STAT 339
Nonparametric Clustering and Density Estimation

3 May 2017

Outline

Outline

Selecting K in a Mixture Model

- Mixture density form

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{y} \mid \theta_k), \qquad \sum_{k=1}^{K} \pi_k = 1$$

  where p_k are simple densities (e.g., Normal / Product of Bernoullis)
- One of the main challenges: How to choose K?
- Standard approaches:
  1. Cross-Validation using Log Likelihood metric
  2. (Bayesian setting) Marginal Likelihood (averaging out parameters)

Analogy to Polynomial Regression

Polynomial Normal Regression model:

$$
\begin{aligned}
t_n &= f(x) + \varepsilon_n \\
&= w_0 + w_1 x_1 + \cdots + w_D x_D + \varepsilon_n, \quad n = 1, \ldots, N \\
\varepsilon_1, \ldots, \varepsilon_N &\overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)
\end{aligned}
$$

How to choose D?

1. Cross-validation using Mean Squared Prediction Error metric
2. (Bayesian setting) Marginal likelihood (averaging out parameters)

Parametric vs. Nonparametric Prior: Regression

Polynomial Normal Regression model:

$$
\begin{aligned}
t_n &= f(x) + \varepsilon_n \\
&= w_0 + w_1 x_1 + \cdots + w_D x_D + \varepsilon_n, \quad n = 1, \ldots, N \\
\varepsilon_1, \ldots, \varepsilon_N &\overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)
\end{aligned}
$$

Standard prior on $f(x)$ is through a prior on $\mathbf{w}$

$$p(\mathbf{w} \mid \sigma_0^2) = \mathcal{N}(0, \sigma_0^2 \mathbf{I}_{D+1})$$

Induces a (marginal) prior on $\mathbf{t}$:

$$
\begin{aligned}
p(\mathbf{t} \mid \sigma_0^2, \sigma^2) &= \int p(\mathbf{w} \mid \sigma_0^2) p(\mathbf{t} \mid \mathbf{w}, \mathbf{X}) \\
&= \int \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{D+1}) \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}_N)\, d\mathbf{w} \\
&= \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_N + \sigma_0^2 \mathbf{X}\mathbf{X}^\top)
\end{aligned}
$$

Parametric vs. Nonparametric Prior: Regression

GP Normal Regression model:

$$t_n = f(\mathbf{x}) + \varepsilon_n$$

Prior is *directly* on $f(x)$:

$$p(\mathbf{f} \mid \mathbf{X}, \theta) = \mathcal{N}(\mathbf{m}, \mathbf{C} + \sigma^2 \mathbf{I})$$

$$\text{where } \mathbf{m}_n = m(\mathbf{x}_n) \qquad \mathbf{C}_{nn'} = c(\mathbf{x}_n, \mathbf{x}_{n'})$$

where m is a mean function returning the expected t at any x, and c is a covariance function returning the covariance between t_n and $t_{n'}$ values at x_n and $x_{n'}$, respectively.

Note that by setting $m(\mathbf{x}) \equiv 0$ and $c(\mathbf{x}_n, \mathbf{x}_{n'}) = \mathbf{x}_n \mathbf{x}_{n'}^{\mathsf{T}}$, we get standard linear regression.

Parametric vs. Nonparametric Prior: Clustering

- Gaussian Mixture density form

$$p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{y} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad \text{where } \sum_{k=1}^{K} \pi_k = 1$$

- Standard prior (diagonal $\boldsymbol{\Sigma}$ case):

$$\begin{aligned} p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) &= \mathrm{Dir}(\alpha_1, \ldots, \alpha_K) \\ p(\boldsymbol{\mu}_k \mid \boldsymbol{\mu}_{0,k}, \boldsymbol{\Sigma}_{0,k}) &= \mathcal{N}(\boldsymbol{\mu}_{0,k}, \boldsymbol{\Sigma}_{0,k}) \\ p(\sigma^2_{k,d} \mid a_{k,d}, b_{k,d}) &= \mathrm{InverseGamma}(a_{k,d}, b_{k,d}) \end{aligned}$$

- Induces a (marginal) prior on $\mathbf{y}$:

$$p(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, a, b) =$$

$$\sum_{k=1}^{K} \frac{\alpha_k}{\sum_{k'=1}^{K} \alpha_{k'}} \prod_{d=1}^{D} \frac{\Gamma(a_{k,d} + \frac{1}{2})}{\Gamma(a_{k,d})} \sqrt{2\pi b_{k,d}} \left(1 + \frac{(y_d - \mu_{0,k,d})^2}{2b_{k,d}}\right)^{a_{k,d} + \frac{1}{2}}$$

Parametric vs. Nonparametric Prior: Clustering

- An infinite Gaussian mixture model

$$p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(\mathbf{y} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad \text{where } \sum_{k=1}^{\infty} \pi_k = 1$$

- Analogous to the GP regression model, we can put a prior *directly* on the mixture density, G.

$$p(\mathbf{y} \mid \alpha, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, a, b) = G, \quad G \sim \mathrm{DP}(\alpha, G_0)$$

  where $\mathrm{DP}(\alpha, G_0)$ is a **Dirichlet Process** with *concentration parameter* α and *base measure* G
- The *concentration parameter*, α, governs the mixing weights, as in the finite mixture model
- The *base measure*, G_0, is the prior distribution over any particular $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$; e.g., the conjugate prior parameterized by $\boldsymbol{\mu}_0, \boldsymbol{\Sigma}, a, b$.

Outline

Outline

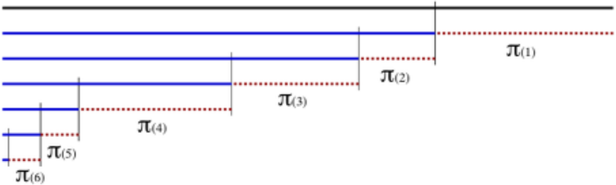
The Dirichlet Prior on Mixing Weights

- Gaussian mixture density

$$p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{y} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad \text{where } \sum_{k=1}^{K} \pi_k = 1$$

- The prior on $\boldsymbol{\pi}$ distributes a unit mass across K weights.
- In the Dirichlet prior, the prior expectation is that the weight on component k is $\frac{\alpha_k}{\sum_{k'} \alpha}$.
- For larger α the strength of this belief is greater.
- For smaller α that is the mean case, but individual distributions drawn from the Dirichlet tend to put most mass on one component.

Generating Samples from a Dirichlet



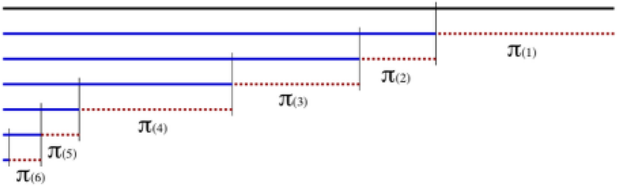
- Many methods, but one is iterative and illustrative to understand the DP.

To generate $\pi_1, \ldots, \pi_K$ from a Dir$(\alpha_1, \ldots, \alpha_K)$:

For $k = 1, \ldots, K$

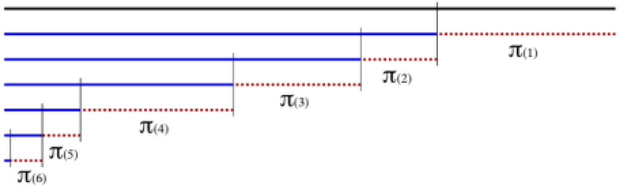
1. Draw $\tilde{\pi}_k \sim \text{Beta}(\alpha_k, \sum_{k'=k+1}^{K} \alpha_{k'})$
2. Set $\pi_k := \tilde{\pi}_k \prod_{k'=1}^{k-1} (1 - \tilde{\pi}_{k'})$

Stick-Breaking Process



- Idea: We start with a "stick" of length 1, and break off a random piece for $k = 1$; then repeat the process with the remaining stick, until we have K pieces.

Infinite Stick-Breaking Process



We can construct an infinite version of this process by breaking off sticks forever: "Zeno's random breadstick"

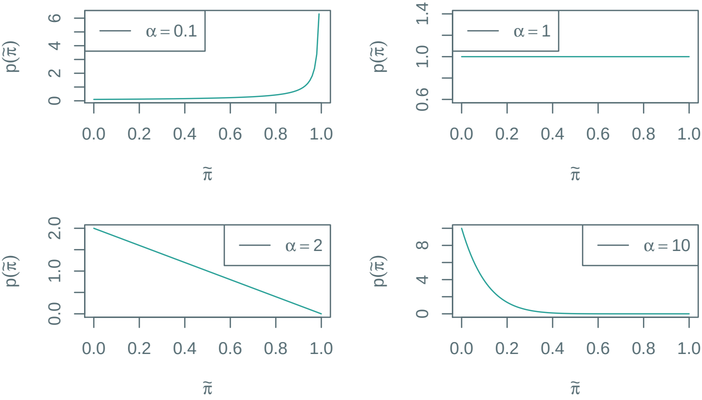
To generate infinitely many mixing weights $\pi_1, \pi_2, \ldots$ from a Dirichlet Process with concentration parameter α:

For $k = 1, 2, \ldots$

1. Draw $\tilde{\pi}_k \sim \text{Beta}(1, \alpha)$
2. Set $\pi_k := \tilde{\pi}_k \prod_{k'=1}^{k-1} (1 - \tilde{\pi}_{k'})$

Stick-Breaking Process: Interpreting α

- Suppose we stop when we've broken off probability 0.999
- How does the choice of α affect the number of clusters we get before this happens?



Outline

Completing the DP Prior

Recall that we said that the DP put a prior directly on the infinite mixture density of $\mathbf{y}$:

$$p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(\mathbf{y} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad \text{where } \sum_{k=1}^{\infty} \pi_k = 1$$

$$p(\mathbf{y} \mid \alpha, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, a, b) = \mathsf{DP}(\alpha, G_0)$$

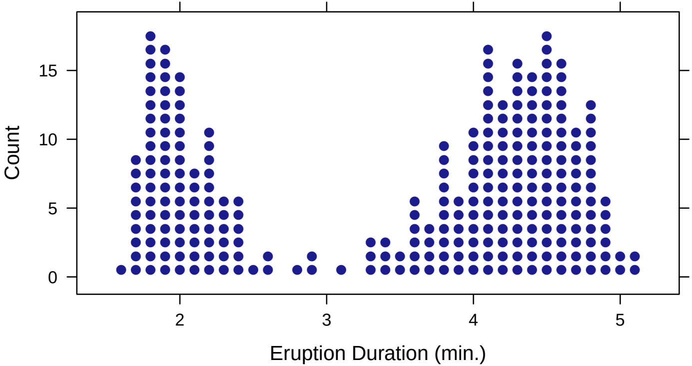
What is the role of G_0?

- G_0 is the prior on each set of component parameters.
- Generatively: after breaking off a "stick" with weight π_k, draw $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ from G_0

Outline

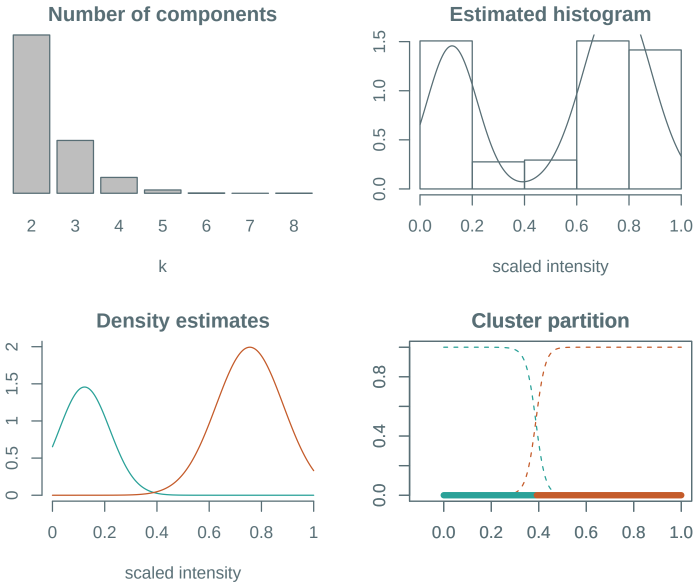
Outline

Old Faithful Eruption Durations

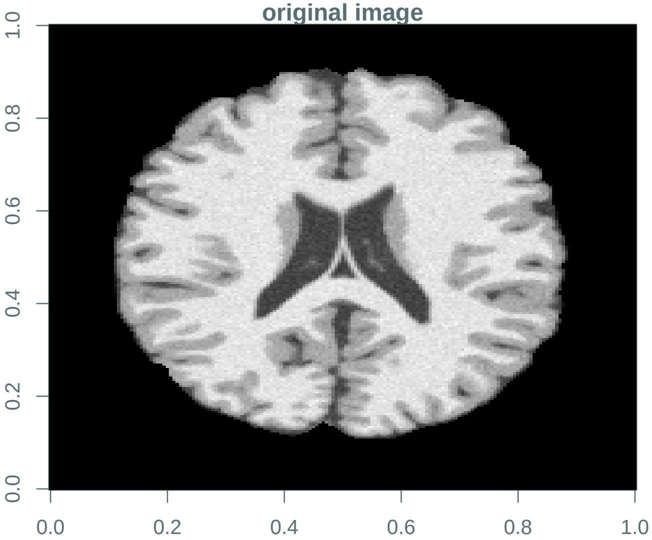


Goal: Use a DP infinite mixture model with Gibbs sampling to find clusters in this data.

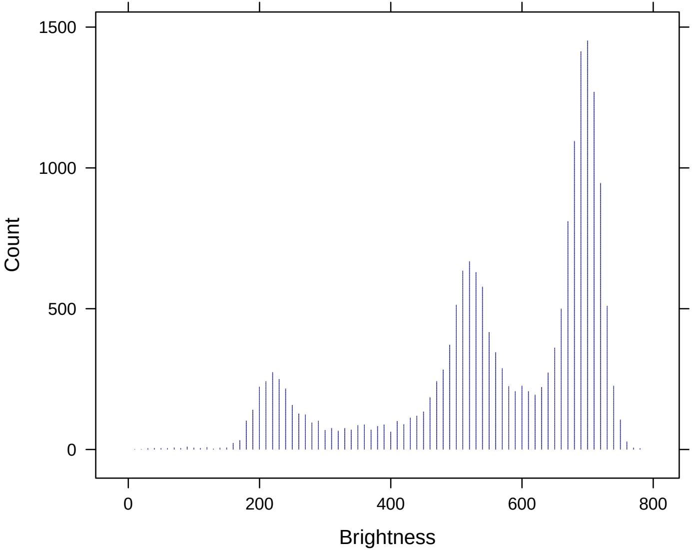
Gibbs Final Results



Outline



Goal: Cluster pixels by brightness using a DP-GMM



Goal: Cluster pixels by brightness using a DP-GMM

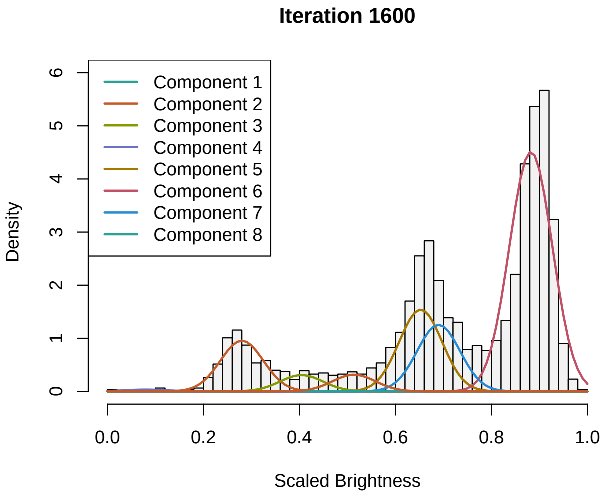


Figure: Cluster Estimates at Selected Gibbs Iterations for the MRI data

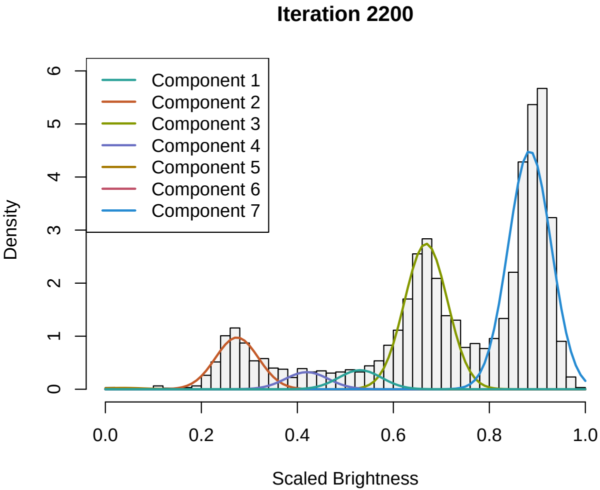


Figure: Cluster Estimates at Selected Gibbs Iterations for the MRI data

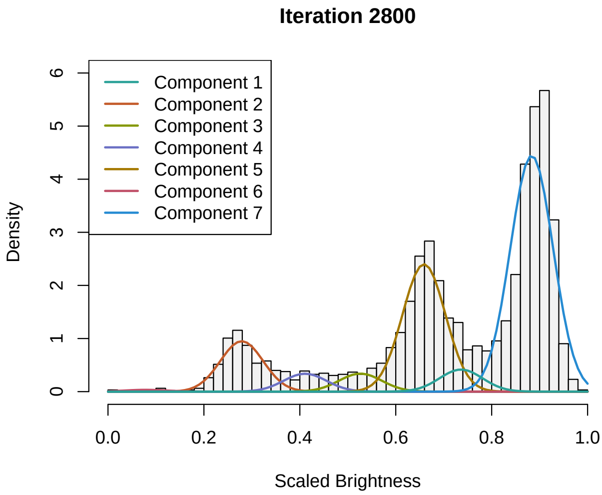


Figure: Cluster Estimates at Selected Gibbs Iterations for the MRI data

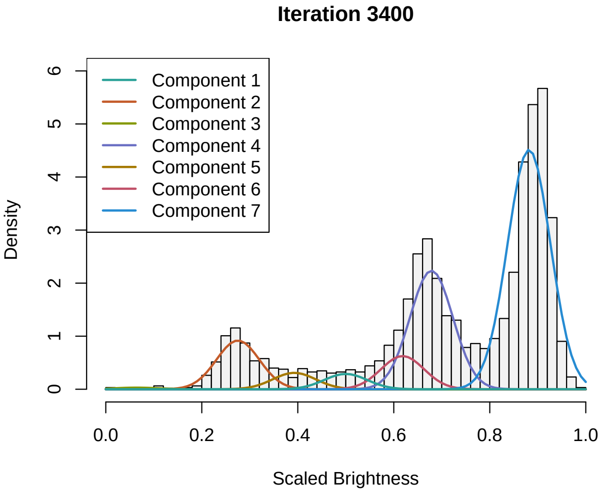


Figure: Cluster Estimates at Selected Gibbs Iterations for the MRI data

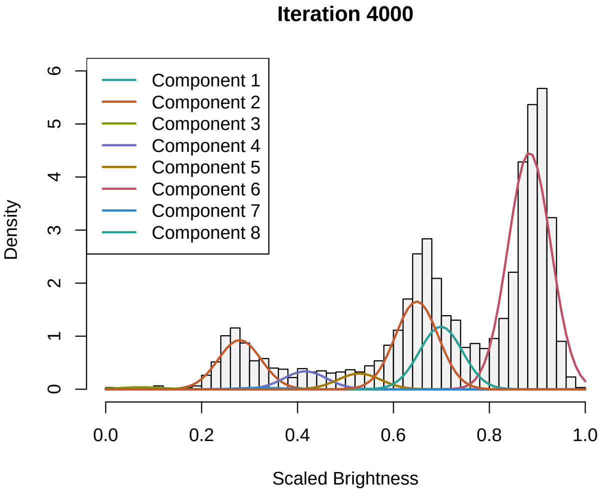
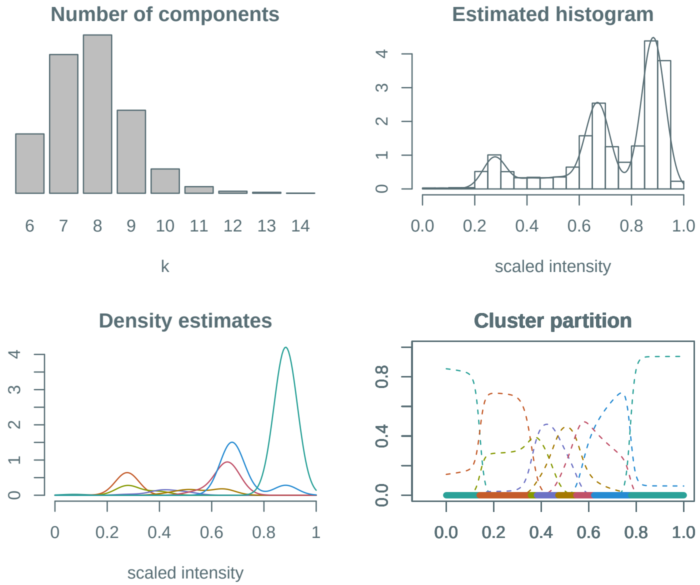


Figure: Cluster Estimates at Selected Gibbs Iterations for the MRI data

Gibbs Final Results



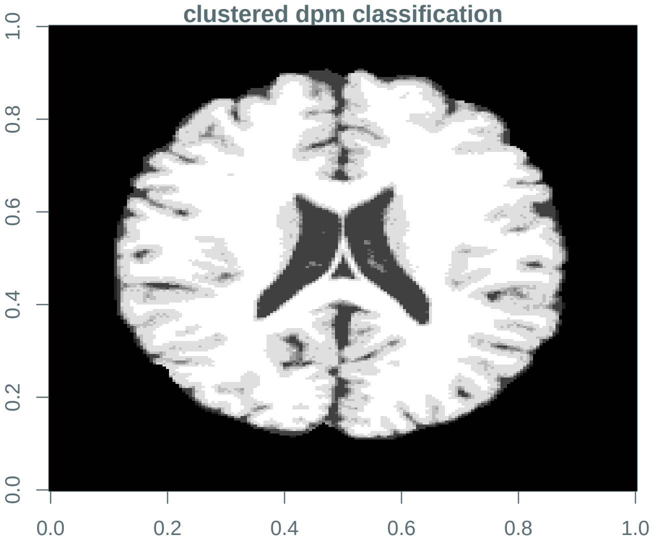
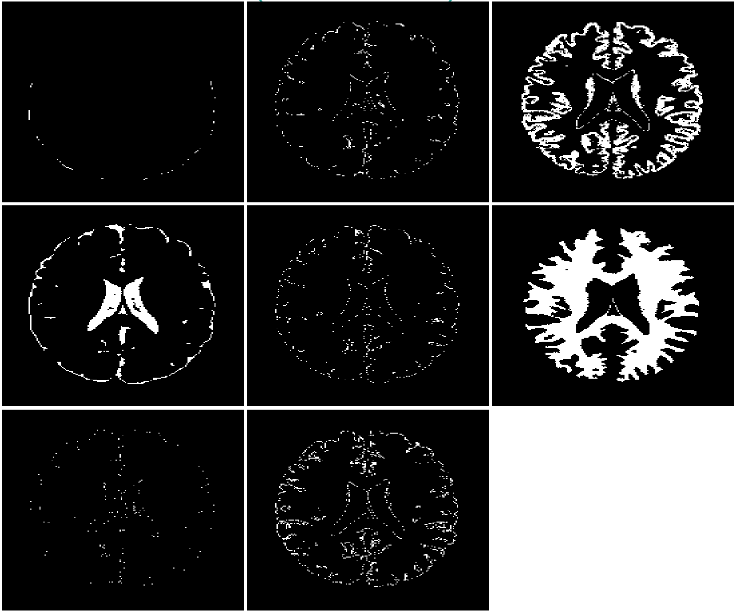
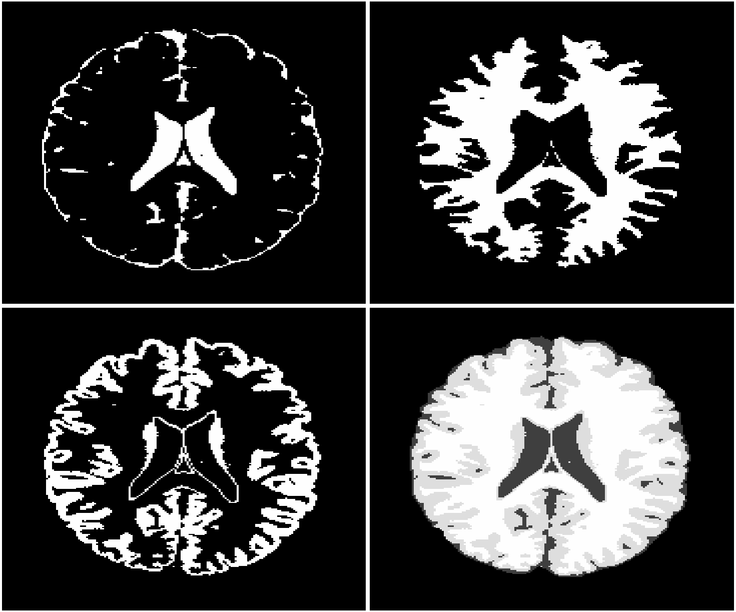


Figure: Original image with each pixel assigned to the mean brightness of its

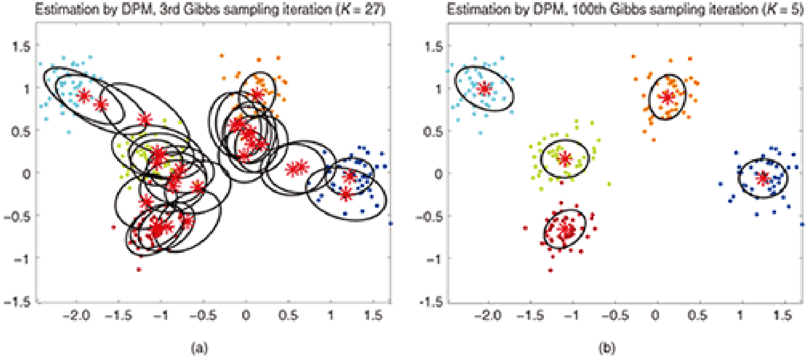
Individual Clusters (7 Clusters)



Top 3 clusters



2D Data



Estimation by DPM, 3rd Gibbs sampling iteration (K = 27)
Estimation by DPM, 100th Gibbs sampling iteration (K = 5)
1.5
1
0.5
0
−0.5
−1
−1.5
−2.0
−1.5
−1
−0.5
0
0.5
1
1.5
(a)
(b)

Outline

The Full Model So Far

We have defined our (Gaussian, for concreteness) infinite mixture model as follows:

$$\mathbf{x}_n \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \sum_{k=1}^{\infty} \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\boldsymbol{\pi} \sim \mathrm{Stick}(\alpha) \qquad \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \overset{i.i.d}{\sim} G_0 \quad k = 1, 2, \ldots$$

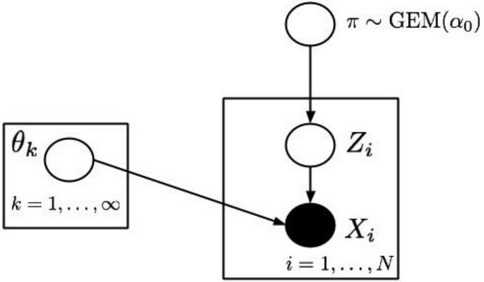
where

- $\mathrm{Stick}(\alpha)$ is the "infinite stick-breaking process" with parameter α that returns a random infinite sequence of weights that sum to 1
- G_0 is a joint prior distribution for all component parameters; here $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$

An Expanded Model

To generate data we can sample cluster indicators $z_n, n = 1 \ldots, N$ from the $\boldsymbol{\pi}$ distribution over the cluster labels; then generate $\mathbf{x}_n$ from $\mathcal{N}(\boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}_{z_n})$.

$$\boldsymbol{\pi} \sim \text{Stick}(\alpha) \qquad \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \overset{i.i.d.}{\sim} G_0$$
$$z_n \sim \text{Categorical}(\boldsymbol{\pi})$$
$$\mathbf{x}_n \mid z_n, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



- GEM is another notation for Stick
- Here $\theta_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and i indexes observations.

Outline of a Gibbs Sampler

At iteration s, given $\{bz, \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}\}^{(s-1)}$

1. Assign data points to clusters: sample $\mathbf{z}^{(s)}$
2. Using updated $\mathbf{z}^{(s)}$, update $\boldsymbol{\pi}^{(s)}$
3. Using updated $\mathbf{z}^{(s)}$ (hence, partition of data into clusters), update $\theta_k, k = 1, \ldots, \infty$

Seems elegant enough, abstractly, but.... requires infinitely many variables!

A Collapsed Model

- Instead of sampling the full (infinte) $\boldsymbol{\pi}$ vector of cluster weights, we can collapse all "unrepresented" clusters into a single one.
- Then, only update params for components represented in $\mathbf{z}$, $1, \ldots, K$, and approximate likelihood for "something new" by sampling parameters from the prior.
- Turns out we will be able to calculate

  $$p(z_n \mid \mathbf{z}_{-n}, \theta_1, \ldots, \theta_K, \theta_{new}) = \int p(z_n \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid \alpha, \mathbf{z}_{-n}) \, d\boldsymbol{\pi}$$

  integrating out (averaging over) all possible "stick weights", $\boldsymbol{\pi}$.
- Then we can put each z_n in its own Gibbs block and sample it conditioned on all the others.

Integrating out $\boldsymbol{\pi}$ in the finite model

Recall from our Naive Bayes text classifier that when we put a Dirichlet prior on a (finite) set of category weights, we can find the predictive distribution analytically. If

$$p(\boldsymbol{\pi}) = \mathsf{Dir}(\alpha_1, \ldots, \alpha_K) \qquad p(z = k \mid \boldsymbol{\pi}) = \pi_k$$

Then

$$p(\boldsymbol{\pi} \mid \mathbf{z}) = \mathsf{Dir}(\alpha_1 + N_1, \ldots, \alpha_K + N_K)$$

where N_k counts the number of n for which $z_n = k$, and

$$\begin{aligned} p(z_{N+1} = k \mid \mathbf{z}) &= \int p(z_{new} = k \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid \mathbf{z})\, d\boldsymbol{\pi} \\ &= \int \pi_k \mathsf{Dir}(\boldsymbol{\pi} \mid \alpha_1 + N_1, \ldots, \alpha_K + N_K)\, d\boldsymbol{\pi} \\ &= \mathbb{E}_{\mathsf{Dir}(\boldsymbol{\pi} \mid \alpha_1 + N_1, \ldots, \alpha_K + N_K)} \{\pi_k\} \\ &= \frac{\alpha_k + N_k}{(\sum_k \alpha_k) + N} \end{aligned}$$

Integrating out $\boldsymbol{\pi}$ in the infinite model

So with finite K, the conditional distribution of any z_n given all the others is defined by

$$p(z_{N+1} = k \mid \mathbf{z}) = \frac{\alpha_k + N_k}{\left(\sum_{k=1}^{K} \alpha_k\right) + N}$$

What happens if we hold $\alpha := \sum_{k=1}^{K} \alpha_k$ constant, set α_k to be constant at α/K, and let $K \to \infty$?

$$p(z_{N+1} = k \mid \mathbf{z}) = \lim_{K \to \infty} \frac{\alpha/K + N_k}{\alpha + N} = \frac{N_k}{\alpha + N}$$

So z_{N+1} will be assigned to an existing cluster proportionally to the number of other cases assigned to that cluster. How much proability is left over?

Prior Probability of a New Cluster

If we number represented clusters as $1, \ldots, L$, then the total probability that z_{N+1} is in an existing cluster is

$$\sum_{l=1}^{L} \frac{N_l}{\alpha + N} = \frac{N}{\alpha + N}$$

which means that with probability

$$\frac{\alpha}{\alpha + N}$$

z_{N+1} belongs to some "new" cluster.

Outline

The "Chinese Restaurant Process"

- The process outlined here is often described using the metaphor of a Chinese Restaurant with infinitely many tables, each with infinite capacity.

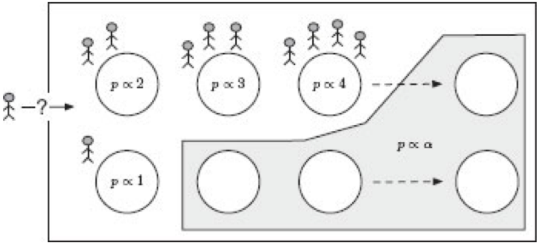


FIGURE 10.6 A cartoon depiction of the Chinese restaurant process. A new diner sits at a non-empty table with probability proportional to the number of diners and sits at a new table with probability proportonal to α.

- Defines a probability distribution over partitions into arbitrarily many components.

Outline

Posterior Distribution for z_n

Having defined a (conditional) prior for z_n (given all other zs), finding the posterior is simply a matter of multiplying by the likelihood:

$$p(z_n = k \mid \mathbf{x}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \begin{cases} \left(\frac{N_l}{N_k+\alpha}\right)\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), & 1 \leq k \leq L \\ \left(\frac{\alpha}{N_k+\alpha}\right)\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new}), & k = k_{new} \end{cases}$$

Posterior Distribution for $\boldsymbol{\theta}$

Having fixed all the $\mathbf{z}$s (and thus partitioned the data), we can update each $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ as in the finite mixture model:

$$p(\boldsymbol{\mu}_k, \Sigma_k \mid \mathbf{z}, \mathbf{X}) \propto G_0 \cdot \mathcal{N}(\mathbf{X}_k \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$
$$p(\boldsymbol{\mu}_{new}, \Sigma_{new} \mid \mathbf{z}, \mathbf{X}) \propto G_0$$

where G_0 is the prior (base measure of the DP) and $\mathbf{X}_k$ represents the data matrix for those observations currently assigned to cluster k.