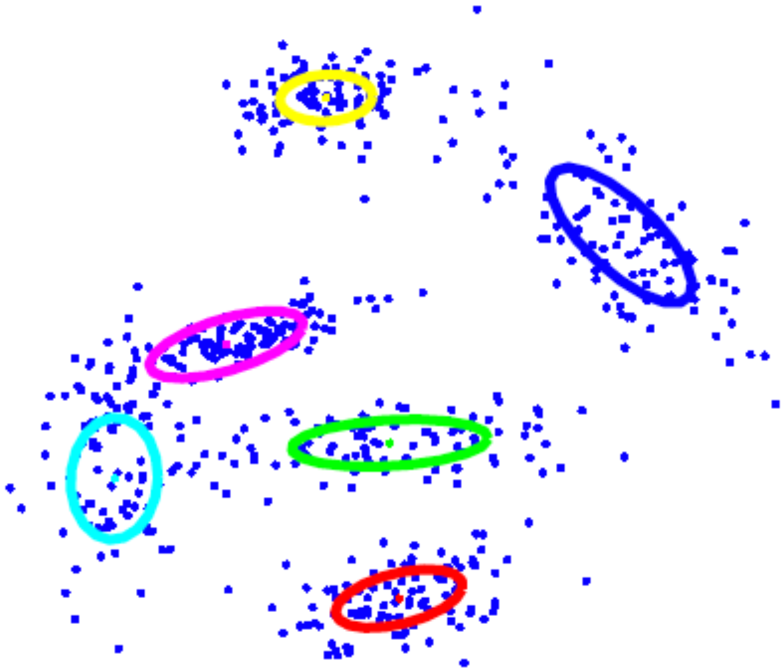
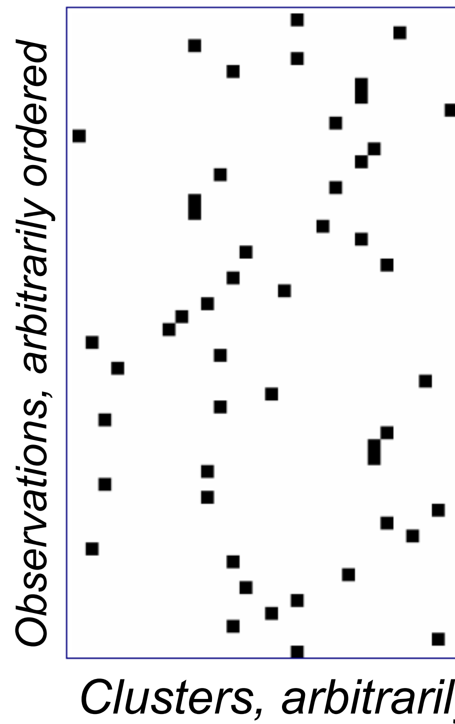
Applied Bayesian Nonparametrics

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

December 6: Course Review & Outlook

Nonparametric Clustering

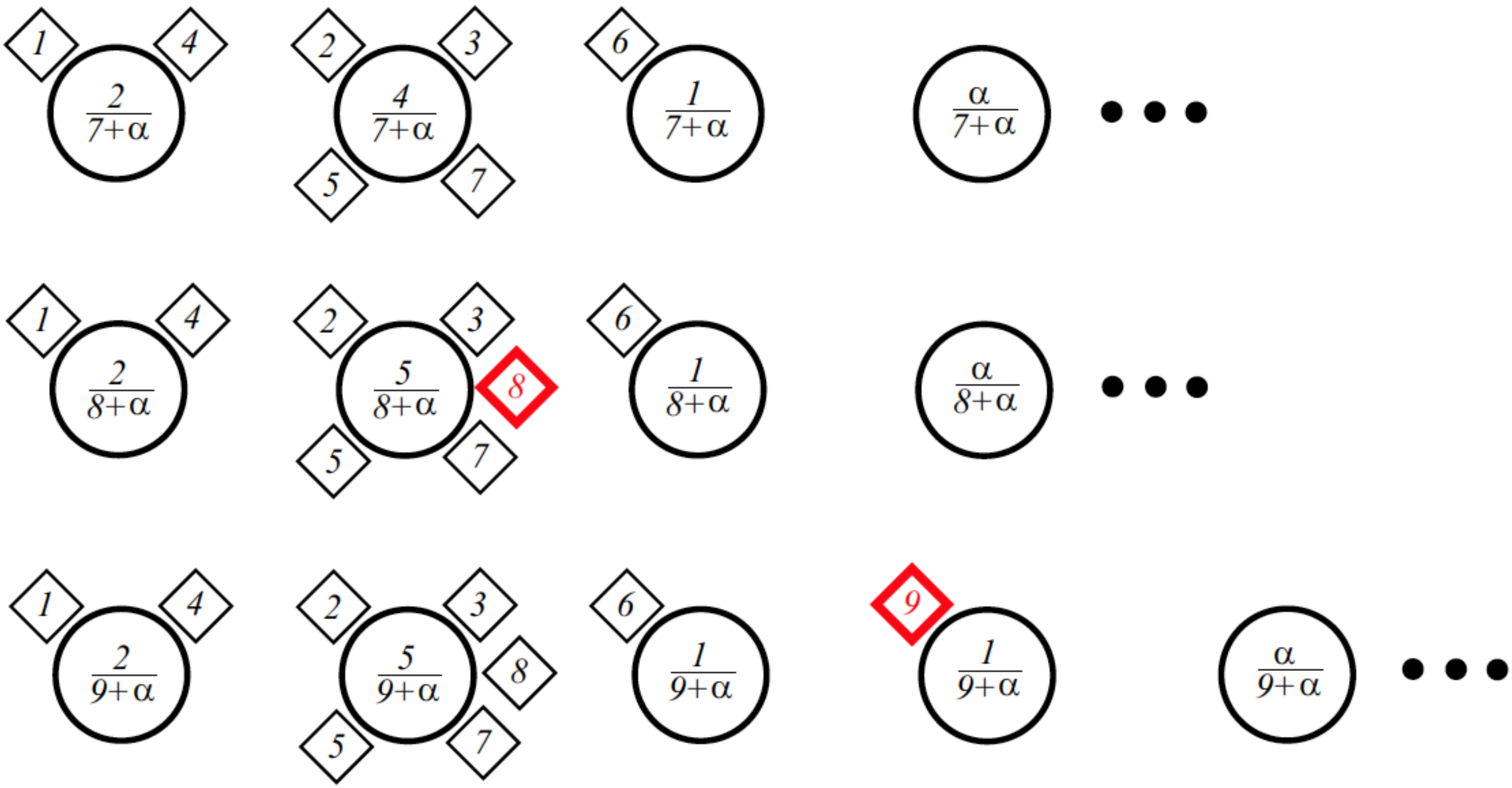




Ghahramani, BNP 2009

- *Large Support:* All partitions of the data, from one giant cluster to N singletons, have positive probability under prior
- *Exchangeable:* Partition probabilities are invariant to permutations of the data
- *Desirable:* Good asymptotics, computational tractability, flexibility and ease of generalization…

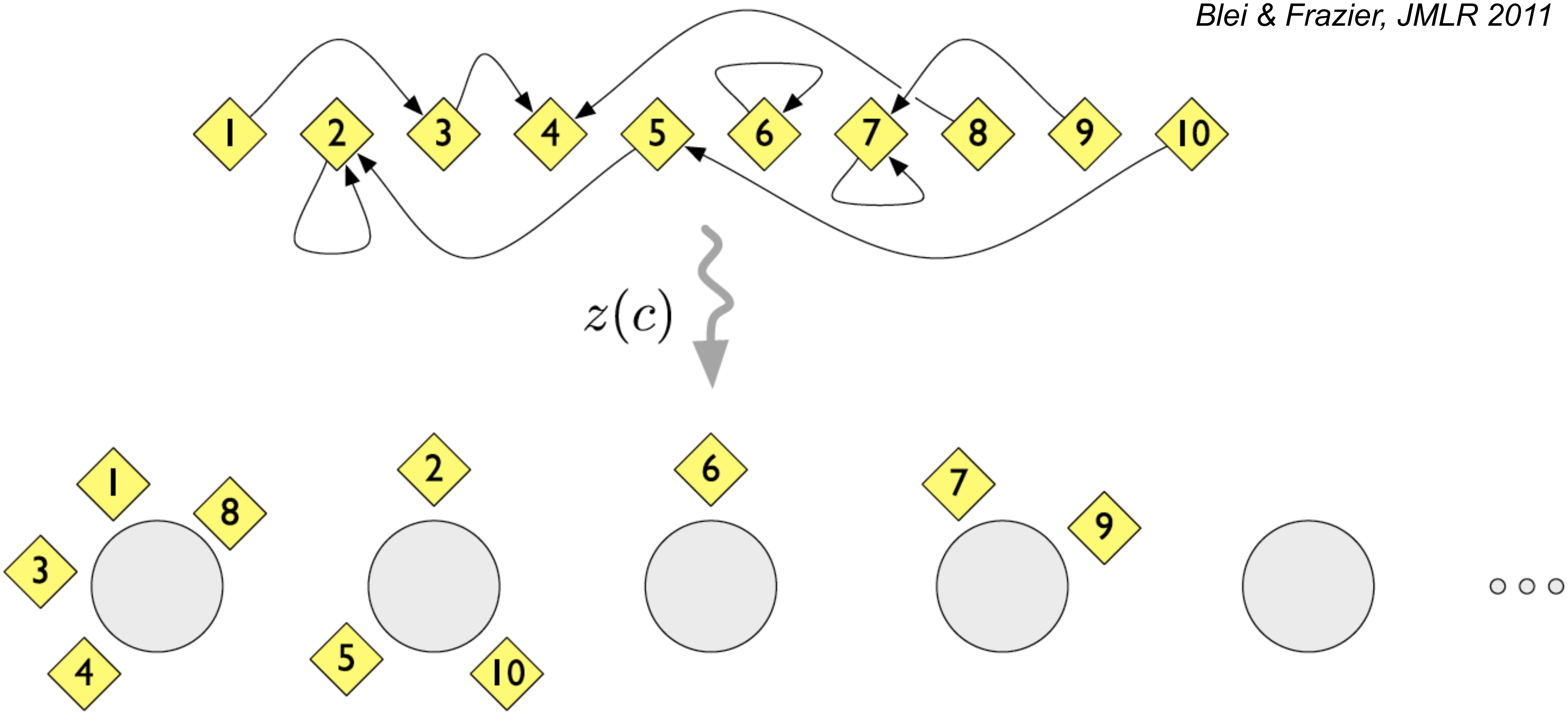
Chinese Restaurant Process (CRP)



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

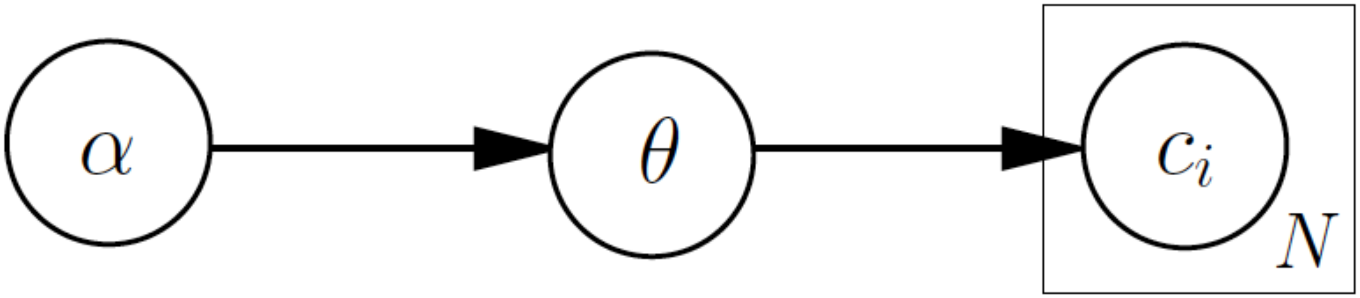
Distance Dependent CRP

Blei & Frazier, JMLR 2011



- *Good:* Simple, computationally easy generalization which can make clustering depend on any feature: time, space, …
- *Tricky:* Relationship between local distance and global clustering behavior hard to analyze (no marginal invariance)

Finite Dirichlet Mixtures



$$p(\theta) = \frac{\prod_{k=1}^{K} \theta_k^{\alpha_k - 1}}{D(\alpha_1, \alpha_2, \ldots, \alpha_K)} \qquad P(\mathbf{c}|\theta) = \prod_{i=1}^{N} P(c_i|\theta) = \prod_{i=1}^{N} \theta_{c_i}$$

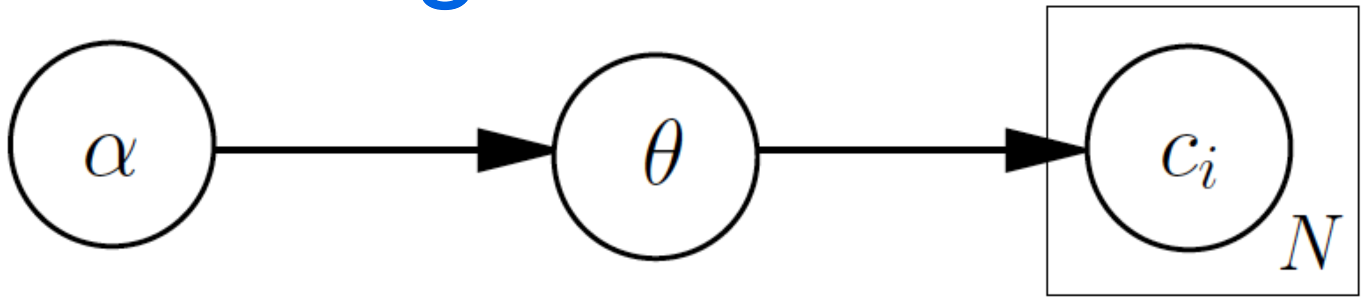
$$D(\alpha_1, \alpha_2, \ldots, \alpha_K) = \int_{\Delta_K} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\, d\theta = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$$

$$D(\tfrac{\alpha}{K}, \tfrac{\alpha}{K}, \ldots, \tfrac{\alpha}{K}) = \frac{\Gamma(\frac{\alpha}{K})^K}{\Gamma(\alpha)}$$

$$P(\mathbf{c}) = \int_{\Delta_K} \prod_{i=1}^{n} P(c_i|\theta) p(\theta)\, d\theta = \frac{\prod_{k=1}^{K} \Gamma(m_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K} \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)}$$

Marginal likelihoods generally expressed as ratios of normalizers

From Assignments to Partitions



$$P(\mathbf{c}) = \int_{\Delta_K} \prod_{i=1}^{n} P(c_i|\theta)p(\theta)\,d\theta = \frac{\prod_{k=1}^{K}\Gamma(m_k+\frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K}\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$$

$$= \left(\frac{\alpha}{K}\right)^{K_+}\left(\prod_{k=1}^{K^+}\prod_{j=1}^{m_k-1}(j+\tfrac{\alpha}{K})\right)\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$$

K_+ is the number of classes for which $m_k > 0$

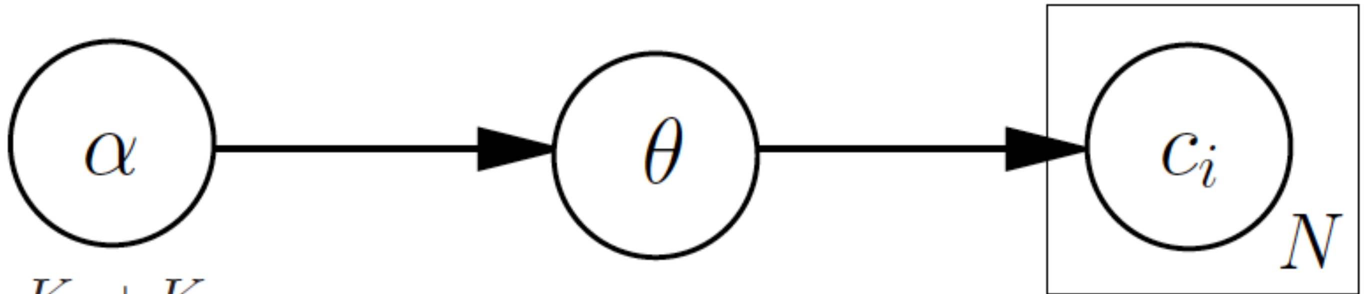
There are K^N possible values for $\mathbf{c}$

$P(\mathbf{c}) \rightarrow 0$ as $K \rightarrow \infty$

Instead look at label equivalence classes: $K = K_0 + K_+$

$$P([\mathbf{c}]) = \sum_{\mathbf{c}\in[\mathbf{c}]} P(\mathbf{c}) = \frac{K!}{K_0!}\left(\frac{\alpha}{K}\right)^{K_+}\left(\prod_{k=1}^{K^+}\prod_{j=1}^{m_k-1}(j+\tfrac{\alpha}{K})\right)\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$$

An Infinite Limit



$$K = K_0 + K_+$$

$$P([\mathbf{c}]) = \sum_{\mathbf{c}\in[\mathbf{c}]} P(\mathbf{c}) = \frac{K!}{K_0!}\left(\frac{\alpha}{K}\right)^{K_+}\left(\prod_{k=1}^{K^+}\prod_{j=1}^{m_k-1}(j+\tfrac{\alpha}{K})\right)\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$$

$$\lim_{K\to\infty} \alpha^{K_+}\cdot\frac{K!}{K_0!\,K^{K_+}}\cdot\left(\prod_{k=1}^{K_+}\prod_{j=1}^{m_k-1}(j+\tfrac{\alpha}{K})\right)\cdot\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$$

$$= \alpha^{K_+}\cdot \quad 1 \quad \cdot\left(\prod_{k=1}^{K_+}(m_k-1)!\right)\cdot\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$$

- *Good:* Recover the CRP as an infinite limit of a standard, widely studied parametric model.
- *Tricky:* Dealing with equivalence classes in the infinite limit, which becomes harder for more complex models

De Finetti's Theorem

- Finitely exchangeable random variables satisfy:

$$p(x_1, \ldots, x_N) = p\big(x_{\tau(1)}, \ldots, x_{\tau(N)}\big) \qquad \text{for any permutation } \tau(\cdot)$$

- A sequence is infinitely exchangeable if every finite subsequence is exchangeable
- Exchangeable variables need not be independent, but always have a representation with conditional independencies:

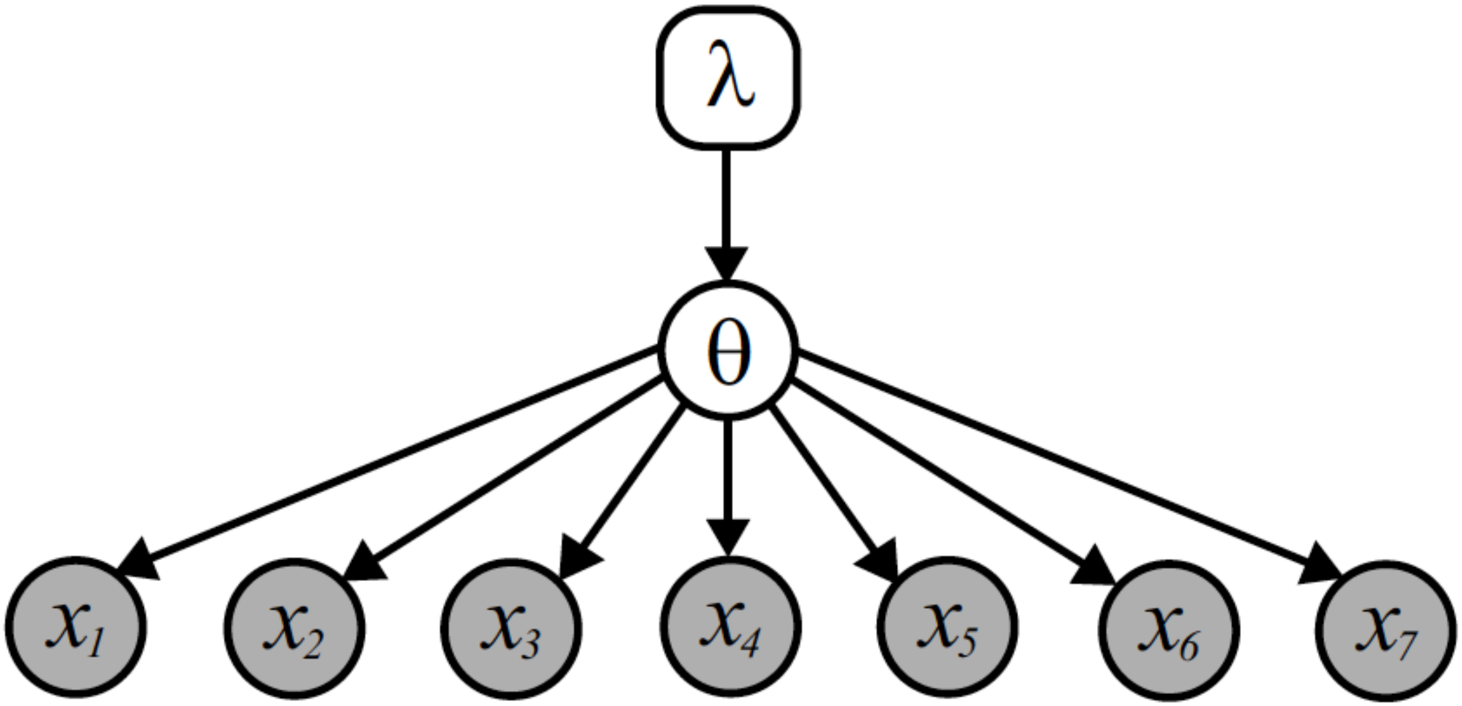
Theorem 2.2.2 (De Finetti). *For any infinitely exchangeable sequence of random variables* $\{x_i\}_{i=1}^{\infty}$, $x_i \in \mathcal{X}$, *there exists some space* Θ, *and corresponding density* $p(\theta)$, *such that the joint probability of any* N *observations has a mixture representation:*

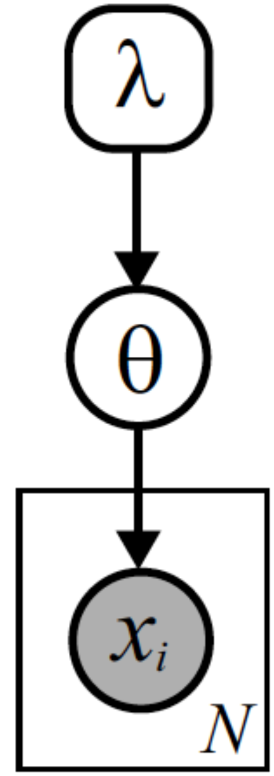
$$p(x_1, x_2, \ldots, x_N) = \int_{\Theta} p(\theta) \prod_{i=1}^{N} p(x_i \mid \theta)\, d\theta \tag{2.77}$$

When $\mathcal{X}$ *is a* K*–dimensional discrete space,* Θ *may be chosen as the* $(K-1)$*–simplex. For Euclidean* $\mathcal{X}$, Θ *is an infinite–dimensional space of probability measures.*

An explicit construction is useful in hierarchical modeling…

De Finetti's Directed Graph

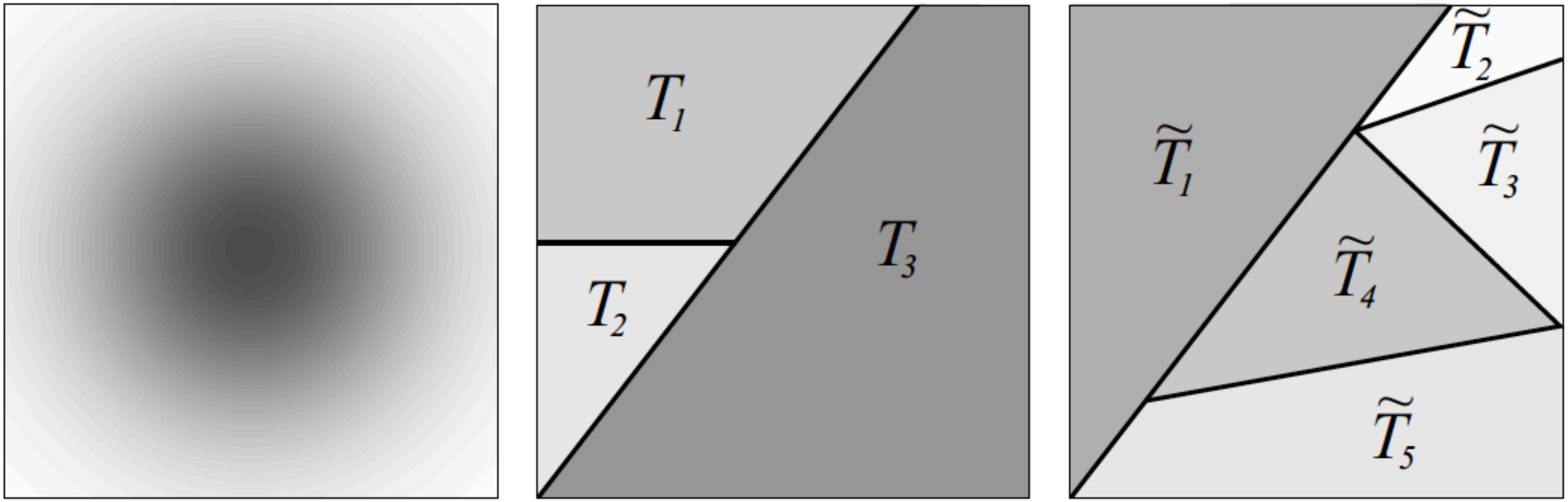




$$p(x_1, \ldots, x_N, \theta \mid \lambda) = p(\theta \mid \lambda) \prod_{i=1}^{N} p(x_i \mid \theta)$$

What distribution underlies the infinitely exchangeable CRP?

Dirichlet Processes



$$\mathbb{E}[G(T)] = H(T) \qquad\qquad G \sim \mathrm{DP}(\alpha, H)$$

For any finite partition

$$\bigcup_{k=1}^{K} T_k = \Theta \qquad\qquad T_k \cap T_\ell = \emptyset \qquad k \neq \ell$$

the distribution of the measure of those cells is Dirichlet:

$$(G(T_1), \ldots, G(T_K)) \sim \mathrm{Dir}(\alpha H(T_1), \ldots, \alpha H(T_K))$$

DP Posteriors and Conjugacy

Proposition 2.5.1. *Let $G \sim \mathrm{DP}(\alpha, H)$ be a random measure distributed according to a Dirichlet process. Given N independent observations $\bar{\theta}_i \sim G$, the posterior measure also follows a Dirichlet process:*

$$p(G \mid \bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H) = \mathrm{DP}\left(\alpha + N, \frac{1}{\alpha + N}\left(\alpha H + \sum_{i=1}^{N} \delta_{\bar{\theta}_i}\right)\right) \qquad (2.169)$$

Proof Hint: For any finite partition, we have

$$p\big((G(T_1), \ldots, G(T_K)) \mid \bar{\theta} \in T_k\big) = \mathrm{Dir}(\alpha H(T_1), \ldots, \alpha H(T_k) + 1, \ldots, \alpha H(T_K))$$

An observation must be of one of the countably infinite atoms which compose the random Dirichlet measure

DPs and Polya Urns

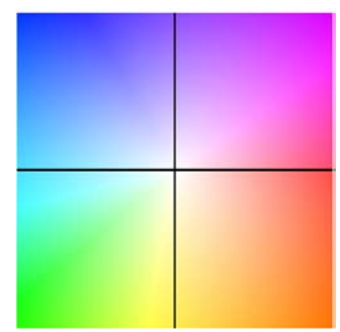
Theorem 2.5.4. *Let $G \sim \mathrm{DP}(\alpha, H)$ be distributed according to a Dirichlet process, where the base measure H has corresponding density $h(\theta)$. Consider a set of N observations $\bar{\theta}_i \sim G$ taking K distinct values $\{\theta_k\}_{k=1}^{K}$. The predictive distribution of the next observation then equals*

$$p(\bar{\theta}_{N+1} = \theta \mid \bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H) = \frac{1}{\alpha + N} \left(\alpha h(\theta) + \sum_{k=1}^{K} N_k \delta(\theta, \theta_k) \right) \qquad (2.180)$$

where N_k is the number of previous observations of θ_k, as in eq. (2.179).

My variation on the classical balls in urns analogy:

- Consider an urn containing α pounds of very tiny, colored sand (the space of possible colors is Θ)
- Take out one grain of sand, record its color as $\bar{\theta}_1$
- Put that grain back, add 1 extra pound of that color
- Repeat this process…



DPs are Neutral: "Almost" independent

The distribution of a random probability measure G is neutral with respect to a finite partition $(T_1, \ldots, T_K)$ *iff*

$$G(T_k) \qquad \text{is independent of} \qquad \left\{ \frac{G(T_\ell)}{1 - G(T_k)} \,\middle|\, \ell \neq k \right\}$$

$$\text{given that } G(T_k) < 1.$$

Theorem 2.5.2. *Consider a distribution $\mathcal{P}$ on probability measures G for some space Θ. Assume that $\mathcal{P}$ assigns positive probability to more than one measure G, and that with probability one samples $G \sim \mathcal{P}$ assign positive measure to at least three distinct points $\theta \in \Theta$. The following conditions are then equivalent:*

(i) $\mathcal{P} = \mathrm{DP}(\alpha, H)$ *is a Dirichlet process for some base measure H on Θ.*

(ii) $\mathcal{P}$ *is neutral with respect to every finite, measurable partition of Θ.*

(iii) For every measurable $T \subset \Theta$, and any N observations $\bar{\theta}_i \sim G$, the posterior distribution $p\big(G(T) \mid \bar{\theta}_1, \ldots, \bar{\theta}_N\big)$ depends only on the number of observations that fall within T (and not their particular locations).

The Stick-Breaking Construction: DP Realizations are Discrete

Theorem 2.5.3. *Let $\pi = \{\pi_k\}_{k=1}^{\infty}$ be an infinite sequence of mixture weights derived from the following stick–breaking process, with parameter $\alpha > 0$:*

$$\beta_k \sim \mathrm{Beta}(1, \alpha) \qquad k = 1, 2, \ldots \tag{2.174}$$

$$\pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell) = \beta_k \left(1 - \sum_{\ell=1}^{k-1} \pi_\ell\right) \tag{2.175}$$

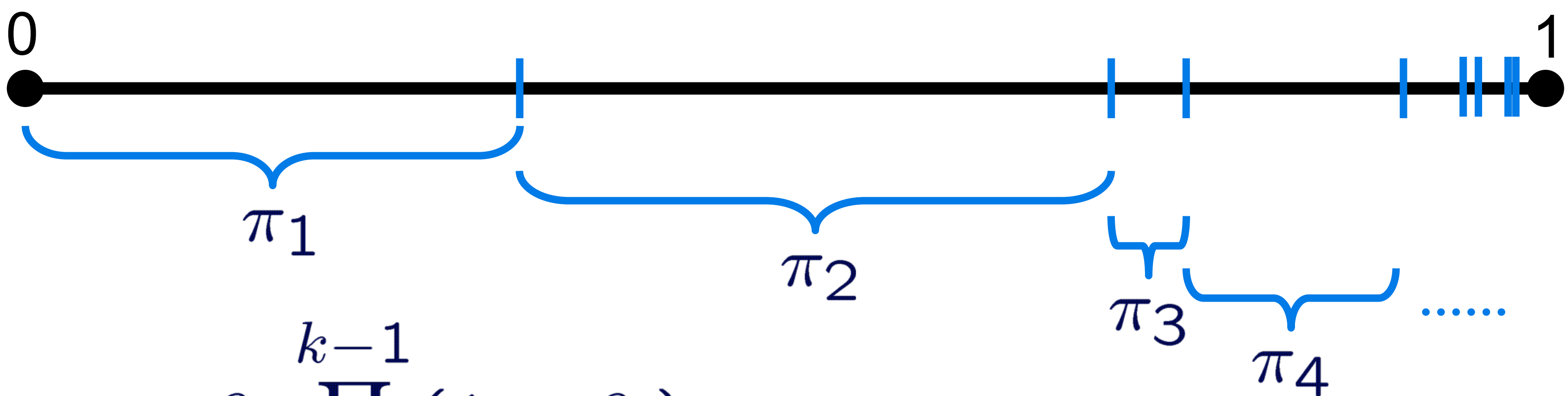
Given a base measure H on Θ, consider the following discrete random measure:

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k) \qquad \theta_k \sim H \tag{2.176}$$

This construction guarantees that $G \sim \mathrm{DP}(\alpha, H)$. Conversely, samples from a Dirichlet process are discrete with probability one, and have a representation as in eq. (2.176).

DP Stick-Breaking Construction

$$p(x) = \sum_{k=1}^{\infty} \pi_k f(x \mid \theta_k)$$



$$\pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell)$$

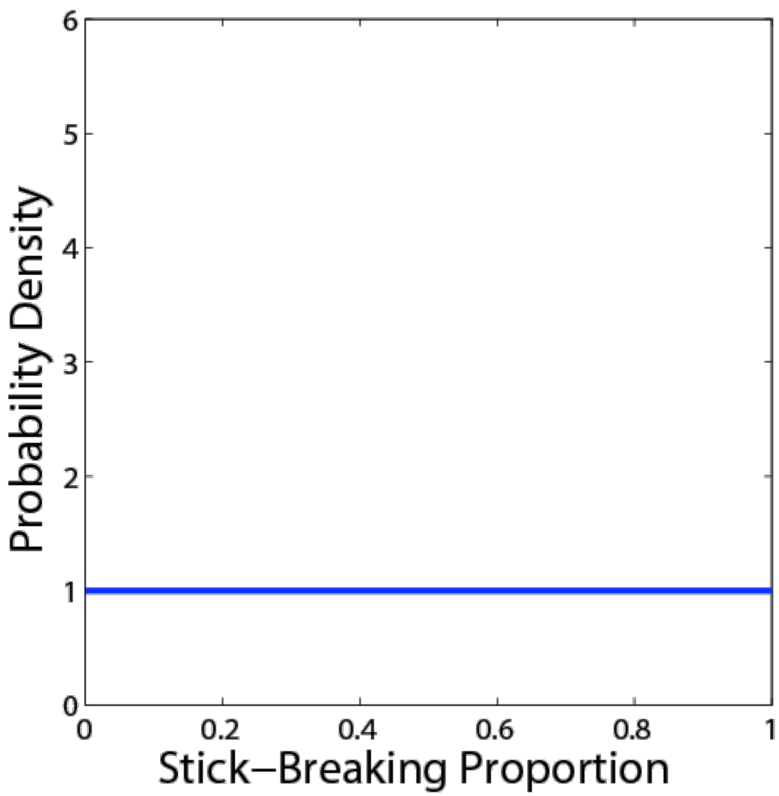
$$\beta_k \sim \text{Beta}(1, \alpha)$$

$\alpha \longrightarrow$ concentration parameter

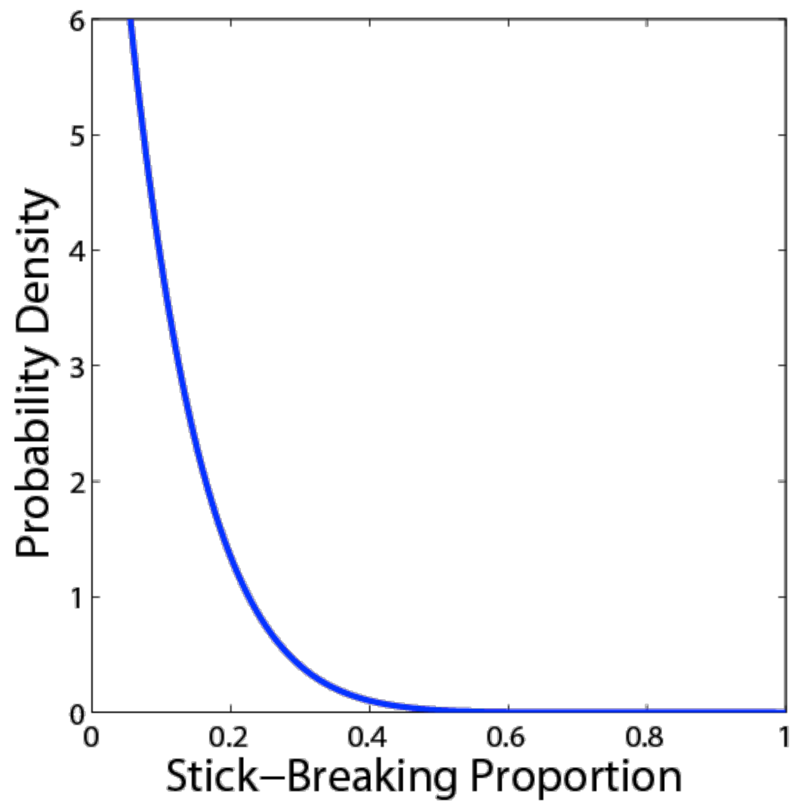
Stick-Breaking Construction: *Sethuraman, 1994*

Dirichlet Stick-Breaking

$$v_k \sim \text{Beta}(1, \alpha) \qquad E[v_k] = \frac{1}{1+\alpha}$$



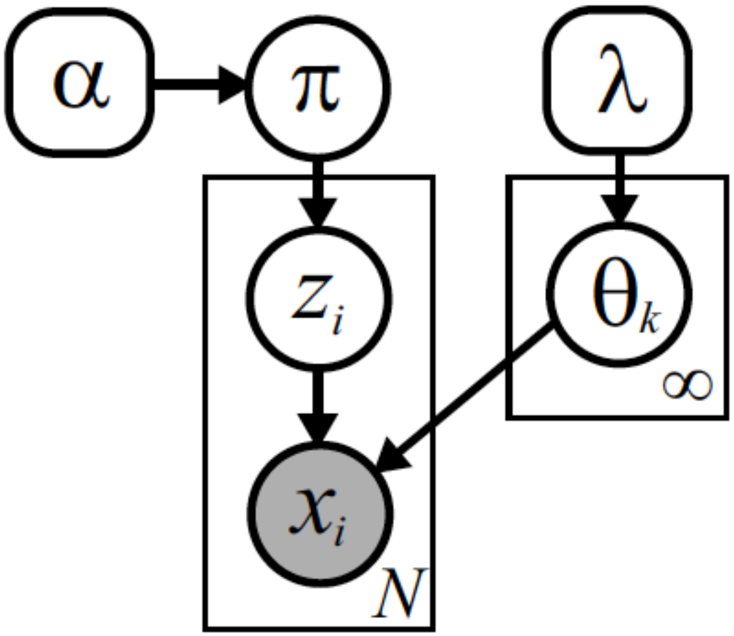
$\alpha = 1$

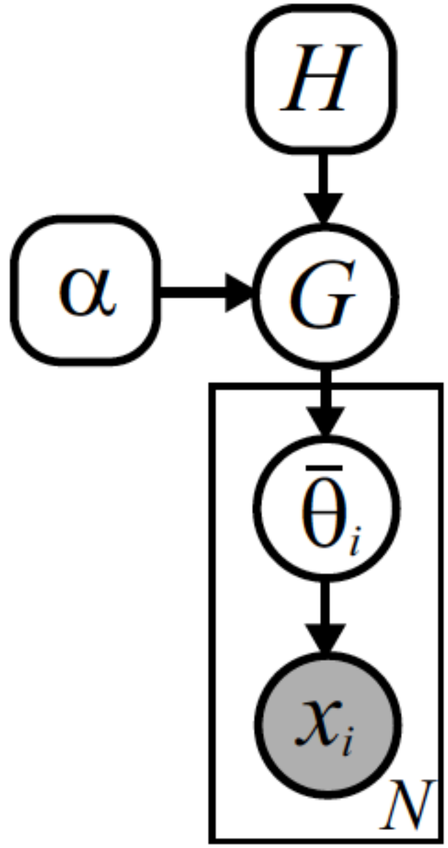


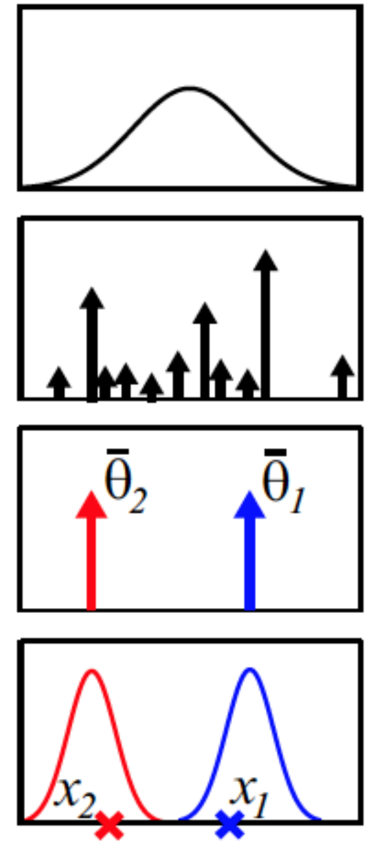
$\alpha = 10$

DP Mixture Models

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







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

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

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

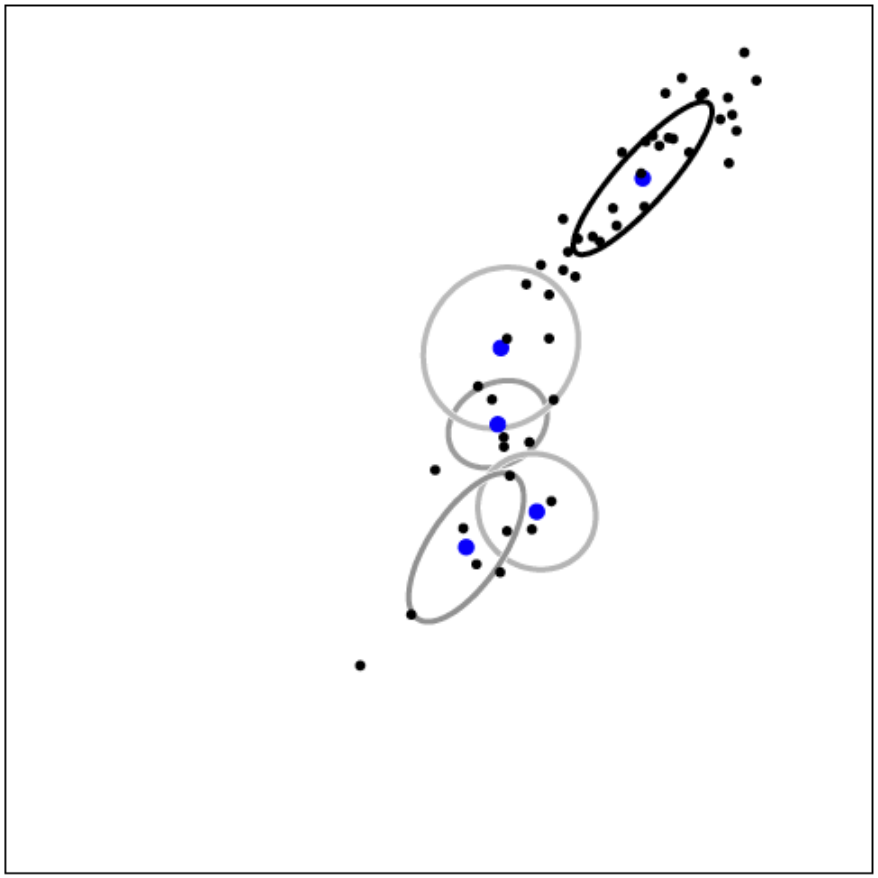
$$\pi \sim \mathrm{GEM}(\alpha)$$

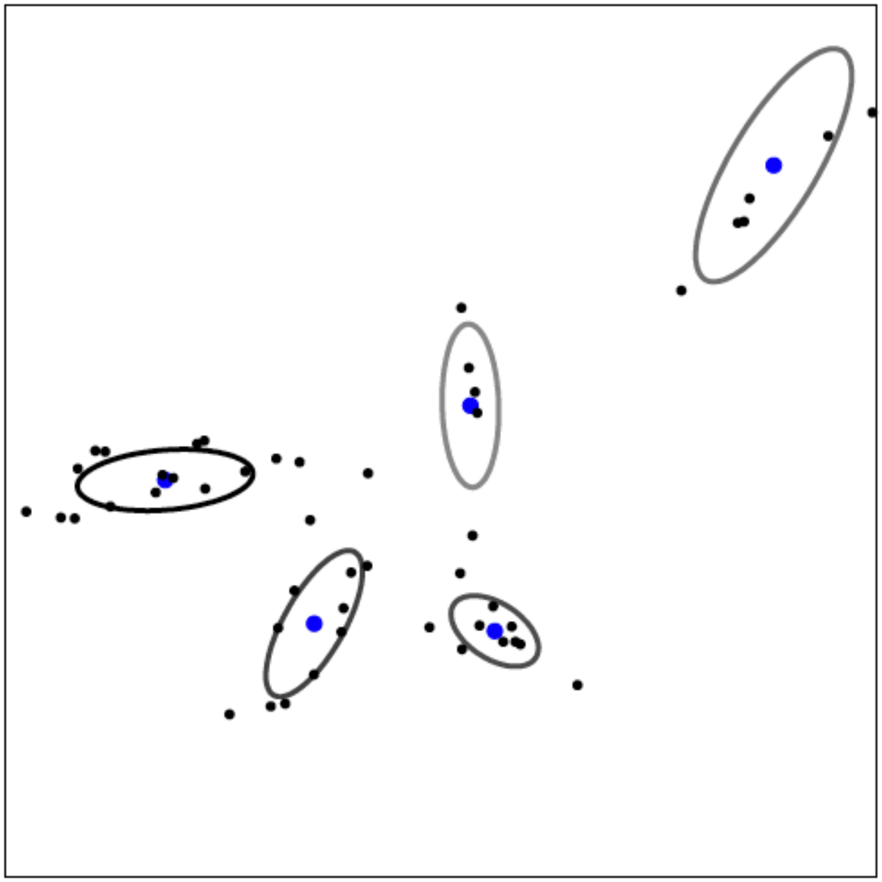
$$\theta_k \sim H(\lambda) \qquad k = 1, 2, \ldots$$

$$z_i \sim \pi$$

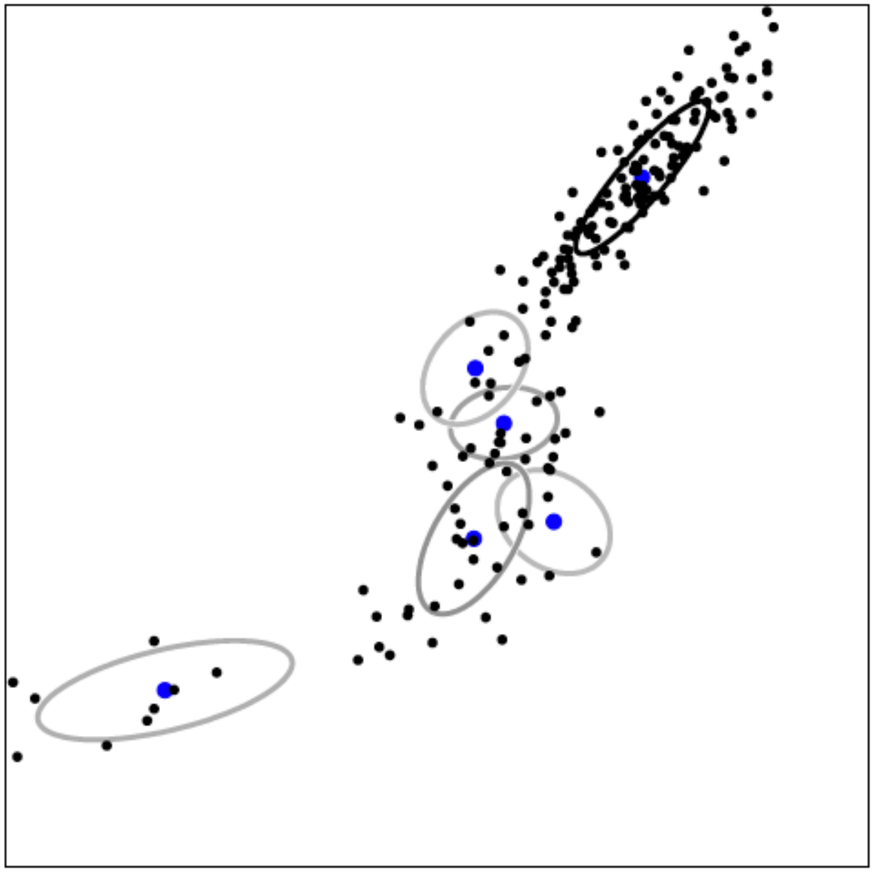
$$x_i \sim F(\theta_{z_i})$$

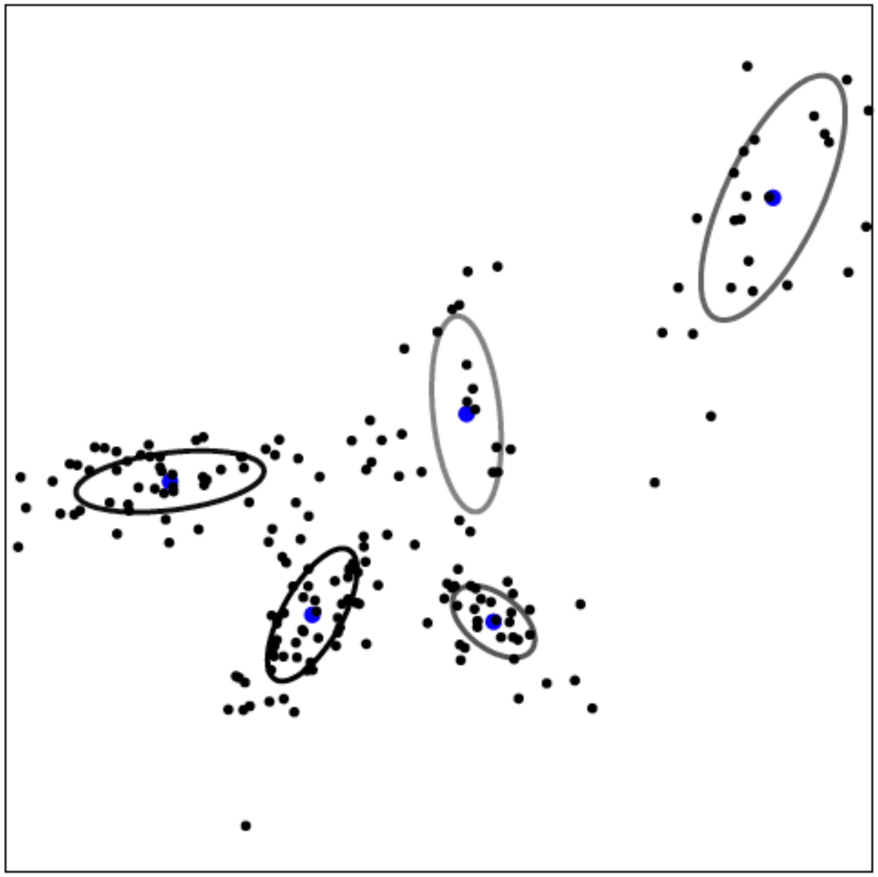
Samples from DP Mixture Priors



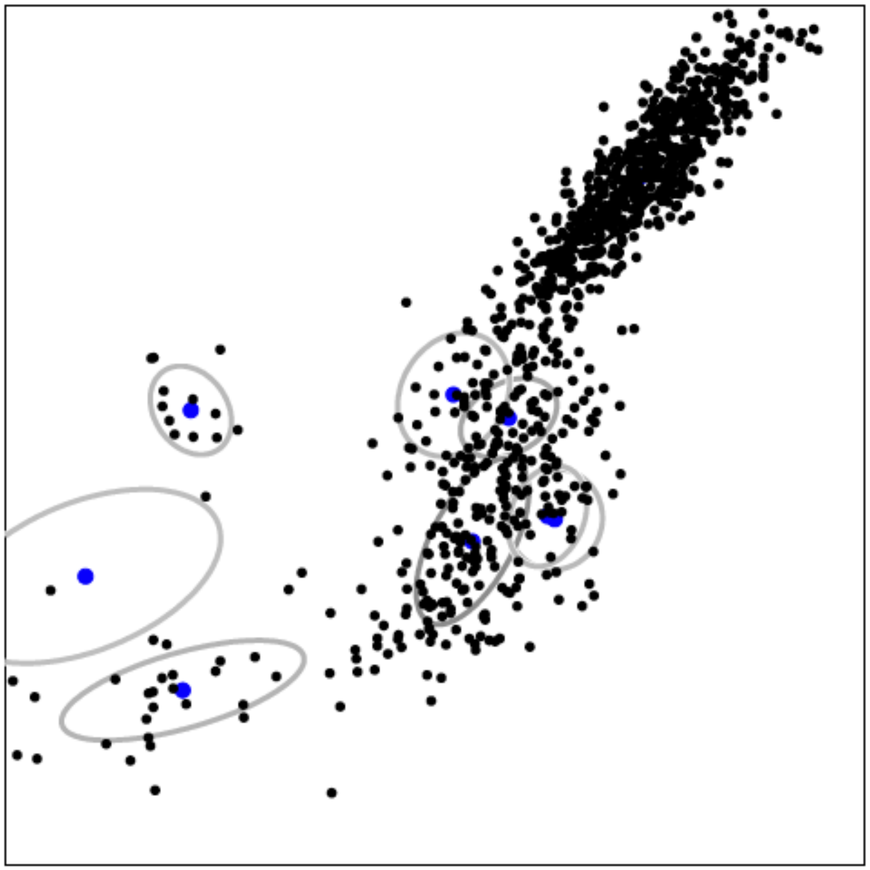


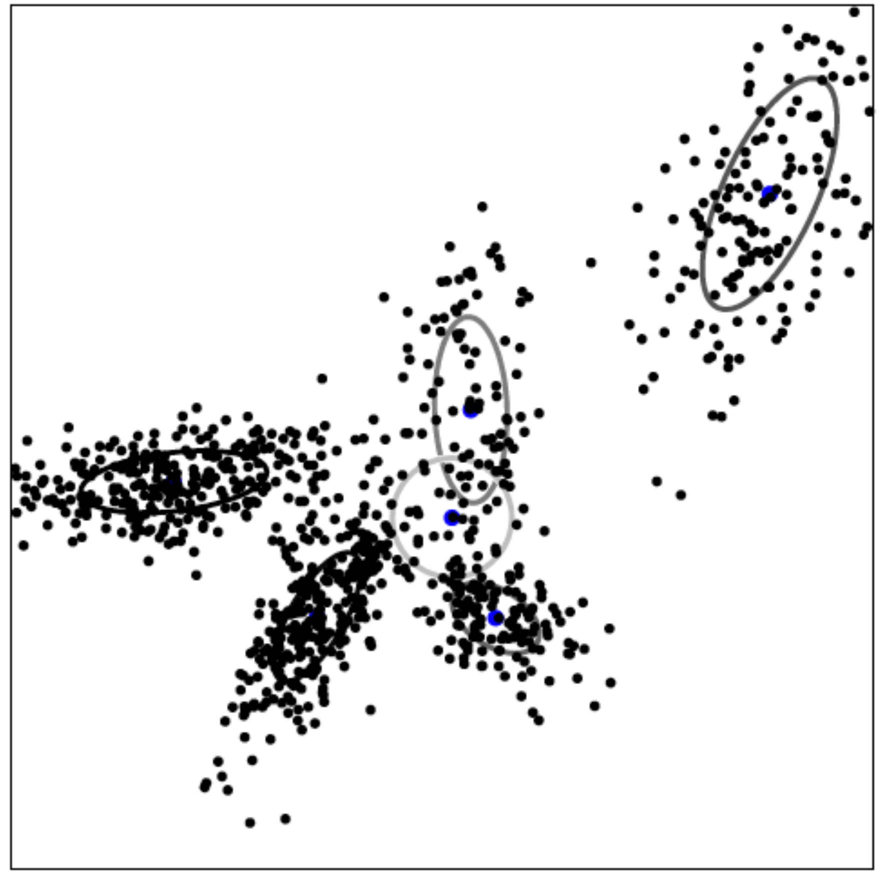
Samples from DP Mixture Priors





Samples from DP Mixture Priors





Views of the Dirichlet Process

- Implicit stochastic process: Finite Dirichlet marginals
- Implicit stochastic process: Neutrality
- Explicit stochastic process: Normalized gamma process
- Explicit stochastic process: Stick-breaking construction
- Marginalized predictions: Polya urn and the CRP
- Infinite limit of finite Dirichlet mixture model

Pitman-Yor Processes

Generalizing the Dirichlet Process

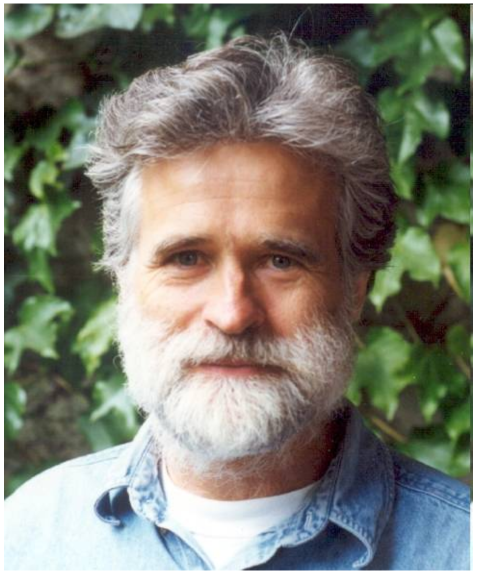
- Distribution on partitions leads to a generalized *Chinese restaurant process*
- Special cases arise as excursion lengths for Markov chains, Brownian motions, …

Power Law Distributions

	DP	PY
Number of unique clusters in N observations	$\mathcal{O}(b \log N)$	***Heaps' Law:*** $\mathcal{O}(bN^a)$
Size of sorted cluster weight k	$\mathcal{O}\left(\alpha_b \left(\frac{1+b}{b}\right)^{-k}\right)$	***Zipf's Law:*** $\mathcal{O}\left(\alpha_{ab} k^{-\frac{1}{a}}\right)$

Natural Language Statistics

Goldwater, Griffiths, & Johnson, 2005
Teh, 2006



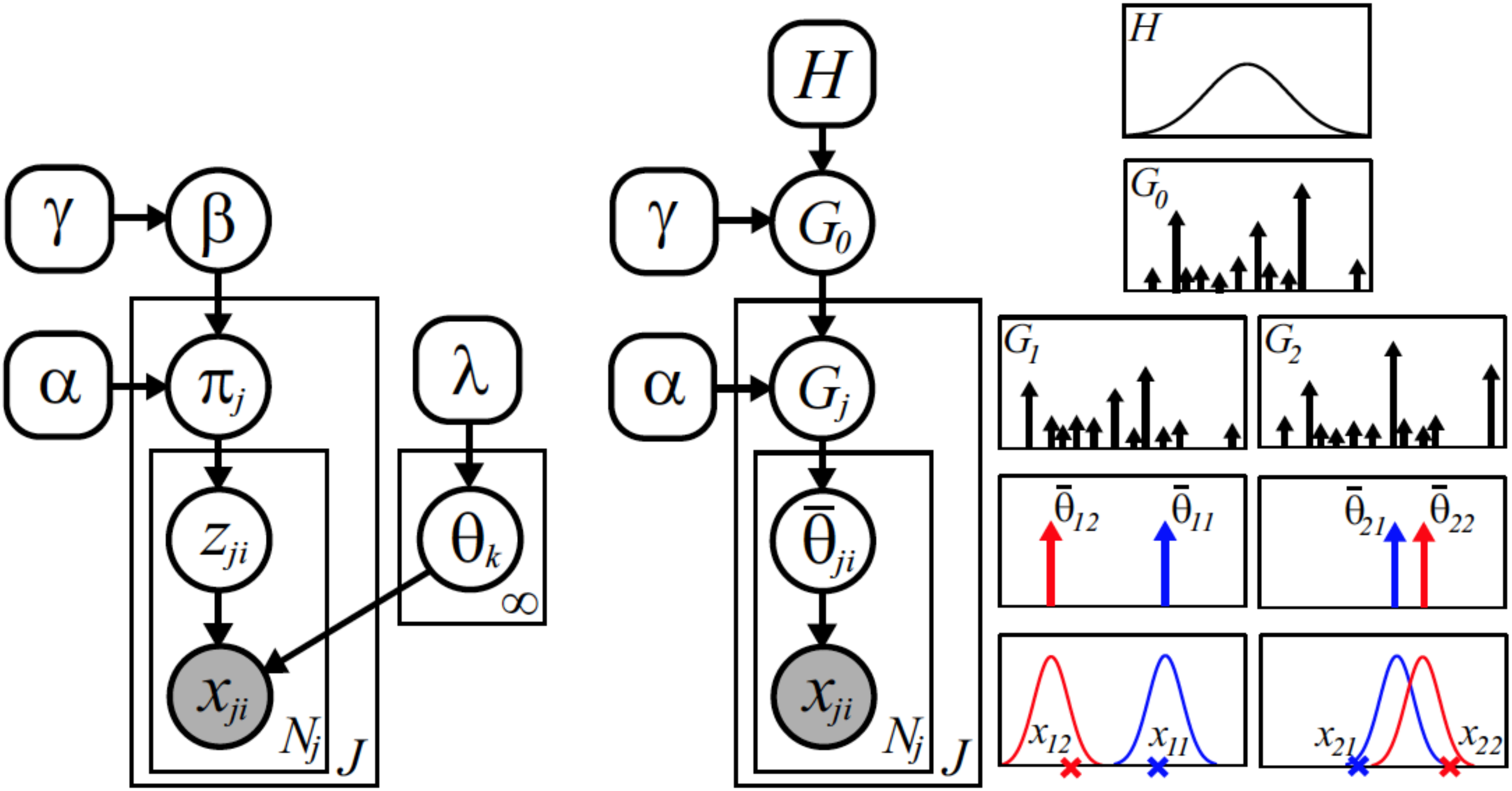
Jim Pitman

Marc Yor

Hierarchical and Dependent DP Models

- Hierarchical DP and the Chinese restaurant franchise
- Nested DP and the nested Chinese restaurant process
- Hierarchical DP hidden Markov models, switching LDS
- Hierarchical DP hidden Markov trees
- Gaussian processes and correlated mixture models
- …

Hierarchical Dirichlet Process



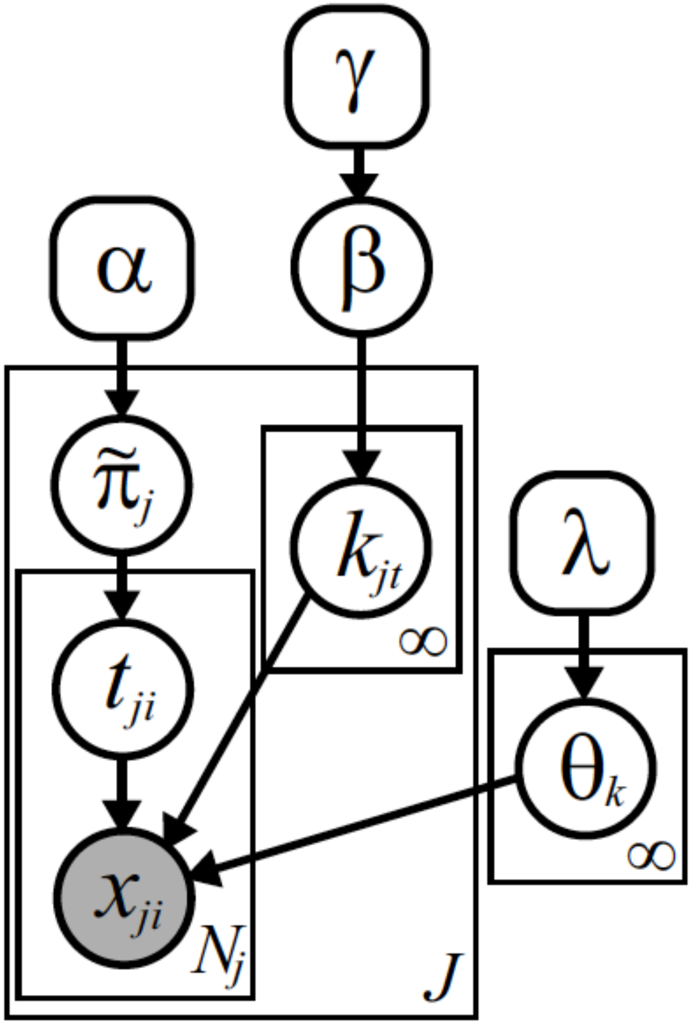
$$G_0(\theta) = \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k)$$

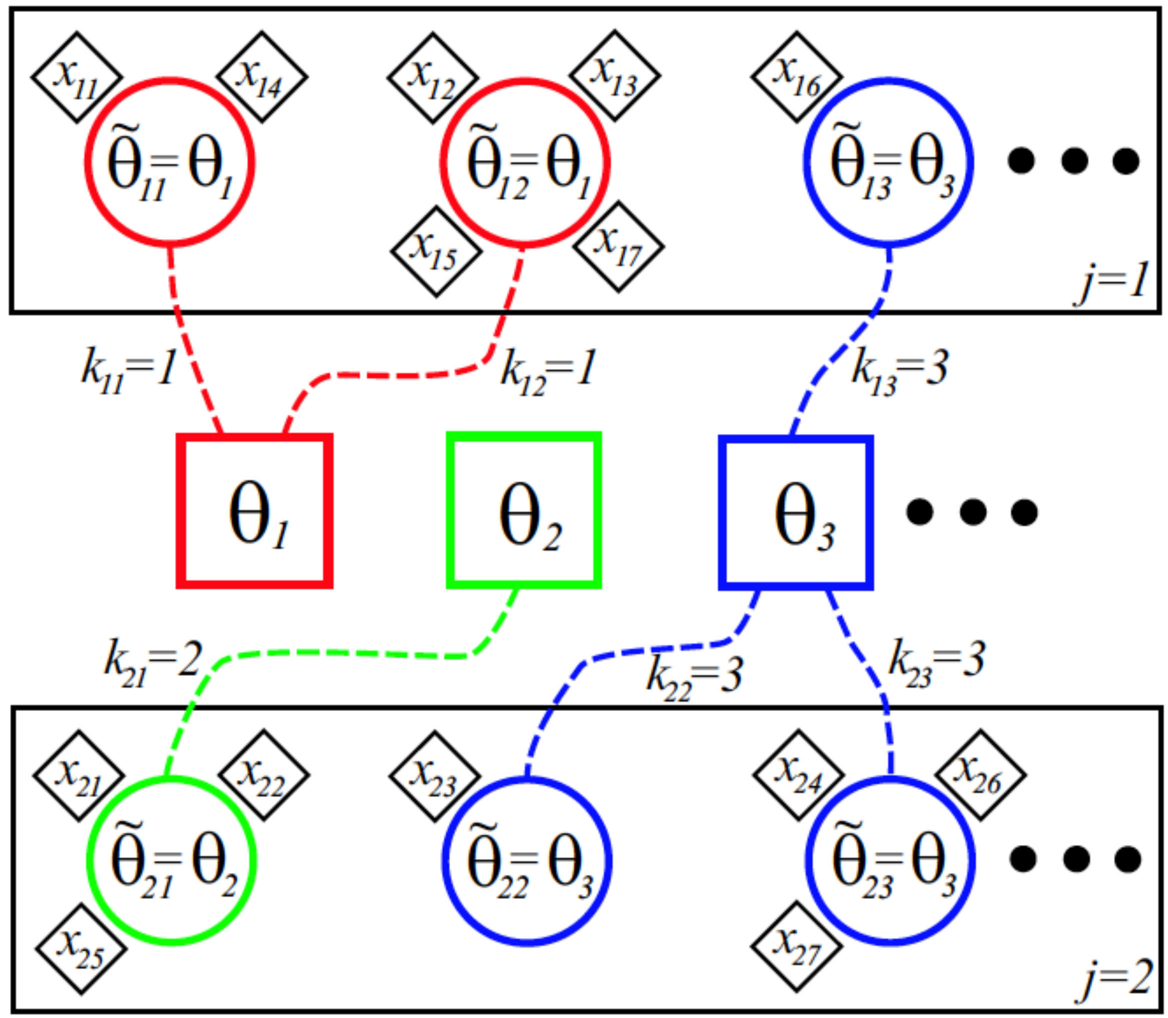
$$\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma)$$

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

$$G_j(\theta) = \sum_{k=1}^{\infty} \pi_{jk} \delta(\theta, \theta_k)$$

Chinese Restaurant Franchise





$$p(t_{ji} \mid t_{j1}, \ldots, t_{ji-1}, \alpha) \propto \sum_t N_{jt}\delta(t_{ji}, t) + \alpha\delta(t_{ji}, \bar{t})$$

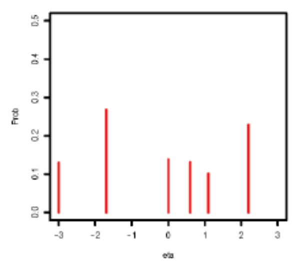
$$p(k_{jt} \mid \mathbf{k}_1, \ldots, \mathbf{k}_{j-1}, k_{j1}, \ldots, k_{jt-1}, \gamma) \propto \sum_k M_k\delta(k_{jt}, k) + \gamma\delta(k_{jt}, \bar{k})$$

Nested Dirichlet Process

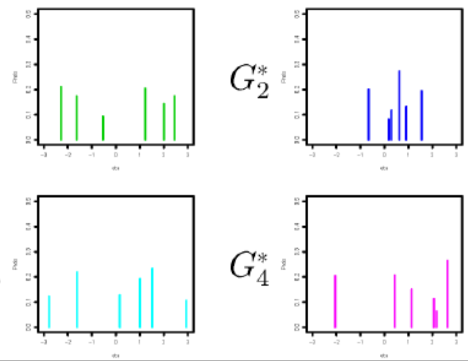
HDP

$$G_j \sim \mathrm{DP}(\alpha G_0)$$
$$G_0 \sim \mathrm{DP}(\beta H)$$

G_0



G_1^* G_2^* G_3^* G_4^* $\dots$

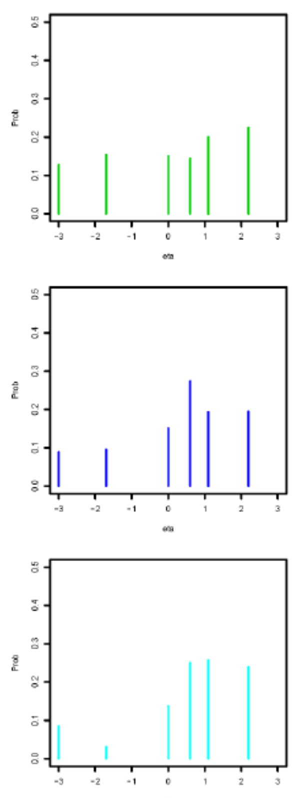


NDP

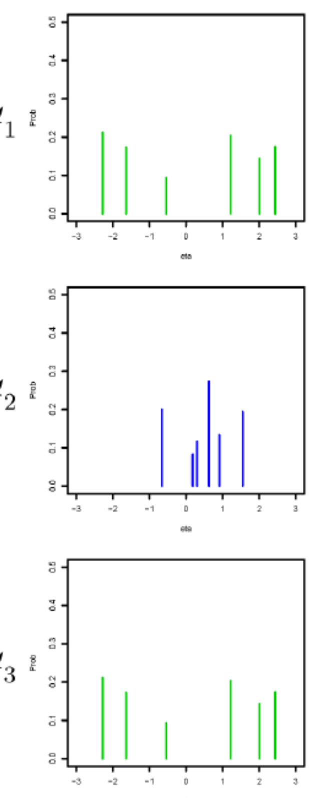
$$G_j \sim Q$$
$$Q \sim \mathrm{DP}(\alpha \mathrm{DP}(\beta H))$$

$$G_0(\theta) = \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k)$$
$$G_j(\theta) = \sum_{k=1}^{\infty} \pi_{jk} \delta(\theta, \theta_k)$$

G_1 G_2 G_3

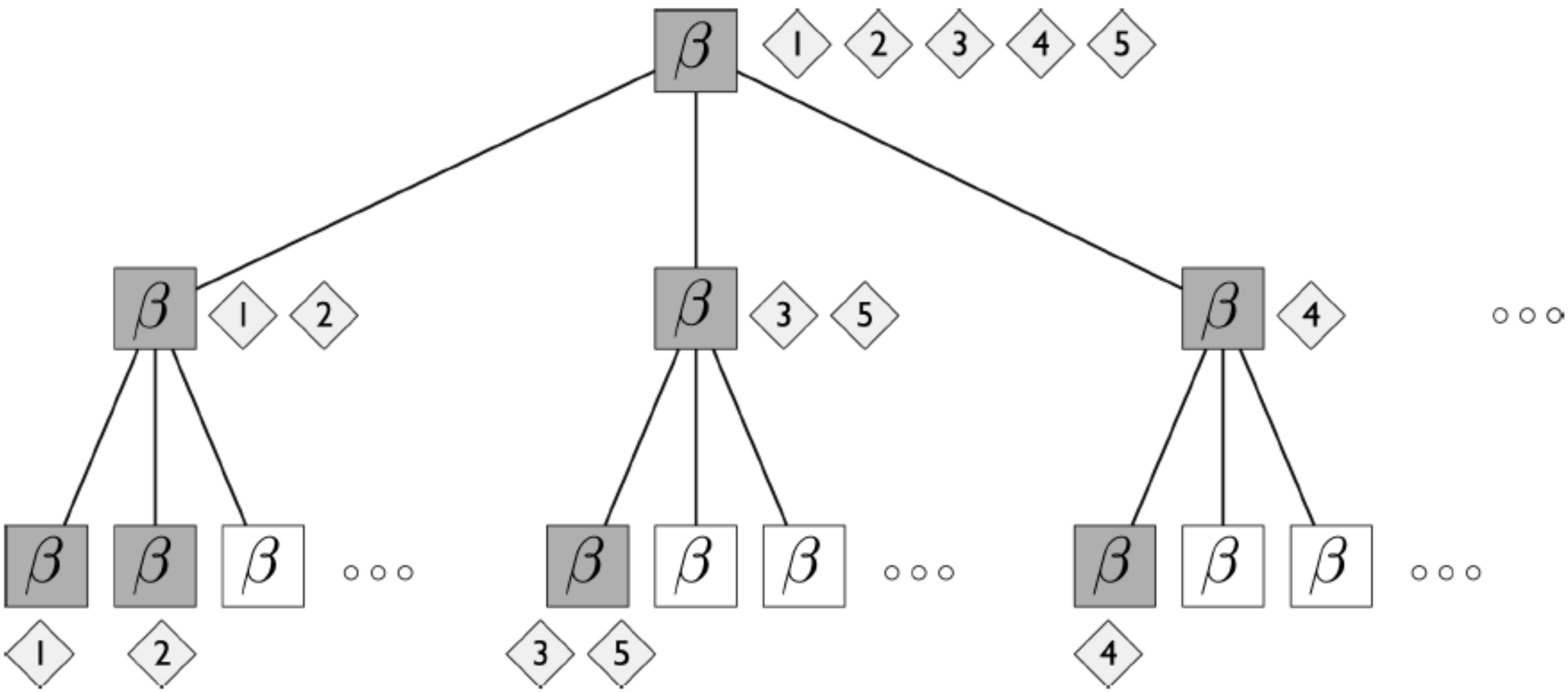


G_1 G_2 G_3



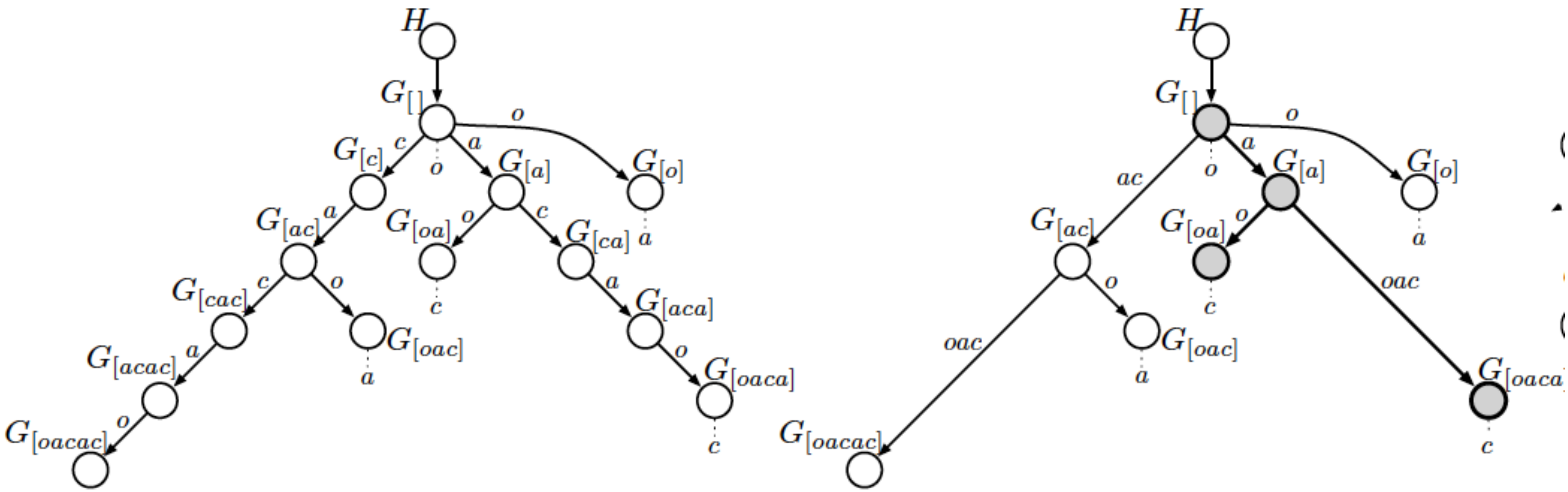
$$G_j(\cdot) \sim Q \equiv \sum_{k=1}^{\infty} \pi_k^* \delta_{G_k^*(\cdot)}$$
$$G_k^*(\cdot) \equiv \sum_{l=1}^{\infty} w_{lk}^* \delta_{\theta_{lk}^*}(\cdot)$$
$$\boldsymbol{\theta}_{lk}^* \sim H$$

Hierarchical LDA and the Nested CRP



- *Good:* Topics are arranged in a hierarchy of unknown structure, results tend to be more semantically interpretable
- *Limiting:* Each document is generated by a single path through the tree, cannot combine disparate topics

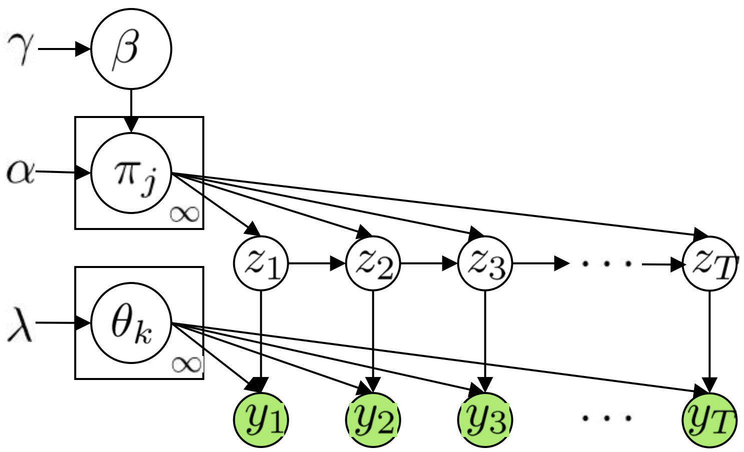
Infinite Markov Models & The Sequence Memoizer

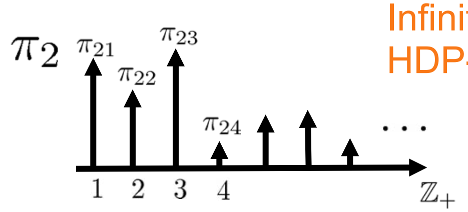


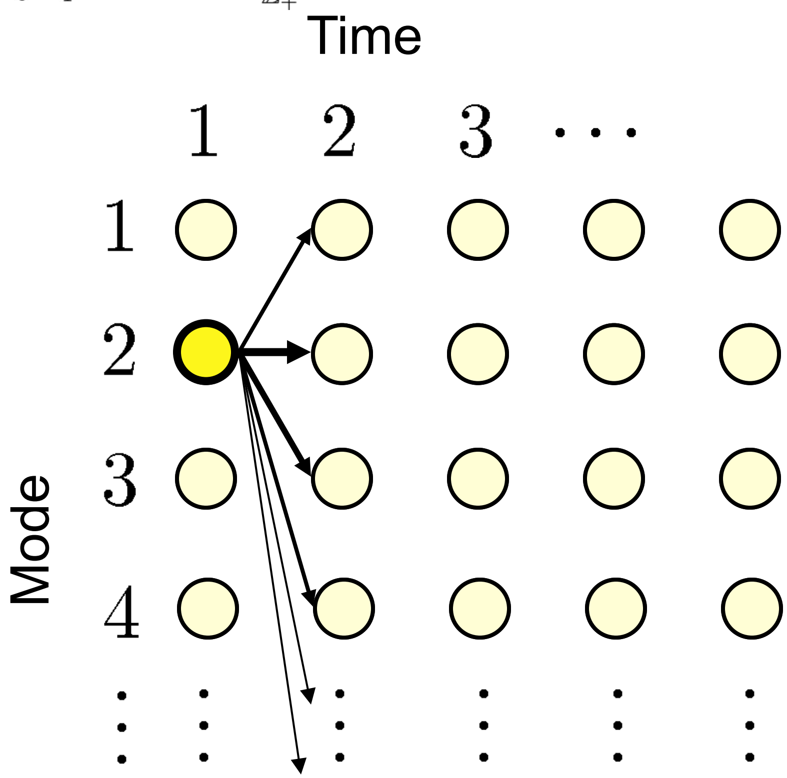
- *Good:* Tractably learn and predict with Markov models of infinite depth: Only finite contexts observed in training
- *Limiting:* Structure of tree depends on assumption that modeling sequences of discrete characters

HDP-HMM

Infinite HMM: Beal, et.al., NIPS 2002
HDP-HMM: Teh, et. al., JASA 2006



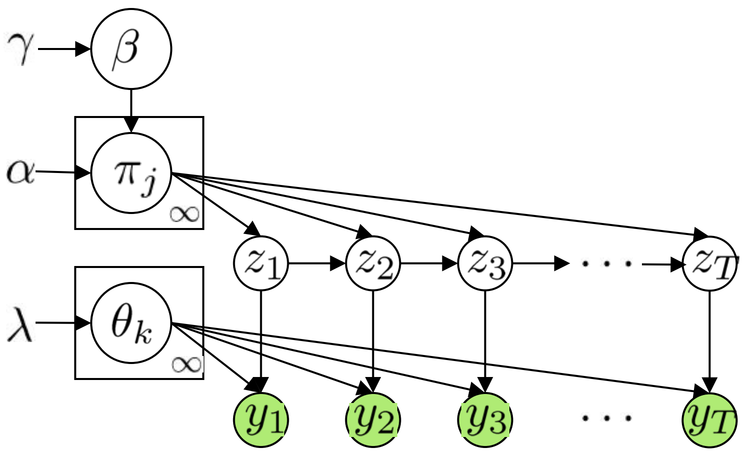


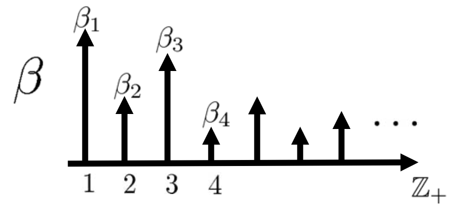


Hierarchical Dirichlet Process HMM

- Dirichlet process (DP):
  - Mode space of unbounded size
  - Model complexity adapts to observations
- Hierarchical:
  - Ties mode transition distributions
  - *Shared* sparsity

HDP-HMM





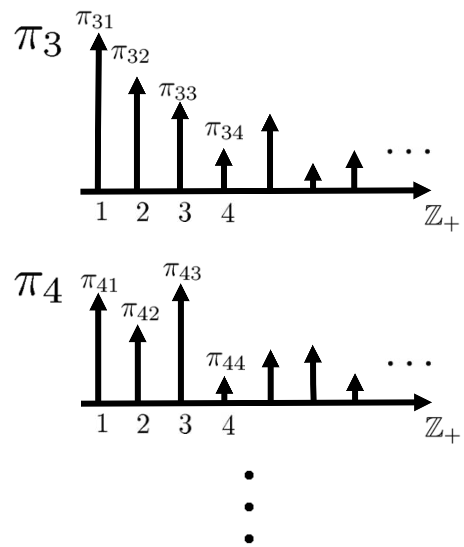
Hierarchical Dirichlet Process HMM

- Global transition distribution:

$$\beta \sim \mathrm{Stick}(\gamma)$$

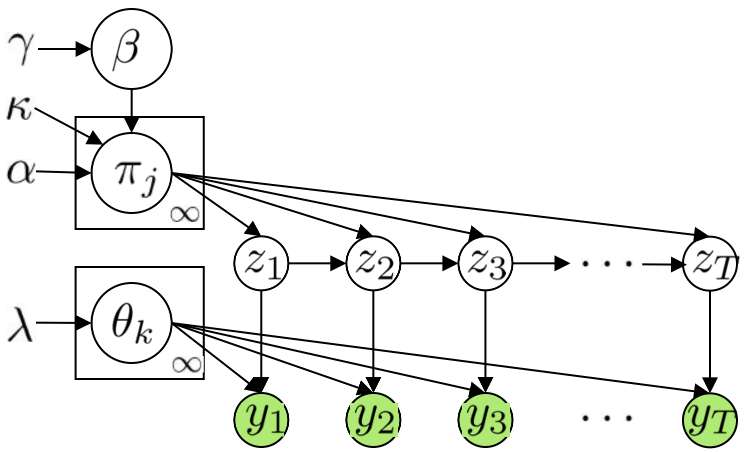
- Mode-specific transition distributions:

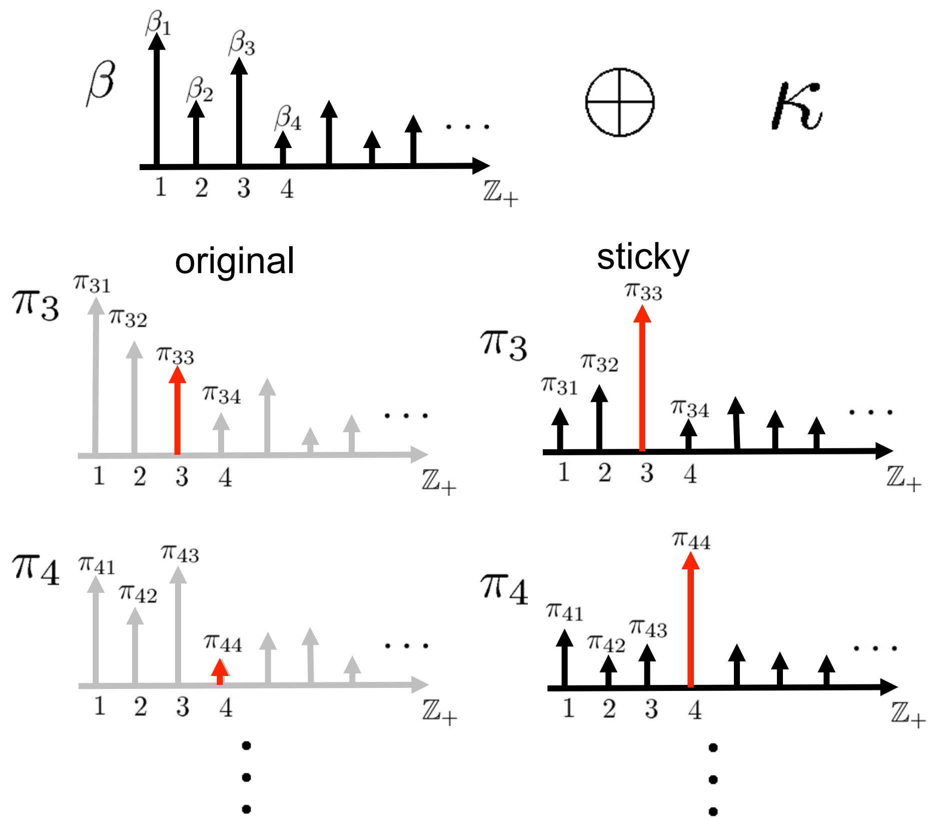
$$\pi_j \sim \mathrm{DP}(\alpha\beta) \quad j = 1, 2, 3, \ldots$$



sparsity of β is shared $\longrightarrow$ $\boxed{E[\pi_{jk}] = \beta_k}$

Sticky HDP-HMM





$$\begin{aligned}\beta &\sim \text{Stick}(\gamma)\\ \pi_j &\sim \text{DP}\left(\alpha\beta + \kappa\delta_j\right)\end{aligned}$$

mode-specific base measure

$$\boxed{E[\pi_{jk}] = \beta_k}$$

Increased probability of self-transition

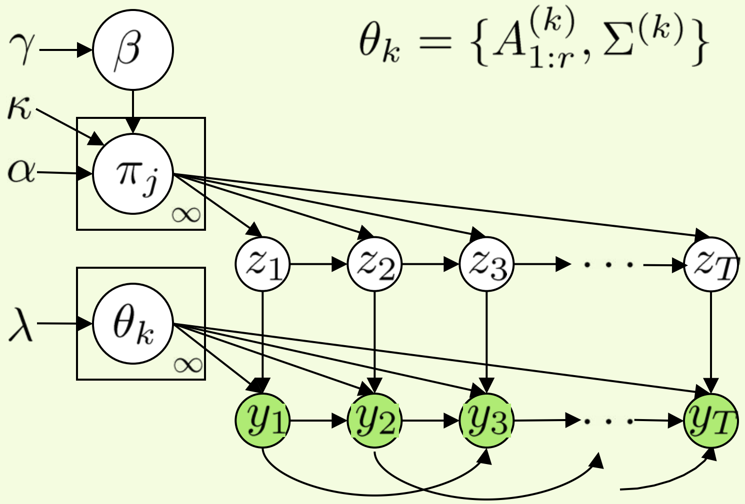
$$\boxed{E[\pi_{jk}] = \frac{\alpha\beta_k + \kappa\delta(j,k)}{\alpha + \kappa}}$$

Infinite HMM: Beal, et.al., NIPS 2002

HDP-AR-HMM and HDP-SLDS

HDP-AR-HMM

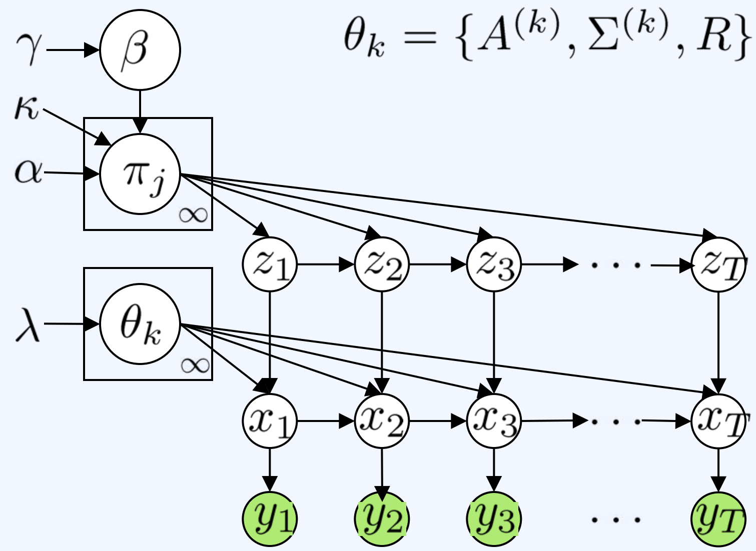
$$\theta_k = \{A_{1:r}^{(k)}, \Sigma^{(k)}\}$$



$$\begin{aligned} z_t &\sim \pi_{z_{t-1}} \\ y_t &= \sum_{i=1}^{r} A_i^{(z_t)} y_{t-i} + e_t(z_t) \end{aligned}$$

HDP-SLDS

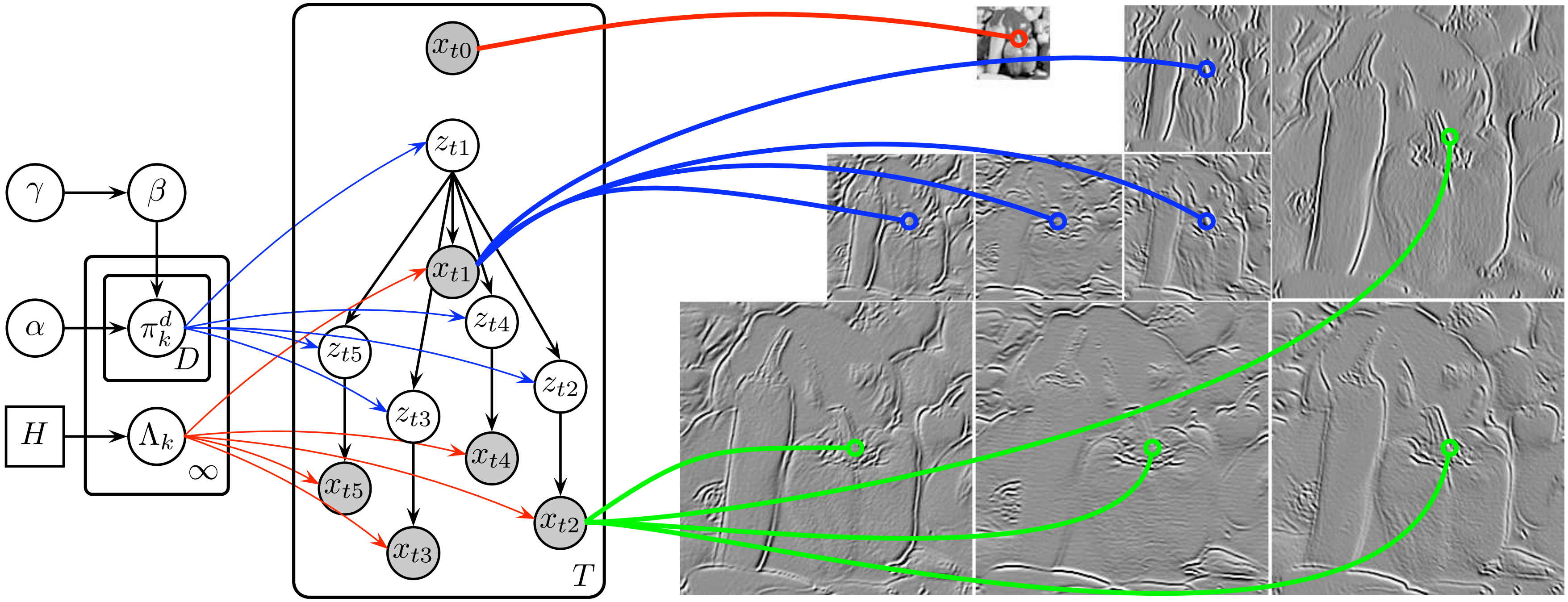
$$\theta_k = \{A^{(k)}, \Sigma^{(k)}, R\}$$



$$\begin{aligned} z_t &\sim \pi_{z_{t-1}} \\ x_t &= A^{(z_t)} x_{t-1} + e_t(z_t) \\ y_t &= C x_t + w_t \end{aligned}$$

$$C = [I \;\; 0]$$

HDP Hidden Markov Trees



$z_{ti} \longrightarrow$ indexes infinite set of hidden states

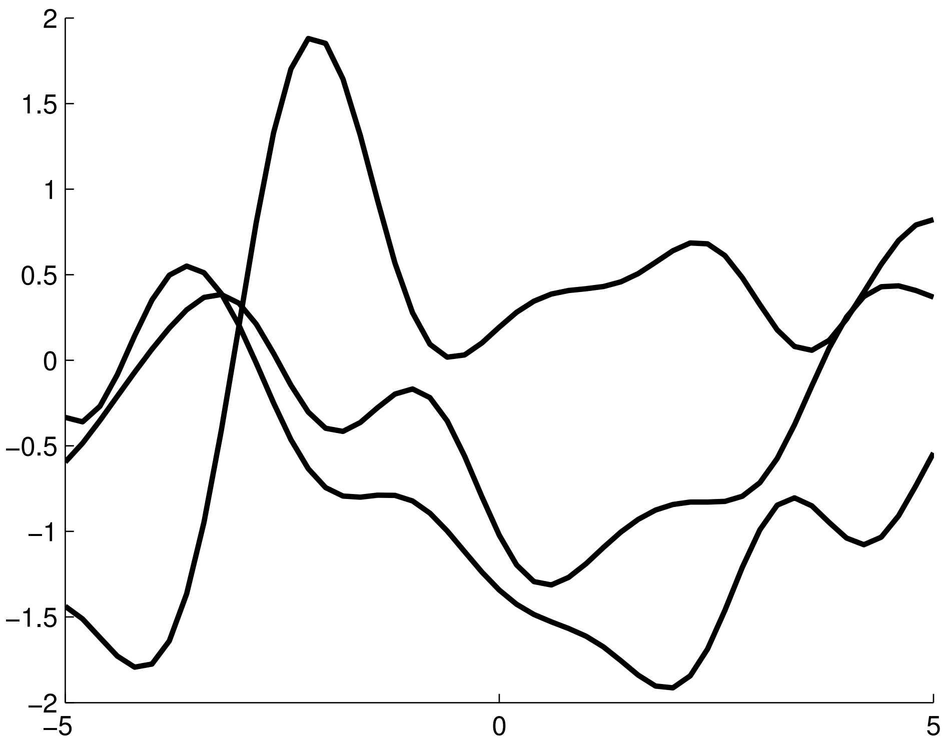
$$z_{ti} \in \{1, 2, 3, \ldots\}$$

$\pi_k \longrightarrow$ infinite set of state transition distributions

$$z_{ti} \sim \pi^{d_{ti}}_{z_{\mathrm{Pa(ti)}}}$$

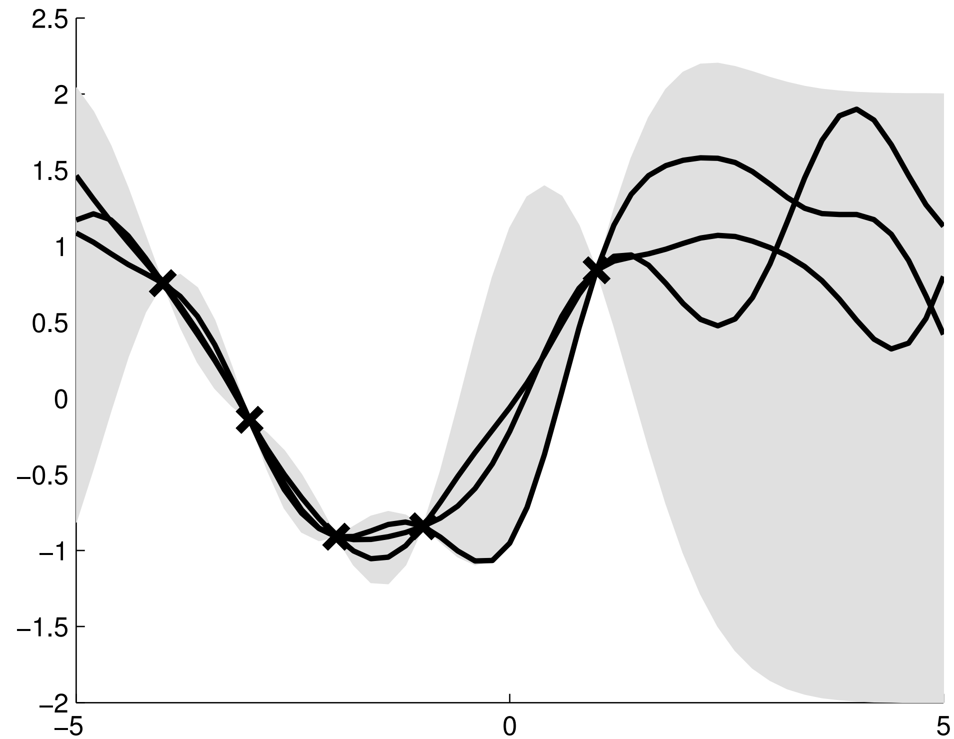
HDP-HMM & HDP-HMT cleanly deal with problem of choosing state space size, but retain other Markov model assumptions

1D Gaussian Processes



Samples from Prior

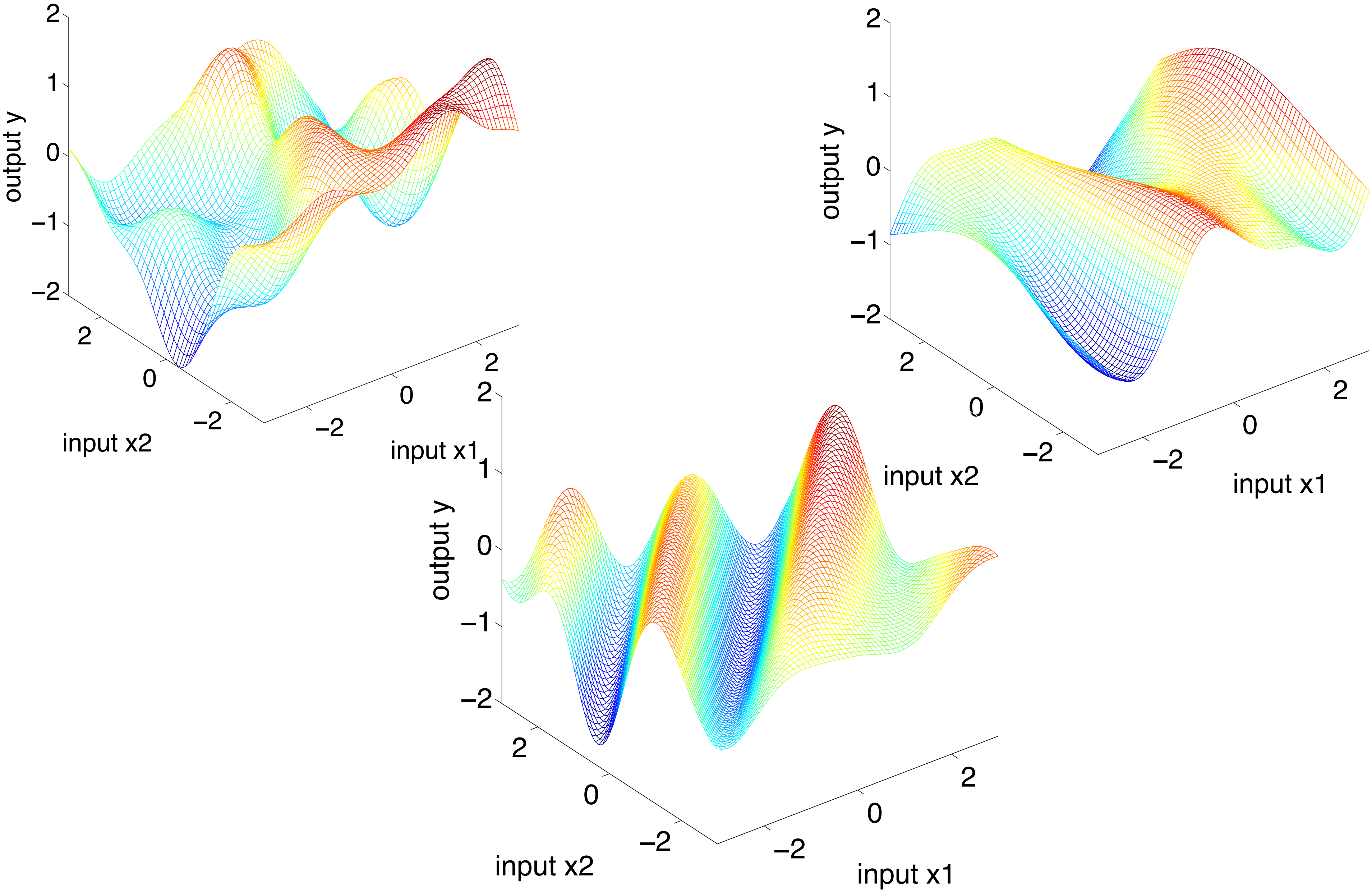
$$\kappa(x, x') = \sigma_f^2 \exp(-\frac{1}{2\ell^2}(x - x')^2)$$



Posterior Given 5 Noise-Free Observations

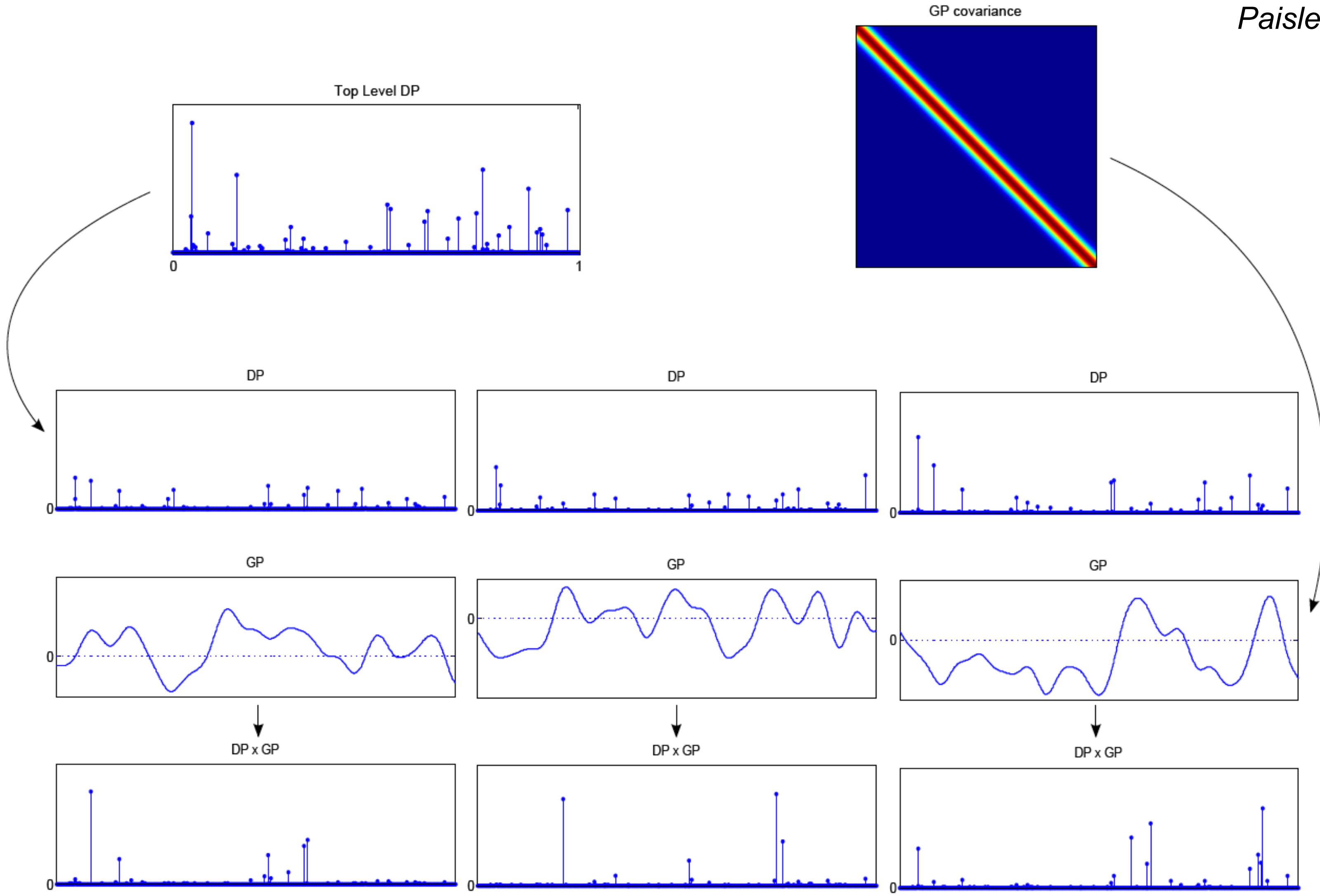
Squared exponential kernel or radial basis function (RBF) kernel has a countably *infinite* set of underlying feature functions

2D Gaussian Processes



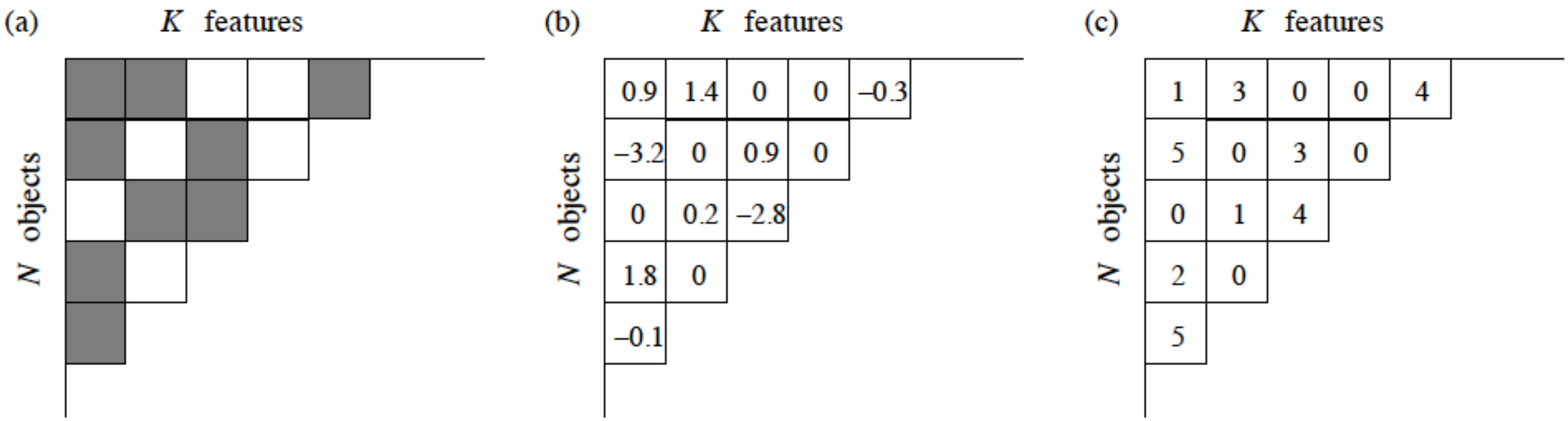
Correlated Mixtures via GPs

Paisley 2011



Why the GP? *Provides functions which are smooth, allow flexible correlation modeling, and computationally tractable*

Latent Feature Models



Distributions on binary matrices indicating feature presence/absence

Depending on application, features can be associated with any parameter value of interest

- ***Latent Feature*** model: Each group of observations is associated with a *subset* of the possible latent features
- Factorial power: There are 2^K combinations of K features
- Question: What is the analog of the DP for feature modeling?

From Clustering to Factorial Modeling

Dirichlet Process & Chinese Restaurant Process

- Implicit stochastic process: Finite Dirichlet marginals
- Implicit stochastic process: Neutrality
- Explicit stochastic process: Normalized gamma process
- Explicit stochastic process: Stick-breaking construction
- Marginalized predictions: Polya urn and the CRP
- Infinite limit of finite Dirichlet mixture model

Beta Process & Indian Buffet Process

- Implicit stochastic process: Poisson feature counts
- Implicit stochastic process: Completely random measure
- Explicit stochastic process: Un-normalized beta process
- Explicit stochastic process: Stick-breaking construction(s)
- Marginalized predictions: Indian buffet process
- Infinite limit of finite beta-Bernoulli feature model

Every temporal/spatial/hierarchical DP model should generalize…

Big Challenge: Learning & Inference

Collapsed or marginalize infinite model

- Chinese restaurant process and ddCRP
- Indian buffet process
- *Powerful but limited applicability*

Fixed truncation of true infinite model

- Truncated stick breaking
- Finite Dirichlet-multinomial
- Finite beta-Bernoulli
- Starting point for most variational methods

Dynamic truncations which avoid approximation

- Slice sampling
- Retrospective MCMC and reversible jump MCMC
- Local search for posterior modes

For the hardest problems, none are satisfactory…