Gibbs Sampling for (Coupled) Infinite Mixture Models in the Stick Breaking Representation

Ian Porteous, Alex Ihler, Padhraic Smyth, Max Welling
Department of Computer Science
UC Irvine, Irvine CA 92697-3425
{*iporteou,ihler,smyth,welling*}*@ics.uci.edu*

Presented by Layla Oesper
Department of Computer Science
Brown University
September 27,2011
CSCS 2950-P

Main Idea

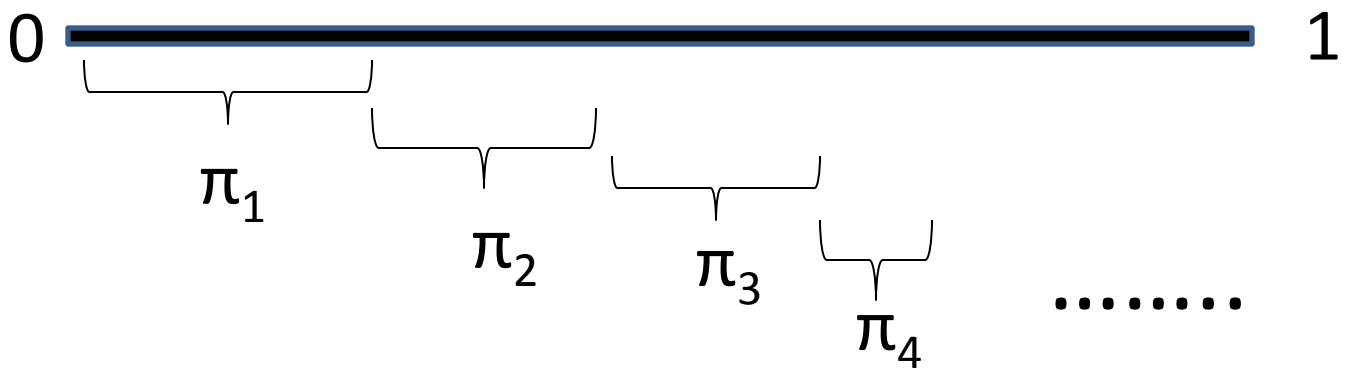
- Problem: Recently introduced Gibbs samplers for infinite mixtures do not mix well over cluster labels.

- Solution: Introduction of particular mixing moves to Gibbs samplers.

Notation

- X = observed data $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$
- Z = assignment variables $\mathbf{Z} = \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$
- V = stick variables $V = V_1, V_2, \ldots, V_i, \ldots$
- π = cluster weights $\pi = \pi_1, \pi_2, \ldots, \pi_i, \ldots$

Stick Breaking Construction



$$V_i \sim Beta(a_i, b_i)$$

$$\pi_i = \pi_i(V) = V_i \prod_{j=1}^{i-1} (1 - V_j)$$

Changing these parameters allows us to realize different types of processes

Stick Breaking Construction

$$\pi_i = \pi_i(V) = V_i \prod_{j=1}^{i-1}(1 - V_j)$$

$$V_i \sim Beta(1, \alpha)$$

Dirichlet Process

$$\pi_i = \pi_i(V) = V_i \prod_{j=1}^{i-1}(1 - V_j)$$

$$V_i \sim Beta(1 - a, b + i \times a)$$

$$a \in [0, 1), b > -a$$

Pitman-Yor Process

Stick Breaking Construction

$$\pi_i = \pi_i(V) = V_i \prod_{j=1}^{i-1}(1 - V_j)$$

$$V_i \sim Beta(1, \alpha)$$

Exponential distribution of cluster sizes

$$\pi_i = \pi_i(V) = V_i \prod_{j=1}^{i-1}(1 - V_j)$$

$$V_i \sim Beta(1 - a, b + i \times a)$$

$$a \in [0, 1), b > -a$$

If you set :

$$a = \beta, b = 0$$

Then, there is a power law distribution of cluster sizes.

Stick Breaking Construction

$$\pi_i = \pi_i(V) = V_i \prod_{j=1}^{i-1}(1 - V_j)$$

$$V_i \sim Beta(a_i, b_i)$$

$$a_i = \gamma_i$$

$$b_i = \sum_{j=1}^{\infty} \gamma_j$$

Another formulation

Posterior Expected Cluster Weights

$$\mathbb{E}[\pi_i | Z] = \frac{\gamma_i + N_i}{\gamma + N}$$

$$\gamma = \sum_{i=1}^{\infty} \gamma_i$$

Ishwaran & James 2003

Stick Breaking Construction

WARNING!

Stick breaking representation operates over the space of explicit cluster assignments, not equivalence classes.

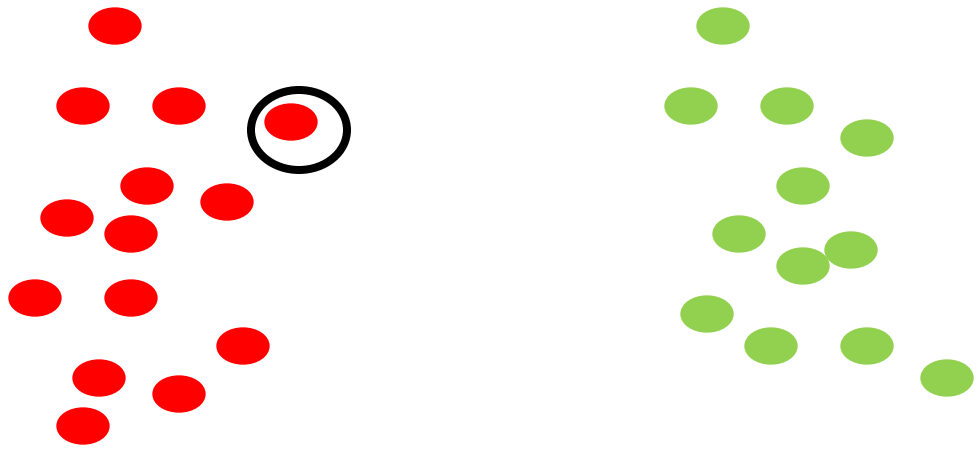
Mixing over Cluster Labels

- Claim: Clusters with lower indexes have high prior probability than clusters with higher indexes (due to stick breaking). This may results in slow mixing over cluster labels.

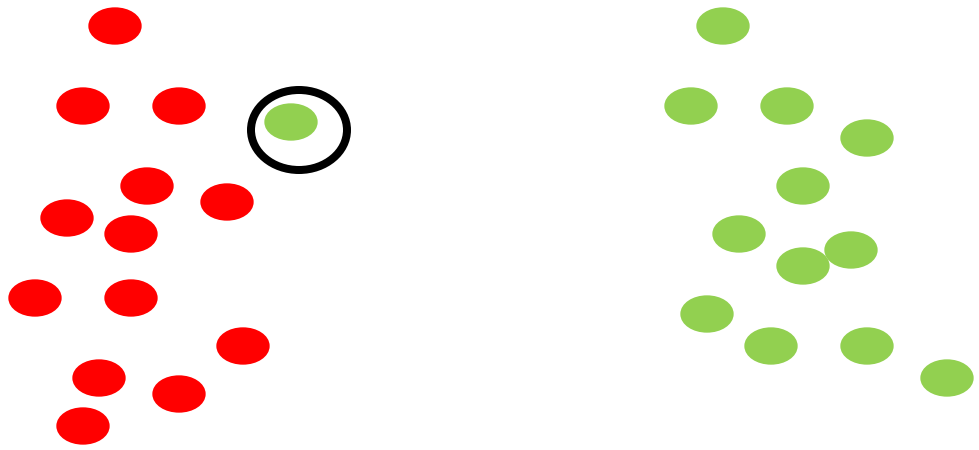
- Solution: Explicit mixing over cluster labels needs to occur.

Intuition Behind Slow Mixing

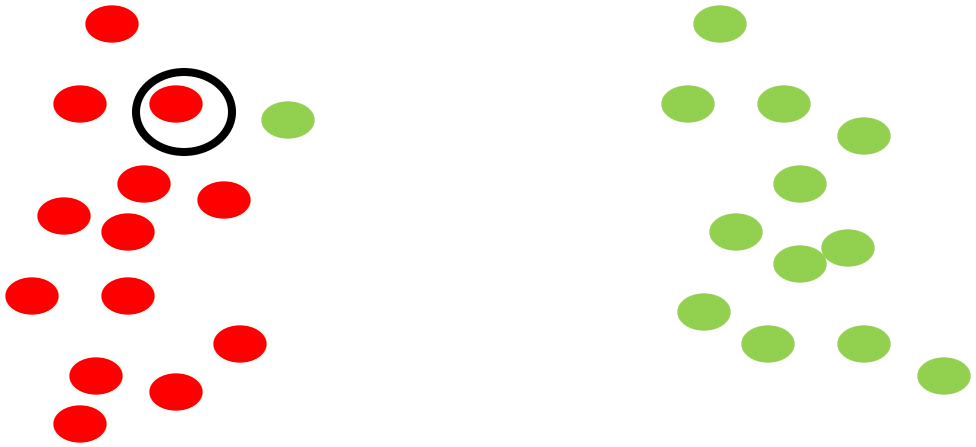
Intuition Behind Slow Mixing



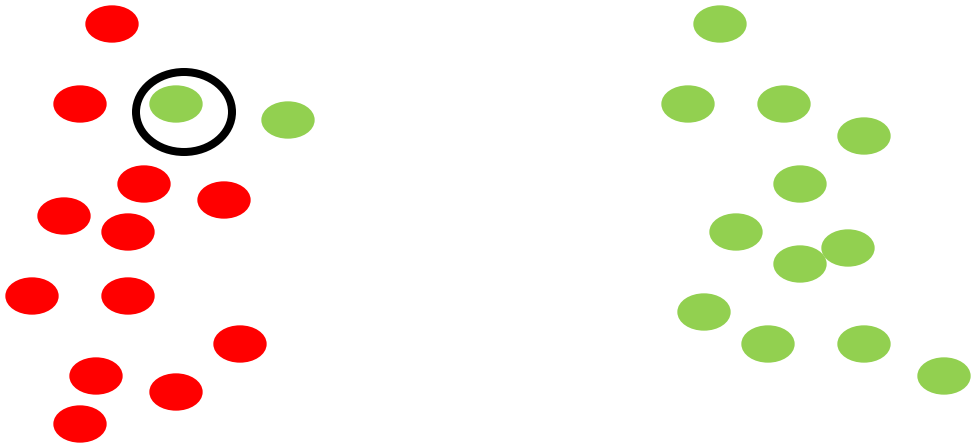
Intuition Behind Slow Mixing



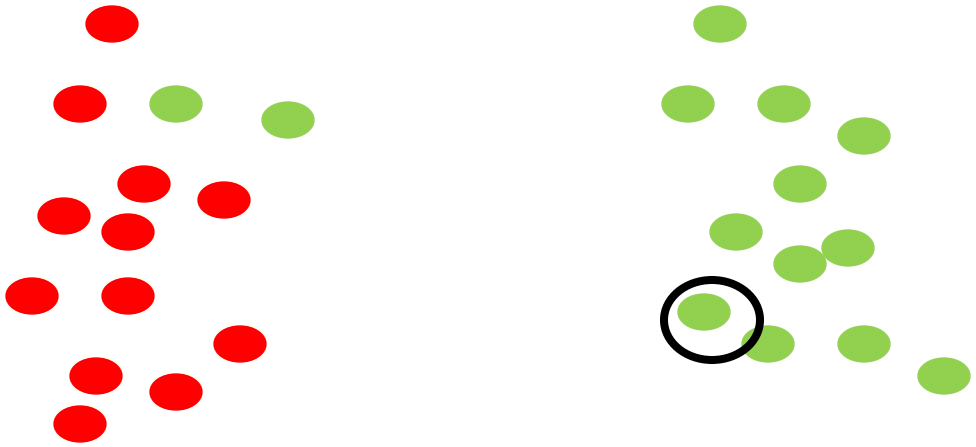
Intuition Behind Slow Mixing



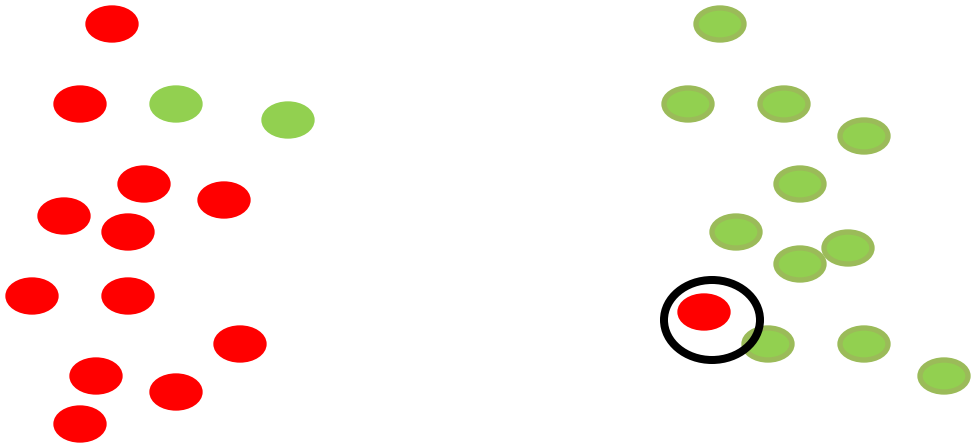
Intuition Behind Slow Mixing



Intuition Behind Slow Mixing



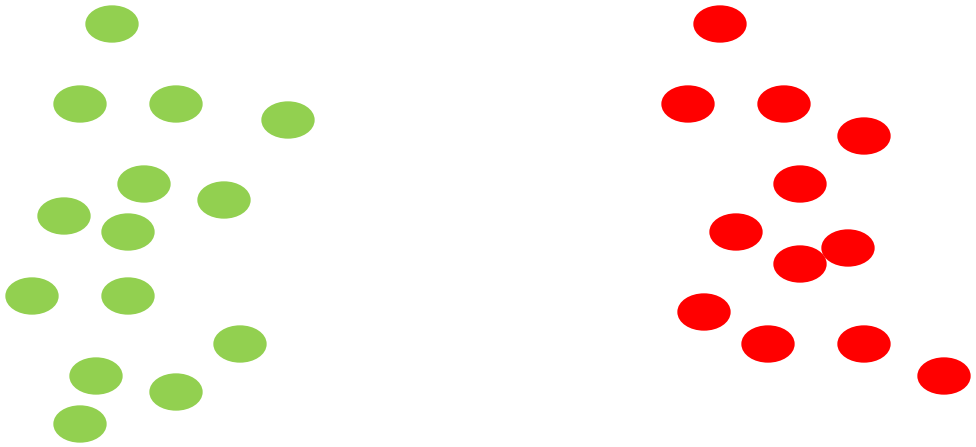
Intuition Behind Slow Mixing



Intuition Behind Slow Mixing

Repeat until….

Intuition Behind Slow Mixing



This may take a long time to happen.

Monte Carlo Model

The Setup:

$$P(X|Z,\boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}[\mathbf{x}_n; \boldsymbol{\mu}_{z_n}, \Sigma_{z_n}]$$ ← Gaussian Clusters

$$P(Z|V) = \prod_{n=1}^{N} \pi_{z_n}(V)$$ ← Stick Breaking Construction

$$P(V) = \prod_{i=1}^{\infty} \mathcal{B}(V_i\,;\, a_i, b_i)$$ ← Stick Lengths are Beta Distributed

$$P(\boldsymbol{\theta}) = \prod_{i=1}^{\infty} \mathcal{NW}[\boldsymbol{\mu}_i, \Sigma_i^{-1}]$$ ← Normal-Wishart prior for cluster means and inverse covariances

Monte Carlo Model

The Original Method:

1. Iterate through all data points x_n
   A. Draw u uniformly at random from [0,1]
   B. Find the first cluster index i^* such that:

$$\sum_{i=1}^{i^*} P(z_n = i | Z_{(-n)}, X) \geq u$$

   C. Set $z_n = i^*$
2. Repeat.

Metropolis Hastings Algorithm

1. Propose a move from state x to state x' using a proposal distribution:

$$q(x'|x)$$

2. Decide whether or not to accept this move with acceptance probability:

$$r = min(1, \alpha) \qquad \alpha = \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)}$$

Additional Moves

1. **Label-Swap:**
   a) Choose 2 clusters with prior probability : $p_i = \gamma_i / \gamma$
   b) Propose swap using Metropolis-Hastings accept/reject rule

$$P_{\text{accept}} = \min\left[1, \frac{\prod_{n=0}^{N_j-1}(\gamma_i + n)\prod_{n=0}^{N_i-1}(\gamma_j + n)}{\prod_{n=0}^{N_i-1}(\gamma_i + n)\prod_{n=0}^{N_j-1}(\gamma_j + n)}\right]$$

Additional Moves

2. Label-Permute:

a) Sample some index i from the prior: $p_i = \gamma_i / \gamma$

b) Randomly permute all the cluster labels with an index <= i

Probability of accept is similar to label-swap, but considers larger range of clusters permuted when creating the Metropolis-Hastings accept rule

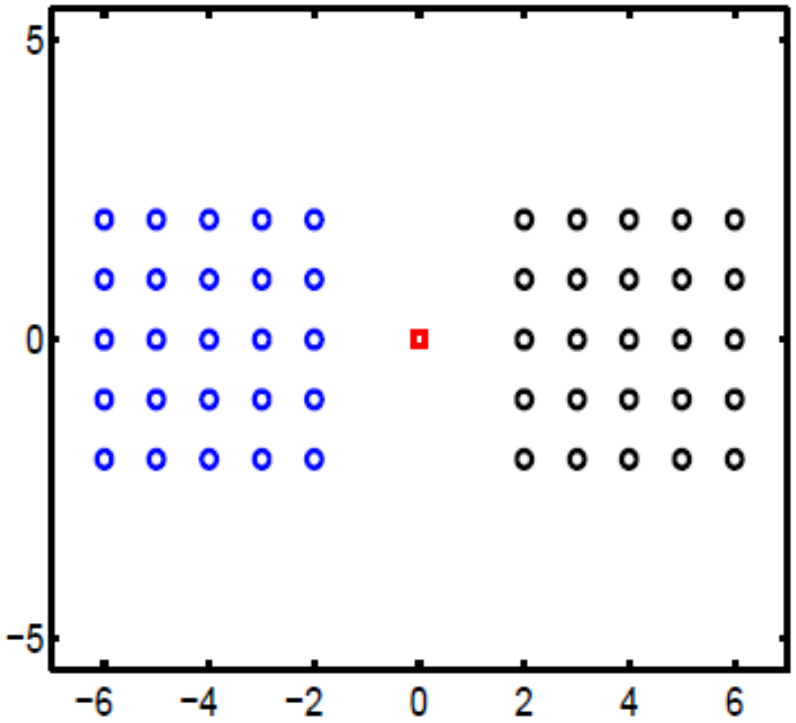
Monte Carlo Model

The Updated Method:

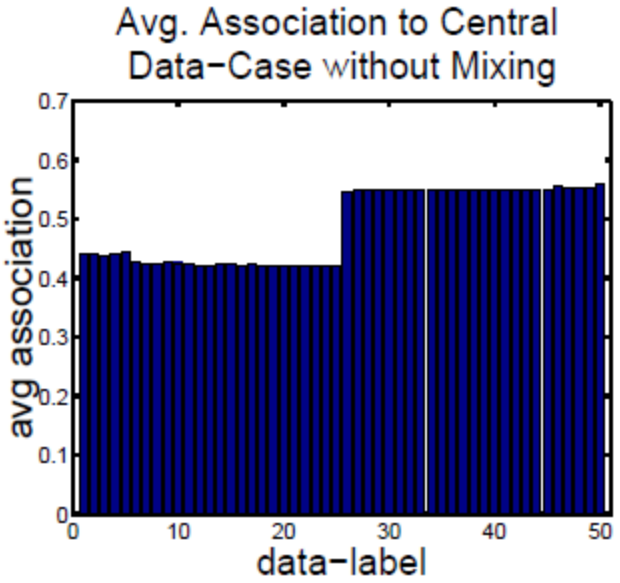
1. Iterate through all data points x_n
   A. Draw u uniformly at random from [0,1]
   B. Find the first cluster index i^* such that:

$$\sum_{i=1}^{i^*} P(z_n = i | Z_{(-n)}, X) \geq u$$

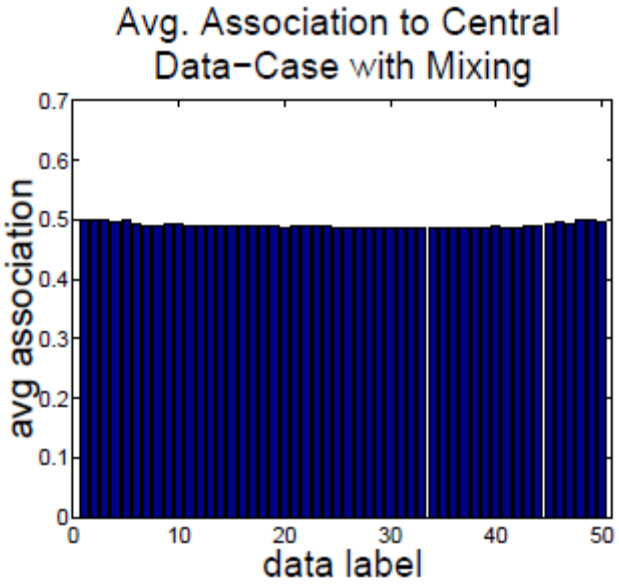
   C. Set $z_n = i^*$
   D. With probability M_i make mixing move i.
2. Repeat until convergence.


5
0
−5
−6
−4
−2
0
2
4
6

(a)


Avg. Association to Central
Data−Case without Mixing
avg association
data−label
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
10
20
30
40
50

(b)


Avg. Association to Central
Data−Case with Mixing
avg association
data label
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
10
20
30
40
50

(c)

Swap Moves

- Why just these two moves?
- Why do these moves work?
- We saw an example where there were 2 clusters best explained the data – how well does the method work when a larger number of clusters are necessary?
- How do these moves work on different sized clusters?

Motivation Behind the Dependent Dirichlet Process

Table 1: Girls, weight for age. Age is in months, LSD, USD and RSD are the "lower standard deviation", "upper standard deviation", and ratio of the USD to LSD. All weights are in kilograms.

Age	LSD	Mean	USD	RSD
0	0.5	3.2	0.4	0.80
1	0.6	4.0	0.5	0.83
20	1.2	11.2	1.2	1.00
40	1.6	14.8	2.1	1.31
60	1.9	17.7	2.8	1.47
100	3.8	26.0	5.8	1.53

Changes from a left skewed distribution to a right skewed distribution

*We'd like a model that is allowed to change over time, where the conditional distribution of weight, at similar ages is similar.

MacEachern 2000

Dependent Dirichlet Process

- Take a finite number of the previously described model, indexed by t (time).
- Couple the models at the level of the cluster parameters θ
  - Can impose “smoothness” over time between
    - Cluster means
    - Cluster covariances
    - Cluster size

The Model

Family of T joint distributions where:

$$P(X, Z, V, \boldsymbol{\theta}) = P(\boldsymbol{\theta}) \prod_t P(X_t|Z_t, \boldsymbol{\theta}_t) P(Z_t|V_t) P(V_t)$$

Individual terms are from the previous model, and $P(\boldsymbol{\theta})$ is coupled via a joint prior distribution.

Alternate between updating class assignments and updating cluster parameters for the different time steps by sampling from the following conditional distributions:

$$P(z_{n_t}|X_t, Z_{-n_t}, \boldsymbol{\theta}_t) \qquad P(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{-t}, X_t, Z_t)$$

Additional Moves

Label-Swap:

1. Swap labels i and j for all time slices in an interval $[t_1, t_2]$ where these boundaries are picked using:
   a. $t_1 = t_2$
   b. Uniformly at random in [1, T] where T is last time slice
   c. $t_1=1$ and $t_2=T$
2. Propose swap using Metropolis-Hastings accept/reject rule

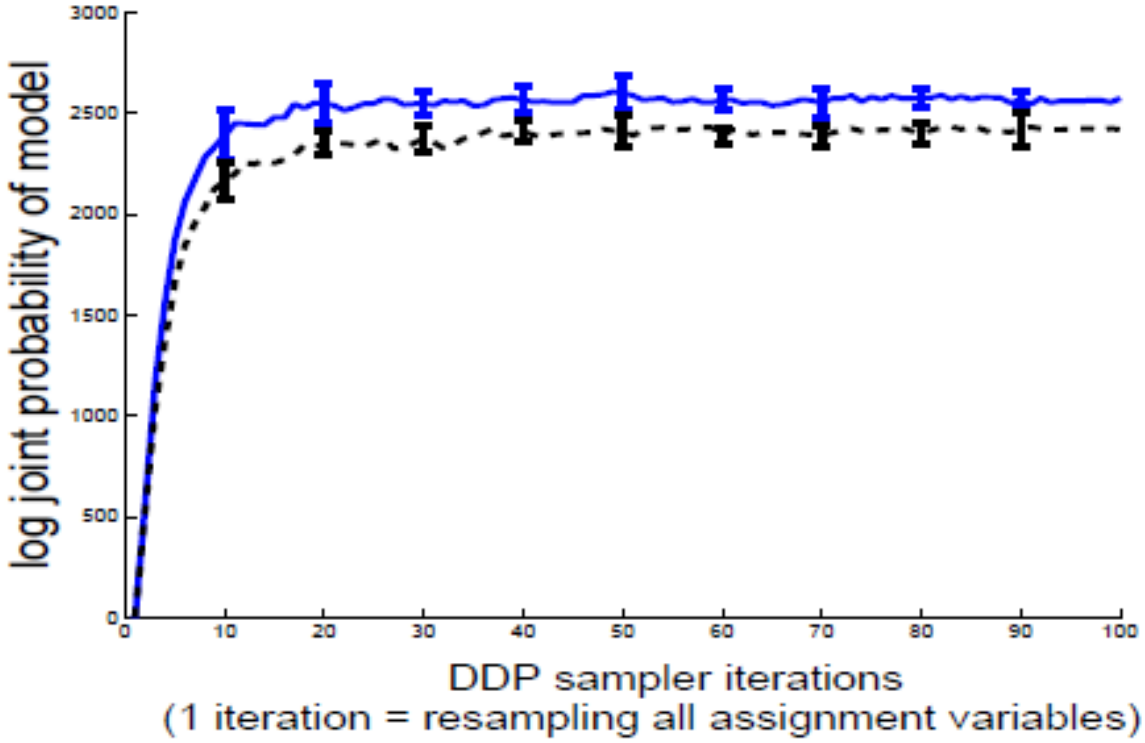


Figure 2: Log joint probability for the DDP sampler with extra mixing moves (solid) and without extra mixing moves (dashed). Curve averaged over 5 runs. The DDP with mixing moves ends up in a region of higher probability.

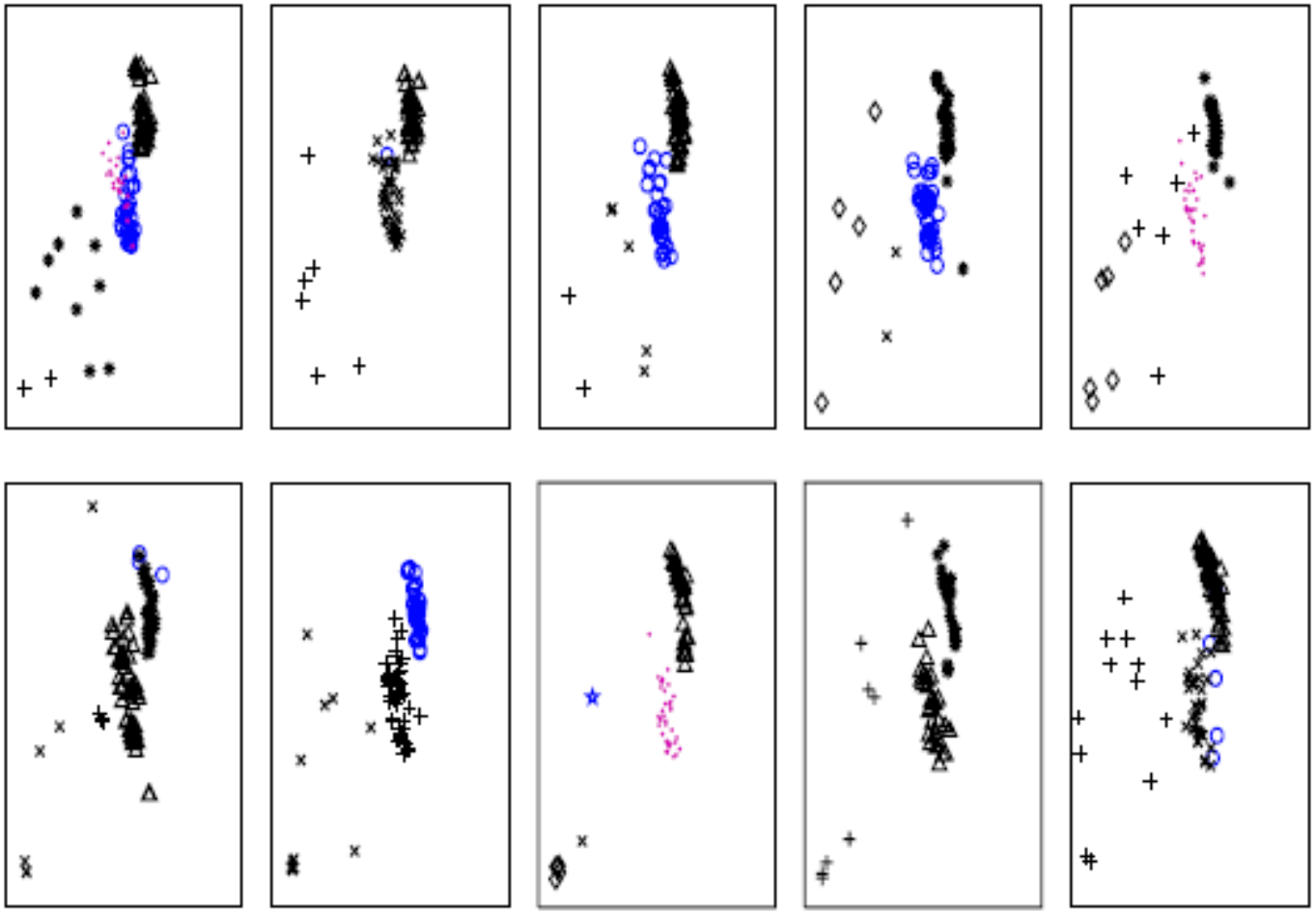


Figure 3: Cluster assignment with highest joint probability for DDP sampler without mixing moves. Points with different markers have different cluster labels. Clusters are not in correspondence.

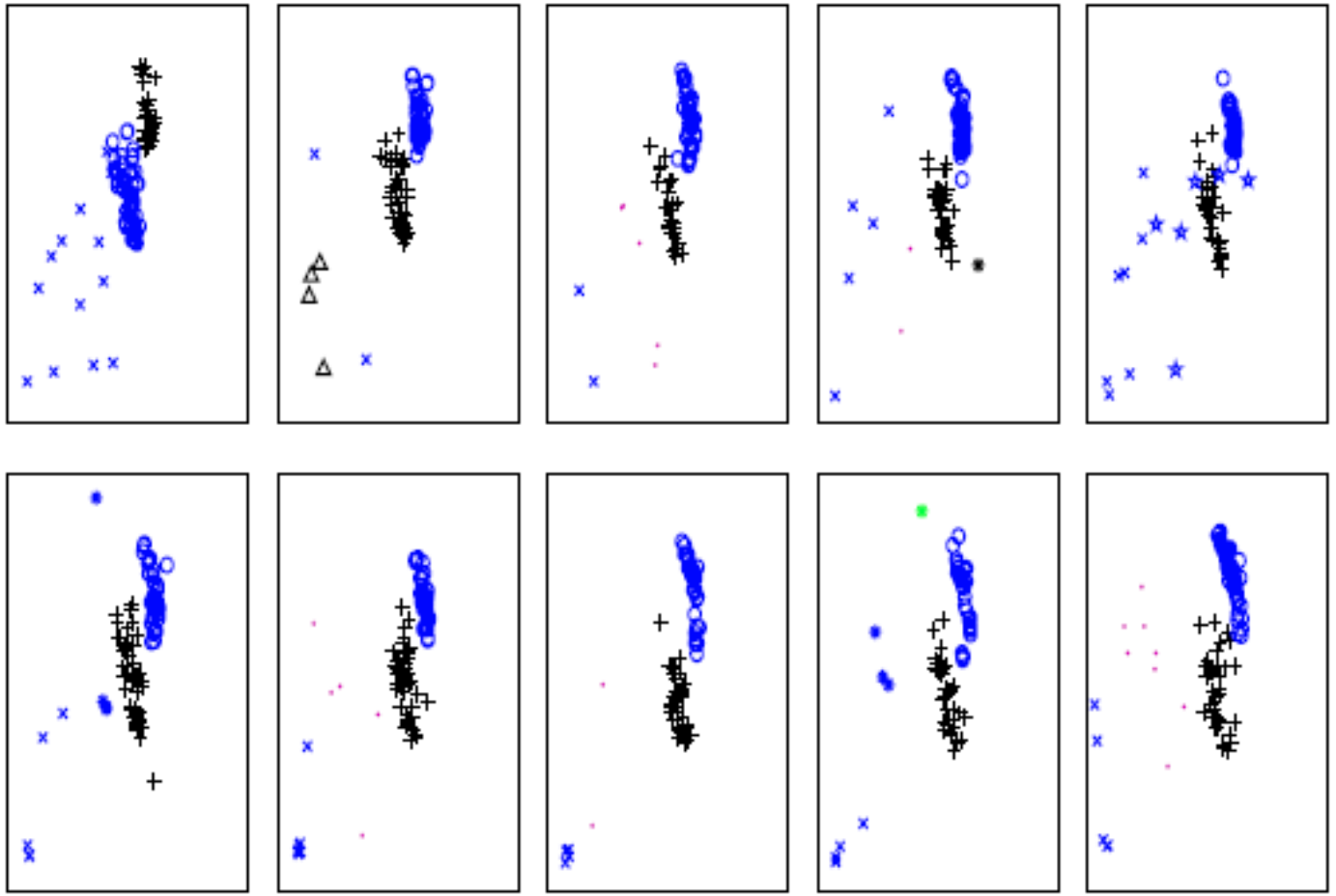


Figure 4: Same as in figure 3 but for DDP sampler with mixing moves. Almost all clusters have been brought into correspondence.

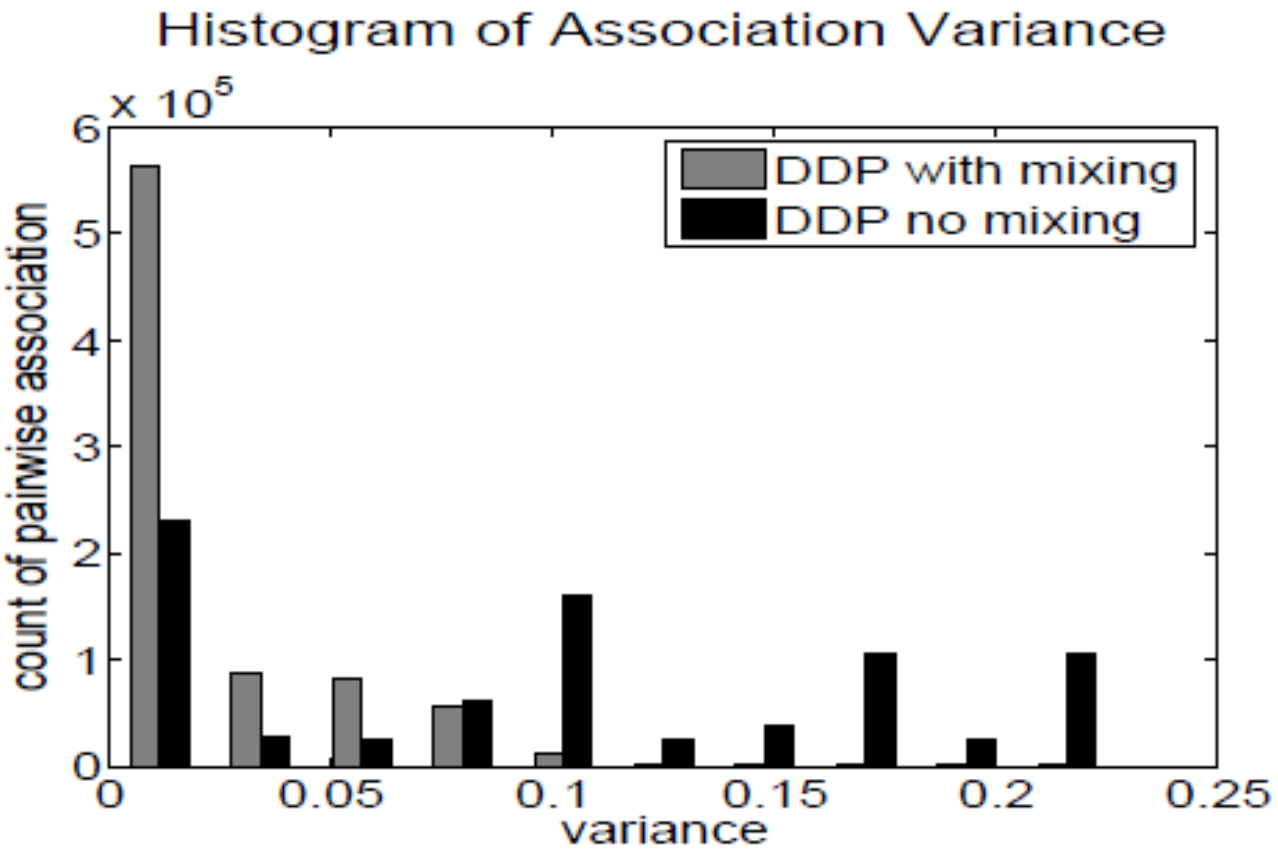


Figure 5: Histogram of the variance of association matrices computed across different sampling runs. The smaller variance of the DDP with mixing moves indicates that it consistently finds a good clustering and does not get stuck in local modes.

DDP Swapping Moves

- Why was the label-permute move not effective for the doppler data?
- How do the different time swap choices affect the results?