Modern Bayesian Nonparametrics

Peter Orbanz
Yee Whye Teh

Cambridge University and Columbia University
Gatsby Computational Neuroscience Unit, UCL

NIPS 2011

OVERVIEW

1. Nonparametric Bayesian models
2. Regression
3. Clustering
4. Applications

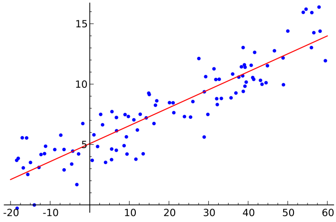
Coffee refill break

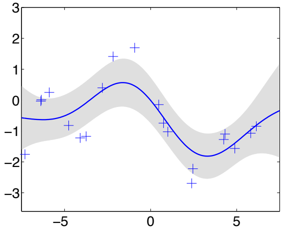
5. Asymptotics
6. Exchangeability
7. Latent feature models
8. Dirichlet process
9. Completely random measures
10. Summary

Parameters and Patterns

Parameters

$$P(X|\theta) \quad = \quad \text{Probability}[\text{data}|\text{pattern}]$$





Inference idea

$$\text{data} \quad = \quad \text{underlying pattern} \quad + \quad \text{independent noise}$$

TERMINOLOGY

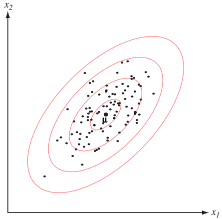
Parametric model

- Number of parameters fixed (or constantly bounded) w.r.t. sample size

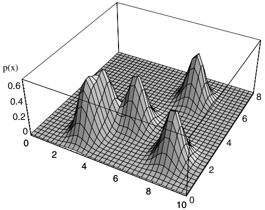
Nonparametric model

- Number of parameters grows with sample size
- ∞-dimensional parameter space

Example: Density estimation



Parametric



Nonparametric

Definition

A nonparametric Bayesian model is a Bayesian model on an ∞-dimensional parameter space.

Interpretation

Parameter space $\mathcal{T}$ = set of possible patterns, for example:

Problem	$\mathcal{T}$
Density estimation	Probability distributions
Regression	Smooth functions
Clustering	Partitions

Solution to Bayesian problem = posterior distribution on patterns

[Sch95]

REGRESSION

Nonparametric regression

Patterns = continuous functions, say on interval $[a, b]$:

$$\theta : [a, b] \to \mathbb{R} \qquad\qquad \mathcal{T} = C[a, b]$$

Gaussian process prior

- Hyperparameters: Mean function and covariance function

$$m \in C[a, b] \qquad \text{and} \qquad k : [a, b] \times [a, b] \to \mathbb{R}$$

- Plug in finite set $\mathbf{s} = \{s_1, \dots, s_n\} \subset [a, b]$:

$$m(\mathbf{s}) = \begin{pmatrix} m(s_1) \\ \vdots \\ m(s_n) \end{pmatrix} \qquad \text{and} \qquad k(\mathbf{s}, \mathbf{s}) = \begin{pmatrix} k(s_1, s_1) & \dots & k(s_1, s_n) \\ \vdots & & \vdots \\ k(s_n, s_1) & \dots & k(s_n, s_n) \end{pmatrix}$$

- Distribution of θ is Gaussian process if

$$(\theta(s_1), \dots, \theta(s_n)) \sim \mathcal{N}\big(m(\mathbf{s}), k(\mathbf{s}, \mathbf{s})\big) \qquad \text{for any } \mathbf{s} \subset [a, b]^n$$

[RW06]

Observation model

- Inputs $\mathbf{s} = (s_1, \ldots, s_n)$
- Outputs $\mathbf{t} = (t_1, \ldots, t_n)$

$$t_i \sim \mathcal{N}\big(\theta(s_i), \sigma_{\text{noise}}\big)$$

Posterior distribution

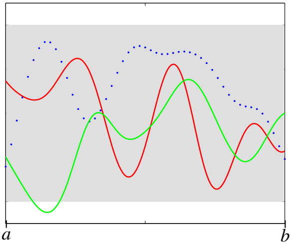
- Posterior is again a Gaussian Process
- Quantifies prediction uncertainty

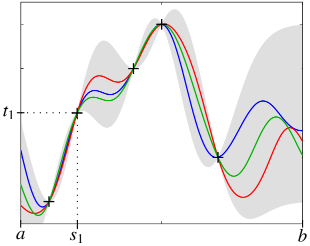
Predictions at test points $\mathbf{s}_*$

Test inputs $\mathbf{s}_* = (s_{*1}, \ldots, s_{*m})$

$$\hat{\mathbf{m}} = k(\mathbf{s}_*, \mathbf{s})\big(k(\mathbf{s}, \mathbf{s}) + \sigma^2_{\text{noise}}\mathbf{I}\big)^{-1}\mathbf{t}$$

$$\hat{\mathbf{k}} = k(\mathbf{s}_*, \mathbf{s}_*) - k(\mathbf{s}_*, \mathbf{s})\big(k(\mathbf{s}, \mathbf{s}) + \sigma^2_{\text{noise}}\mathbf{I}\big)^{-1}k(\mathbf{s}, \mathbf{s}_*)$$





[RW06]

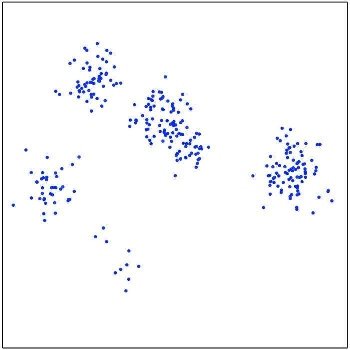
CLUSTERING

CLUSTERING









Standard probabilistic model for clustering

- For each observation $i = 1, \ldots, n$:

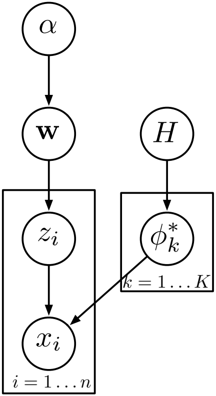
  Data: $x_i | z_i = k \sim F(\phi_k)$

  Cluster indicator: $z_i \sim \mathbf{w}$

- Parameters:

  Mixing proportions: $\mathbf{w} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$

  Cluster parameters: $\phi_k^* \sim H$



Learning and model selection

- For each $K = 1, 2, 3, \ldots$:
  - While learning not converged:
    - Update latent variables;
    - Update parameter.
  - Determine fit of model with K clusters.

Natural object of inference in clustering problems

- A cluster c is a subset of indices $[n] = \{1, \dots, n\}$.
- A partition π is a set of clusters.
  - Clusters are non-empty and disjoint;
  - Union of clusters is $[n]$.

x_1 x_2 x_3 x_4 x_5 x_6 x_7

$$\pi = \{\{1,6,7\},\{2\},\{3\},\{4,5\}\}$$

- Denote set of partitions of $[n]$ by $\mathcal{P}_{[n]}$.

Bayesian nonparametric model for clustering

- Prior distribution over $\mathcal{P}_{[n]}$.
- Likelihood model for data.

EXCHANGEABILITY

Data set 1:

Data set 2:

- Exchangeability:

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_1 = x_{\sigma(1)}, \ldots, X_n = x_{\sigma(n)})$$

$$\begin{aligned} &\mathbb{P}(\boldsymbol{\pi} = \{\{1,6,7\},\{2\},\{3\},\{4,5,8\}\}) \\ =&\mathbb{P}(\boldsymbol{\pi} = \{\{4,6,3\},\{8\},\{7\},\{1,5,2\}\}) \end{aligned}$$

Uniform distribution over $\mathcal{P}_{[n]}$

- Exchangeable.
- Not self-consistent.

1/15	{{1 2 3 4}}	{{1 2 3}}	2/15
1/15	{{1 2 3} {4}}		
1/15	{{1 2 4} {3}}		
1/15	{{1 2} {3 4}}	{{1 2} {3}}	3/15
1/15	{{1 2} {3} {4}}		
1/15	{{1 3 4} {2}}		
1/15	{{1 3} {2 4}}	{{1 3} {2}}	3/15
1/15	{{1 3} {2} {4}}		
1/15	{{1 4} {2 3}}		
1/15	{{1} {2 3} {4}}	{{1} {2 3}}	3/15
1/15	{{1} {2 3 4}}		
1/15	{{1 4} {2} {3}}		
1/15	{{1} {2 4} {3}}	{{1} {2} {3}}	4/15
1/15	{{1} {2} {3 4}}		
1/15	{{1} {2} {3} {4}}		

Preferential attachment

- Elements inserted into partition one at a time:
  - Inserted into an existing cluster, or
  - Into a new cluster.
- Example:

$$
\begin{aligned}
\mathbb{P}(8 \to \{1,6,7\}) &= (1-\delta)\tfrac{3}{7} \\
\mathbb{P}(8 \to \{2\}) &= (1-\delta)\tfrac{1}{7} \\
\mathbb{P}(8 \to \{3\}) &= (1-\delta)\tfrac{1}{7} \\
\mathbb{P}(8 \to \{4,5\}) &= (1-\delta)\tfrac{2}{7} \\
\mathbb{P}(8 \to \text{ new }) &= \delta
\end{aligned}
$$

- Typically not exchangeable.

$\pi = \{\{1\}\}$

2 new

$\pi = \{\{1\},\{2\}\}$

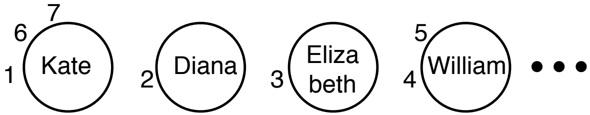
3 new

•
•
•

$\pi = \{\{1,6,7\},\{2\},\{3\},\{4,5\}\}$

8 new

Chinese Restaurant Process



$$\pi = \{\{1, 6, 7\}, \{2\}, \{3\}, \{4, 5\}\}$$

- One customer enters the restaurant at a time:
  - The first customer sits at the first table.
  - Subsequent customer $n + 1$:
    - Joins table c with probability $\dfrac{|c|}{n + \alpha}$.
    - Starts a new table with probability $\dfrac{\alpha}{n + \alpha}$.
- Distribution over partitions that is exchangeable and self-consistent.

$$\boldsymbol{\pi} \sim \mathrm{CRP}(\alpha)$$

$$\pi = \{\{1,6,7\},\{2\},\{3\},\{4,5\}\}$$

For $c \in \pi$: $$\phi_c^* \mid \boldsymbol{\pi} \sim H$$

For $i \in c$: $$x_i \mid \boldsymbol{\pi}, \boldsymbol{\phi}^* \sim F(\phi_c^*)$$

THE GENERATIVE PROCESS

$$\boldsymbol{\pi} \sim \mathrm{CRP}(\alpha)$$

$$\pi = \{\{1,6,7\},\{2\},\{3\},\{4,5\}\}$$

For $c \in \pi$: $\phi_c^* \mid \boldsymbol{\pi} \sim H$

$$\phi_{\mathrm{K}} \quad \phi_{\mathrm{D}} \quad \phi_{\mathrm{E}} \quad \phi_{\mathrm{W}}$$

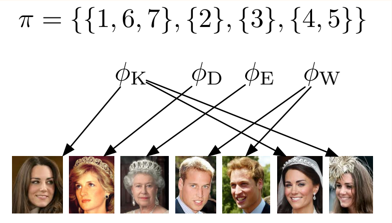
For $i \in c$: $x_i \mid \boldsymbol{\pi}, \boldsymbol{\phi}^* \sim F(\phi_c^*)$

THE GENERATIVE PROCESS

$$\boldsymbol{\pi} \sim \mathrm{CRP}(\alpha)$$

For $c \in \pi$: $\qquad \phi^*_c \mid \boldsymbol{\pi} \sim H$

For $i \in c$: $\qquad x_i \mid \boldsymbol{\pi}, \boldsymbol{\phi}^* \sim F(\phi^*_c)$



Gibbs sampling

- Update cluster parameters:

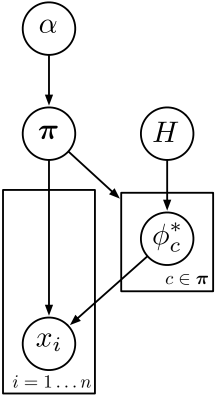
$$\text{For } c \in \boldsymbol{\pi}: \qquad p(\phi_c^*) = h(\phi_c^*) \prod_{i \in c} f(x_i | \phi_c^*)$$

- Update partition:

$$\text{For } i \in [n]: \qquad p(i \in c_{-i}) \propto \frac{|c_{-i}|}{n-1+\alpha} f(x_i|\phi_c^*)$$

$$p(i \text{ in new cluster }) \propto \frac{\alpha}{n-1+\alpha} f(x_i|\phi_{\text{new}}^*)$$

- Other samplers: split-merge [?], conditional sampling [?, ?, ?], variational inference [?, ?].



[Nea00]

Finite mixture model

- For each observation $i = 1, \ldots, n$:

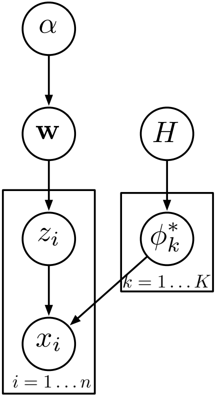
  Data: $x_i | z_i = k \sim F(\theta_k)$

  Cluster indicator: $z_i \sim \mathbf{w}$

- Parameters:

  Mixing proportions: $\mathbf{w} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$

  Cluster parameters: $\phi_k^* \sim H$



Infinite limit

- Derive the induced distribution over partitions.

$$\mathbb{P}(\boldsymbol{\pi}_K = \pi) = \frac{\Gamma(K+1)\Gamma(\alpha)}{\Gamma(K - |\pi| + 1)} \prod_{c \in \pi} \frac{\Gamma(|c| + \alpha/K)}{\Gamma(\alpha/K)}$$

- Take $K \to \infty$.

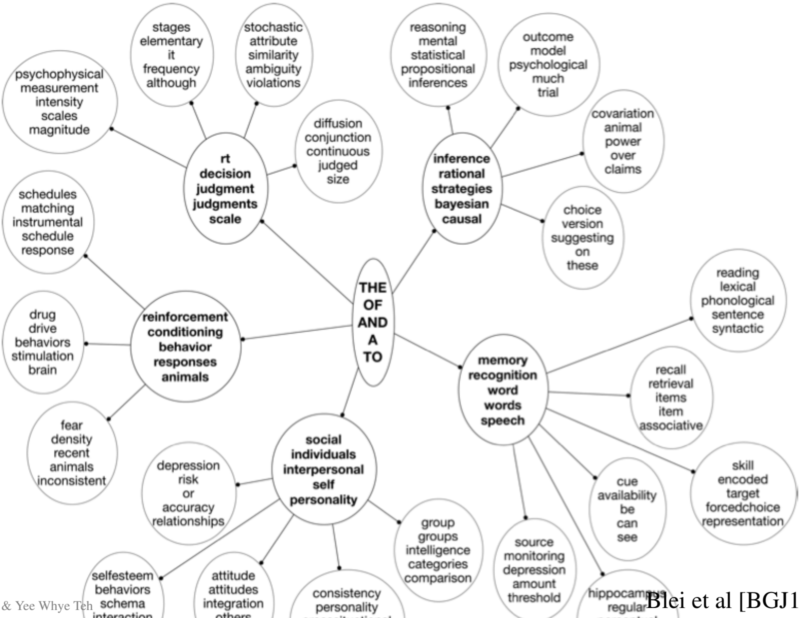
[Nea92, Ras00, IZ02]

APPLICATIONS

APPLICATIONS

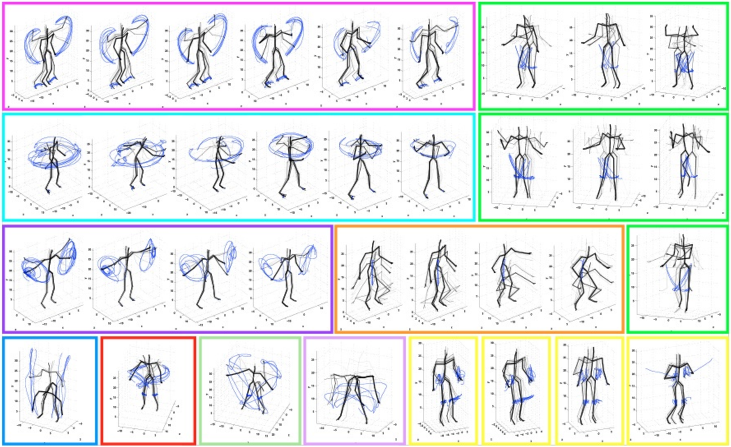
Applications	Object of interest	Bayesian nonparametric model
Classification & regression	Function	Gaussian process
Clustering	Partition	Chinese restaurant process
Density estimation	Density	Dirichlet process mixture
Hierarchical clustering	Hierarchical partition	Dirichlet/Pitman-Yor diffusion tree, Kingman's coalescent, Nested CRP
Latent variable modelling	Features	Beta process/Indian buffet process
Survival analysis	Hazard	Beta process, Neutral-to-the-right process
Power-law behaviour		Pitman-Yor process, Stable-beta process
Dictionary learning	Dictionary	Beta process/Indian buffet process
Dimensionality reduction	Manifold	Gaussian process latent variable model
Deep learning	Features	Cascading/nested Indian buffet process
Topic models	Atomic distribution	Hierarchical Dirichlet process
Time series		Infinite HMM
Sequence prediction	Conditional probs	Sequence memoizer
Reinforcement learning	Conditional probs	infinite POMDP
Spatial modelling	Functions	Gaussian process, dependent Dirichlet process
Relational modelling		Infinite relational model, infinite hidden relational model, Mondrian process
⋮	⋮	⋮

LEARNING TOPIC HIERARCHIES



Blei et al [BGJ10]

Motion Capture Segmentation



Fox et al [FSJW10]

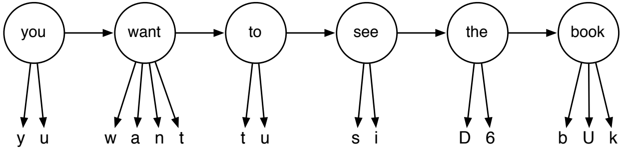
Word Segmentation

山花貞夫・新民連会長は十六日の記者会見で、村山富市首相ら 社会党 執行部とさきがけが連携強化をめざした問題について「私たちの行動が新しい政界の動きを作ったといえる。統一会派を超えて将来の日本の...

今后一段时期,不但居民会更多地选择国债,而且一些金融机构在准备金利率调低后,出于安全性方面的考虑,也会将部分资金用来购买 国债 。

yuwanttusiD6bUk?

Goldwater et al [GGJ06a], Mochihashi et al [MYU09]



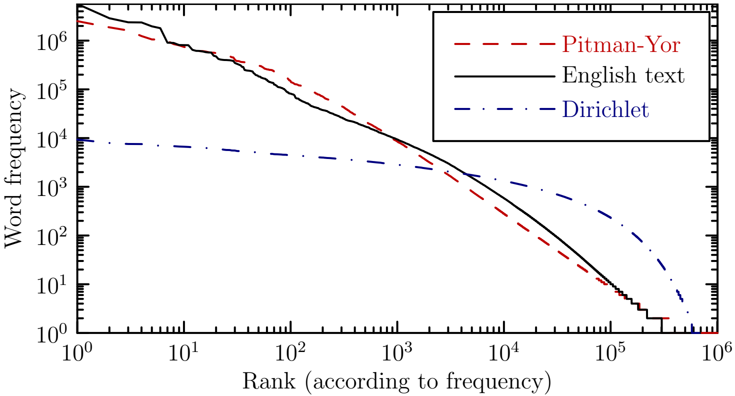
yuwanttusiD6bUk

	P	R	F	BP	BR	BF	LP	LR	LF
NGS-u	67.7	70.2	68.9	80.6	84.8	82.6	52.9	51.3	52.0
MBDP-1	67.0	69.4	68.2	80.3	84.3	82.3	53.6	51.3	52.4
DP	61.9	47.6	53.8	92.4	62.2	74.3	57.0	57.5	57.2
NGS-b	68.1	68.6	68.3	81.7	82.5	82.1	54.5	57.0	55.7
HDP	**79.4**	**74.0**	**76.6**	**92.4**	**83.5**	**87.7**	**67.9**	**58.9**	**63.1**

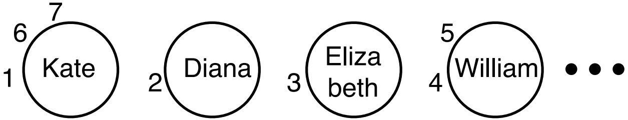
Model	MSR	CITYU	Kyoto
NPY(2)	80.2 (51.9)	**82.4** (**126.5**)	62.1 (23.1)
NPY(3)	**80.7** (**48.8**)	81.7 (128.3)	**66.6** (**20.6**)
ZK08	66.7 (—)	69.2 (—)	—

Goldwater et al [GGJ06a], Mochihashi et al [MYU09]

POWER-LAW BEHAVIOUR



TWO-PARAMETER CHINESE RESTAURANT PROCESS



- One customer enters the restaurant at a time:
  - The first customer sits at the first table.
  - Subsequent customer $n + 1$:
    - Joins table c with probability $\dfrac{|c| - d}{n + \alpha}$.
    - Starts a new table with probability $\dfrac{\alpha + |\pi| d}{n + \alpha}$.
- Distribution over partitions is still exchangeable, and has power-law properties.

[PPY92, PY97, GGJ06b, Teh06]

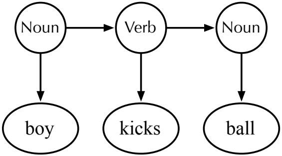
Language Modelling

T	N-1	IKN	MKN	HDLM	HPYLM
2×10^6	2	148.8	**144.1**	191.2	144.3
4×10^6	2	137.1	**132.7**	172.7	**132.7**
6×10^6	2	130.6	126.7	162.3	**126.4**
8×10^6	2	125.9	122.3	154.7	**121.9**
10×10^6	2	122.0	118.6	148.7	**118.2**
12×10^6	2	119.0	115.8	144.0	**115.4**
14×10^6	2	116.7	113.6	140.5	**113.2**
14×10^6	1	169.9	**169.2**	180.6	169.3
14×10^6	3	106.1	102.4	136.6	**101.9**

Compression

Algorithm	bits/byte
gzip	2.61
bzip2	2.11
CTW	1.99
PPM	1.93
SM	**1.89**

[Teh06, GWT10]

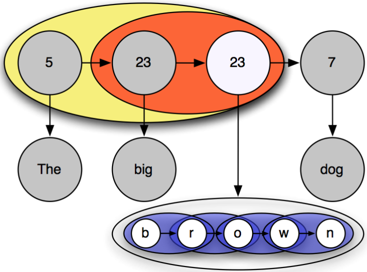


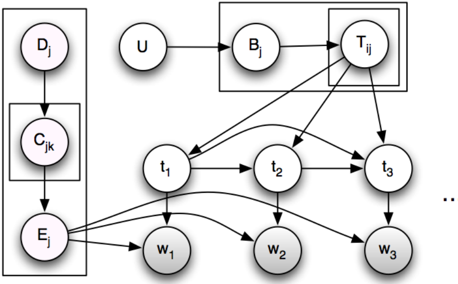
Language	mkcls	HMM	1HMM	1HMM-LM	Best pub.	Tokens	Tag types
Arabic	58.5	57.1	62.7	**67.5**	-	54,379	20
Bulgarian	66.8	67.8	69.7	**73.2**	-	190,217	54
Czech	59.6	62.0	66.3	**70.1**	-	1,249,408	12[c]
Danish	62.7	69.9	73.9	**76.2**	66.7⋆	94,386	25
Dutch	64.3	66.6	68.7	**70.4**	67.3†	195,069	13[c]
Hungarian	54.3	65.9	69.0	**73.0**	-	131,799	43
Portuguese	68.5	72.1	73.5	**78.5**	75.3⋆	206,678	22
Spanish	63.8	71.6	74.7	**78.8**	73.2⋆	89,334	47
Swedish	64.3	66.6	67.0	**68.6**	60.6†	191,467	41

Blunsom & Cohn [BC11]

Construction of complex Bayesian nonparametric models

- Graphical models.
- Hierarchical Bayesian models [TJ10].
- Dependent stochastic processes [GKM05, Dun10].



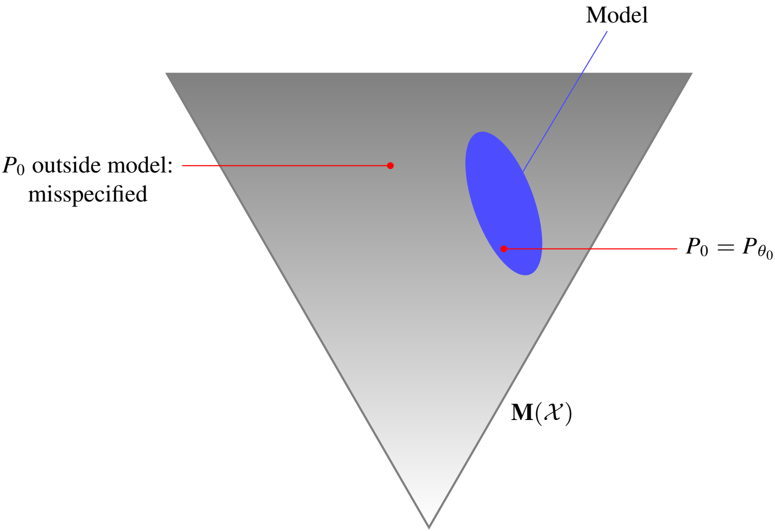


Blunsom & Cohn [BC11]

5 Minutes Break

ASYMPTOTICS

Coverage of Priors



[Gho10, KvdV06]

Coverage of Nonparametric Priors

Large coverage

- Support of nonparametric priors is larger (∞-dimensional) than of parametric priors (finite-dimensional).
- However: No uniform prior (or even "neutral" improper prior) exists on $\mathbf{M}(\mathcal{X})$.

Interpretation of nonparametric prior assumptions

Concentration of nonparametric prior on subset of $\mathbf{M}(\mathcal{X})$ typically represents structural prior assumption.

- GP regression with unknown bandwidth:
  - Any continuous function possible
  - Prior can express e.g. "very smooth functions are more probable"
- Clustering: Expected number of clusters is...
  - ...small $\longrightarrow$ CRP prior
  - ...power law $\longrightarrow$ two-parameter CRP

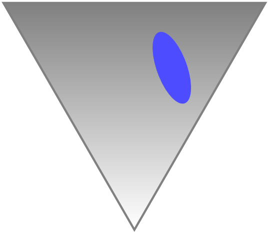
POSTERIOR CONSISTENCY

Definition 1 (weak consistency of Bayesian models)

Suppose we sample $P_0 = P_{\theta_0}$ from the prior and generate data from P_0. If the posterior converges to δ_{θ_0} for $n \to \infty$ *with probability one under the prior*, the model is called *consistent*.

Doob's Theorem

Under very mild conditions, Bayesian models are consistent in the weak sense.



Problem

- Definition holds up to a set of probability zero under the prior.
- This set can be huge and is a prior assumption.

Definition 2 (frequentist consistency of Bayesian models)

A Bayesian model is *consistent at P_0* if the posterior converges to δ_{P_0} with growing sample size.

[Gho10]

CONVERGENCE RATES

Objective

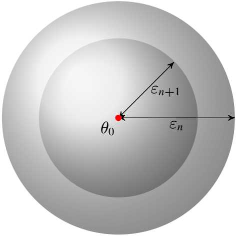
How quickly does posterior concentrate at θ_0 as $n \to \infty$?

Measure: Convergence rate

- Find smallest balls $B_{\varepsilon_n}(\theta_0)$ for which

$$Q(B_{\varepsilon_n}(\theta_0)|X_1, \dots, X_n) \xrightarrow{n\to\infty} 1$$

- Rate = sequence $\varepsilon_1, \varepsilon_2, \dots$



The best we can hope for

- Optimal rate is $\varepsilon_n \propto n^{-1/2}$
- Given by optimal convergence of estimators
- Achieved in smooth parametric models

Technical tools

Sieves, covering number, metric entropies... $\longrightarrow$ familiar from learning theory!

[Gho10, vdV98]

Asymptotics: Sample Results

Consistency

- DP mixtures: Consistent in many cases. No blanket statements.
- Range of consistency results for GP regression

Convergence rates: Example

Bandwidth adaptation with GPs:

- True parameter $\theta_0 \in C^{\alpha}[0,1]^d$, smoothness α unknown
- With gamma prior on GP bandwidth:

$$\text{Convergence rate is } n^{-\alpha/(2\alpha+d)}$$

Bernstein-von Mises Theorems

- Class of theorems establishing that posterior is asymptotically normal.
- Available for Gaussian processes and various regression settings.

[Sch65, GvdV07, vdVvZ08a, vdVvZ08b]

EXCHANGEABILITY

MOTIVATION

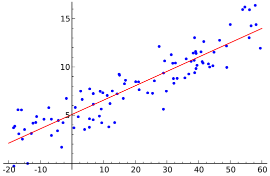
Can we justify our assumptions?

Recall:

$$\text{data} = \text{pattern} + \text{noise}$$

In Bayes' theorem:

$$Q(d\theta|x_1,\ldots,x_n) = \frac{\prod_{j=1}^n p(x_j|\theta)}{p(x_1,\ldots,x_n)} Q(d\theta)$$



Exchangeability

$X_1, X_2, \ldots$ are *exchangeable* if $P(X_1, X_2, \ldots)$ is invariant under any permutation σ:

$$P(X_1 = x_1, X_2 = x_2, \ldots) = P(X_1 = x_{\sigma(1)}, X_2 = x_{\sigma(2)}, \ldots)$$

In words:

Order of observations does not matter.

De Finetti's Theorem

$$P(X_1 = x_1, X_2 = x_2, \ldots) = \int_{\mathbf{M}(\mathcal{X})} \Big(\prod_{j=1}^{\infty} \theta(X_j = x_j) \Big) Q(d\theta)$$

$$\Updownarrow$$

$$X_1, X_2, \ldots \text{ exchangeable}$$

where:

- $\mathbf{M}(\mathcal{X})$ is the set of probability measures on $\mathcal{X}$
- θ are values of a random probability measure Θ with distribution Q

Implications

- Exchangeable data decomposes into pattern and noise
- More general than i.i.d.-assumption
- Caution: θ is in general an ∞-dimensional quantity

[Sch95, Kal05]

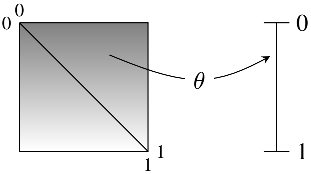
Random graph with independent edges

Given: $\theta : [0,1]^2 \to [0,1]$ symmetric function

- $U_1, U_2, \ldots \sim \text{Uniform}[0,1]$
- Edge (i,j) present:

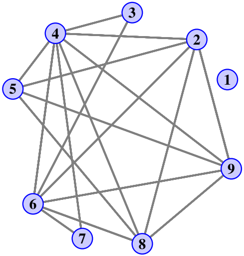
$$(i,j) \sim \text{Bernoulli}(\theta(U_i, U_j))$$

Call this distribution $P(\mathcal{G}|\theta)$.



Aldous-Hoover Theorem

$$\text{Random graph } \mathcal{G} \text{ exchangeable}$$
$$\Updownarrow$$
$$P(\mathcal{G}) = \int_{\mathcal{T}} P(\mathcal{G}|\theta) Q(d\theta)$$



[Ald81, Kal05]

Exchangeability: Random Graphs

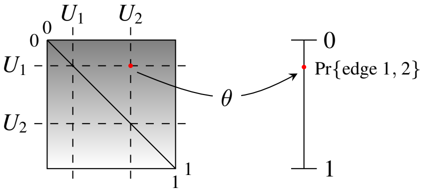
Random graph with independent edges

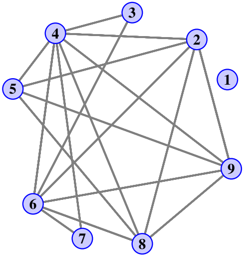
Given: $\theta : [0,1]^2 \to [0,1]$ symmetric function

- $U_1, U_2, \ldots \sim \text{Uniform}[0,1]$
- Edge (i,j) present:

$$(i,j) \sim \text{Bernoulli}(\theta(U_i, U_j))$$

Call this distribution $P(\mathcal{G}|\theta)$.





Aldous-Hoover Theorem

$$\text{Random graph } \mathcal{G} \text{ exchangeable}$$
$$\Updownarrow$$
$$P(\mathcal{G}) = \int_{\mathcal{T}} P(\mathcal{G}|\theta) Q(d\theta)$$

[Ald81, Kal05]

GENERAL THEME: SYMMETRY

Other types of exchangeable data

Data	Theorem	Mixture of...	Applications
Points	de Finetti	I.i.d. point sequences	“Standard” models
Sequences	Diaconis-Freedman	Markov chains	Time series
Partition	Kingman	"Paint-box" partitions	Clustering
Graphs	Aldous-Hoover	Graphs with independent edges	Networks
Arrays	Aldous-Hoover	Arrays with independent entries	Collaborative filtering

Ergodic decomposition theorems

$$\mu(X) = \int_{\Omega} \mu[X|\Phi = \phi]\nu(\phi)$$

- Symmetry (group invariance) on lhs $\longrightarrow$ Integral decomposition on rhs
- Permutation invariance on lhs $\longrightarrow$ Independence on rhs

[Kal05]

LATENT FEATURE MODELS

Latent feature models

- Grouping problem with overlapping clusters.
- Encode as binary matrix: Observation n in cluster $k \quad \Leftrightarrow \quad X_{nk} = 1$
- Alternatively: Item n possesses feature $k \quad \Leftrightarrow \quad X_{nk} = 1$

Indian buffet process (IBP)

1. Customer 1 tries $\text{Poisson}(\alpha)$ dishes.
2. Subsequent customer $n + 1$:
   - tries a previously tried dish k with probability $\frac{n_k}{n+1}$,
   - tries Poisson $\left(\frac{\alpha}{n+1}\right)$ new dishes.

Properties

- An exchangeable distribution over finite sets (of dishes).
- Intepretation:
  Observation (= customer) n in cluster (= dish) k if customer "tries dish k"

[GG06, GG11]

De Finetti Representation

Alternative description

$$\begin{array}{ccc} w_1 & \dots & w_K \\ \end{array}$$
$$\begin{pmatrix} X_{11} & \dots & X_{1K} \\ \vdots & & \vdots \\ X_{N1} & \dots & X_{NK} \end{pmatrix}$$

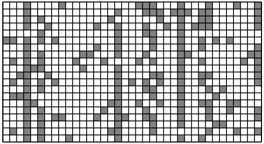
1. Sample $w_1, \dots, w_K \sim_{\text{iid}} \text{Beta}(1, \alpha/K)$
2. Sample $X_{1k}, \dots, X_{nk} \sim_{\text{iid}} \text{Bernoulli}(w_k)$

We need some form of limit object for $\text{Beta}(1, \alpha/K)$ for $K \to \infty$.

Beta Process (BP)

Distribution on objects of the form

$$\theta = \sum_{k=1}^{\infty} w_k \delta_{\phi_k} \qquad \text{with } w_k \in [0,1] \; .$$



- IBP matrix entries are sampled as $X_{nk} \sim_{\text{iid}} \text{Bernoulli}(w_k)$.
- Beta process is the de Finetti measure of the IBP, that is, $Q = \text{BP}$.
- θ is a random measure (but not normalized)

[GG06, TJ07, Hjo90a]

DIRICHLET PROCESS

- Set $[n] = \{1, 2, \ldots, n\}$.
- Partition: $\pi = \{\{1,2,5\},\{3,4\},\{6\},\{7,8,9\}\}$.

Kingman's representation

Exchangeable partitions $\Leftrightarrow$ Random probability measures

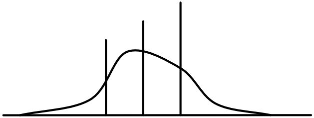
$$\theta = \text{Probability measure}$$

$$\text{For } i \in [n]: \quad \phi_i \,|\, \theta \sim \theta$$

$$i, j \text{ in the same cluster} \quad \Leftrightarrow \quad \phi_i = \phi_j$$

$$\mathbb{P}(\boldsymbol{\pi} = \pi) = \int_{\mathbf{M}(\Phi)} \mathbb{P}(\boldsymbol{\pi} = \pi | \theta) Q(d\theta)$$

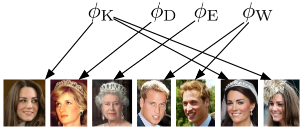
- Atoms in θ: clusters with more than one element.
- Smooth part of θ: clusters with exactly one element.



[Kin75, Pit06]

Chinese Restaurant Process for Clustering

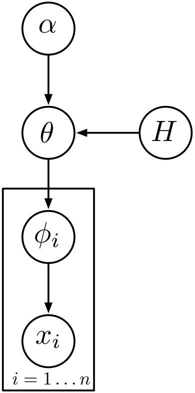
$$\pi = \{\{1,6,7\},\{2\},\{3\},\{4,5\}\}$$



- Full generative model:

$$\begin{aligned} \theta &\sim Q \\ \phi_i \,|\, \theta &\sim \theta \\ x_i \,|\, \phi_i &\sim F(\phi_i) \end{aligned}$$

- Prior Q is a Dirichlet process (DP) with mass parameter α and base distribution H.
- Two-parameter CRP: Pitman-Yor process (PYP) with additional discount parameter d.

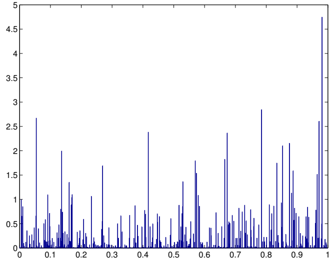


[Fer73, BM73, PPY92, PY97]

Dirichlet Process

- All clusters can contain more than one element $\Rightarrow$ θ only contains atoms:

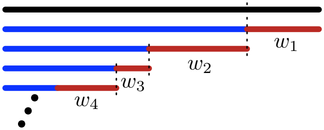
$$\theta = \sum_{j=1}^{\infty} w_j \delta_{\phi_j^*}$$



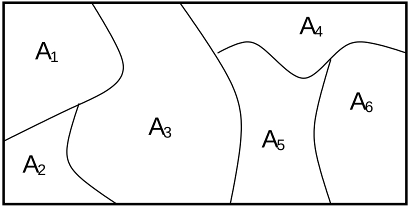
- What is the prior on $\{w_j, \phi_j^*\}$?
- Stick-breaking representation:

$$\begin{aligned} \phi_j^* &\sim H \\ v_j &\sim \mathrm{Beta}(1, \alpha) \end{aligned} \qquad w_j = v_j \prod_{i=1}^{j-1} (1 - v_j)$$

  Masses decreasing on average: GEM distribution.



- Strictly decreasing masses: Poisson-Dirichlet distribution.



- Random probability measure with Dirichlet marginals:

$$(\theta(A_1), \ldots, \theta(A_k)) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_k))$$

for $A_1, \ldots, A_k$ partition of the space.

 [Fer73, Orb11]

Completely Random Measures

$$\theta = \sum_{j=1}^{\infty} w_j \delta_{\phi_j^*}$$

Measure

- $\theta(S)$ – mass in set S.
- A function $\theta : \Omega \to \mathbb{R}_+$ with certain properties, e.g. if S, S' disjoint sets,

$$\theta(S \cup S') = \theta(S) + \theta(S')$$

Random Measure

- A random function $\theta : \Omega \to \mathbb{R}_+$.

Completely Random Measure (CRM)

- If S, S' are disjoint sets, then

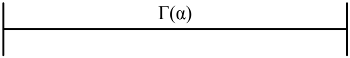
$$\theta(S) \perp\!\!\!\perp \theta(S')$$

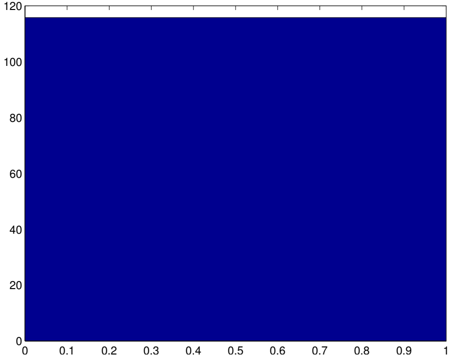
[Kin67]

Infinitely Divisible Distributions

- Random variable X is infinitely divisible if for every n, there exists n iid random variables $X_1, \ldots, X_n$ such that $\sum_{i=1}^{n} X_i = X$.
- Examples: Gaussian, gamma, Poisson, negative-binomial, Cauchy, stable.

Example: Gamma CRM

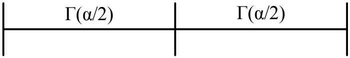


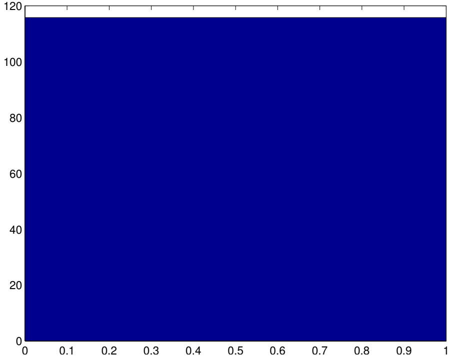


Infinitely Divisible Distributions

- Random variable X is infinitely divisible if for every n, there exists n iid random variables $X_1, \ldots, X_n$ such that $\sum_{i=1}^{n} X_i = X$.
- Examples: Gaussian, gamma, Poisson, negative-binomial, Cauchy, stable.

Example: Gamma CRM



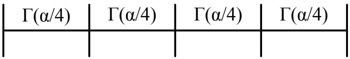


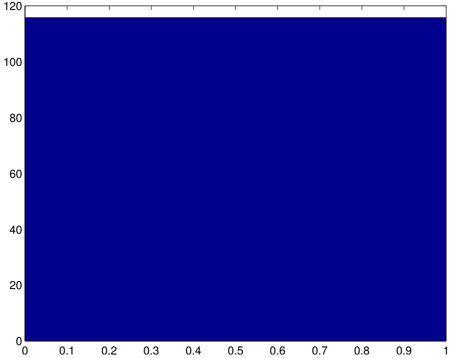
Completely Random Measures

Infinitely Divisible Distributions

- Random variable X is infinitely divisible if for every n, there exists n iid random variables $X_1, \ldots, X_n$ such that $\sum_{i=1}^{n} X_i = X$.
- Examples: Gaussian, gamma, Poisson, negative-binomial, Cauchy, stable.

Example: Gamma CRM

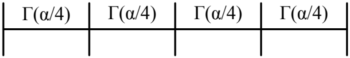


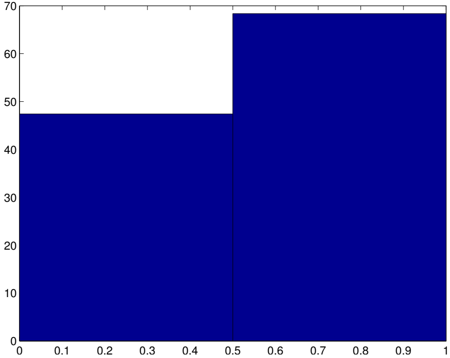


Infinitely Divisible Distributions

- Random variable X is infinitely divisible if for every n, there exists n iid random variables $X_1, \ldots, X_n$ such that $\sum_{i=1}^n X_i = X$.
- Examples: Gaussian, gamma, Poisson, negative-binomial, Cauchy, stable.

Example: Gamma CRM



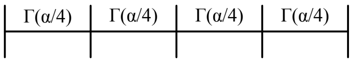


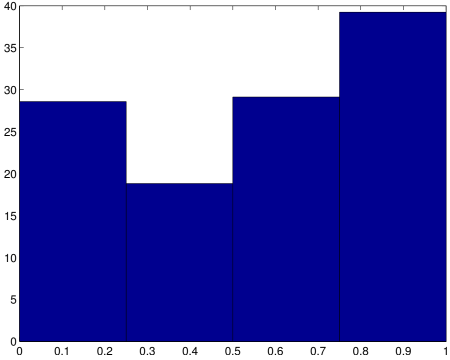
Completely Random Measures

Infinitely Divisible Distributions

- Random variable X is infinitely divisible if for every n, there exists n iid random variables $X_1, \ldots, X_n$ such that $\sum_{i=1}^{n} X_i = X$.
- Examples: Gaussian, gamma, Poisson, negative-binomial, Cauchy, stable.

Example: Gamma CRM

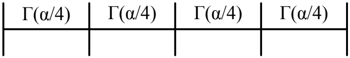


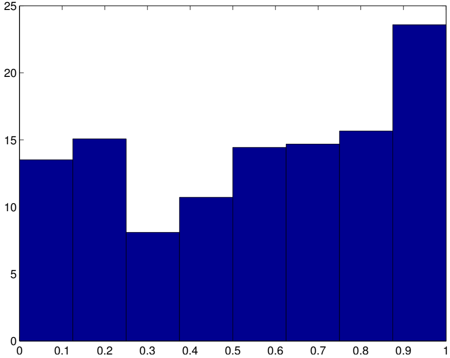


Infinitely Divisible Distributions

- Random variable X is infinitely divisible if for every n, there exists n iid random variables $X_1, \ldots, X_n$ such that $\sum_{i=1}^{n} X_i = X$.
- Examples: Gaussian, gamma, Poisson, negative-binomial, Cauchy, stable.

Example: Gamma CRM

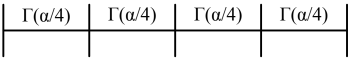


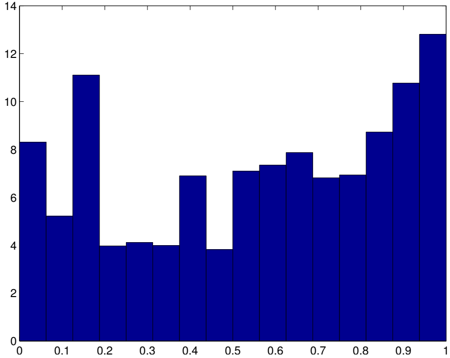


Infinitely Divisible Distributions

- Random variable X is infinitely divisible if for every n, there exists n iid random variables $X_1, \ldots, X_n$ such that $\sum_{i=1}^{n} X_i = X$.
- Examples: Gaussian, gamma, Poisson, negative-binomial, Cauchy, stable.

Example: Gamma CRM



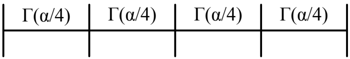


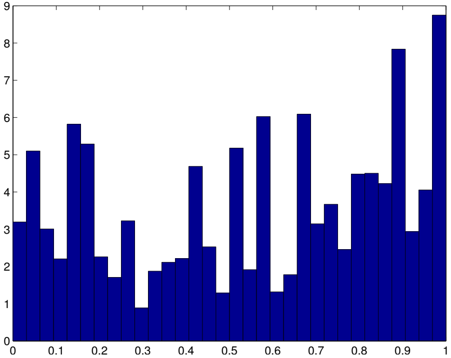
Completely Random Measures

Infinitely Divisible Distributions

- Random variable X is infinitely divisible if for every n, there exists n iid random variables $X_1, \ldots, X_n$ such that $\sum_{i=1}^{n} X_i = X$.
- Examples: Gaussian, gamma, Poisson, negative-binomial, Cauchy, stable.

Example: Gamma CRM



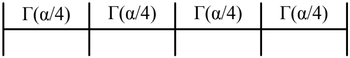


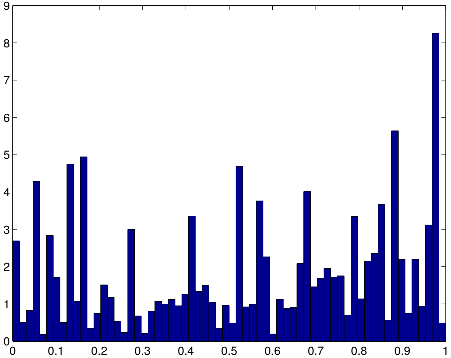
Completely Random Measures

Infinitely Divisible Distributions

- Random variable X is infinitely divisible if for every n, there exists n iid random variables $X_1, \ldots, X_n$ such that $\sum_{i=1}^{n} X_i = X$.
- Examples: Gaussian, gamma, Poisson, negative-binomial, Cauchy, stable.

Example: Gamma CRM



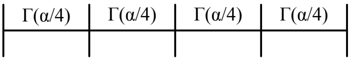


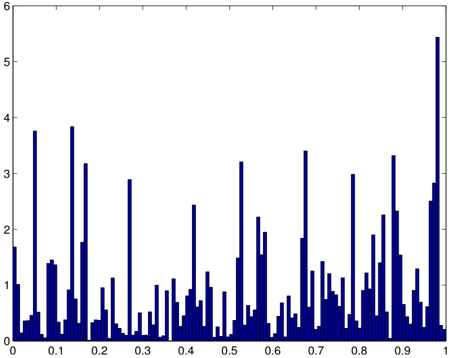
COMPLETELY RANDOM MEASURES

Infinitely Divisible Distributions

- Random variable X is infinitely divisible if for every n, there exists n iid random variables $X_1, \ldots, X_n$ such that $\sum_{i=1}^{n} X_i = X$.
- Examples: Gaussian, gamma, Poisson, negative-binomial, Cauchy, stable.

Example: Gamma CRM

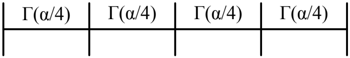


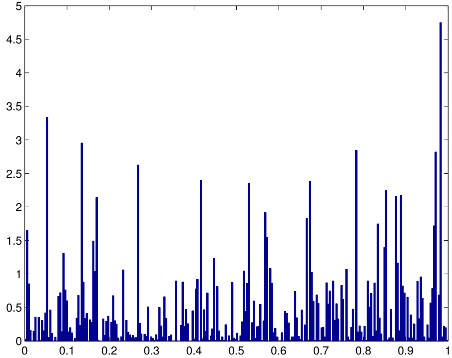


Infinitely Divisible Distributions

- Random variable X is infinitely divisible if for every n, there exists n iid random variables $X_1, \ldots, X_n$ such that $\sum_{i=1}^{n} X_i = X$.
- Examples: Gaussian, gamma, Poisson, negative-binomial, Cauchy, stable.

Example: Gamma CRM

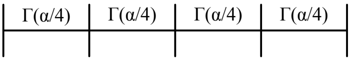


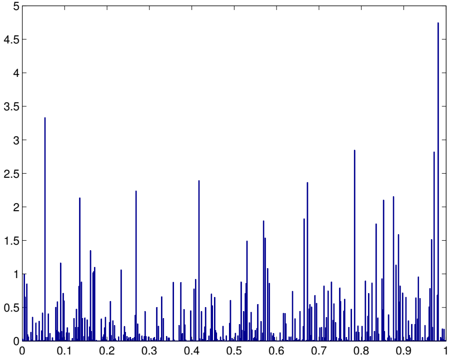


Infinitely Divisible Distributions

- Random variable X is infinitely divisible if for every n, there exists n iid random variables $X_1, \ldots, X_n$ such that $\sum_{i=1}^{n} X_i = X$.
- Examples: Gaussian, gamma, Poisson, negative-binomial, Cauchy, stable.

Example: Gamma CRM

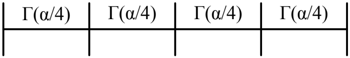


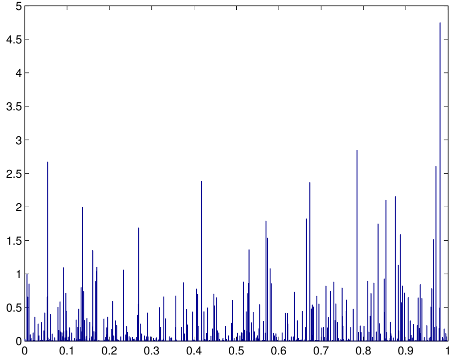


Infinitely Divisible Distributions

- Random variable X is infinitely divisible if for every n, there exists n iid random variables $X_1, \ldots, X_n$ such that $\sum_{i=1}^{n} X_i = X$.
- Examples: Gaussian, gamma, Poisson, negative-binomial, Cauchy, stable.

Example: Gamma CRM



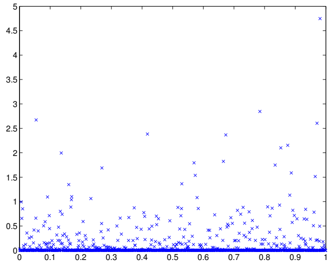


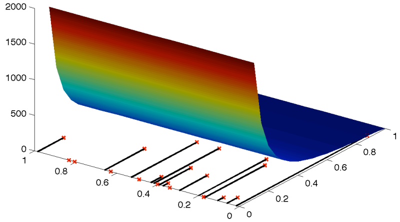
Completely Random Measures

A CRM can always be decomposed into 3 components:

$$\mu = \mu_0 + \sum_{i=1}^{\infty} v_i \delta_{\psi_i^*} + \sum_{j=1}^{\infty} w_j \delta_{\phi_j^*}$$

- μ_0 is measure that is not random.
- Locations $\{\psi_i^*\}$ are fixed, masses $\{v_i\}$ are mutually independent and independent of $\{w_j, \phi_j^*\}$,
- $\{(w_j, \phi_j^*)\}$ is drawn from a Poisson process over $\mathbb{R}_+ \times \Phi$ with rate $\rho(w, \phi)$ (the Lévy measure).





[Kin67]

Completely Random Measures

- Gamma Process

$$\rho(w,\phi) = \alpha w^{-1} e^{-w} h(\phi)$$

  - Normalizing a Gamma process $\Rightarrow$ Dirichlet process.

- Beta Process [Hjo90b]

$$\rho(w,\phi) = \alpha w^{-1} \mathbf{1}(0 \le w \le 1) h(\phi)$$

- Stable process [Kin75]

$$\rho(w,\phi) = \frac{\alpha}{\Gamma(1-d)} w^{-d-1} h(\phi)$$

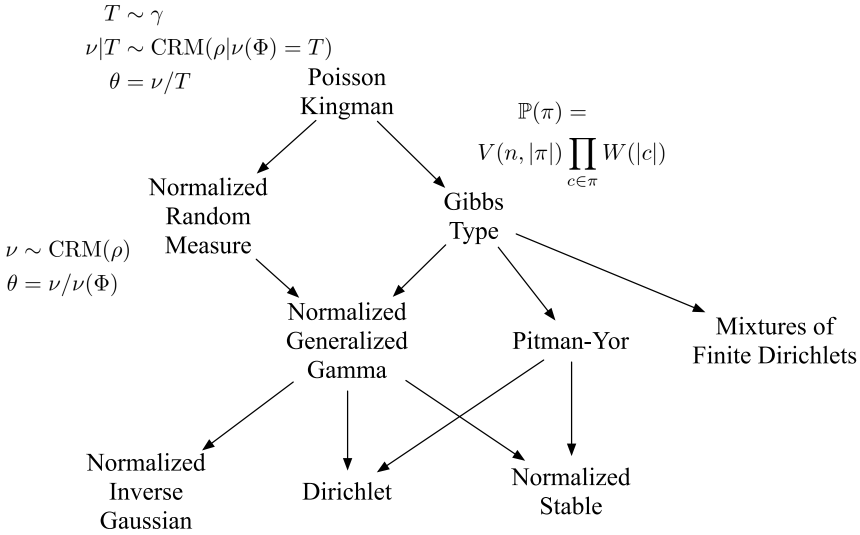
- Stable-beta process [KL01, TG09, BJPar]

$$\rho(w,\phi) = \frac{\alpha\Gamma(1+\beta)}{\Gamma(1-d)\Gamma(\beta+d)} w^{-d-1}(1-w)^{\beta+d-1} \mathbf{1}(0 \le w \le 1) h(\phi)$$

- Generalized gamma process [**?**]

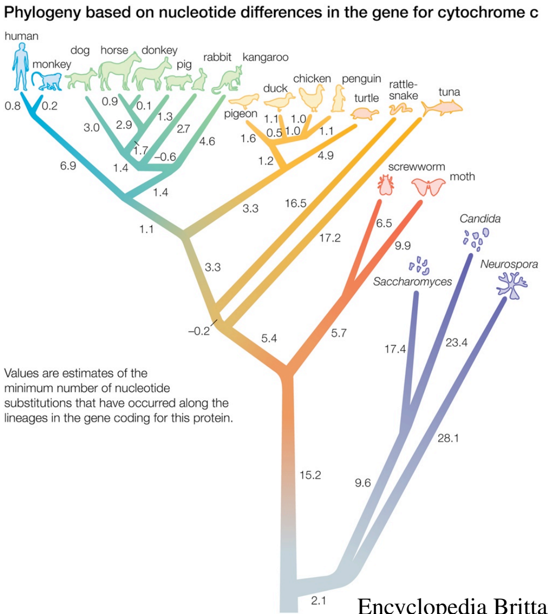
$$\rho(w,\phi) = \frac{\alpha}{\Gamma(1-d)} w^{-d-1} e^{-\tau w} h(\phi)$$

FAMILIES OF EXCHANGEABLE RANDOM PARTITIONS



[Kin75, Pit03, LMP05, GP06, JLP09, FW11]

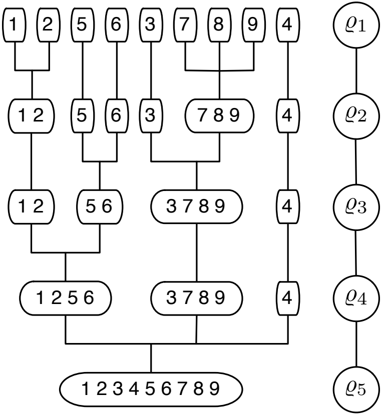
Hierarchical Partitions



Encyclopedia Brittanica (Heredity)

Bayesian Hierarchical Clustering

- Bayesian approach to hierarchical clustering:
  - Prior over hierarchies T.
  - Likelihood model for data.
- Necessarily nonparametric.
- Prior can be described by Markov chain of partitions.



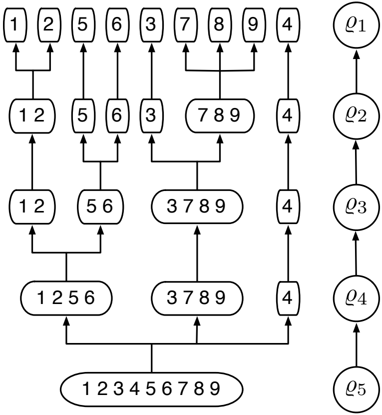
Fragmentation Processes

- Start with $\varrho_L = \{[n]\}$.
- At each stage, fragment each cluster into smaller clusters.
- A fragmentation can be described by independent partitionings of clusters at previous stage.

For each $c \in \varrho_i$: $\quad F_c \sim \mathrm{CRP}(\alpha, d, c)$

$$\varrho_{i-1} = \bigcup_{c \in \varrho_i} F_c$$

- Nested Chinese restaurant process [BGJ10], tree-structured stick-breaking [AGJ10].

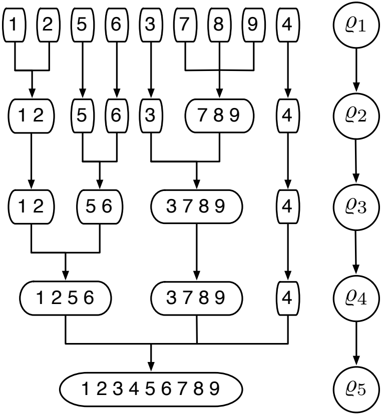


- Start with $\varrho_1 = [n]$.
- At each stage, coagulate clusters to form larger clusters.
- A coagulation can be described by a partitioning of clusters at previous stage.

$$C \sim \mathrm{CRP}(\alpha, d, \varrho_i)$$

$$\varrho_{i+1} = \left\{ \bigcup_{c' \in c} c' : c \in C \right\}$$

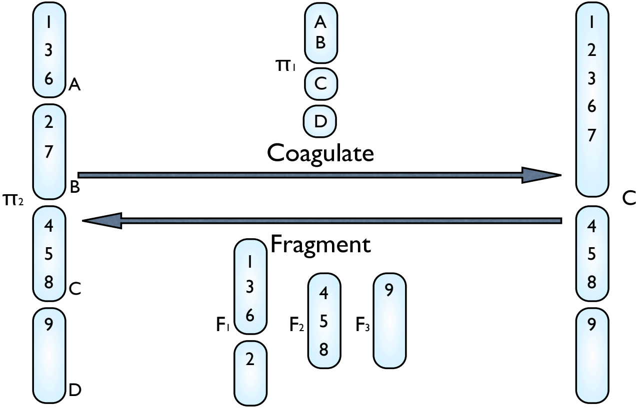
- Chinese restaurant franchise [TJBB06].



Random Hierarchical Partitions

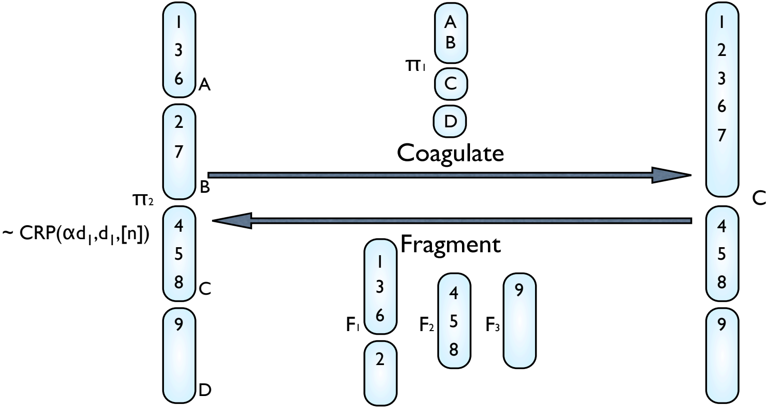
	Discrete iterations	Continuum limit
Fragmentation	Nested CRP, tree-structured stick-breaking Gibbs fragmentation tree [BGJ10, AGJ10, MPW08]	Dirichlet diffusion tree, Pitman-Yor diffusion tree [Nea03, KG11]
Coagulation	Chinese restaurant franchise [TJBB06]	Kingman's coalescent, Λ-coalescent [Kin82, Pit99, TDR08]

Fragmentation-Coagulation Duality

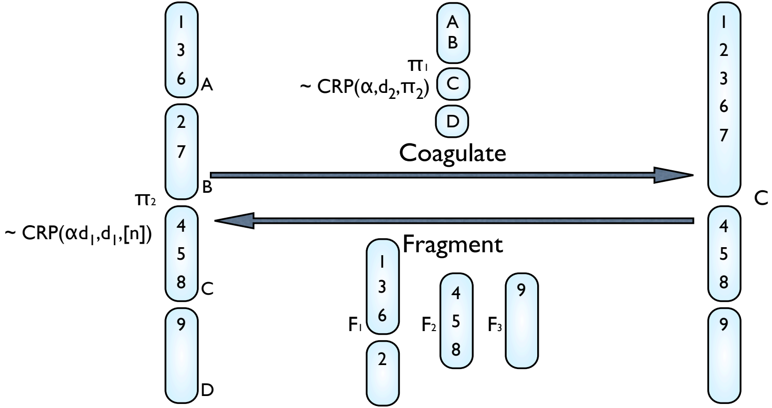


[Pit99, Jam10]

FRAGMENTATION-COAGULATION DUALITY

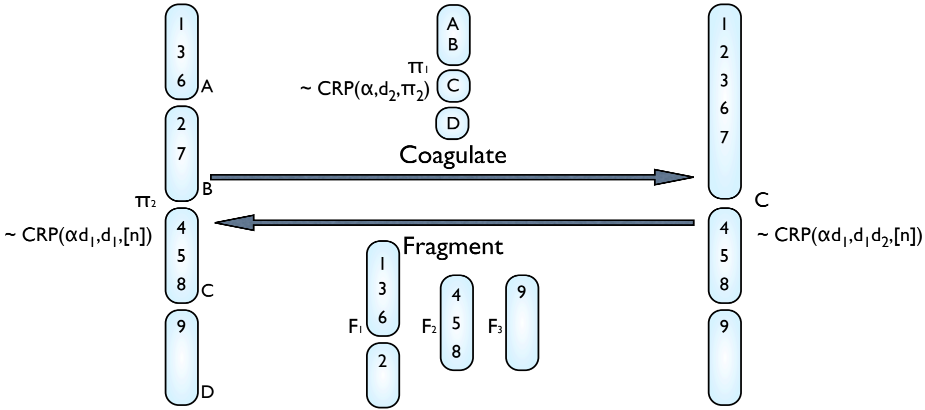


FRAGMENTATION-COAGULATION DUALITY



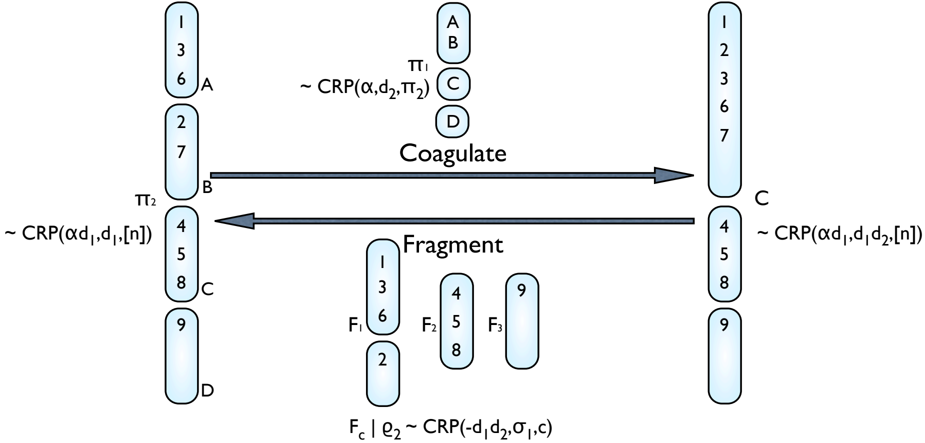
 [Pit99, Jam10]

FRAGMENTATION-COAGULATION DUALITY



[Pit99, Jam10]

FRAGMENTATION-COAGULATION DUALITY



[Pit99, Jam10]

Hierarchical Pitman-Yor Process

- Fragmentation-coagulation duality implies:

$$\begin{aligned} G_1|G_0 &\sim \mathrm{PYP}(\alpha, d_2, G_0) \\ G_2|G_1 &\sim \mathrm{PYP}(\alpha d_1, d_1, G_1) \end{aligned} \quad \Rightarrow \quad G_2|G_0 \sim \mathrm{PYP}(\alpha d_1, d_1 d_2, G_0)$$

  - Computational implication: sequence memoizer [WAG$^+$09].
- Dirichlet Process case:

$$\begin{aligned} G_1|G_0 &\sim \mathrm{DP}(\alpha/d, G_0) \\ G_2|G_1 &\sim \mathrm{PYP}(\alpha, d_1, G_1) \end{aligned} \quad \Rightarrow \quad G_2|G_0 \sim \mathrm{DP}(\alpha, G_0)$$

  - Modelling implication: hierarchical Dirichlet process (HDP) [TJBB06].

Concluding Remarks

SUMMARY

Why Bayesian Nonparametrics?

- World is complicated.
- Objects of interest often infinite dimensional.
- Alternative to model selection.
- Flexible modelling language with interesting properties.
- Works well with finite data while enjoying asymptotic guarantees.

Technical Tools

- Stochastic processes.
- Exchangeability.
- Graphical, hierarchical and dependent models.

Open Challenges

- Novel models and useful applications.
- Better inference and flexible software packages.
- Learning theory for Bayesian nonparametric models.

References I

[AGJ10] Ryan Prescott Adams, Zoubin Ghahramani, and Michael I. Jordan. Tree-structured stick breaking for hierarchical data. In J. Shawe-Taylor, R. Zemel, J. Lafferty, and C. Williams, editors, *Advances in Neural Information Processing (NIPS) 23*, 2010.

[Ald81] David J. Aldous. Representations for Partially Exchangeable Arrays of Random Variables. *Journal of Multivariate Analysis*, 11:581–598, 1981.

[BC11] P. Blunsom and T. Cohn. A hierarchical Pitman-Yor process HMM for unsupervised part of speech induction. In *Proceedings of ACL HLT*, 2011.

[BGJ10] D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the Association for Computing Machines*, 57(2):1–30, 2010.

[BJPar] T. Broderick, M. I. Jordan, and J. Pitman. Beta processes, stick-breaking, and power laws. *Bayesian Analysis*, to appear.

[BM73] D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1:353–355, 1973.

[Dun10] D. B. Dunson. Nonparametric Bayes applications to biostatistics. In N. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics*. Cambridge University Press, 2010.

[Fer73] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230, 1973.

[FSJW10] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Sharing features among dynamical systems with beta processes. In *Neural Information Processing Systems 22*. MIT Press, 2010.

[FW11] S. Favaro and S. G. Walker. Slice sampling σ–stable Poisson–Kingman mixture models, 2011.

[GG06] T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, volume 18, 2006.

[GG11] T. L. Griffiths and Z. Ghahramani. The Indian buffet process: An introduction and review. *J. Mach. Learn. Res.*, 12:1185–1224, 2011.

[GGJ06a] S. Goldwater, T. L. Griffiths, and M. Johnson. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 2006.

[GGJ06b] S. Goldwater, T.L. Griffiths, and M. Johnson. Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems*, volume 18, 2006.

REFERENCES II

[Gho10] S. Ghosal. Dirichlet process, related priors and posterior asymptotics. In N. L. Hjort et al., editors, *Bayesian Nonparametrics*, pages 36–83. Cambridge University Press, 2010.

[GKM05] A. E. Gelfand, A. Kottas, and S. N. MacEachern. Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035, 2005.

[GP06] A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5684, 2006.

[GvdV07] Subhashis Ghosal and Aad van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723, 2007.

[GWT10] J. Gasthaus, F. Wood, and Y. W. Teh. Lossless compression based on the sequence memoizer. In *Data Compression Conference*, 2010.

[Hjo90a] N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.*, 18:1259–1294, 1990.

[Hjo90b] N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, 18(3):1259–1294, 1990.

[IZ02] H. Ishwaran and M. Zarepour. Exact and approximate sum-representations for the Dirichlet process. *Canadian Journal of Statistics*, 30:269–283, 2002.

[Jam10] L. F. James. Coag-frag duality for a class of stable Poisson-Kingman mixtures. http://arxiv.org/abs/1008.2420, 2010.

[JLP09] L. F. James, A. Lijoi, and I. Prüenster. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36:76–97, 2009.

[Kal05] Olav Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Springer, 2005.

[KG11] D. Knowles and Z. Ghahramani. Pitman-Yor diffusion trees. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2011.

[Kin67] J. F. C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.

[Kin75] J. F. C. Kingman. Random discrete distributions. *Journal of the Royal Statistical Society*, 37:1–22, 1975.

[Kin82] J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43, 1982. Essays in Statistical Science.

REFERENCES III

[KL01] Y. Kim and J. Lee. On posterior consistency of survival models. *Annals of Statistics*, 29(3):666–686, 2001.

[KvdV06] B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*, 34(2):837–877, 2006.

[LMP05] A. Lijoi, R. H. Mena, and I. Prüenster. Hierarchical mixture modelling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, 100:1278–1291, 2005.

[MPW08] P. McCullagh, J. Pitman, and M. Winkel. Gibbs fragmentation trees. *Bernoulli*, 14(4):988–1002, 2008.

[MYU09] D. Mochihashi, T. Yamada, and N. Ueda. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of ACL-IJCNLP*, 2009.

[Nea92] R. M. Neal. Bayesian mixture modeling. In *Proceedings of the Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, volume 11, pages 197–211, 1992.

[Nea00] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

[Nea03] R. M. Neal. Density modeling and clustering using Dirichlet diffusion trees. In *Bayesian Statistics*, volume 7, pages 619–629, 2003.

[Orb11] Peter Orbanz. Projective limit random probabilities on Polish spaces. *Electronic Journal of Statistics*, 5:1354–1373, 2011.

[Pit99] J. Pitman. Coalescents with multiple collisions. *Annals of Probability*, 27:1870–1902, 1999.

[Pit03] J. Pitman. Poisson-Kingman partitions. In D. R. Goldstein, editor, *Statistics and Science: a Festschrift for Terry Speed*, pages 1–34. Institute of Mathematical Statistics, 2003.

[Pit06] J. Pitman. *Combinatorial Stochastic Processes*. Lecture Notes in Mathematics. Springer-Verlag, Berlin, 2006.

[PPY92] M. Perman, J. Pitman, and M. Yor. Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39, 1992.

[PY97] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900, 1997.

[Ras00] C. E. Rasmussen. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems*, volume 12, 2000.

REFERENCES IV

[RW06] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[Sch65] L. Schwartz. On Bayes procedures. *Z. Wahr. Verw. Gebiete*, 4:10–26, 1965.

[Sch95] M. J. Schervish. *Theory of Statistics*. Springer, 1995.

[Set94] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[TDR08] Y. W. Teh, H. Daume III, and D. M. Roy. Bayesian agglomerative clustering with coalescents. In *Advances in Neural Information Processing Systems*, volume 20, pages 1473–1480, 2008.

[Teh06] Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992, 2006.

[TG09] Y. W. Teh and D. Görür. Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems*, volume 22, pages 1838–1846, 2009.

[TJ07] R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In *J. Mach. Learn. Res. Proceedings (AISTATS)*, volume 2, pages 564–571. 2007.

[TJ10] Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. In N. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics*. Cambridge University Press, 2010.

[TJBB06] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[vdV98] A. van der Vaart. *Asymptotic Statistics*. Cambrdige University Press, 1998.

[vdVvZ08a] A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463, 2008.

[vdVvZ08b] A. W. van der Vaart and J. H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, volume 3 of *Inst. Math. Stat. Collect.*, pages 200–222. Inst. Math. Statist., Beachwood, OH, 2008.

[WAG+09] F. Wood, C. Archambeau, J. Gasthaus, L. F. James, and Y. W. Teh. A stochastic memoizer for sequence data. In *Proceedings of the International Conference on Machine Learning*, volume 26, pages 1129–1136, 2009.