Supplemental Material
From Patches to Images: A Nonparametric Generative Model

Geng Ji, Michael C. Hughes, Erik B. Sudderth

Abstract

We provide two key pieces of supplementary material: (1) high-resolution result images shown in the main paper, and (2) mathematical and algorithmic details of our variational algorithms. Image files are inside the images/ folder included in supplement.zip. We encourage the readers to check those results back and forth on an image viewer to see the difference between methods. The remainder of this document provides further details for our variational posterior inference method. Our open-source Python code is available online at github.com/bnpy/hdp-grid-image-restoration.

Contents

A DP Grid: Variational Inference Details

As in the main text, our goal is to best explain many observed noisy images y_m with the DP Grid model. Specifically, we wish to estimate the posterior distribution $p(\beta, \Lambda, x_m, w_m, \Psi_m^{\text{patch}} | y_m)$. In each subsection below, we look at a subset of random variables and derive: (1) the chosen approximate posterior family, (2) useful expectations for computing terms of the variational objective $\mathcal{L}$, and (3) the coordinate ascent update equations that will improve $\mathcal{L}$.

A.1 Approximate Posterior for Global Random Variables

The DP mixture model has two global random variables which are shared across all images: the per-cluster stick-breaking frequencies β_k and the per-cluster precision matrix Λ_k. Our chosen approximate posterior factors for these quantities have standard exponential-family forms, where the associated free parameters are marked with hats:

$$q(\Lambda_k) = \text{Wish}\big(\hat{\nu}_k, \hat{W}_k\big)$$
$$q(\beta_k) = \text{Beta}\big(\hat{\rho}_k \hat{\omega}_k, (1-\hat{\rho}_k)\hat{\omega}_k\big)$$

The Wishart approximate posterior $q(\Lambda_k)$ has a positive scalar $\hat{\nu}_k \in \mathbb{R}^+$ and a $G \times G$ positive definite matrix $\hat{W}_k$.

The Beta posterior $q(\beta_k)$ has a positive scalar parameter $\rho_k \in [0, 1]$ which defines the mean of β_k, and another positive scalar $\omega_k \in \mathbb{R}^+$ which controls the variance.

A.1.1 Useful Expectations

Expectations for cluster-specific precision matrices. Under the chosen $q(\Lambda_k)$, we have the expectations:

$$\mathbb{E}_q[\Lambda_k] = \hat{\nu}_k \hat{W}_k^{-1} \tag{1}$$
$$\mathbb{E}_q[\log |\Lambda_k|] = \sum_{g=1}^{G} \psi\Big(\frac{\hat{\nu}_k + 1 - g}{2}\Big) + G \log 2 - \log |\hat{W}_k|$$

in which ψ stands for the derivative of the logarithm of gamma function, often called the *digamma* function.

Expectations for cluster frequencies. Under our chosen family for $q(\beta)$ we have closed-form expressions for key expectations of the cluster frequencies π_{0k} for active clusters $k \leq K$:

$$\mathbb{E}_q[\pi_{0k}] = \hat{\rho}_k \prod_{\ell=1}^{k-1} (1 - \hat{\rho}_\ell) \tag{2}$$
$$\mathbb{E}_q[\log \pi_{0k}] = \sum_{\ell=1}^{k-1} \mathbb{E}[\log(1 - \beta_\ell)] + \mathbb{E}[\log \beta_k]$$

The remaining mass above some cluster index K is also known:

$$\mathbb{E}_q[\pi_{0>K}] = \prod_{\ell=1}^{K}(1-\hat{\rho}_\ell) \tag{3}$$

Closed-form expectations of direct functions of β_k:

$$\begin{aligned}\mathbb{E}_q[\log \beta_k] &= \psi(\hat{\rho}_k\hat{\omega}_k) - \psi(\hat{\omega}_k) \\ \mathbb{E}_q[\log(1-\beta_k)] &= \psi((1-\hat{\rho}_k)\hat{\omega}_k) - \psi(\hat{\omega}_k)\end{aligned} \tag{4}$$

A.1.2 Coordinate Ascent Updates

Following Hughes and Sudderth (2013), we only explicitly compute posterior statistics for the K "active" clusters that have been assigned to at least one patch. All clusters with index $> K$ are by definition independent of the data. Thus, their posterior factors are simply equal to their priors (Hughes and Sudderth, 2013) and need not be instantiated.

Update for $q(\Lambda_k)$. The Wishart posterior for the corpus-wide cluster precision matrix Λ_k enjoys standard exponential family additive updates where the relevant sufficient statistics are N_k, an aggregated usage count, and S_k, an aggregated outer-product. These statistics are averaged across all G grid alignments:

$$\begin{aligned}\hat{\nu}_k &= \nu + \frac{1}{G}N_k, \quad N_k = \sum_{m=1}^{M}\sum_{g=1}^{G}\sum_{n=1}^{N_{mg}} \hat{r}_{mgnk} \\ \hat{W}_k &= W + \frac{1}{G}S_k, \quad S_k = \sum_{m=1}^{M}\sum_{g=1}^{G}\sum_{n=1}^{N_{mg}} \mathbb{E}_q\left[\mathbb{1}_k(z_{mgn}) v_{mgn} v_{mgn}^T\right]\end{aligned} \tag{5}$$

Update for $q(\beta_k)$. For the DP-mixture, the optimal update to each cluster's stick-breaking weight $q(\beta_k)$ also has a standard closed form, as described in Hughes and Sudderth (2013).

$$\hat{\rho}_k\hat{\omega}_k \leftarrow N_k + 1, \quad (1-\hat{\rho}_k)\hat{\omega}_k \leftarrow N_{>k} + \gamma \tag{6}$$

Here, the count $N_{>k}$ represents the aggregated statistic for all clusters with index larger than k: $N_{>k} = \sum_{\ell=k+1}^{K} N_k$

A.2 Approximate Posterior for Patch Random Variables

Recall that from the main paper, we have the follow approximate posterior family for the patch-specific random variables: u, v, z.

$$\begin{aligned}q(z_{mgn}|w_m = g) &= \mathrm{Cat}\big(\hat{r}_{mgn1}, ..., \hat{r}_{mgnK}\big) \\ q(u_{mgn}|w_m = g) &= \mathrm{Norm}\big(\hat{u}_{mgn}, \hat{\phi}^u_{mgn}\big) \\ q(v_{mgn}|w_m = g, z_{mgn} = k) &= \mathrm{Norm}\big(\hat{v}_{mgnk}, \hat{\phi}^v_{mgnk}\big)\end{aligned}$$

We interpret the responsibility parameter $\hat{r}_{mgnk}$ as the posterior probability of assigning the n-th patch in grid g to the k-th cluster. The vector $\hat{r}_{mgn}$ must have K positive entries that sum to one. The posterior for scalar DC offset u has a simple Gaussian distribution with free mean and variance parameters. Similarly, the posterior for vector v has a Gaussian form with mean and covariance matrix.

Note that each of these factors conditions on the value of the grid indicator w_m for the current image m. This conditioning provides flexible posterior structures and elegant update equations not possible with naive mean-field methods.

A.2.1 Useful Expectations

Under our structured approximate posterior, we have the following expectations:

$$\mathbb{E}_q[\mathbb{1}_k(z_{mgn})v_{mgn}] = \hat{r}_{mgnk}\hat{v}_{mgnk} \tag{7}$$

$$\mathbb{E}_q[v_{mgn}] = \sum_{k=1}^{K} \hat{r}_{mgnk}\hat{v}_{mgnk} \tag{8}$$

Similarly, we have the following outer-product expectations:

$$\mathbb{E}_q[\mathbb{1}_k(z_{gmn})v_{mgn}v_{mgn}^T] = \hat{r}_{mgnk}(\hat{v}_{mgnk}\hat{v}_{mgnk}^T + \hat{\phi}^v_{mgnk}) \tag{9}$$

$$\mathbb{E}_q[v_{mgn}v_{mgn}^T] = \sum_{k=1}^{K} \hat{r}_{mgnk}(\hat{v}_{mgnk}\hat{v}_{mgnk}^T + \hat{\phi}^v_{mgnk}) \tag{10}$$

A.2.2 Coordinate Ascent Updates

Updating $q(z|w)$. Within image m, we update the $n-$th patch inside the $g-$th grid by computing a scalar positive weight for each active cluster $k = 1, 2, \ldots K$:

$$\hat{r}_{mgnk} \propto \exp\Big(\mathbb{E}_q[\log \pi_{0k}] + \frac{1}{2}\big(\mathbb{E}_q[\log|\Lambda_k|] + \log|\hat{\phi}^v_{mgnk}| + F_{mgn}^T\hat{\phi}^v_{mgnk}F_{mgn}\big)\Big) \tag{11}$$

in which $F_{mgn} \triangleq \frac{1}{\delta^2}C_{mgn}^T(P_{mgn}\hat{x}_m - \hat{u}_{mgn})$. The entire vector $\hat{r}_{mgn}$ is then normalized to sum to one. Each entry k defines the posterior probability (or *responsibility*) that cluster k explains this patch. The required expectations have known closed-form due to our exponential family assumptions. We provide closed-form expressions for $E_q[\pi_{0k}]$ and $E_q[\log|\Lambda_k|]$ in Eq. (2) and Eq. (1).

Updating $q(v|w,z)$. We update the approximate posterior over the vector $v_{mgn} \in \mathbb{R}^G$ by computing its mean and covariance via closed-form updates:

$$\hat{v}_{mgnk} = \frac{1}{\delta^2}\hat{\phi}^v_{mgnk}C_{mgn}^T(P_{mgn}\hat{x}_m - \hat{u}_{mgn}), \quad \hat{\phi}^v_{mgnk} = \Big(\frac{1}{\delta^2}C_{mgn}^TC_{mgn} + \mathbb{E}_q[\Lambda_k]\Big)^{-1} \tag{12}$$

A closed-form expression for $\mathbb{E}_q[\Lambda_k]$ is given in Eq. (1). For most patches that are fully-observed, matrix C_{mgn} would just reduce to an identity matrix and the updates simplify accordingly.

Updating $q(u|w)$. Similarly, the update for the mean and variance of the scalar offset u_{mgn} is:

$$\hat{u}_{mgn} = \hat{\phi}^u_{mgn}\Big(\frac{r}{s^2} + \frac{1}{\delta^2}\mathbf{1}^T\big(P_{mgn}\hat{x}_m - C_{mgn}\mathbb{E}_q[v_{mgn}]\big)\Big), \quad \hat{\phi}^u_{mgn} = 1\Big/\Big(\frac{1}{s^2} + \frac{D_{mgn}}{\delta^2}\Big) \tag{13}$$

The required expectation $\mathbb{E}[v_{mgn}]$ is defined in Eq. (8). $D_{mgn} \in (0, G]$ is the number of observable pixels of patch n in the g-th grid of image m.

A.3 Approximate Posterior for Image Random Variable

As in the main paper, the posterior $q(w_m)$ for alignment indicator w_m is assumed uniform. Thus we only need to focus on the approximate posterior $q(x_m)$ for the clean image x_m:

$$q(x_m) = \mathrm{Norm}(x_m|\hat{x}_m, \hat{\phi}^x_m) \tag{14}$$

A.3.1 Coordinate Ascent Updates

The mean and covariance of posterior for the whole-image vector x_m both have closed-form updates:

$$\hat{x}_m = \hat{\phi}_m\Big(\frac{y_m}{\sigma^2} + \frac{h_m}{\delta^2}\Big), \quad \hat{\phi}^x_m = \frac{\delta^2\sigma^2}{\delta^2 + \sigma^2}I \tag{15}$$

The covariance update conveniently yields a diagonal matrix. The mean depends on the average image vector across all patches in all grids, denoted h_m:

$$h_m \triangleq \frac{1}{G}\sum_{g=1}^{G}\sum_{n=1}^{N_{mg}} P^T_{mgn}(C_{mgn}\mathbb{E}_q[v_{mgn}] + \hat{u}_{mgn}). \tag{16}$$

Recall that the expectation $\mathbb{E}_q[v_{mgn}]$ is defined in Eq. (8)

B HDP Grid: Variational Inference Details

While the DP Grid model above assumes the same cluster probability vector π_0 for each image m, our HDP Grid model allows image-specific cluster probabilities π_m to be learned from data. These are tied together via the hierarchical Dirichlet process prior.

B.1 Approximate Posterior for HDP Random Variables

Our revised approximate posterior family $\mathcal{Q}$ now includes the HDP factors:

$$q(\beta) = \prod_{k=1}^{\infty} \mathrm{Beta}\left(\beta_k|\hat{\rho}_k\hat{\omega}_k, (1-\hat{\rho}_k)\hat{\omega}_k\right), \tag{17}$$

$$q([\pi_{m1}\ldots\ \pi_{mK}\ \pi_{m>K}]) = \mathrm{Dir}(\hat{\theta}_{m1}, \ldots, \hat{\theta}_{mK}, \hat{\theta}_{m>K})$$

Here, the image-specific free parameter $\hat{\theta}_m$ is a vector of length $K+1$, where the last dimension represents all inactive clusters. Its optimal update is:

$$\hat{\theta}_{mk} = \begin{cases} \alpha \mathbb{E}_q[\pi_{0k}] + \frac{1}{G} \sum_{g=1}^{G} \sum_{n=1}^{N_{mg}} \hat{r}_{mgnk}, & k \leq K \\ \\ \alpha \mathbb{E}_q[\pi_{0>K}], & k = K+1 \end{cases} \tag{18}$$

in which $\mathbb{E}_q[\pi_{0k}]$ follows from Eq. (2) and $\mathbb{E}_q[\pi_{0>K}]$ from Eq. (3). The update for $\hat{\rho}_k$ and $\hat{\omega}_k$ has no closed form but can be executed easily via gradient descent. Details can be found in Appendix D of the supplement of Hughes et al. (2015), which is available online.[1]

Other factors remain unchanged from the DP Grid model. Their respective updates remain unchanged as well, except that we substitute $\mathbb{E}_q[\log \pi_{mk}]$ for $\mathbb{E}_q[\log \pi_{0k}]$ in the patch-cluster responsibility update in Eq. (11):

$$\hat{r}_{mgnk} \propto \exp\left(\mathbb{E}_q[\log \pi_{mk}] + \frac{1}{2}\big(\mathbb{E}_q[\log |\Lambda_k|] + \log |\hat{\phi}^v_{mgnk}| + F_{mgn}^T \hat{\phi}^v_{mgnk} F_{mgn}\big)\right) \tag{19}$$

Sparse responsibilities. In practice, we optimize downstream computations by enforcing $\hat{r}_{mgn}$ to be a one-hot vector rather than a dense vector of K entries. To do this, after computing the dense $\hat{r}_{mgn}$ vector as before, we place probability mass one on its maximum entry k'. The advantage of restricting to sparse $\hat{r}$ vectors is that we need only compute and store $\hat{v}_{mgnk'}$ rather than all $k \in \{1, ..., K\}$. Using sparse posteriors significantly reduces memory and computational costs but does not noticeably impact inference quality.

B.2 HDP Denoising Algorithm

In Alg. 1, we describe the procedure used to perform our HDP denoising algorithm, which combines K' novel clusters from the noisy test image with the original K clusters learned from a training dataset of clean images. The annealing schedule used to decay δ over 8 iterations from the initial value of the noise-level σ to a final value of $0.5/255$ is equivalent to the schedule used in the public EPLL code.

[1] www.michaelchughes.com/papers/HughesKimSudderth_AISTATS_2015_supplement.pdf

Algorithm 1 HDP denoising algorithm for single image given pretrained external model.

Input:

y_m : noisy image
σ : standard deviation of noise
K' : number of internal clusters to learn from provided image

Output:

$\hat{x}_m$: restored image

function DENOISEIMAGE(y_m)
  Extend $q(\beta)$ and $q(\Lambda)$ to contain $K + K'$ clusters
  Initialize $\mathbb{E}[q(\pi_m)]$ as uniform, $\mathbb{E}[q(x_m)]$ the noisy image, and $\mathbb{E}[q(u_{mgn})]$ the patch mean
  for iteration $t := 1 \to 8$ **do**
    if $t = 1$ **then**
      $\delta := \sigma$
    else
      $\delta := \max\left\{\frac{\sigma}{2^{t/2}}, \frac{0.5}{255}\right\}$
    end if
    for grid $g := 1 \to G$ **do**
      for patch $n := 1 \to N_{mg}$ **do**
        Update $q(z_{mgn})$ using Eq. (19)
        Update $q(v_{mgn})$ using Eq. (12)
        Update $q(u_{mgn})$ using Eq. (13)
      end for
    end for
    Update $q(x_m)$ using Eq. (15)
    Update $q(\pi_m)$ using Eq. (18)
    Delete unused image-specific clusters
  end for
  return $\hat{x}_m$
end function

References

Michael C Hughes and Erik B Sudderth. Memoized online variational inference for Dirichlet process mixture models. In *Neural Information Processing Systems*, 2013.

Michael C Hughes, Dae Il Kim, and Erik B Sudderth. Reliable and scalable variational inference for the hierarchical Dirichlet process. In *Artificial Intelligence and Statistics*, 2015.