}$. Furthermore, we have $d = 3$, with the components of x_ℓ denoting red, green, blue intensities in $[0, 1]$.

Modifying (9) by fixing $\beta = \mathbf{0}$ reduces it to a simple Gaussian mixture model with *a priori* independent labels $\mathbf{z}$. In contrast, the Potts-Markov prior on $\mathbf{z} \mid \beta$ induces the full conditionals

$$p(z_\ell = k \mid \mathbf{z}_{-\ell}, \beta_k, \mu_k, \Sigma_k, x_\ell) \propto \exp(\beta_k U_{k\ell}(\mathbf{z})) \cdot p(x_\ell \mid \mu_k, \Sigma_k, z_\ell = k), \tag{11}$$

where the weights $\exp(\beta_k U_{k\ell})$ encourage each z_ℓ to conform with its neighbors.

3 Data and Implementation

We select color images from the Berkeley Segmentation Dataset [Martin et al., 2001] and compare the segmentations produced by our GMM-PM model (9) with those from K -Means algorithms and simple Gaussian mixture models. All images contain $|\mathcal{L}| = 154,401$ pixels (size 321×481 or 481×321); the resulting spatial graph $\mathcal{G}$ has $|\mathcal{E}| = 615,200$ edges.

We use a Gibbs sampler, incorporating the exchange algorithm (1.3) to perform posterior inference on GMM-PM (9); given the sheer volume of data, most hyperparameters are chosen such that the priors are non-informative. See Section A for complete details on posterior

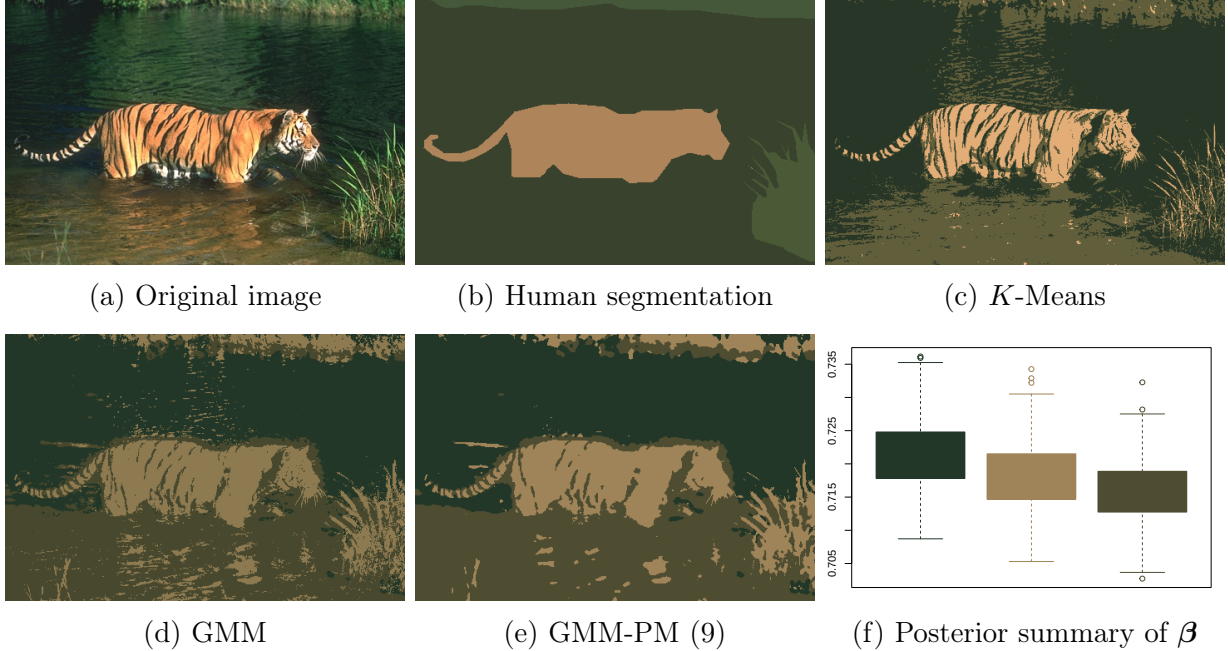


Figure 3: Segmentations of an image (3a, 108073 in the Berkeley Segmentation Dataset), with 3c to 3e using $K = 3$ segments. While K -Means and the GMM retain a lot of detail (the tiger’s stripes, ripples in the water), GMM-PM attempts to smooth them out in favor of high-level structure.

inference, Section B for hyperparameter selection. The Gibbs sampler for GMM-PM has been implemented in C++ with an R wrapper (using Rcpp). The K -means algorithm is run using a native R implementation.

4 Evaluation and Results

The dataset contains segmentations performed by humans, which we use as a reference to compute Adjusted Rand Index (ARI) scores [Rand, 1971] for the models of interest; see Table 1. These scores are a measure of similarity between two segmentations (not necessarily having the same number of clusters): 1 indicates a perfect match, 0 indicates a completely random segmentation. Figures 1, 3 and 4 illustrate segmentations of particular images.

Table 1 indicates that although GMM-PM with fixed β can outperform GMM-PM in full generality, the discrepancy is rarely large, and the former loses out on information from the posterior distribution of the smoothness parameter β . In general, human segmentations identify many more segments than our mixture models, as they can tell apart different objects with the same color, which may explain why the performances aren’t particularly high. To illustrate this point, note that in Figure 1, our models cannot separate the bird and the branch which have very similar colors.



(a) Original image



(b) Human segmentation



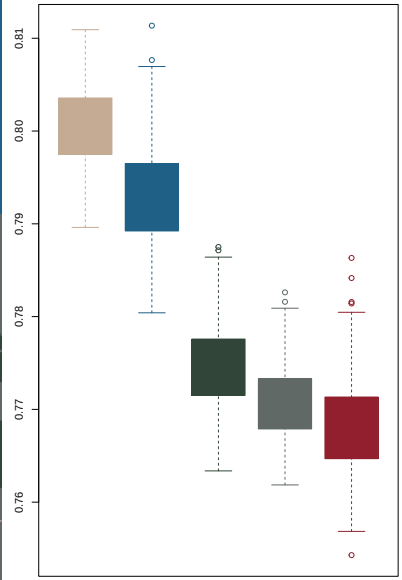
(c) K -Means



(d) GMM



(e) GMM-PM (9)



(f) Posterior summary of β

Figure 4: Segmentations of an image (4a, 86000 in the Berkeley Segmentation Dataset), with 4c to 4e using $K = 5$ segments. While K -Means and GMM retain more detail, the GMM-PM smooths them out and focuses on large-scale features. The posterior distribution of β (4f) seems to indicate that the building and sky segments are smoother than the flower segments.

Table 1: Adjusted Rand Index (ARI) scores of segmentations against human references; higher is better. The last four columns report scores for the following models: K -Means, simple Gaussian Mixture (GMM), model (9) with fixed $\beta = 1$ (GMM-PM- $(\beta = 1)$), and model (9) (GMM-PM).

Image ID	K	K -Means	GMM	GMM-PM- $(\beta = 1)$	GMM-PM
108073	3	0.2395	0.2241	0.2767	0.2678
198054	7	0.3603	0.4081	0.6521	0.4444
253027	3	0.1707	0.3470	0.4481	0.4408
253036	5	0.4154	0.6150	0.4444	0.4459
295087	8	0.2591	0.2693	0.2967	0.2982
300091	5	0.1563	0.1404	0.1892	0.2648
37073	5	0.7151	0.4705	0.5428	0.5757
38092	7	0.4471	0.5342	0.5453	0.5381
42049	4	0.3009	0.3187	0.3412	0.3403
86000	5	0.4587	0.4913	0.6341	0.6270

5 Discussion

We have presented a spatially aware Gaussian mixture model (GMM-MP), and demonstrated that it frequently outperforms naïve Gaussian mixtures in the context of image segmentation, especially when the focus is not on fine detail but on broad strokes. While the use of the Potts-Markov prior incorporates spatial information for the purpose of smoothing (by rewarding cohesion), one can imagine priors which encourage other kinds of behavior. A generalization of (3) would be priors of the form

$$p(\mathbf{z} \mid \beta) \propto \exp \left(\sum_{\ell \in \mathcal{L}} \alpha(z_{\ell}; \ell) + \sum_{\ell, \ell' \in \mathcal{L}} \beta(z_{\ell}, z_{\ell'}; \ell, \ell') \right) \quad (12)$$

for potentials α, β that could model a variety of interactions between labels. Priors like these could be plugged directly into this project’s setup with little change to the core sampling algorithm.

Segmentation results such as those in Figures 1, 3 and 4 show that a Gaussian mixture is not always good at modeling colors that appear at the boundaries of the RGB color space $[0, 1]^3$, including bright highlights and very dark areas. This suggests that mixtures of Gaussians (which are supported on all of $\mathbb{R}^3$) may not be the best choice for modeling color distributions. We may swap out $\{f(\cdot; \theta)\}_{\theta \in \Theta}$ in (1) for any convenient family of likelihoods with minimal change to the spatial prior. Another solution may be to transform our color features $x_{\ell} \in [0, 1]^3$ to range over $\mathbb{R}^3$, say using logit links, before plugging them into our model.

References

Peter D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics. Springer New York, New York, NY, 2009.

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 *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.

Iain Murray, Zoubin Ghahramani, and David J. C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 359–366. AUAI Press, 2006.

J. Møller, A. N. Pettitt, R. Reeves, and K. K. Berthelsen. An Efficient Markov Chain Monte Carlo Method for Distributions with Intractable Normalising Constants. *Biometrika*, 93(2):451–458, 2006.

William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

A Posterior Inference

We produce posterior samples of $\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{x}$ using a Gibbs sampling scheme. This involves updating a state $(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ by iteratively cycling through the following full-conditionals.

- (a) Sample new component means and covariances $\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid -$ as follows: for each $1 \leq k \leq K$, set

$$\begin{aligned} n_k &:= \sum_{\ell \in \mathcal{L}} \mathbf{1}(z_\ell = k), & \bar{x}_k &:= \frac{1}{n_k} \sum_{\ell \in \mathcal{L}} x_\ell \mathbf{1}(z_\ell = k), \\ \mu'_k &:= \frac{\kappa_0 m_0 + n_k \bar{x}_k}{\kappa_0 + n_k}, & S_k &:= \sum_{\ell \in \mathcal{L}} (x_\ell - \bar{x}_k)(x_\ell - \bar{x}_k)^\top \mathbf{1}(z_\ell = k), \\ \Sigma'_k &:= \Sigma_0 + S_k + \frac{\kappa_0 n_k}{\kappa_0 + n_k} (\bar{x}_k - m_0)(\bar{x}_k - m_0)^\top. \end{aligned}$$

Then, replace the current $\boldsymbol{\mu}, \boldsymbol{\Sigma}$, by sampling from

$$\begin{aligned} \Sigma_k \mid - &\sim \text{Inverse-Wishart}(\Sigma'_k, \nu_0 + n_k), \\ \mu_k \mid \Sigma_k, - &\sim \text{Normal}(\mu'_k, (\kappa_0 + n_k)^{-1} \Sigma_k). \end{aligned}$$

These are standard Gaussian-Inverse-Wishart conjugate prior updates (see Hoff [2009]).

- (b) Sample new labels $\mathbf{z} \mid -$ as follows: for each $\ell \in \mathcal{L}$ (we choose a randomized order), sample z_ℓ from a categorical distribution with unnormalized probabilities

$$p(z_\ell = k \mid -) \propto \exp(\beta_k U_{k\ell}) \cdot p(x_\ell \mid \mu_k, \Sigma_k, z_\ell = k)$$

for $1 \leq k \leq K$. The second term is simply a (multivariate) Gaussian likelihood

$$p(x_\ell \mid \mu_k, \Sigma_k, z_\ell = k) = (2\pi)^{-d/2} \det(\Sigma_k)^{-1/2} \exp\left(-\frac{1}{2}(x_\ell - \mu_k)^\top \Sigma_k^{-1}(x_\ell - \mu_k)\right).$$

These can be pre-computed and stored before looping over $\ell \in \mathcal{L}$. The statistic $\mathbf{U}$ can be locally updated each time a new z_ℓ is sampled.

(c) Sample new Potts-Markov smoothness terms β | – using an auxiliary variable Metropolis-Hastings algorithm as follows: for $1 \leq t \leq T_\beta$, iterate the following.

(i) Propose $\beta^* \sim q(\cdot; \beta)$. Here, we use the proposal

$$\log \beta^* = \log \beta + \epsilon, \quad \epsilon \sim \text{Normal}(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_K),$$

to automatically keep $\beta \succeq 0$. This makes the proposal ratio

$$\frac{q(\beta; \beta^*)}{q(\beta^*; \beta)} = \prod_{k=1}^K \exp(\epsilon_k).$$

(ii) Draw an auxiliary set of labels $\mathbf{z}^* \sim p(\mathbf{z}^* | \beta^*)$ from the Potts-Markov prior. This is accomplished by performing a few Gibbs sampling sweeps (with a hot-start initialization at the current $\mathbf{z}$), repeatedly looping through $\ell \in \mathcal{L}$ and resampling z_ℓ^* with unnormalized probabilities

$$p(z_\ell^* | -) \propto \exp(\beta_k^* U_{k\ell}(\mathbf{z}^*)) =: \exp\left(\beta_k^* \sum_{\ell' \in \mathcal{N}_\ell} \mathbf{1}(z_{\ell'}^* = k)\right).$$

(iii) Update β to β^* with probability

$$\frac{q(\beta; \beta^*) p(\beta^*)}{q(\beta^*; \beta) p(\beta)} \cdot \frac{p(\mathbf{z} | \beta^*)}{p(\mathbf{z} | \beta)} \cdot \frac{p(\mathbf{z}^* | \beta)}{p(\mathbf{z}^* | \beta^*)},$$

keep the old value of β otherwise. Note that $p(\beta), p(\beta^*)$ are $\text{Gamma}(a_\beta, b_\beta)$ likelihoods, and the ratio

$$\frac{p(\mathbf{z} | \beta^*)}{p(\mathbf{z} | \beta)} \cdot \frac{p(\mathbf{z}^* | \beta)}{p(\mathbf{z}^* | \beta^*)} = \exp((\beta^* - \beta)^\top (\mathbf{T}(\mathbf{z}) - \mathbf{T}(\mathbf{z}^*)))$$

is free of the intractable partition function values $Z(\beta), Z(\beta^*)!$

This describes a Gibbs within Metropolis-Hastings within Gibbs MCMC sampling scheme. In practice, we choose 5 inner Gibbs loops (step (c)(ii)), $T_\beta = 10$ Metropolis-Hastings loops (step (c)), and 750 outer Gibbs loops with the first burn-in 250 samples discarded. Convergence is assessed by examining trace plots of parameters.

One practical consideration is tuning the proposal step size σ_β . This is adjusted dynamically during the burn-in period so that the acceptance rate for β^* (step (c)) settles around 30%.

We use the same sampling scheme with $\beta = \mathbf{0}$ and skipping step (c) to perform posterior inference on a Gaussian mixture model.

B Hyperparameter Selection

Since the joint density (10) is intractable and there is no convenient way of holding out image data once a spatial structure $\mathcal{G}$ is imposed, we estimate the required number of components K by maximizing an expected log predictive density score from a simpler Gaussian mixture (fixing $\beta = \mathbf{0}$ in 9), with the idea that this reduction already models the color distribution of an image well and is thus a good proxy. The remaining hyperparameters in (9) are chosen to be weakly informative, with $a_\beta = b_\beta = 2$, $m_0 = \frac{1}{2}\mathbf{1}$, $\Sigma_0 = \frac{1}{K}\mathbf{I}_d$, $\nu_0 = 10^{-2}$, $\kappa_0 = 5$; given the huge volume of data, we do not spend much effort tuning these. We also use a hot-start initialization of $(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ using parameters obtained from a run of K -Means, while β is initialized to $\mathbf{1}$.