Bayesian Nonparametrics

Part I

Peter Orbanz

OVERVIEW

Today

1. Basic terminology
2. Clustering
3. Latent feature models

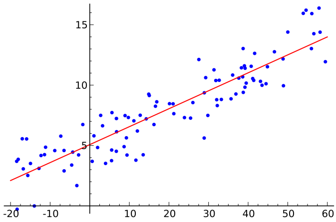
Tomorrow

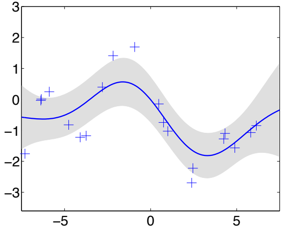
5. Constructing nonparametric Bayesian models
6. Exchangeability
7. Asymptotics

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

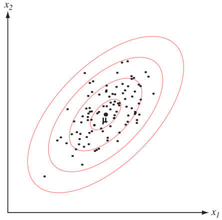
Parametric model

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

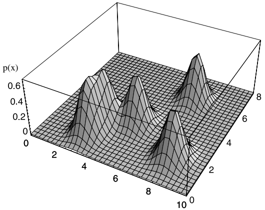
Nonparametric model

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

Example: Density estimation



Parametric



Nonparametric

Nonparametric Bayesian model

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:

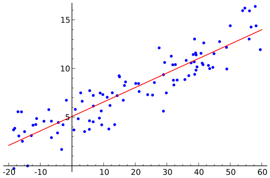
Problem	$\mathcal{T}$
Density estimation	Probability distributions
Regression	Smooth functions
Clustering	Partitions

Solution to Bayesian problem = posterior distribution on patterns

Can we justify our assumptions?

Recall:

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



In Bayes' theorem:

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

Definition

$X_1, X_2, \ldots$ 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

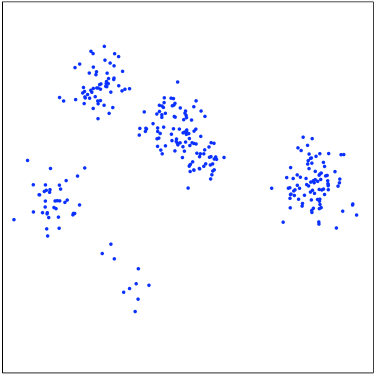
Implications

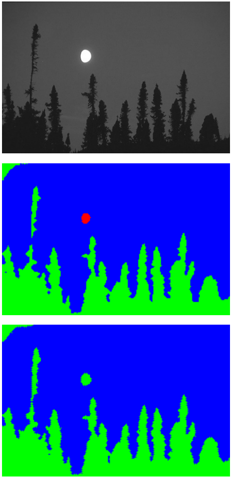
- Exchangeable data decomposes into pattern and noise
- More general than i.i.d.-assumption
- Caution: θ is in general an ∞-dimensional quantity

CLUSTERING

CLUSTERING

- Observations $X_1, X_2, \ldots$
- Each observation belongs to exactly one cluster
- Unknown pattern = partition of $\{1, \ldots, n\}$ or $\mathbb{N}$





Mixture models

$$p(x|m) = \int_{\Omega_\theta} p(x|\theta)m(d\theta)$$

m is called the *mixing measure*

Two-stage sampling

Sample $X \sim p(\,.\,|m)$ as:

1. $\Theta \sim m$
2. $X \sim p(\,.\,|\theta)$

Finite mixture model

$$p(x|\boldsymbol{\theta}, \mathbf{c}) = \int_{\Omega_\theta} p(x|\theta)m(d\theta) \qquad \text{with} \qquad m(\,.\,) = \sum_{k=1}^{K} c_k \delta_{\theta_k}(\,.\,)$$

Random mixing measure

$$M(\,.\,) = \sum_{k=1}^{K} C_k \delta_{\Theta_k}(\,.\,)$$

Conjugate priors

A Bayesian model is *conjugate* if the posterior is an element of the same class of distributions as the prior ("closure under sampling").

$p(x\|\theta)$	conjugate prior
$\frac{1}{Z(\theta)} h(x) \exp(\langle S(x), \theta \rangle)$	$\frac{1}{K(\lambda, y)} \exp(\langle \theta, y \rangle - \lambda \log Z(\theta))$
Gaussian	Gaussian/inverse Wishart
multinomial	Dirichlet
...	...

Choice of priors in BMM

- Choose conjugate prior for each parameter
- In particular: Dirichlet prior on $(C_1, \ldots, C_k)$

Dirichlet process

A Dirichlet process is a distribution on random probability measures of the form

$$M(\,.\,) = \sum_{k=1}^{\infty} C_k \delta_{\Theta_k}(\,.\,) \qquad \text{where} \qquad \sum_{k=1}^{\infty} C_k = 1$$

Constructive definition of DP (α, G_0)

$$\begin{aligned} \Theta_k &\sim_{\text{iid}} G_0 \\ V_k &\sim_{\text{iid}} \text{Beta}(1, \alpha) \end{aligned}$$

Compute C_k as

$$C_k := V_k \prod_{i=1}^{k-1} (1 - V_i)$$

"Stick-breaking construction"

POSTERIOR DISTRIBUTION

DP Posterior

$$\theta_{n+1}|\theta_1,\ldots,\theta_n \sim \frac{1}{n+\alpha}\sum_{j=1}^{n}\delta_{\theta_j}(\theta_{n+1}) + \frac{\alpha}{n+\alpha}G_0(\theta_{n+1})$$

Mixture Posterior

$$p(x_{n+1}|x_1,\ldots,x_n) = \sum_{k=1}^{K_n}\frac{n_k}{n+\alpha}p(x_{n+1}|\theta_k^*) + \frac{\alpha}{n+\alpha}\int p(x_{n+1}|\theta)G_0(\theta)d\theta$$

Conjugacy

- The posterior of DP (α, G_0) is DP $\left(\alpha+n, \frac{1}{n+\alpha}(\sum_k n_k\delta_{\theta_k^*} + \alpha G_0)\right)$
- Hence: The Dirichlet process is conjugate.

Latent variables

$$p(x_{n+1}|x_1,\dots,x_n) = \sum_{k=1}^{K_n} \frac{n_k}{n+\alpha} p(x_{n+1}|\theta_k^*) + \frac{\alpha}{n+\alpha} \int p(x_{n+1}|\theta) G_0(\theta) d\theta$$

We do not actually observe the Θ_j (they are latent). We observe X_j.

Assignment probabilities

$$\begin{pmatrix} q_{10} & q_{11} & \dots & q_{1K_n} \\ \vdots & \vdots & & \vdots \\ q_{n0} & q_{n1} & \dots & q_{nK_n} \end{pmatrix}$$

Where:

- $q_{jk} \propto n_k p(x_j|\theta_k^*)$
- $q_{j0} \propto \alpha \int p(x_j|\theta) G_0(\theta) d\theta$

Gibbs Sampling

Uses an assignment variable ϕ_j for each observation X_j.

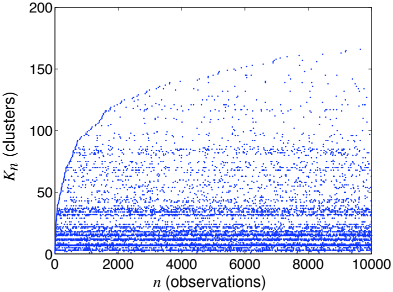
- Assignment step: Sample $\phi_j \sim \text{Multinomial}(q_{j0},\dots,q_{jK_n})$
- Parameter sampling: $\theta_k^* \sim G_0(\theta_k^*) \prod_{x_j \in \text{Cluster } k} p(x_j|\theta_k^*)$

Number of Clusters

Dirichlet process

$K_n = \#$ clusters in sample of size n

$$\mathbb{E}[K_n] = O(\log(n))$$



Modeling assumption

- Parametric clustering: K_∞ is *finite* (possibly unknown, but fixed).
- Nonparametric clustering: K_∞ is *infinite*

Rephrasing the question

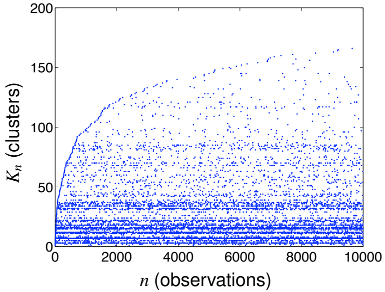
- Estimate of K_n is controlled by distribution of the cluster sizes C_k in $\sum_k C_k \delta_{\Theta_k}$.
- Ask instead: What should we assume about the distribution of C_k?

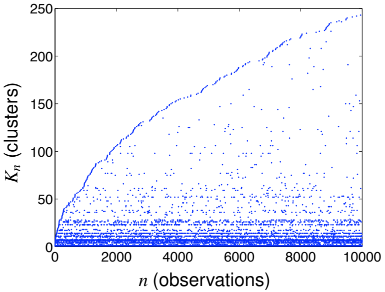
Pitman-Yor process

$$p(x_{n+1}|x_1,\ldots,x_n) = \sum_{k=1}^{K_n} \frac{n_k - d}{n+\alpha} p(x_{n+1}|\theta_k^*) + \frac{\alpha + K_n \cdot d}{n+\alpha} \int p(x_{n+1}|\theta) G_0(\theta) d\theta$$

Discount parameter $d \in [0, 1]$.

Cluster sizes





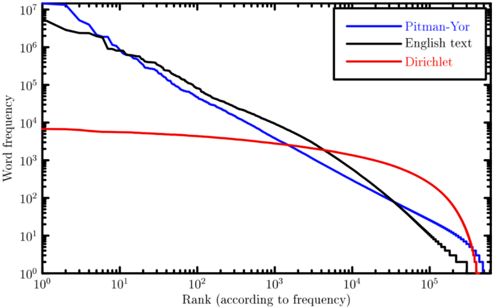
The distribution of cluster sizes is called a *power law* if

$$C_j \sim \gamma(\beta) \cdot j^{-\beta} \qquad \text{for some } \beta \in [0,1] \; .$$

Examples of power laws

- Word frequencies
- Popularity (number of friends) in social networks

Pitman-Yor language model



Discrete measures and partitions

Sampling from a discrete measure determines a *partition* of $\mathbb{N}$ into blocks b_k:

$$\Theta_n \sim_{\text{iid}} \sum_{k=1}^{\infty} c_k \delta_{\theta_k^*} \qquad \text{and set} \qquad n \in b_k \quad \Leftrightarrow \quad \Theta_n = \theta_k^*$$

As $n \longrightarrow \infty$, the block proportions converge: $\frac{|b_k|}{n} \longrightarrow c_k$

Induced random partition

The distribution of a random discrete measure $M = \sum_{k=1}^{\infty} C_k \delta_{\Theta_k}$ induces the distribution of a *random partition* $\Pi = (B_1, B_2, \dots)$.

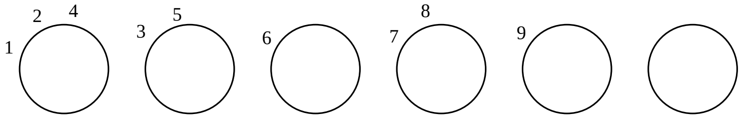
Exchangeable random partitions

- Π is called *exchangeable* if its distribution depends only on the sizes of its blocks.
- All exchangeable random paritiitions, and only those, can be represented by a random discrete distribution as above (Kingman's theorem).

Chinese Restaurant Process

The distribution of the random partition induced by the Dirichlet process is called the *Chinese Restaurant Process*.

"Customers and tables" analogy

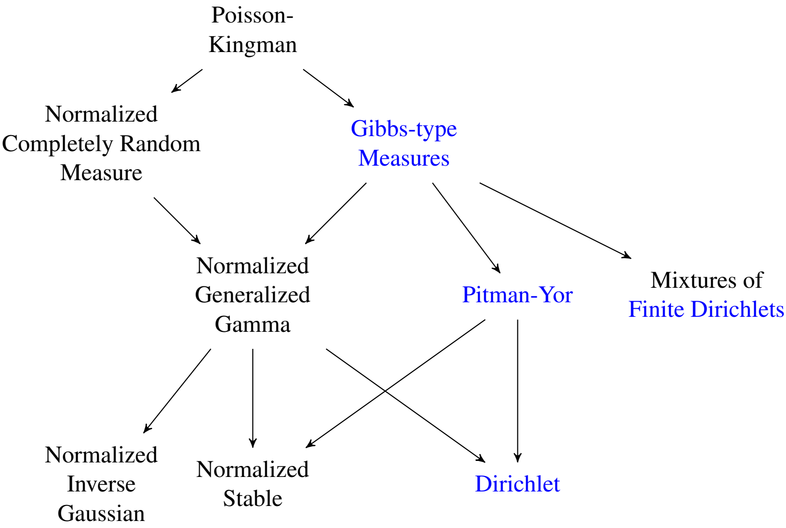


Customers = observations (indices in $\mathbb{N}$)
Tables = clusters (blocks)

Historical remark

- Originally introduced by Dubins & Pitman as a distribution on infinite permutations
- A permutation of n items defines a partition of $\{1, \ldots, n\}$ (regard cycles of permutation as blocks of partition)
- The induced distribution on partitions is the CRP we use in clustering

Families of Exchangeable Random Partitions



[Kin75, Pit03, LMP05a, GP06a, JLP09]

Classification (due to Prünster)

class	probability of new cluster	prior class
I	$\mathbb{P}\{\Theta_{n+1} \in \text{new cluster} \mid \Theta^{(n)}\} = f(n)$	Dirichlet processes
II	$\mathbb{P}\{\Theta_{n+1} \in \text{new cluster} \mid \Theta^{(n)}\} = f(n, K_n)$	Gibbs-type measures
III	$\mathbb{P}\{\Theta_{n+1} \in \text{new cluster} \mid \Theta^{(n)}\} = f(n, K_n, \mathbf{n})$	

General partition priors

- Gibbs-type measures are completely classified [GP06b]
- Properties of some cases well-studied, e.g.:
  - Dirichlet process
  - Pitman-Yor process
  - Normalized inverse Gaussian process [LMP05b]
- In the future: We will have a range of models which express different prior assumptions on the distribution of cluster sizes.

Summary: Clustering

Nonparametric Bayesian clustering

- Infinite number of clusters, $K_n \leq n$ of which are observed.
- If partition exchangeable, it can be represented by a random discrete distribution.

Inference

Latent variable algorithms, since assignments ($\equiv$ partition) not observed.

- Gibbs sampling
- Variational algorithms

Prior assumption

- Distribution of cluster sizes.
- Implies prior assumption on number K_n of clusters.

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

Properties

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

DE FINETTI REPRESENTATION

Alternative description

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

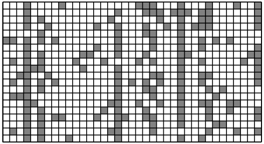
$$\begin{array}{ccc} w_1 & \dots & w_K \\ \end{array}$$
$$\begin{pmatrix} X_{11} & \dots & X_{1K} \\ \vdots & & \vdots \\ X_{N1} & \dots & X_{NK} \end{pmatrix}$$

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)

REFERENCES I

[FLP12] S. Favaro, A. Lijoi, and I. Prünster. Conditional formulae for Gibbs-type exchangeable random partitions. *Ann. Appl. Probab.* To appear., 2012.

[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.

[GHP07] A. V. Gnedin, B. Hansen, and J. Pitman. Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws. *Probability Surveys*, 4:146–171, 2007.

[GP06a] A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5684, 2006.

[GP06b] A. Gnedin and J. Pitman. Exchangeable Gibbs partititions and Stirling triangles. *J. Math. Sci.*, 138(3):5674–5685, 2006.

[Hjo90] N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.*, 18:1259–1294, 1990.

[IJ01] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[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.

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

[LMP05a] 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.

[LMP05b] A. Lijoi, R. H. Mena, and I. Prünster. Hierarchical mixture modeling with normalized inverse-Gaussian priors. *J. Amer. Statist. Assoc.*, 100:1278–1291, 2005.

[LP10] A. Lijoi and I. Prünster. Models beyond the Dirichlet process. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics*. Cambrdige University Press, 2010.

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

References II

[Pem07] R. Pemantle. A survey of random processes with reinforcement. *Probab. Surv.*, 4:1–79, 2007.

[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.

[Rob95] C. P. Robert. Mixtures of distributions: inference and estimation. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1995.

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

[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.

[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.

Bayesian Nonparametrics

Part II

Peter Orbanz

OVERVIEW

1. Constructing nonparametric Bayesian models
    - Hierarchical and dependent models
    - Representations
    - Exchangeability
2. Asymptotics

New models from old ones

Apply Bayesian representation recursively

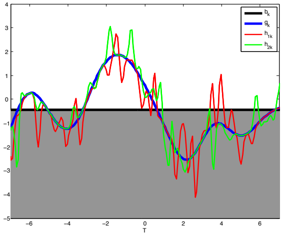
Split parameter Θ:

$$\Theta \quad \to \quad \Psi \;\text{ and }\; \Theta|\Psi$$

$$\begin{tikzcd}[column sep=small]
& P(X_1, X_2, \dots) \arrow[dl] \arrow[dr] & & \\
P(X_1, X_2, \dots |\Theta) & & Q(\Theta) \arrow[dl] \arrow[dr] & \\
& Q(\Theta|\Psi) & & R(\Psi)
\end{tikzcd}$$

Example: Hierarchical Gaussian process

- Sample $\Psi \sim R$
  (large length-scale, mean 0)
- Sample $\Theta|\Psi \sim Q(\,.\, |\Psi)$
  (smaller length scale, mean Ψ)



Decomposes underlying pattern:

- Low-frequency component Ψ
- High-frequency component Θ

Sampling scheme

- Sample $G_0 \sim \mathrm{DP}(\gamma, H)$
- Sample $G_1, G_2, \ldots \sim \mathrm{DP}(\alpha, G_0)$
- Sample $x_{ij} \sim G_j$

$G_1, G_2, \ldots$ have common "vocabulary" of atoms

Application: Nonparametric LDA

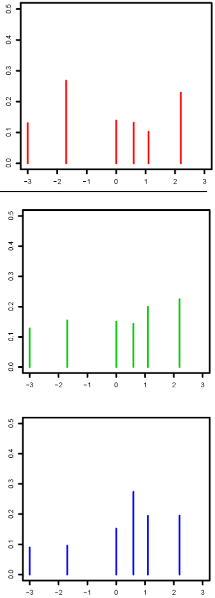
$$G_0 = \sum_{k=1}^{\infty} C_k \delta_{\Theta_k^*} \qquad G_j = \sum_{l=1}^{\infty} D_l^j \delta_{\Phi_l^j}$$

- Θ_k = finite probability (="topic")
- C_k = occurence probability of topic k
- Document j drawn from weighted combination of topics, with proportions D_l^j ("admixture model")

G_0

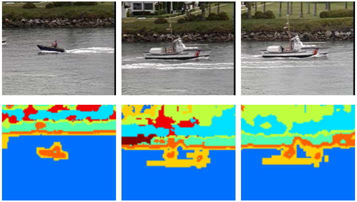
G_1

G_2



Setting

- Solution (= pattern) depends on a *covariate*, e.g. time, space,...
- Example: Video segmentation



For each frame: Solution is a segmentation, i.e. a clustering

Covariate-dependent clustering

$$M(\,.\,,t) = \sum_{k=1}^{\infty} C_k(t)\delta_{\Theta_k(t)}(\,.\,)$$

for each covariate value t.

Dependent Dirichlet process

Model functions $C : T \rightarrow [0,1]$ and $\Theta : T \rightarrow \Omega_\theta$ with Gaussian processes.

1. Transform GP to have $\text{Beta}(1, \alpha(t))$ marginal distribution for each t.
2. Sample functions $V_1(t), V_2(t), \ldots$ from this process.
3. $C_k(t) := V_k(t) \prod_{i=1}^{k-1}(1 - V_i(t))$

Properties

- Marginal at t is DP $(\alpha(t), G_t)$ with Gaussian base measure G_t.
- Clustering solutions vary smoothly in t.

Covariate-dependent models: General theme

- Random object $\Psi \in \Omega_\psi$ with distribution P, covariate space T.
- Covariate-dependent P: Distribution of random mapping $\hat{\Psi} : T \rightarrow \Omega_\psi$.

EXAMPLES

Applications	Pattern	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
...	...	...

REPRESENTATIONS

Density Representations

Densities

$$P(dx) = p(x)\lambda(dx) \qquad\qquad P(A) = \int_A p(x)\lambda(dx)$$

We call λ the *carrier measure* and p the *density* of P w.r.t. λ.

Useful carrier measures

- λ should be translation-invariant.
- Such measures exist only on certain spaces, roughly speaking: On finite-dimensional spaces.

Consequence: Representation problem 1

- Nonparametric models: No useful carrier measure on parameter space.
- We have to find alternatives to density representation.

Bayesian model: General case

Prior distribution Q, likelihood $P[X \in \,.\,|\Theta]$, posterior $Q[\Theta \in \,.\,|X = x]$

Bayes' Theorem

If the posterior has a density w.r.t. the prior for each x, then

$$Q[d\theta|X = x] \quad = \quad \frac{dQ[\,.\,|X = x]}{dQ(\,.\,)}Q(d\theta) \quad = \quad \frac{dP[X \in \,.\,|\theta]}{dP(X \in \,.\,)}(x)Q(d\theta)$$

The "Bayes equation" is a density of the posterior with respect to the prior.

Representation Problem 2

- For many nonparametric models, this density cannot exist for all x.
- Such models are called *undominated*.
- Random discrete measure models are generally undominated.

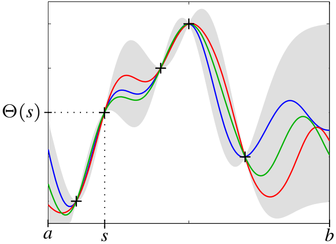
In other words:

NPB models do not generally satisfy Bayes' theorem.

Nonparametric regression

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

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



Recall definition

$$\Theta \sim \mathrm{GP} \qquad \Leftrightarrow \qquad (\Theta(s_1), \ldots, \Theta(s_d)) \quad \text{is } d\text{-dimensional Gaussian}$$

for *any* finite set $S \subset [a, b]$.

Construction: Intuition

- The marginal of the GP for any finite $S \subset [a, b]$ is a Gaussian.
- All these Gaussians are marginals of each other.
- Conversely: If we start with such Gaussians for all S, do they define a GP?

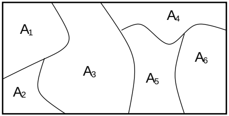
They do. The theorems which guarantee this are called *extension theorems* or *projective limit theorems*.

Idea

- GP: We have constructed a *random function* Θ.
- If Θ is a *random measure*, can we construct it in a similar way?

Extension theorem

- For a finite partition $I = (A_1, \ldots, A_d)$ of V, suppose we know the distribution P_{I} of $(\Theta(A_1), \ldots, \Theta(A_d))$.
- If the P_{I} for all partitions I are projective (= are marginals of each other), they define a unique random measure Θ on V.



Example: DP

Choose P_{I} as Dirichlet distribution with parameters α and $(G_0(A_1), \ldots, G_0(A_d))$.
Then $\Theta \sim \mathrm{DP}\,(\alpha, G_0)$.

REPRESENTATIONS

Stick-breaking

- Simple; most widely used where applicable.
- Constructive.
- Available only for few models (DP, Pitman-Yor process, normalized inverse Gaussian process, beta process).

Projective limits

- Generally applicable.
- Mathematically more challenging, many open problems.

Representations by known stochastic processes

- E.g. Lévy process and Poisson process representations.
- Often come with a useful set of theoretical tools.

Conjugate models

- How can we compute a posterior without a Bayes equation?
- Virtually all NPB models (DP, GP, etc) are conjugate.

Functional vs structural conjugacy

Functional conjugacy: There is a mapping

$$\text{prior hyperparameter} \times \text{data} \mapsto \text{posterior hyperparameter}$$

Structural conjugacy: Closure under sampling, but no functional conjugacy.

Example

Neutral-to-the-right processes are structurally but not functionally conjugate.

Constructing conjugate models

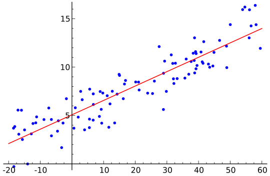
- In hierarchical models: Use conjugate components.
- Roughly: Projective limits of fct. conjugate marginals are fct. conjugate.

EXCHANGEABILITY

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

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

$$\Updownarrow$$

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

where:

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

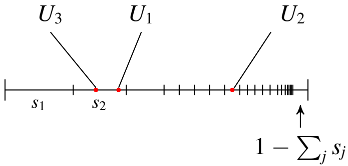
Implications

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

EXCHANGEABILITY: RANDOM PARTITIONS

Paint-box distribution

- Weights $s_1, s_2, \dots \geq 0$ with $\sum s_j \leq 1$
- $U_1, U_2, \dots \sim \text{Uniform}[0,1]$



Random partition of $\mathbb{N}$:

$$i, j \in \mathbb{N} \text{ in same block} \quad \Leftrightarrow \quad U_i, U_j \text{ in same interval}$$

$$\{i\} \text{ separate block} \quad \Leftrightarrow \quad U_i \text{ in interval } 1 - \sum s_j$$

Kingman's Theorem

$$\text{Random partition } \pi \text{ of } \mathbb{N} \text{ exchangeable}$$

$$\Updownarrow$$

$$\text{Mixture of paint-boxes } \beta(\,.\,|\mathbf{s}): \quad P(\pi) = \int \beta(\pi|\mathbf{s}) Q(d\mathbf{s})$$

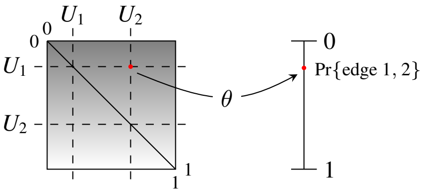
Random graph with independent edges

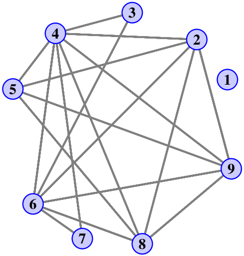
Given: $\theta : [0,1]^2 \to [0,1]$ symmetric function

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

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

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





Aldous-Hoover Theorem

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

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

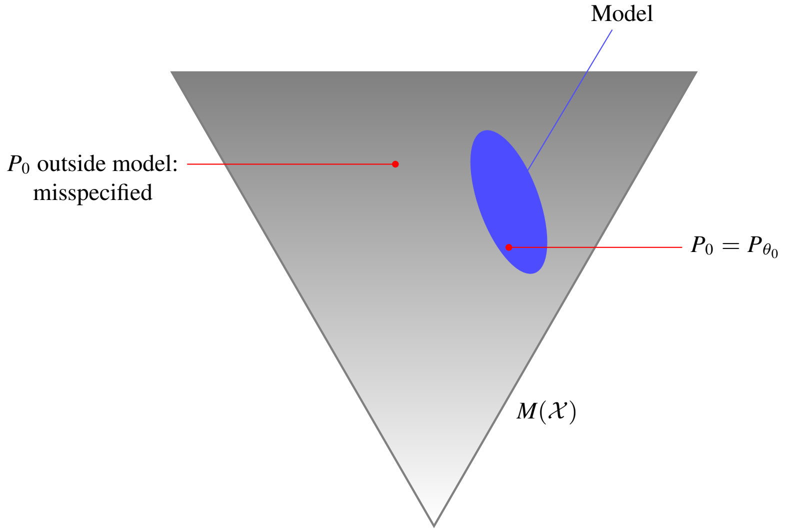
Ergodic decomposition theorems

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

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

ASYMPTOTICS

Support of Priors



Large support

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

Interpretation of nonparametric prior assumptions

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

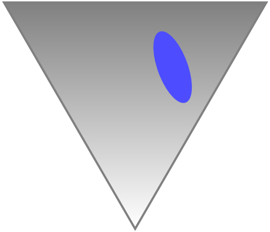
- GP regression with unknown bandwidth:
  - Any continuous function possible
  - Prior can express e.g. "very smooth functions are more probable"
- Clustering: Expected number of clusters is...
  - ...small $\longrightarrow$ CRP prior
  - ...power law $\longrightarrow$ two-parameter CRP

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.

Objective

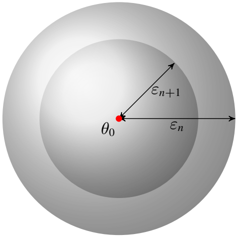
How quickly does posterior concentrate at θ_0 as $n \to \infty$?

Measure: Convergence rate

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

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

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



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!

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]

REFERENCES I

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

[Fer73] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1(2), 1973.

[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.

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

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

[Kin78] J. F. C. Kingman. The representation of partition structures. *Journal of the London Mathematical Society*, 18:374–380, 1978.

[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.

[LP10] A. Lijoi and I. Prünster. Models beyond the Dirichlet process. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics*. Cambrdige University Press, 2010.

[Mac00] S. N. MacEachern. Dependent Dirichlet processes. Technical report, Ohio State University, 2000.

[Orb09] P. Orbanz. Construction of nonparametric Bayesian models from parametric bayes equations. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

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

[Orb12] P. Orbanz. Nonparametric priors on complete separable metric spaces. 2012.

[RT09] D.M. Roy and Y.-W. Teh. The Mondrian process, 2009.

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

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

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

References II

[TJBB06] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, (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.