December 1: Spatially Dependent Pitman-Yor Processes via Gaussian Processes
Are Images Bags of Features?

Inspired by the successes of *topic models* for text data, some have proposed learning from *local image features*.
Are Images Bags of Features?

Inspired by the successes of *topic models* for text data, some have proposed learning from *local image features*

**First Approach:** Fei-Fei & Perona 2005, Sivic et. al. 2005
- Ignore spatial structure entirely (bag of “*visual words*”)

**Second Approach:** Russell et. al. 2006, Todorovic & Ahuja 2007
- Cluster features via one or more *bottom-up segmentations*
Segmentation: Mean Shift

**EDISON:** Comaniciu & Meer, 2002

- Cluster by modes of *Parzen window* density estimator in space of appearance features
- Very *sensitive* to bandwidth parameter
Outline

Natural Scene Statistics
- Counts, partitions, and power laws
- Hierarchical *Pitman-Yor* processes

Spatial Priors for Image Partitions
- What’s wrong with Potts models?
- Spatial dependence via *Gaussian processes*

Unsupervised Image Analysis
- Variational inference
- Image *segmentation*
Priors on Counts & Partitions

Segmentation as Partitioning

- How many regions does this image contain?
- What are the sizes of these regions?

Unsupervised Object Category Discovery

- How many object categories have I observed?
- How frequently does each category appear?
The *Pitman-Yor process* defines a distribution on infinite discrete measures, or *partitions*

\[
\pi_1 = v_1 \\
\pi_2 = v_2(1 - v_1) \\
\pi_3 = v_3(1 - v_2)(1 - v_1) \\
\vdots \\
\pi_k = v_k \left(1 - \sum_{\ell=1}^{k-1} \pi_\ell\right) = v_k \prod_{\ell=1}^{k-1} (1 - v_\ell) \\
v_k \sim \text{Beta}(1 - a, b + ka) \\
\textbf{Dirichlet process:} \quad a = 0
\]
Pitman-Yor Stick-Breaking

\[ v_k \sim \text{Beta}(1 - a, b + ka) \]

\[ E[v_k] = \frac{1 - a}{1 - a + b + ka} \]

\[ a = 0.1, \quad b = 3 \]

\[ a = 0.5, \quad b = 7 \]

\[ k = 1 \quad \text{blue} \quad k = 10 \quad \text{red} \quad k = 20 \quad \text{green} \]
Why Pitman-Yor?

Generalizing the Dirichlet Process

- Distribution on partitions leads to a generalized *Chinese restaurant process*.
- Special cases arise as excursion lengths for Markov chains, Brownian motions, …

Power Law Distributions

<table>
<thead>
<tr>
<th></th>
<th>DP</th>
<th>PY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of unique clusters in N observations</td>
<td>$\mathcal{O}(b \log N)$</td>
<td>$\mathcal{O}(bN^a)$</td>
</tr>
<tr>
<td>Size of sorted cluster weight $k$</td>
<td>$\mathcal{O}\left(\frac{\alpha b}{b} \left(1 + \frac{b}{b}\right)^{-k}\right)$</td>
<td>$\mathcal{O}\left(\alpha ab k^{-\frac{1}{a}}\right)$</td>
</tr>
</tbody>
</table>

*Heaps’ Law:* Natural Language Statistics  
Goldwater, Griffiths, & Johnson, 2005,  
Teh, 2006

*Zipf’s Law:*
Natural Scene Statistics

- Does Pitman-Yor prior match human segmentation?
- How do statistics vary across scene categories?

Oliva & Torralba, 2001
Manual Image Segmentation

Labels for more than 29,000 segments in 2,688 images of natural scenes
Object Size Histograms

**Forest Scenes**

- Small Objects
  - Segment Areas
  - PY(0.02, 2.20)
  - DP(2.40)

- Large Objects

**Inside City Scenes**

- Segment Areas
  - PY(0.32, 0.80)
  - DP(2.90)
Object Counts per Image

- **Forest Scenes**
  - Segment Counts
  - PY(0.02, 2.20)
  - DP(2.40)

- **Inside City Scenes**
  - Segment Counts
  - PY(0.32, 0.80)
  - DP(2.90)
Object Name Frequencies

- **trees**
- **sky**
- **waterfall**
- **person**
- **rainbow**
- **wheelbarrow**
- **lichen**

**forest scenes**

- **trees**
- **sky**
- **waterfall**
- **person**
- **rainbow**
- **wheelbarrow**
- **lichen**

**inside city scenes**
Feature Extraction

• Partition image into ~1,000 superpixels
• Compute texture and color features:
  SIFT Descriptor (Lowe 2004)
  Robust Hue Descriptor (van de Weijer & Schmid, 2006)
• VQ histograms to discrete visual words
PY Mixture Segmentation

LabelMe Segments:
PY Mixture Segmentation

LabelMe Segments:
Outline

**Natural Scene Statistics**
- Counts, partitions, and power laws
- Hierarchical *Pitman-Yor* processes

**Spatial Priors for Image Partitions**
- What’s wrong with Potts models?
- Spatial dependence via *Gaussian processes*

**Unsupervised Image Analysis**
- Variational inference
- Image *segmentation*
Discrete Markov Random Fields

Ising and Potts Models

\[ p(z) = \frac{1}{Z(\beta)} \prod_{(s,t) \in E} \psi_{st}(z_s, z_t) \]

\[ \log \psi_{st}(z_s, z_t) = \begin{cases} 
\beta_{st} > 0 & z_s = z_t \\
0 & \text{otherwise}
\end{cases} \]

Previous Applications

- Interactive foreground segmentation
- Supervised training for known categories

…but very little success at segmentation of unconstrained natural scenes.
Region Classification with Markov Field Aspect Models

Verbeek & Triggs, CVPR 2007
10-State Potts Samples

States sorted by size: largest in blue, smallest in red
The Ising/Potts model is not well suited to segmentation tasks

R.D. Morris  X. Descombes
J. Zerubia
INRIA, 2004, route des Lucioles, BP93, Sophia Antipolis Cedex, France.

Even within the phase transition region, samples lack the size distribution and spatial coherence of real image segments.
Geman & Geman, 1984

128 x128 grid
8 nearest neighbor edges
K = 5 states
Potts potentials: $\beta = 2/3$

200 Iterations

10,000 Iterations
Product of Potts and DP?

Orbanz & Buhmann 2006

\[ p(z) = \frac{1}{Z(\beta, \pi)} \prod_{(s,t) \in E} \psi_{st}(z_s, z_t) \prod_{s \in V} \pi(z_s) \]

Potts Potentials

DP Bias: \( \pi \sim \text{Stick}(\alpha) \)
Spatially Dependent Pitman-Yor

- Cut random **surfaces** (samples from a GP) with **thresholds** (as in Level Set Methods)
- Assign each pixel to the **first** surface which exceeds threshold (as in Layered Models)

Duan, Guindani, & Gelfand, *Generalized Spatial DP*, 2007
Spatially Dependent Pitman-Yor

- Cut random *surfaces* (samples from a GP) with *thresholds* (as in Level Set Methods)

- Assign each pixel to the *first* surface which exceeds threshold (as in Layered Models)

- Retains *Pitman-Yor marginals* while jointly modeling rich *spatial dependencies* (as in Copula Models)
Stick-Breaking Revisited

\[ \pi_k = v_k \prod_{\ell=1}^{k-1} (1 - v_\ell) \quad v_k \sim \text{Beta}(1 - a, b + k \alpha) \]

\[ \begin{align*}
\pi_1 &= v_1 \\
\pi_2 &= v_2(1 - v_1) \\
\pi_3 &= v_3(1 - v_2)(1 - v_1) \\
&\vdots
\end{align*} \]

\[ v_k = \mathbb{P}(z_i = k \mid z_i \neq k-1, \ldots, 1) \]

Multinomial Sampler:
\[ u_i \sim \text{Unif}(0, 1) \]
\[ z_i = \text{CDF}^{-1}_\pi(u_i) \]

Sequential Binary Sampler:
\[ b_{ki} \sim \text{Bernoulli}(v_k) \]
\[ z_i = \min\{k \mid b_{ki} = 1\} \]
PY Gaussian Thresholds

\[
\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-s^2/2} ds
\]

\[P[\Phi(u_{ki}) < v_k] = v_k\]

because

\[\Phi(u_{ki}) \sim \text{Unif}(0, 1)\]

Gaussian Sampler:

\[u_{ki} \sim \mathcal{N}(0, 1)\]

Sequential Binary Sampler:

\[b_{ki} \sim \text{Bernoulli}(v_k)\]

\[z_i = \min\{k \mid u_{ki} < \Phi^{-1}(v_k)\}\]

\[z_i = \min\{k \mid b_{ki} = 1\}\]
PY Gaussian Thresholds

Gaussian Sampler:
\[ u_{ki} \sim \mathcal{N}(0, 1) \]
\[ z_i = \min\{k \mid u_{ki} < \Phi^{-1}(v_k)\} \]

Sequential Binary Sampler:
\[ v_k \sim \text{Beta}(1 - a, b + ka) \]
\[ b_{ki} \sim \text{Bernoulli}(v_k) \]
\[ z_i = \min\{k \mid b_{ki} = 1\} \]
Spatially Dependent Pitman-Yor

Non-Markov Gaussian Processes:

\[ u_{ki} \sim \mathcal{N}(0, 1) \]

\[ u_{ki} \perp u_{\ell i} \]

PY prior:

\[ \psi_k \sim \text{Beta}(1 - a, b + ka) \]

Feature Assignments

\[ z_i = \min \{ k \mid u_{ki} < \Phi^{-1}(\psi_k) \} \]

\[ x_i \sim \text{Mult}(\theta_{z_i}) \]
Preservation of PY Marginals

**Why Ordered Layer Assignments?**

$$\pi_k = v_k \prod_{\ell=1}^{k-1} (1 - v_{\ell})$$

$$v_k = \mathbb{P}(z_i = k \mid z_i \neq k - 1, \ldots, 1)$$

**Stick Size Prior ➔ Random Thresholds**

$$v_k \sim \text{Beta}(1 - a, b + ka)$$

$$\tilde{v}_k = \Phi^{-1}(v_k)$$

$$\mathcal{N}(0, 1)$$
Samples from Spatial Prior

Comparison: Potts Markov Random Field
Logistic of Gaussians?

- Pass set of Gaussian processes through softmax to get *probabilities of independent* segment assignments
- Like adding *white noise* to GP before thresholding

Fernandez & Green, 2002  
Figueiredo et. al., 2005, 2007  
Woolrich & Behrens, 2006  
Blei & Lafferty, 2006
Outline

**Natural Scene Statistics**
- Counts, partitions, and power laws
- Hierarchical *Pitman-Yor* processes

**Spatial Priors for Image Partitions**
- What’s wrong with Potts models?
- Spatial dependence via *Gaussian processes*

**Unsupervised Image Analysis**
- Variational inference
- Image *segmentation*
Covariance Kernels

- Thresholds determine segment **size**: Pitman-Yor
- Covariance determines segment **shape**:

\[ C(y_i, y_j) \Longleftrightarrow \text{probability that features at locations } (y_i, y_j) \text{ are in the same segment} \]

**Bag of Features:**

\[ C(y_i, y_j) = \delta(y_i - y_j) \]

**Image Distance:**

\[ C(y_i, y_j) = e^{-\lambda(y_i-y_j)^2} \]

**Intervening Contours:**

*Discriminative dependence on maximum boundary probability along straight lines connecting feature pairs*

*Berkeley Pb (probability of boundary) detector*
HPY Variational Inference

\[ q(k, t, v, w, \theta) = \]
\[ \prod_{k=1}^{K} q(w_k | \omega_k)q(\theta_k | \eta_k) \]
\[ \prod_{j=1}^{J} \prod_{t=1}^{T} q(v_{jt} | v_{jt})q(k_{jt} | \kappa_{jt}) \]
\[ \prod_{j=1}^{J} \prod_{i=1}^{N_j} q(t_{ji} | \tau_{ji}) \]

\[ \log p(x | \alpha, \gamma, \rho) \geq H(q) + \mathbb{E}_q[\log p(x, k, t, v, w, \theta | \alpha, \gamma, \rho)] \]

- Marginal likelihood of observed features
- Entropy
- Expected values of sufficient statistics
  "negative average energy"
HPY Variational Implementation

Latent Dirichlet Allocation: Blei, Ng, & Jordan 2003
DP Mixtures: Blei & Jordan 2006; Kurihara, Welling, & Teh 2007

Desirable Properties

- Closed form, coordinate ascent updates implemented by sparse matrix operations (faster than collapsed Gibbs)
- Likelihood bound for convergence diagnosis
- Avoid multiple restarts via deterministic annealing

Why Not Collapsed Variational Methods?

- Computational cost: $O(NT + TK)$ versus $O(NK)$

  Thousands of object categories, but only a few are in each image…

- Generalization to Gaussian coupling of PY processes…
Variational for Dependent PY

Factorized Gaussian Posteriors

\[ q(u) = \prod_{k=1}^{K} \prod_{i=1}^{N} \mathcal{N}(u_{ki} \mid \mu_{ki}, \lambda_{ki}) \]

\[ q(\bar{v}) = \prod_{k=1}^{K} \mathcal{N}(\bar{v}_k \mid \nu_k, \delta_k) \]

Sufficient Statistics

\[ z_i = \min \{k \mid u_{ik} < \bar{v}_k\} \]

Allows closed form update of \( q(\theta_k) \) via

\[ P_q[u_{ki} < \bar{v}_k] = \Phi \left( \frac{\nu_k - \mu_{ki}}{\sqrt{\delta_k + \lambda_{ki}}} \right) \]

\[ \log p(x \mid \alpha, \rho) \geq H(q) + \mathbb{E}_q[\log p(u, \bar{v}, \theta \mid \alpha, \rho)] \]
Variational for Dependent PY

Updating Layered Partitions

Evaluation of beta normalization constants:

\[
E_q[\log \Phi(\bar{v}_k)] \leq \log E_q[\Phi(\bar{v}_k)] = \log \Phi \left(\frac{\nu_k}{\sqrt{1 + \delta_k}}\right)
\]

Jointly optimize each layer’s threshold and Gaussian assignment surface, fixing all other layers, via backtracking conjugate gradient with line search

Reducing Local Optima

Place factorized posterior on eigenfunctions of Gaussian process, not single features

\[
\log p(x \mid \alpha, \rho) \geq H(q) + E_q[\log p(u, \bar{v}, \theta \mid \alpha, \rho)]
\]
Robustness and Initialization

Log-likelihood bounds versus iteration, for many random initializations of mean field variational inference on a single image.
Human Image Segmentation
BSDS: Spatial PY Inference
Our Gaussian process layer representation, and low rank covariance, can create some small disconnected regions. Can further polish results by giving connected components their own layers, & possibly merging with spatial neighbors.
Comparing Spatial Models

Image | PY Learned | PY Heuristic | Multiscale NCut
---|---|---|---
with S. Ghosh
Sometimes mean shift’s kernel density estimator is effective in feature space clustering

But it can be unstable in more ambiguous images…
BSDS: Spatial PY & Mean Shift

with S. Ghosh
BSDS: Spatial PY & Mean Shift

with S. Ghosh
Multiple Spatial PY Modes

- Collected in a single uphill search sequence
- Currently exploring ways of getting more diversity…
Conclusions

Hierarchical Pitman-Yor Processes allow…

- efficient variational parsing of scenes into unknown numbers of segments
- empirically justified power law priors
- potential for learning shared appearance models from related images & scenes

Future Directions

- parallelized, scalable learning from extremely large image databases
- nonparametric models of dependency in other application domains