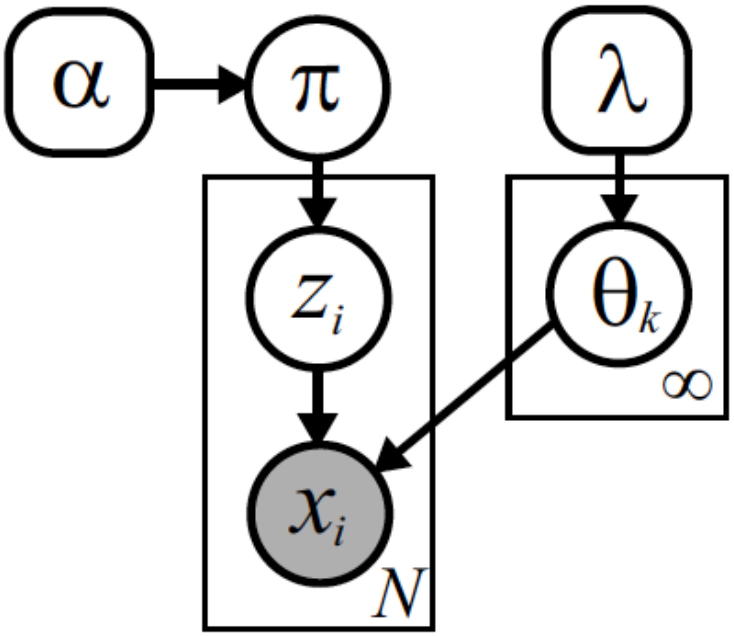
Applied Bayesian Nonparametrics

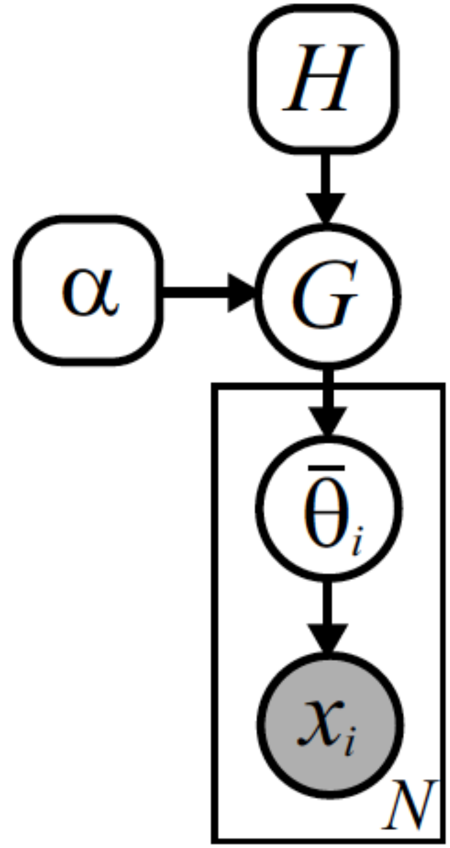
Special Topics in Machine Learning
Brown University CSCI 2950-P, Fall 2011

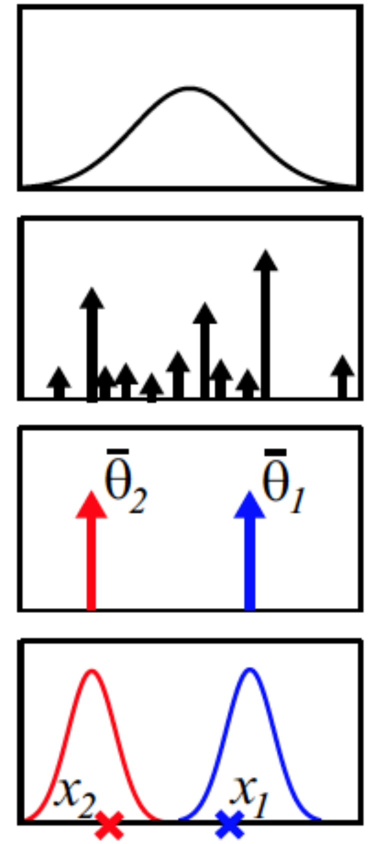
September 18: Exchangeability and the CRP, Infinite Mixtures of GP Experts

DP Mixture Models

$$p(x \mid \pi, \theta_1, \theta_2, \ldots) = \sum_{k=1}^{\infty} \pi_k f(x \mid \theta_k)$$







$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k)$$

$$\pi \sim \mathrm{GEM}(\alpha)$$

$$\theta_k \sim H(\lambda) \qquad k = 1, 2, \ldots$$

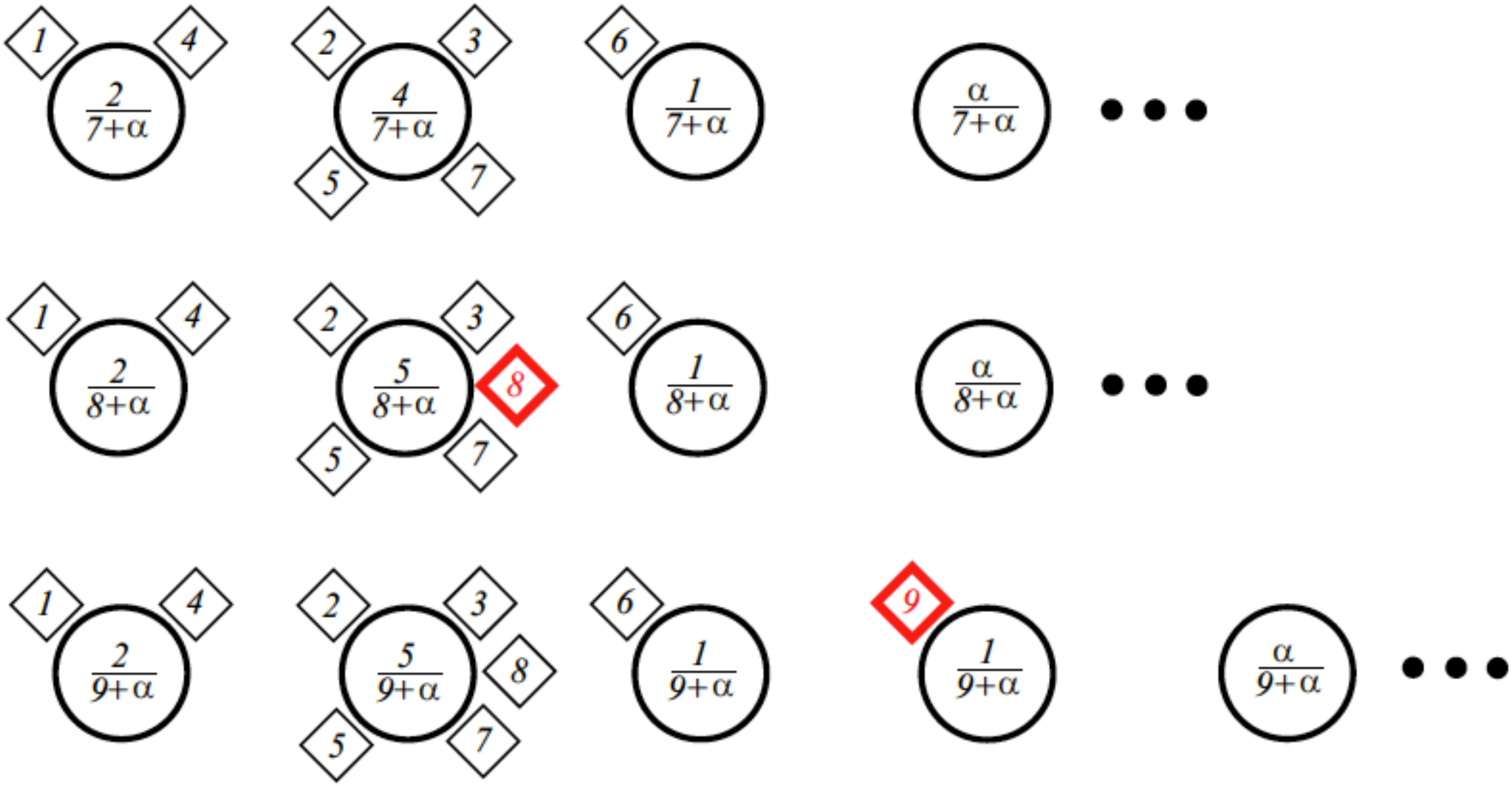
$$\bar{\theta}_i \sim G$$

$$x_i \sim F(\bar{\theta}_i)$$

$$z_i \sim \pi$$

$$x_i \sim F(\theta_{z_i})$$

Chinese Restaurant Process



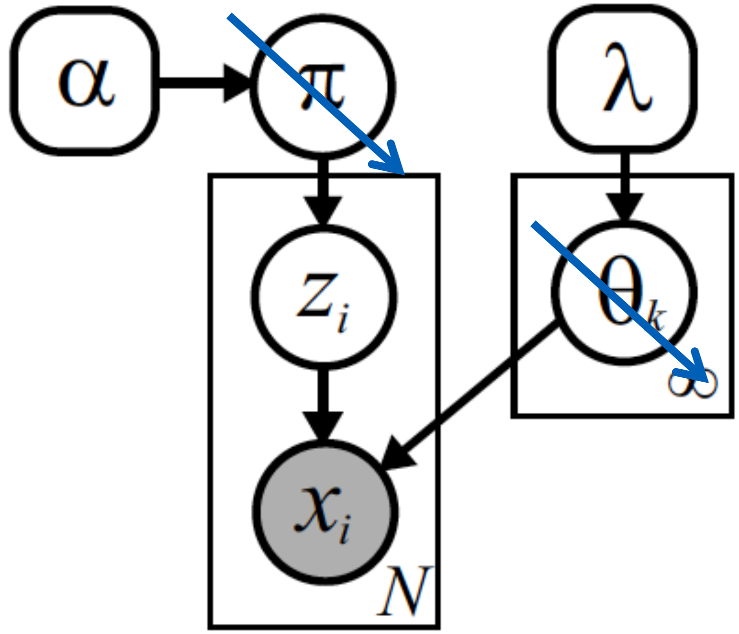
$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N} \left(\sum_{k=1}^{K} N_k \delta(z, k) + \alpha \delta(z, \bar{k}) \right)$$

DP Mixture: CRP Sampler

- Conceptually separates cluster allocations and parameters
- Marginalize cluster sizes to give Chinese restaurant process prior on data partitions

Exchangeability

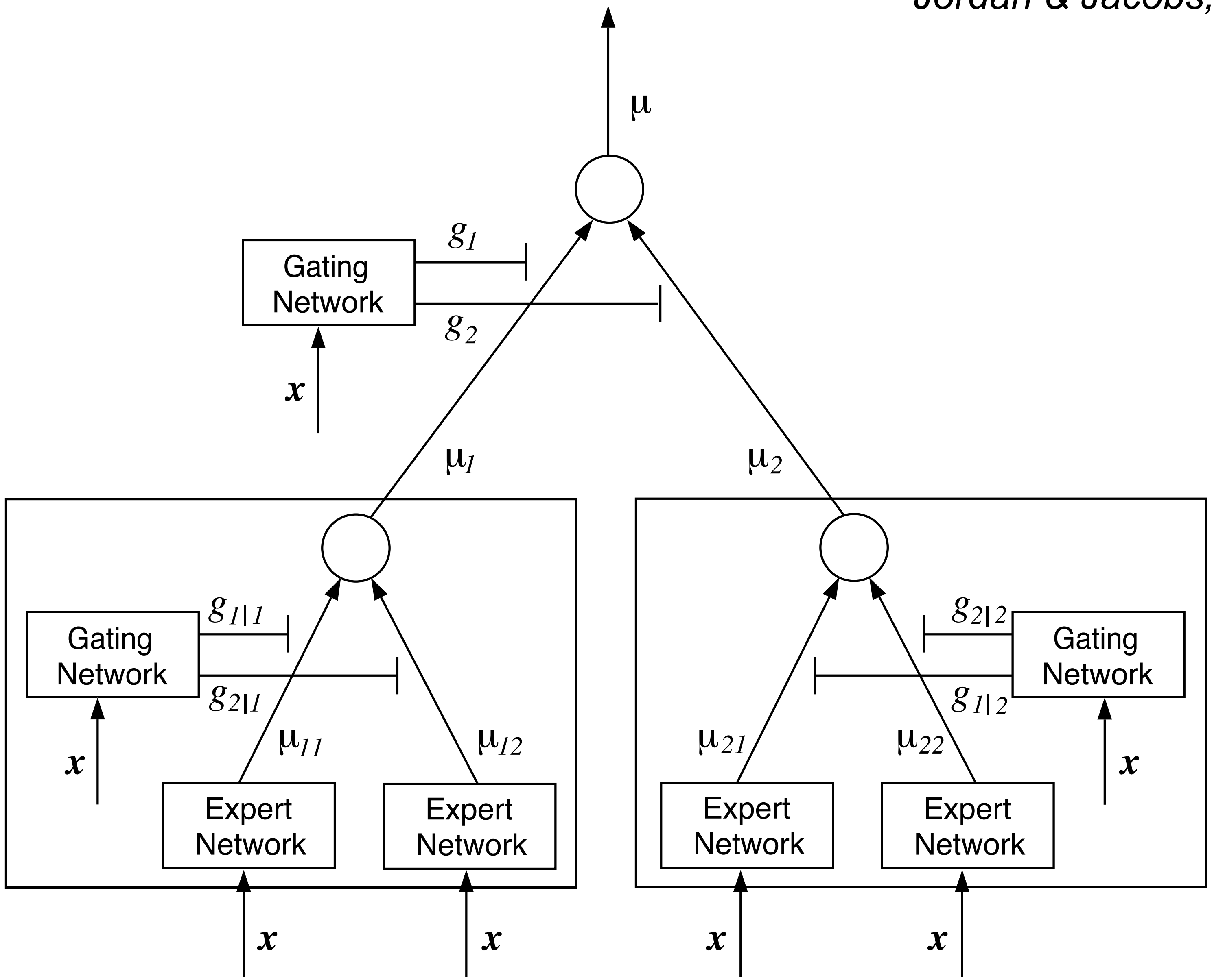
- Under CRP prior, all sequential data orderings give the same distribution on partitions
- Obvious from relationship to underlying DP sampling rule
- Convenient for Gibbs samplers: can think of each observation as the *last* when resampling



$$\pi \sim \mathrm{GEM}(\alpha)$$
$$\theta_k \sim H(\lambda) \qquad k = 1, 2, \ldots$$
$$z_i \sim \pi$$
$$x_i \sim F(\theta_{z_i})$$

Hierarchical Mixtures of Experts

Jordan & Jacobs, 1994



Infinite Mixture of GP Experts

Rasmussen & Williams, 2002

Standard DP Mixture of Gaussian Processes (GP correlations within clusters; expert/cluster assignments are not input-dependent)

Derive CRP Gibbs sampler conditional distributions

components where $n_{-i,j} > 0$: $$p(c_i = j | \mathbf{c}_{-i}, \alpha) = \frac{n_{-i,j}}{n-1+\alpha}$$

all other components combined: $$p(c_i \neq c_{i'} \text{ for all } i' \neq i | \mathbf{c}_{-i}, \alpha) = \frac{\alpha}{n-1+\alpha}$$

$$p(y_i | \mathbf{y}_{-i}, \mathbf{x}, \theta) \sim \mathcal{N}(\mu, \sigma^2), \qquad \begin{cases} \mu = Q(x_i, \mathbf{x})^\top Q^{-1} \mathbf{y}_{-i} \\ \sigma^2 = Q(x_i, x_i) - Q(x_i, \mathbf{x})^\top Q^{-1} Q(x_i, \mathbf{x}) \end{cases}$$

$$Q(x_i, x_{i'}) = v_0 \exp\left(-\frac{1}{2} \sum_d (x_{id} - x_{i'd})^2 / w_d^2\right) + v_1 \delta(i, i')$$

Infinite Mixture of GP Experts

Rasmussen & Williams, 2002

Standard DP Mixture of Gaussian Processes (GP correlations within clusters; expert/cluster assignments are not input-dependent)

Derive CRP Gibbs sampler conditional distributions

Replace by input-dependent pseudo-CRP conditionals

A sampler for what joint distribution??? This kernel-based prediction rule is in general not exchangeable

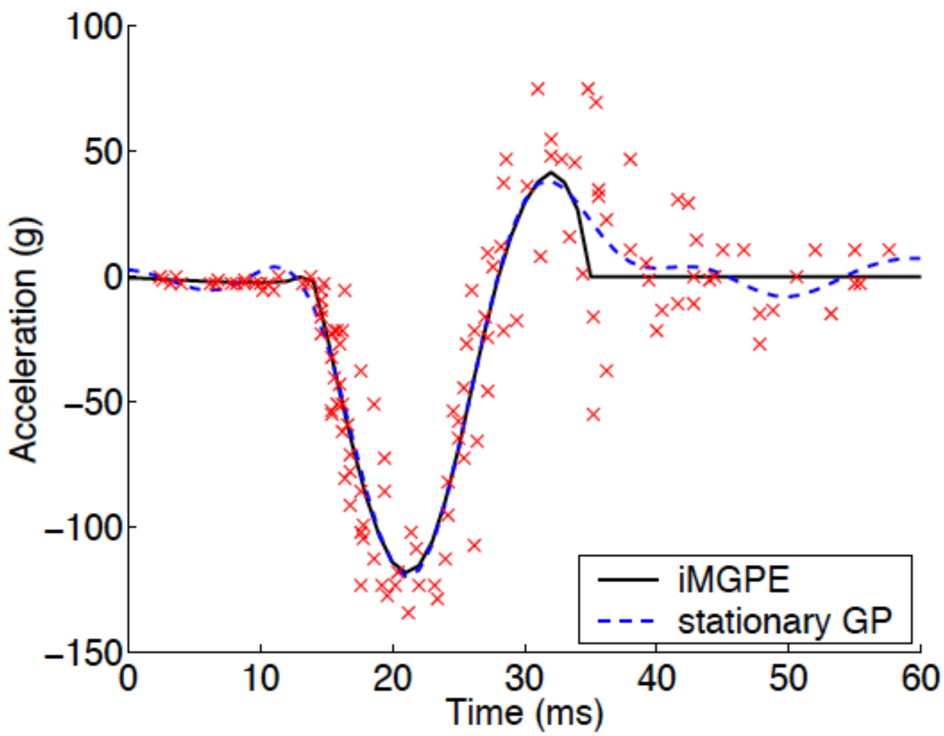
$$n_{-i,j} = (n-1)\frac{\sum_{i' \neq i} K_\phi(x_i, x_{i'})\delta(c_{i'}, j)}{\sum_{i' \neq i} K_\phi(x_i, x_{i'})}$$

$$K_\phi(x_i, x_{i'}) = \exp\big(-\frac{1}{2}\sum_d (x_{id} - x_{i'd})^2/\phi_d^2\big)$$

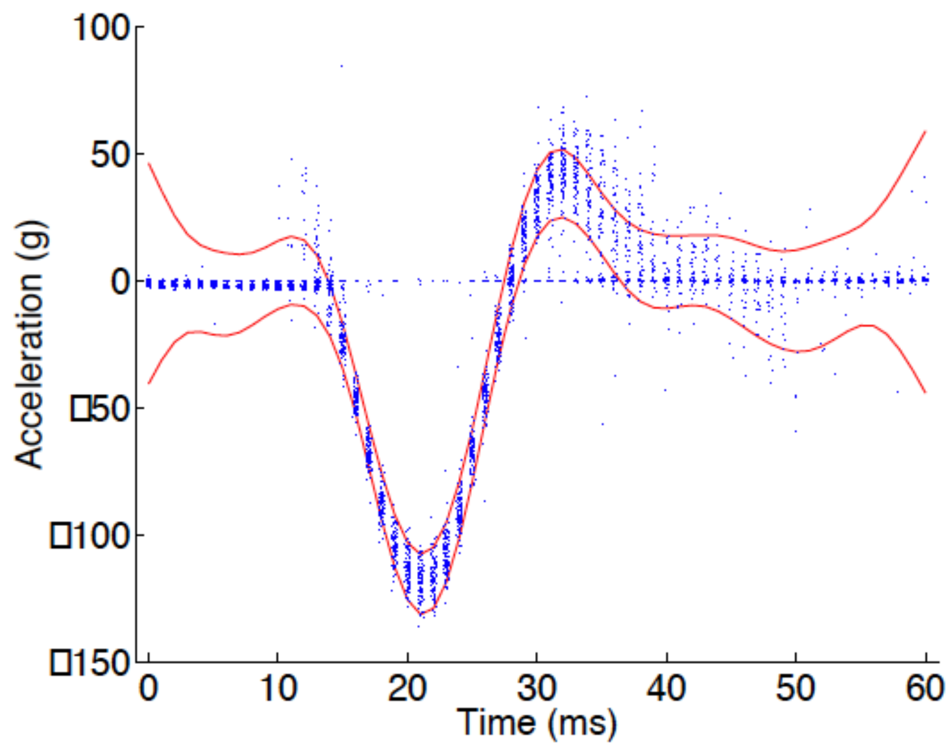
$$p(y_i|\mathbf{y}_{-i}, \mathbf{x}, \theta) \sim \mathcal{N}(\mu, \sigma^2), \qquad \begin{cases} \mu = Q(x_i, \mathbf{x})^\top Q^{-1}\mathbf{y}_{-i} \\ \sigma^2 = Q(x_i, x_i) - Q(x_i, \mathbf{x})^\top Q^{-1} Q(x_i, \mathbf{x}) \end{cases}$$

$$Q(x_i, x_{i'}) = v_0 \exp\big(-\frac{1}{2}\sum_d (x_{id} - x_{i'd})^2/w_d^2\big) + v_1\delta(i, i')$$

Motorcycle Data: Predictions

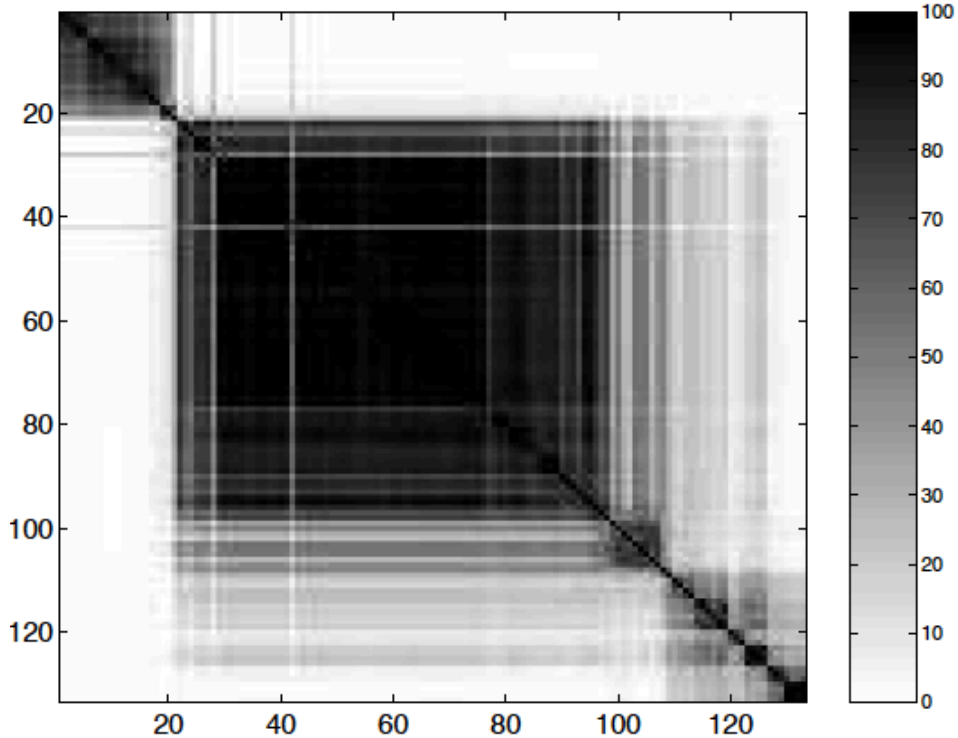


Data, Mean of Stationary GP,
Median of DP mixture of GPs

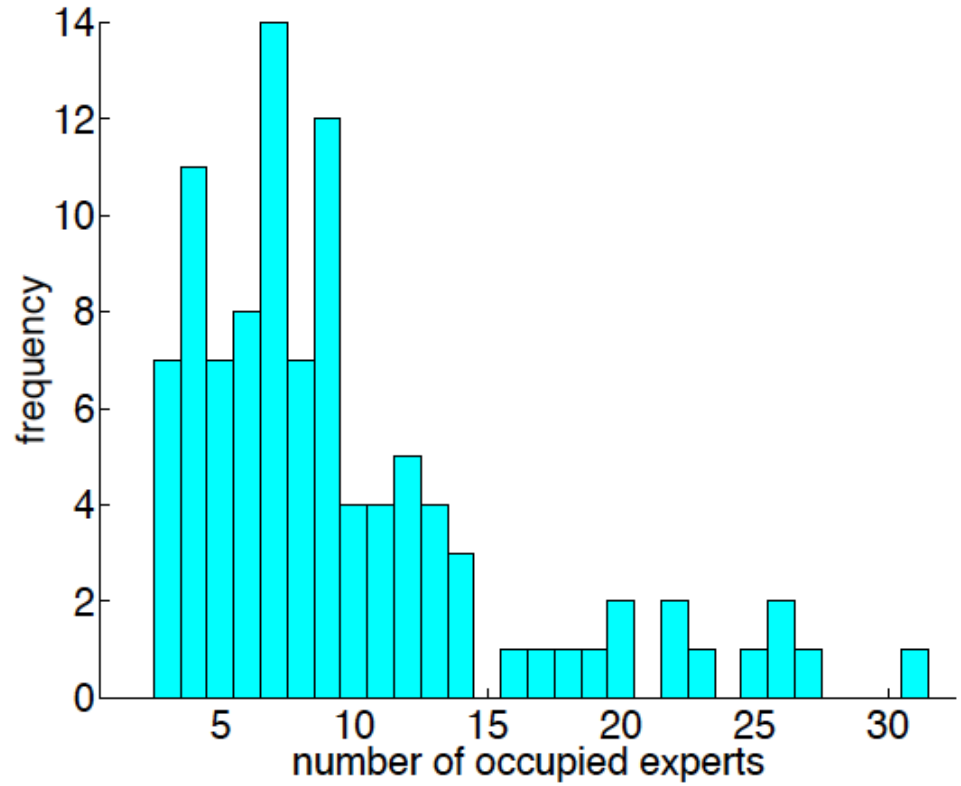


Confidence intervals for GP,
Predictive samples for
DP mixture of GPs

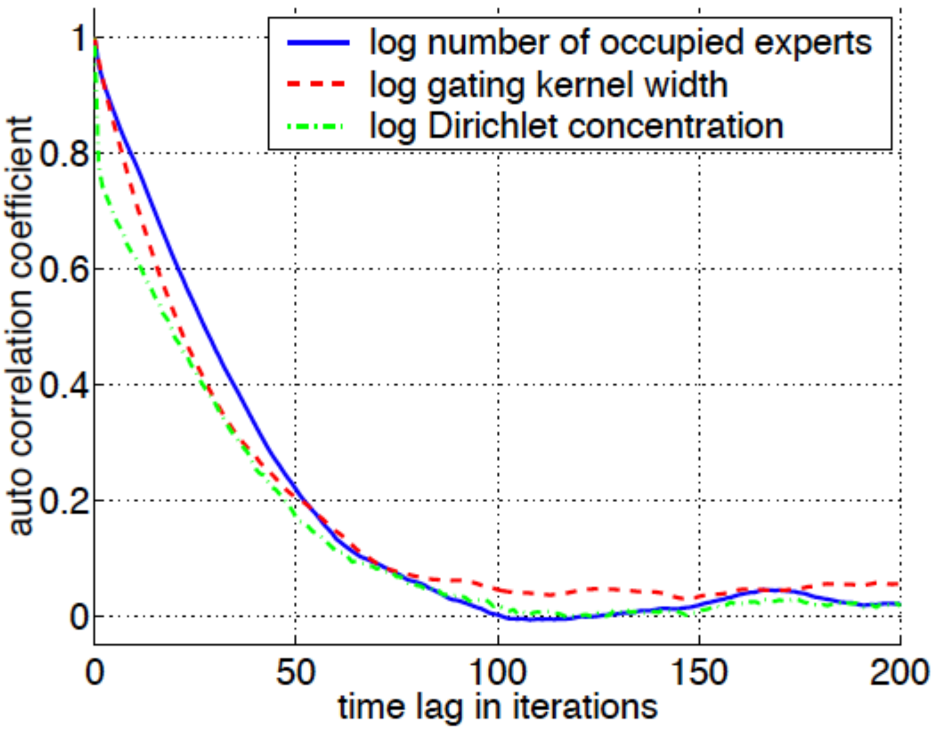
Motorcycle Data: Clustering

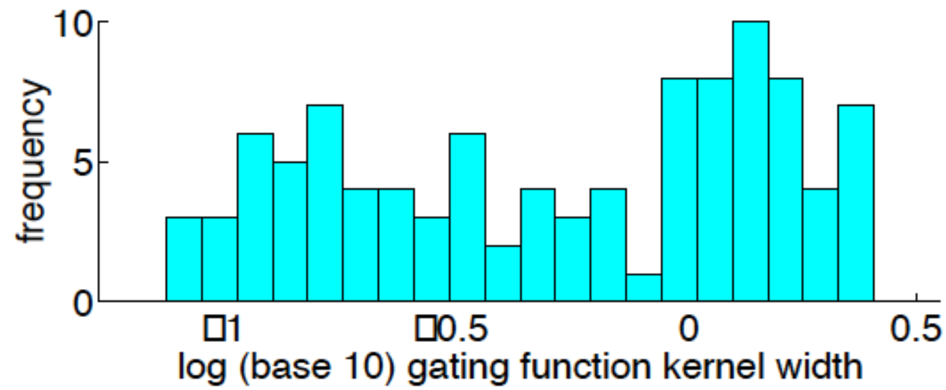


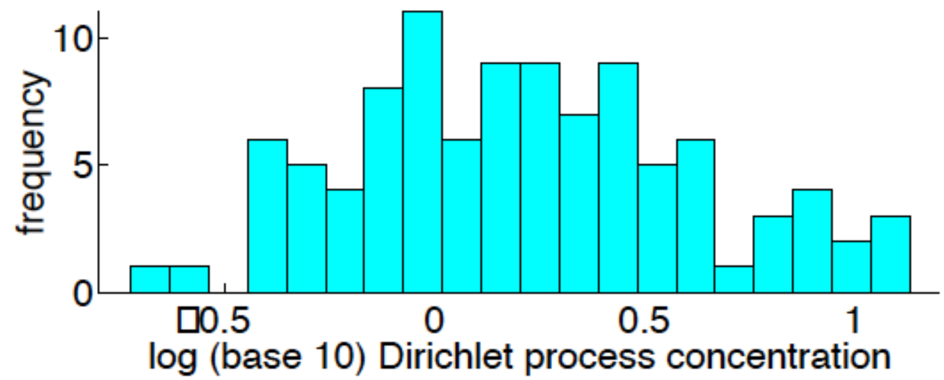
Probability that observation pairs are assigned to the same expert (avoids label switching problems)



Motorcycle Data: Mixing







To What Equilibrium Distribution???

For most kernels the Markov chain
will be irreducible and aperiodic, so…

Fixing the Mixture of GP Experts

components where $n_{-i,j} > 0$:	$p(c_i = j \mid \mathbf{c}_{-i}, \alpha)$	$=$	$\dfrac{n_{-i,j}}{n-1+\alpha}$
all other components combined:	$p(c_i \neq c_{i'} \text{ for all } i' \neq i \mid \mathbf{c}_{-i}, \alpha)$	$=$	$\dfrac{\alpha}{n-1+\alpha}$

$$n_{-i,j} = (n-1)\frac{\sum_{i' \neq i} K_\phi(x_i, x_{i'})\delta(c_{i'}, j)}{\sum_{i' \neq i} K_\phi(x_i, x_{i'})} \qquad K_\phi(x_i, x_{i'}) = \exp\left(-\frac{1}{2}\sum_d (x_{id} - x_{i'd})^2/\phi_d^2\right)$$

1. Treat kernel-dependent prediction rule as defining a true joint distribution, and derive the corresponding sampler
   - Each choice of data ordering yields a different model
   - Resampling variables in the "middle" of the order may be computationally difficult, due to later observations
2. Model assignments to inputs via a joint distribution
   - Meeds & Osindero 2006, Alternative Infinite Mixture GPs
3. Create input-indexed random measures (dependent DP)
4. Define a local similarity-based way of partitioning observations which retains simple conditional distributions
   - Blei & Frazier 2011, Distance Dependent CRP