Dirichlet Processes, Chinese Restaurant Processes and All That

Michael I. Jordan

Department of Statistics and Computer Science Division
University of California, Berkeley

http://www.cs.berkeley.edu/∼jordan

Acknowledgments: David Blei, Yee Whye Teh

Introduction

Some Well-Worn but Still-Very-Useful Distinctions

	Frequentist	Bayesian
Parametric	I	IV
Semiparametric	II	V
Nonparametric	III	VI

I: Logistic regression, ANOVA, Fisher discriminant analysis, ARMA, etc
II: Independent component analysis, Cox model, nonmetric MDS, etc
III: Nearest neighbor, kernel methods, bootstrap, decision trees, neural nets, etc (a focus point for machine learning research)
IV: Graphical models, hierarchical models, etc
V: ?
VI: Gaussian processes, ?

Some Well-Worn but Still-Very-Useful Distinctions

	Frequentist	Bayesian
Parametric	I	IV
Semiparametric	II	V
Nonparametric	III	VI

- What do we mean by "parameters" anyway?
  - note in particular that "nonparametric" doesn't mean "no parameters"
  - it means (very roughly) that the number of parameters grows with the number of data points

The De Finetti Perspective on Parameters

- For Bayesians, the De Finetti theorem is a compelling motivation for both "parameters" and priors on parameters
  - but everyone should know what the De Finetti theorem is, not just Bayesians...
- What is the De Finetti theorem?
- It's the "bag-of-words theorem"

The De Finetti Perspective on Parameters (cont.)

- Suppose that we agree that if our data are reordered, it doesn't matter
  - this is generally **not** an assertion of "independent and identically distributed"; rather, it is an assertion of "exchangeability"
- *Exchangeability*: the joint probability distribution underlying the data is invariant to permutation

  Theorem (De Finetti, 1935). *If* $(x_1, x_2, \ldots)$ *are infinitely exchangeable, then the joint probability* $p(x_1, x_2, \ldots, x_N)$ *has a representation as a mixture*:

$$p(x_1, x_2, \ldots, x_N) = \int \left(\prod_{i=1}^{N} p(x_i \mid \theta) \right) dP(\theta)$$

  for some random variable θ.
- I.e., if you assert exchangeability, it is reasonable to act as if there is an underlying parameter, there is a prior on that parameter, and the data are conditionally IID given that parameter

The De Finetti Perspective on Parameters (cont.)

Theorem (De Finetti, 1935). *If* $(x_1, x_2, \ldots)$ *are infinitely exchangeable, then the joint probability* $p(x_1, x_2, \ldots, x_N)$ *has a representation as a mixture*:

$$p(x_1, x_2, \ldots, x_N) = \int \left(\prod_{i=1}^{N} p(x_i \,|\, \theta) \right) dP(\theta)$$

for some random variable θ.

- The theorem wouldn't be true if we limited ourselves to random variables θ ranging over Euclidean vector spaces
- In particular, we need to allow θ to range over measures, in which case $P(\theta)$ is a distribution on measures
  - the Dirichlet process is an example of a distribution on measures...

Bayesian Nonparametrics

- There are Bayesian nonparametric approaches to many of the main problems in statistics:
  - regression
  - classification
  - clustering
  - survival analysis
  - time series analysis
  - spatial data analysis
  - etc

- These generally involve assumptions of exchangeability or partial exchangeability
  - and corresponding distributions on random objects of various kinds (functions, monotone functions, partitions, measures, etc)

- We'll focus on one problem for concreteness—clustering

Clustering—How to Choose K?

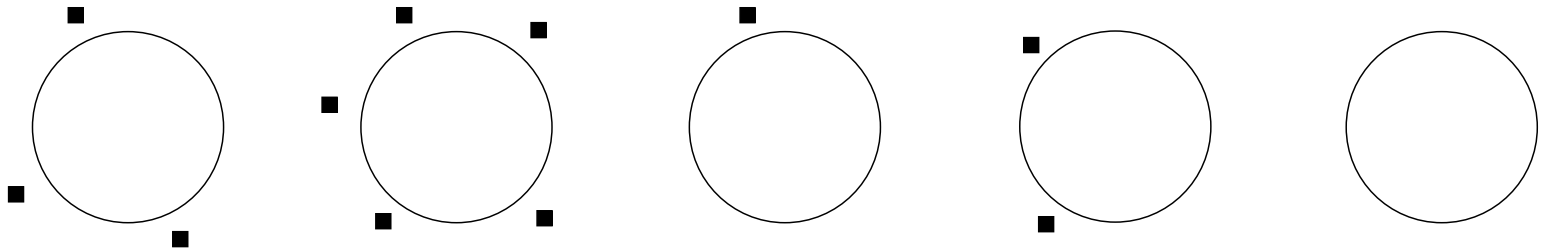
- Adhoc approaches (e.g., hierarchical clustering)
  - they do often yield a data-driven choice of K
  - but there is little understanding of how good these choices are
- Methods based on objective functions (M-estimators)
  - e.g., K-means, spectral clustering
  - do come with some frequentist guarantees
  - but it's hard to turn these into data-driven choices of K
- Parametric likelihood-based approaches
  - finite mixture models, Bayesian variants thereof
  - various model choice methods: hypothesis testing, cross-validation, bootstrap, AIC, BIC, DIC, Laplace, bridge sampling, reversible jump, etc
  - but do the assumptions underlying the method really apply to this setting? (not often)
- Let's try something different...

Chinese Restaurant Process (CRP)

- A random process in which n customers sit down in a Chinese restaurant with an infinite number of tables
  - first customer sits at the first table
  - mth subsequent customer sits at a table drawn from the following distribution:

$$\begin{aligned} P(\text{previously occupied table } i \,|\, \mathcal{F}_{m-1}) &\propto n_i \\ P(\text{the next unoccupied table} \,|\, \mathcal{F}_{m-1}) &\propto \alpha_0 \end{aligned} \tag{1}$$

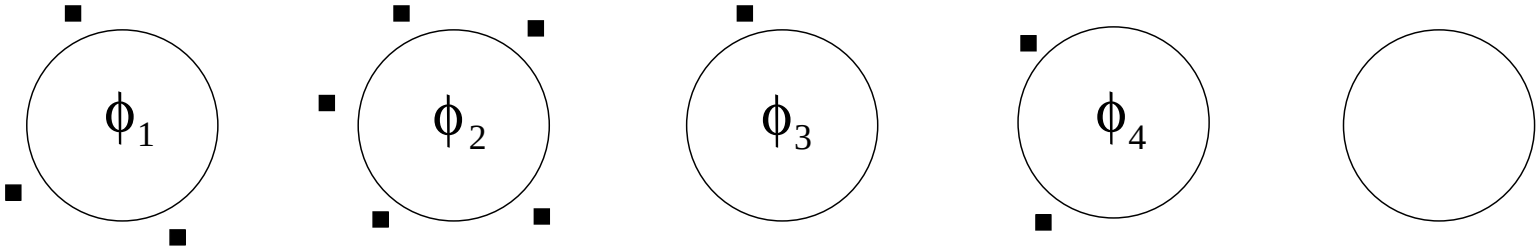
  where n_i is the number of customers currently at table i and where $\mathcal{F}_{m-1}$ denotes the state of the restaurant after $m-1$ customers have been seated



The CRP and Clustering

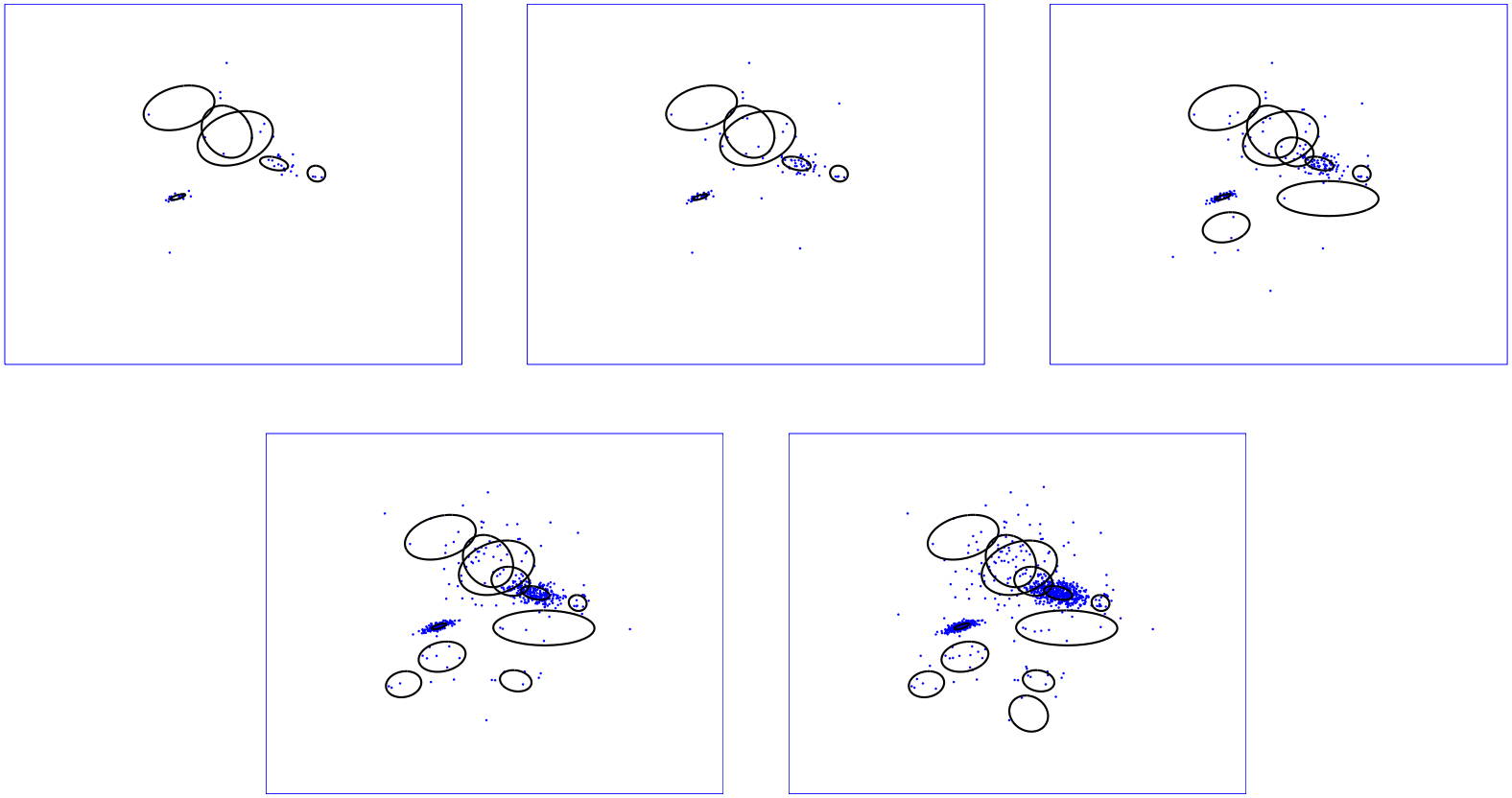
- Data points are customers; tables are clusters
  - the CRP defines a prior distribution on the partitioning of the data and on the number of tables

- This prior can be completed with:
  - a likelihood—e.g., associate a parameterized probability distribution with each table
  - a prior for the parameters—the first customer to sit at table k chooses the parameter vector for that table (ϕ_k) from the prior



- So we now have a distribution—or can obtain one—for any quantity that we might care about in the clustering setting

CRP Prior, Gaussian Likelihood, Conjugate Prior



$$\phi_k = (\mu_k, \Sigma_k) \sim N(a, b) \otimes IW(\alpha, \beta)$$

$$x_i \sim N(\phi_k) \qquad \text{for a data point } i \text{ sitting at table } k$$

Exchangeability

(Blackwell & MacQueen; Kingman; Aldous; Pitman)

- As a prior on the partition of the data, the CRP is exchangeable
- The prior on the parameter vectors associated with the tables is also exchangeable
- The latter probability model is generally called the Pólya urn model. Letting θ_i denote the parameter vector associated with the ith data point, we have:

$$\theta_i \,|\, \theta_1, \ldots, \theta_{i-1} \;\sim\; \alpha_0 G_0 + \sum_{j=1}^{i-1} \delta_{\theta_j}$$

- From these conditionals, a short calculation shows that the joint distribution for $(\theta_1, \ldots, \theta_n)$ is invariant to order (this is the exchangeability proof)
  - De Finetti implies that there is an underlying random "parameter" and a distribution on that parameter. What are they?

The CRP (cont)

- An additional fact about the CRP:
  - as a prior on the number of tables, the CRP is nonparametric—the number of occupied tables grows (roughly) as $O(\log n)$—we're in the world of nonparametric Bayes
- How do we do inference with a CRP?
- How does this relate to more standard model-based clustering?
- Any theory behind this?
- What can we do that's new with this setup?

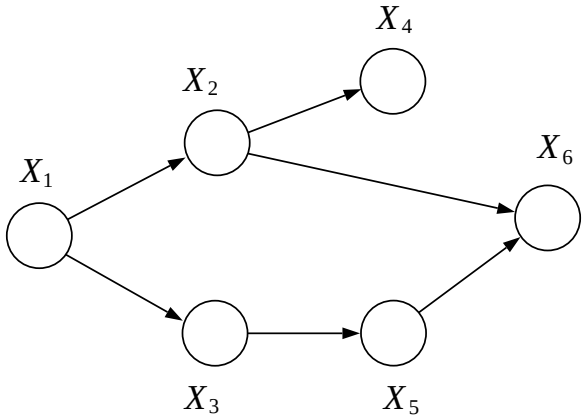
Outline

- Background
- Bayesian mixture models, stick-breaking, Dirichlet processes
- Inference
- Hierarchical and dependent Dirichlet processes
- Semiparametric models
- Applications
- Further directions in nonparametric Bayes: tail-free processes, neutral-to-the-right processes, Polya trees, diffusion trees, Pitman-Yor processes

Background

Directed Graphical Models

- Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v \in \mathcal{V}$ is associated with a random variable X_v:

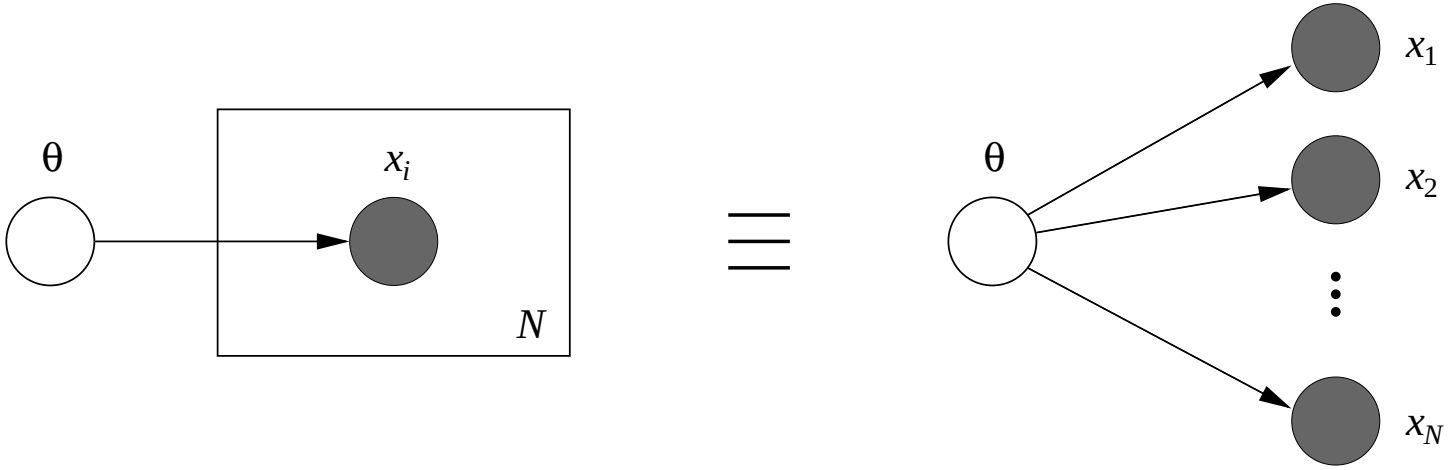


- The joint distribution on $(X_1, X_2, \ldots, X_N)$ factorizes according to the "parent-of" relation defined by the edges $\mathcal{E}$:

$$p(x_1, x_2, x_3, x_4, x_5, x_6; \theta) = p(x_1; \theta_1)\ p(x_2 \,|\, x_1; \theta_2)$$
$$p(x_3 \,|\, x_1; \theta_3)\ p(x_4 \,|\, x_2; \theta_4)\ p(x_5 \,|\, x_3; \theta_5)\ p(x_6 \,|\, x_2, x_5; \theta_6)$$

Plates

- A *plate* is a "macro" that allows subgraphs to be replicated:



- Shading denotes observed variables; i.e., conditioning

- Note that this graph represents the following marginal probability for the observations $(x_1, x_2, \ldots, x_N)$:

$$p(x_1, x_2, \ldots, x_N) = \int \left(\prod_{i=1}^{N} p(x_i \mid \theta) \right) dP(\theta)$$

Gibbs Sampling

- A Markov chain Monte Carlo (MCMC) method
- Consider a set of variables X_V, with distribution $p(x_V)$ (which may be a conditional distribution)
- Set up a Markov chain as follows:
  - initialize the X_i to arbitrary values
  - choose i randomly
  - sample from $p(x_i \mid x_{V \setminus i})$
  - iterate
- Under (usually) easily-checkable conditions, this scheme has $p(x_V)$ as its equilibrium distribution