Nonparametric Bayesian for Sequential Data

Jaesik Choi

Statistical Artificial Intelligence Laboratory (SAIL)*

UNIST

*http://sail.unist.ac.kr

---

1Some slides are courtesy to [Zubin, UAI 2005], [Miller, 2009], and [Orbanz, NIPS 2012]
Overview

1 Introduction

2 Dirichlet Processes - Nonparametric Clustering

3 Gaussian Processes - Nonparametric Regression

4 Case Study: The Relational Automatic Statistician System

5 Appendix
  - Conjugate Prior
  - Dirichlet Processes
  - Chinese Restaurant Processes
  - Beta Processes
  - Indian Buffet Processes
  - Nonparametric Hierarchical Models
Terminology

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

- **Nonparametric model**
  - Number of parameters grows with sample size
  - $\infty$-dimensional parameter space

- **Example**: Density estimation
Definition A nonparametric Bayesian model is a Bayesian model on an \( \infty \)-dimensional parameter space.

Interpretation Parameter space \( \mathcal{T} = \) set of possible model parameters (or pattern), for example:

<table>
<thead>
<tr>
<th>Problem</th>
<th>( \mathcal{T} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density estimation</td>
<td>Probabilistic distributions</td>
</tr>
<tr>
<td>Regression</td>
<td>Smooth functions</td>
</tr>
<tr>
<td>Clustering</td>
<td>Partitions</td>
</tr>
</tbody>
</table>

Solution to Bayesian problem = posterior distribution on model parameters.
Exchangeability

- Can we justify our assumptions?

  Assumption: data = model + noise

  In Bayes’ theorem

  \[
  p(\text{model} | \text{data}) = \frac{p(\text{data} | \text{model})}{p(\text{data})} p(\text{model})
  \]

- Definition \( X_1, \cdots, X_n \) are exchangeable if \( P(X_1, \cdots, X_n) \) is invariant under any permutation \( \pi \):

  \[
  p(X_1 = x_1, \cdots, X_n = x_n) = p(X_1 = x_{\pi(1)}, \cdots, X_n = x_{\pi(n)})
  \]

  e.g.

  \[
  p(1, 1, 0, 0, 0, 1, 1, 0) = p(1, 0, 1, 0, 1, 0, 0, 1)
  \]

  Order of observations does not matter.
De Finetti’s Theorem (binary cases)

Binary sequence case: all exchangeable binary sequences are mixtures of Bernoulli sequences [de Finetti, 1931]

\[ p(x_1, \cdots, x_n) = \int_0^1 \theta^{t_n} (1 - \theta)^{n-t_n} dF(\theta), \]

where \( p(x_1, \cdots, x_n) = p(X_1 = x_1, \cdots, X_n = x_n) \) and \( t_n = \sum_{i=1}^n x_i \).

**Implications**

- Exchangeable data decomposes into mixtures of models with i.i.d. sequences
- Caution: \( \theta \) is in general an \( \infty \)-dimensional quantity
Contents

1 Introduction

2 Dirichlet Processes - Nonparametric Clustering

3 Gaussian Processes - Nonparametric Regression

4 Case Study: The Relational Automatic Statistician System

5 Appendix
  • Conjugate Prior
  • Dirichlet Processes
  • Chinese Restaurant Processes
  • Beta Processes
  • Indian Buffet Processes
  • Nonparametric Hierarchical Models
Clustering

- Observations $X_1, X_2, \cdots$
- Each observation belongs to exactly one cluster
- Unknown pattern $=$ partition of $\{1, \cdots , n\}$
Mixture Models

- Mixture models

\[ p(\text{data} | \text{model}) = \int_{\Omega_{\theta}} p(\text{data} | \theta) m(d\theta) \]

\( m \) is called the mixing measure

- Two-stage sampling
  Sample data \( X \sim p(\cdot | m) \) as:
  1. \( \Theta \sim m \)
  2. \( X \sim p(\cdot | \theta) \)

- Finite mixture model

\[ m(\cdot) = \sum_{k=1}^{K} c_k \delta_{\theta_k}(\cdot) \]
Bayesian Mixture Models (BMMs)

- Random mixing measure
  \[ m(\cdot) = \sum_{k=1}^{K} c_k \delta_{\theta_k}(\cdot) \]

- Conjugate priors
  A Bayesian model is conjugate if the posterior is an element of the same class of distribution as the prior (‘closure under sampling’).

| \(p(data|model)\), likelihood | \(p(model)\) conjugate prior |
|-------------------------------|-----------------------------|
| Multinomial                  | Dirichlet                   |
| Gaussian                     | Smooth functions            |
| Clustering                   | Partitions                  |

- Choice of priors in BMM
  - Choose conjugate prior for each parameter
  - E.g.: Dirichlet prior
Dirichlet Process Mixtures

- Dirichlet process (DP) [Ferguson, 73] [Sethuraman, 94]
  A DP is a distribution on random probability measures of the form
  \[ m(\cdot) = \sum_{k=1}^{\infty} c_k \delta_{\theta_k}(\cdot) \] where \( \sum_{k=1}^{\infty} c_k = 1 \)

- Constructive definition of DP(\( \alpha, G_0 \))
  \[ \theta_k \sim iid G_0 \]
  \[ V_k \sim iid Beta(1, \alpha) \]

  Compute \( c_k \) as

  \[ c_k = V_k \prod_{i=1}^{k-1} (1 - V_i) \]

  This procedure is called ‘Stick-breaking construction’
Posterior Distribution

- **DP Posterior**

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

- **Mixture Posterior**

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

- **Conjugacy**

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

- Latent variables

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

We observe \( x_i \) and do not actually observe \( \theta_k \) (latent).

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

- Gibbs Sampling
  Uses an assignment variable \( \phi_j \) for each observation \( x_j \).
  - Assignment step: Sampling \( \phi_j \sim \text{Multinomial}(q_{j0}, \cdots, q_{jK_n}) \)
  - 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

\[ K_n = \# \text{ of clusters in sample of size } n \]
\[ \mathbb{E}[K_n] = O(\log(n)) \]

- Modeling assumption
  - Parametric clustering: \( K_\infty \) is finite (possibly unknown, but fixed).
  - Nonparametric clustering: \( K_\infty \) is infinite.

- Rephrasing the question
  - Estimate of \( K_n \) is controlled by distribution of the cluster sizes \( c_k \) in
    \[ \sum_k c_k \delta_{\theta_k} \]
  - What should we assume about the distribution of \( c_k \)
Generalizing the DP

- Pitman-Yor process

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

Discount parameter \( d \in [0, 1] \).

- Cluster sizes
Power Laws

The distribution of cluster sizes is called a power law if

\[ c_j \sim \gamma(\beta) \cdot j^{-\beta} \]

for some \( \beta \in [0, 1] \).

- Examples of power laws
  - Word frequencies
  - Popularity (# of friends) in social networks
- Pitman-Yor language model
Random Partitions

- Discrete measures and partitions
  Sampling from a discrete measure determines a partition of $\mathbb{N}$ into blocks $b_k$:
  $$\theta_n \sim iid \sum_{k=1}^{\infty} c_k \delta_{\theta_k^*}$$
  and set $n \in b_k \iff \theta_n = \theta_k^*$
  As $n \to \infty$, the block proportions converge: $\frac{|b_k|}{n} \to c_k$

- 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 $\prod = (B_1, B_2, \cdots)$.

- Exchangeable random partitions
  - $\prod$ is called exchangeable if its distribution depends only on the sizes of its blocks.
  - All exchangeable random partitions, and only those, can be represented by a random discrete distribution (Kingman’s theorem).
Chinese Restaurant Process (CRP)

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

- **‘Customers and tables’ analogy**

![Diagram showing customers and tables analogy](image)

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 \( \{1, \cdots, n\} \) (regard cycles of permutation as blocks of partition)
  - The induced distribution on partitions is the CRP we use in clustering
Nonparametric Bayesian clustering

- Infinite # 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 (i.e., partition) not observed.
- Gibbs sampling
- Variational algorithms

Prior assumption
- Distribution of cluster sizes
- Implies prior assumption on # \( K_n \) of clusters.
Contents

1 Introduction

2 Dirichlet Processes - Nonparametric Clustering

3 Gaussian Processes - Nonparametric Regression

4 Case Study: The Relational Automatic Statistician System

5 Appendix
   - Conjugate Prior
   - Dirichlet Processes
   - Chinese Restaurant Processes
   - Beta Processes
   - Indian Buffet Processes
   - Nonparametric Hierarchical Models
Gaussian Distributions

- **Gaussian**
  \[
  p(x) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right)
  \]

- **Multivariate Gaussian**
  \[
  p(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)
  \]
Gaussian Processes: Definition

- GPs are distributions over functions such that any finite set of function evaluations \([f(x_1), \cdots, f(x_n)]\) have a jointly Gaussian distribution.
- A GP is completely specified by its mean function \(\mu(x) = \mathbf{E}(f(x))\) and covariance kernel function \(k(x, x') = \text{Cov}(f(x), f(x'))\).
- Given data \(x = [x_1, \cdots, x_n]\) and \(y = [f(x_1), \cdots, f(x_n)]\), GP model specified by \(\mu(x)\) and \(k(x, x')\).
- Notation for ‘\(f\) follows the GP’ is:
  \[f \sim \text{GP}(\mu(x), k(x, x'))\]

Likelihood is:

\[
p(y|X) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp \left( -\frac{1}{2} (y - \mu)^T \Sigma^{-1} (y - \mu) \right)
\]

where \(\mu = [\mu(x_1), \cdots, \mu(x_n)]\) and \(\Sigma_{ij} = k(x_i, x_j)\).
Gaussian Processes: Samples with different kernels
Gaussian Processes: Predictions with different kernels

A sample from the prior for each covariance function

Corresponding predictions, mean with two standard deviations:
Gaussian Processes

- Nonparametric regression
  parameter spaces = continuous functions, say on $[a,b]$:

  $$f : [a, b] \rightarrow \mathbb{R} \quad \mathcal{T} = c[a, b]$$

- Gaussian Process

  $$f \sim GP \iff (f(x_1), \cdots, f(x_d))$$ is d-dimensional Gaussian

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

- Construction: Intuition
  - The marginal of the GP for any finite $X \subset [a, b]$ is a Gaussian.
  - All these Gaussians are marginals of each other.
<table>
<thead>
<tr>
<th>Applications</th>
<th>Parameter space</th>
<th>Bayesian Nonparametric model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>Function</td>
<td>Gaussian process</td>
</tr>
<tr>
<td>Clustering</td>
<td>Partition</td>
<td>Chinese restaurant process</td>
</tr>
<tr>
<td>Density estimation</td>
<td>Density</td>
<td>Dirichlet process mixture</td>
</tr>
<tr>
<td>Hierarchical clustering</td>
<td>Hierarchical partition</td>
<td>Dirichlet/Pitmann-or diffusion tree, Nested CRP</td>
</tr>
<tr>
<td>Latent variable modeling</td>
<td>Features</td>
<td>Beta process/Indian buffet process</td>
</tr>
<tr>
<td>Dictionary learning</td>
<td>Dictionary</td>
<td>Beta process/Indian buffet process</td>
</tr>
<tr>
<td>Survival analysis</td>
<td>Hazard</td>
<td>Beta process, Neural-to-the-right process</td>
</tr>
<tr>
<td>Power-law behaviour</td>
<td></td>
<td>Pitman-Yor process, Stable-beta process</td>
</tr>
<tr>
<td>Deep learning</td>
<td>Features</td>
<td>Cascading/nested Indian buffet process</td>
</tr>
<tr>
<td>Dimensionality reduction</td>
<td>Manifold</td>
<td>Gaussian process latent variable model</td>
</tr>
<tr>
<td>Topic models</td>
<td>Atomic distribution</td>
<td>Hierarchical Dirichlet process</td>
</tr>
<tr>
<td>Relational modeling</td>
<td>Relations</td>
<td>Infinite relational model, Mondrian process</td>
</tr>
</tbody>
</table>
Recent advances in Nonparametric Regression
Problem

Descriptive prediction of multiple time series
Problem

Descriptive prediction of multiple time series
Problem

Descriptive prediction of multiple time series

Linear function
decrease x/week

Smooth function
Length scale: y weeks

Rapidly varying smooth function
Length scale: z hours
Problem (this paper)

Descriptive prediction of multiple time series

- Constant function
- Sudden drop btw 9/12/01 ~ 9/15/01
- Smooth function
  - Length scale: y weeks
- Rapidly varying smooth function
  - Length scale: z hours
- Constant function
  - Sudden drop btw 9/12/01 ~ 9/15/01
Models

Automatic Bayesian Covariance Discovery*
[Lloyd et. al. 2014] [Ghahramani. 2015]

* http://www.automaticstatistician.com/
Gaussian Processes

\[ f(x) \sim \mathcal{GP}(\mu(x), k(x, x')) \]

- **Function**
- **Gaussian Process**

**Mean function**

\[ \mu(x) = \mathbb{E}(f(x)) \]

**Covariance kernel function**

\[ k(x, x') = \text{Cov}(f(x), f(x')) \]
Gaussian Processes

\[ f(x) \sim \mathcal{GP}(\mu(x), k(x, x')) \]

- **Function**
- **Gaussian Process**
- **Mean function**
  \[ \mu(x) = \mathbb{E}(f(x)) \]
- **Covariance kernel function**
  \[ k(x, x') = \text{Cov}(f(x), f(x')) \]

- **Mean vector**
  \[ \boldsymbol{\mu} = [\mu(x_1), ..., \mu(x_N)] \]
- **Covariance matrix**
  \[ \Sigma_{ij} = k(x_i, x_j) \]

- **Function evaluations**
- **Multivariate Gaussian**
  \[ [f(x_1), ..., f(x_N)] \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \]
The Automatic Statistician*

(1) Encode characteristic

\[ f(x) \sim \mathcal{GP}(\mu(x), k(x, x')) \]

Find appropriate kernel
The Automatic Statistician*

(1) Encode characteristic

Find appropriate kernel

\[ f(x) \sim \mathcal{GP}(\mu(x), k(x, x')) \]

(2) Compose new kernel

If \( g(x) \sim \mathcal{GP}(0, k_g) \), \( h(x) \sim \mathcal{GP}(0, k_h) \) and \( g(x) \perp h(x) \), then

\[ g(x) + h(x) \sim \mathcal{GP}(0, k_g + k_h) \]

\[ g(x) \times h(x) \sim \mathcal{GP}(0, k_g \times k_h) \]

* Automatic Bayesian Covariance Discovery (http://www.automaticstatistician.com/)
[Lloyd; Duvenaud; Grosse; Tenenbaum; Ghahramani. 2014.] [Ghahramani. Nature. 2015.]
# The Automatic Statistician: Base kernels

<table>
<thead>
<tr>
<th>Base kernel</th>
<th>Encoding function</th>
<th>Kernel function</th>
<th>Parameters</th>
<th>Example kernel function shape</th>
<th>Example encoded functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIN($x, x'$)</td>
<td>Linear function</td>
<td>$\sigma^2 (x - \ell)(x' - \ell)$</td>
<td>$\sigma, \ell$</td>
<td><img src="image1.png" alt="Graph" /></td>
<td><img src="image2.png" alt="Graph" /></td>
</tr>
<tr>
<td>SE($x, x'$)</td>
<td>Smooth function</td>
<td>$\sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right)$</td>
<td>$\sigma, \ell$</td>
<td><img src="image3.png" alt="Graph" /></td>
<td><img src="image4.png" alt="Graph" /></td>
</tr>
<tr>
<td>PER($x, x'$)</td>
<td>Periodic function</td>
<td>In appendix</td>
<td>$\sigma, \ell, p$</td>
<td><img src="image5.png" alt="Graph" /></td>
<td><img src="image6.png" alt="Graph" /></td>
</tr>
</tbody>
</table>
# The Automatic Statistician: Operators

<table>
<thead>
<tr>
<th>Op.</th>
<th>Concept</th>
<th>Params</th>
<th>Example</th>
<th>Example kernel function shape</th>
<th>Example encoded functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>Addition Superposition OR operator</td>
<td>N/A</td>
<td>SE + PER</td>
<td><img src="image" alt="Graph SE + PER" /></td>
<td><img src="image" alt="Encoded Functions" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LIN + PER</td>
<td><img src="image" alt="Graph LIN + PER" /></td>
<td><img src="image" alt="Encoded Functions" /></td>
</tr>
<tr>
<td>×</td>
<td>Multiplication AND operator</td>
<td>N/A</td>
<td>SE × PER</td>
<td><img src="image" alt="Graph SE × PER" /></td>
<td><img src="image" alt="Encoded Functions" /></td>
</tr>
</tbody>
</table>
### The Automatic Statistician: Operators

<table>
<thead>
<tr>
<th>Op.</th>
<th>Concept</th>
<th>Params</th>
<th>Example</th>
<th>Example encoded functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>CP</td>
<td>Divide left vs right</td>
<td>ℓ, s</td>
<td>CP(LIN,LIN)</td>
<td><img src="example1.png" alt="Example" /></td>
</tr>
<tr>
<td>CW</td>
<td>Divide in vs out</td>
<td>ℓ, s, w</td>
<td>CW(LIN,C)</td>
<td><img src="example2.png" alt="Example" /></td>
</tr>
</tbody>
</table>
(1) **Optimization criterion**: Bayesian Information Criterion (BIC)

\[
\text{BIC}(\mathcal{M}) = -2 \log P(D | \mathcal{M}) + |\mathcal{M}| \log |D|
\]

- **Negative log-likelihood**
- **Num. of model parameters**
- **Num. of data points**
- **Model complexity**
The Automatic statistician: Learning

1. Optimization criteria: Bayesian Information Criterion (BIC)

$$\text{BIC}(\mathcal{M}) = -2 \log P(D|\mathcal{M}) + |\mathcal{M}| \log D$$

- Model complexity
- Negative log-likelihood
- Num. of data points
- Num. of model parameters

2. Learning algorithm (Composite Kernel Learning)

- Iteratively select best model (structure $k$, parameter $\theta$)
- **Expand**: the current kernel
- **Optimize**: conjugate gradient descent
- **Select**: the best kernel in the level (greedy)
- **Iterate**: get back to (1) for the next level

(1) Expand

(2) Optimize

(3) Select
Example Result

The Automatic Statistician*

*Automatic Bayesian Covariance Discovery (http://www.automaticstatistician.com/)
[Lloyd; Duvenaud; Grosse; Tenenbaum; Ghahramani. 2014.] [Ghahramani. Nature. 2015.]
Prediction Performance (Extrapolation)

13 famous regression datasets
The Relational Automatic Statistician

[Hwang, Tong and Choi, 2016]

* http://www.automaticstatistician.com/
Motivation

The Automatic Statistician

Adjusted Close of General Electronics

9/11, 2001

Linear function decrease x/week

Smooth function Length scale: y weeks

Rapidly varying smooth function Length scale: z hours
Motivation

The Relational Automatic Statistician

Adjusted Close of General Electronics, Microsoft, ExxonMobil

- Exploit multiple time series
- Find global descriptions
- Hope better predictive performance

9/11, 2001
**Model: Gaussian Process**

\[ P(D|M) = P(D|\mathcal{GP}(0, k(x, x', \theta))) \]

- **Mean function**: \( \mu(x) \)
- **Covariance kernel function**: \( k(x, x') \)
- **Latent function evaluation**: \( f \)
- **Observation**: \( y_i \) for \( i \in 1...N \)
- **Gaussian Process**: \( \mathcal{GP} \)
- **Fixed Given Optimize**: \( k, \theta \)

**Gaussian Processes** → CKL → RKL → SRKL
Model: Composite Kernel Learning (CKL)

\[
P(D|\mathcal{M}) = P(D|\mathcal{GP}(0, k(x, x'; \theta)))
\]

GPs → Composite Kernel Learning → RKL → SRKL

Generalized Multi Kernel Learning
Model: Relational Kernel Learning (RKL)

\[
P(D | \mathcal{M}) = \prod_{j=1}^{M} P(d_j | \mathcal{GP}(0, \sigma_j \times k(x, x'; \theta) + c_j))
\]

GPs → CKL → Relational Kernel Learning → SRKL
Model: Semi-Relational Kernel Learning (SRKL)

\[
P(D|\mathcal{M}) = \prod_{j=1}^{M} P\left( d_j | \mathcal{G}\mathcal{P}(0, \sigma_j \times k(x, x'; \theta) + k_j(x, x'; \theta_j) ) \right)
\]

GPs → CKL → RKL → Semi-Relational Kernel Learning
Semi-Relational Kernel Learning (SRKL)

**Input:** M time series

**Output:** A shared kernel $k$, M spectral mixture (SM) kernels

1. **Expand:** the current shared kernel for all time series
2. **Optimize:** expanded kernels for all M time series (conjugate gradient descent)
   For each series, individual distinction is handled by the SM kernel.
3. **Select:** the best shared kernel for all time series (greedy)
   A shared kernel ($k$) and M SM kernels ($k_j$)
4. **Iterative:** get back to (1) when level s is not reached

Find the best shared and distinctive kernels iteratively!
Experimental Results

Three real-world data sets:

- US top 9 stocks in year 2001
- US top 6 housing markets from 2003 to 2013
- Currency exchange of 4 emerging market
## Data sets

<table>
<thead>
<tr>
<th>Descriptions</th>
<th>Graphs (normalized)</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 adjusted close of stock figures</td>
<td><img src="image1.png" alt="Graph" /></td>
<td>GE, MSFT, XOM, PFE, C, WMT, INTC, BP, AIG</td>
</tr>
<tr>
<td>6 US housing price indices</td>
<td><img src="image2.png" alt="Graph" /></td>
<td>New York, Los Angeles, Chicago, Phoenix, San Diego, San Francisco</td>
</tr>
<tr>
<td>4 emerging market currency exchanges</td>
<td><img src="image3.png" alt="Graph" /></td>
<td>Indonesian - IDR, Malaysian - MYR, South African - ZAR, Russian - RUB</td>
</tr>
</tbody>
</table>
Qualitative Results

US stock market values suddenly drop after US 9/11 attacks.

Currency exchange is affected by FED’s policy change in interest rates around middle Sep 2015.

4 currency exchange rates

Currency exchange is affected by FED’s policy change in interest rates around middle Sep 2015.
An automatically generated report for the dataset: GE
Relational version

2.6 Component 6: A constant. This function applies from 12 Sep 2001 until 15 Sep 2001

This component is constant. This component applies from 12 Sep 2001 until 15 Sep 2001.

This component explains 100.0% of the residual variance; this increases the total variance explained from 95.2% to 100.0%. The addition of this component increases the cross validated MAE by 0.67% from 0.87 to 0.87. This component explains residual variance but does not improve MAE which suggests that this component describes very short term patterns, uncorrelated noise or is an artefact of the model or search procedure.

Figure 1: Raw data (left) and model posterior with extrapolation (right)
## Quantitative Results

<table>
<thead>
<tr>
<th>Data set</th>
<th>Negative log likelihood</th>
<th>Bayesian Information Criteria</th>
<th>Root mean square error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CKL</td>
<td>RKL</td>
<td>SRKL</td>
</tr>
<tr>
<td>STOCK3</td>
<td>332.75</td>
<td>311.84</td>
<td>304.05</td>
</tr>
<tr>
<td>STOCK6</td>
<td>972.00</td>
<td>1007.09</td>
<td>988.14</td>
</tr>
<tr>
<td>STOCK9</td>
<td>1776.31</td>
<td>1763.96</td>
<td>1757.11</td>
</tr>
<tr>
<td>HOUSE2</td>
<td>264.69</td>
<td>304.29</td>
<td>310.38</td>
</tr>
<tr>
<td>HOUSE4</td>
<td>594.79</td>
<td>586.81</td>
<td>1249.82</td>
</tr>
<tr>
<td>HOUSE6</td>
<td>849.64</td>
<td>891.09</td>
<td>1495.40</td>
</tr>
<tr>
<td>CURRENCY4</td>
<td>578.35</td>
<td>617.77</td>
<td>693.76</td>
</tr>
</tbody>
</table>

- STOCK3 = \{GE, MSFT, XOM\}
- STOCK6 = STOCK3 + \{PFE, C, WMT\}
- STOCK9 = STOCK6 + \{INTC, BP, AIG\}
- HOUSE2 = \{NY, LA\}
- HOUSE4 = HOUSE2 + \{Chicago, Phoenix\}
- HOUSE6 = HOUSE4 + \{San Diego, San Francisco\}
- CURRENCY4 = \{IDR, MYR, ZAR, RUB\}
Quantitative Results (box plots)

9 stocks
6 house price indices
4 currency exchanges
Conclusion

• Our research topic is about “Solving descriptive prediction problem of multiple time series”

• We proposed models that can “Exploit both common and distinct changes”

• We found that our models “Show the better qualitative and quantitative performance compared to the state-of-the-art GP regression method”

http://saildemo.unist.ac.kr/automatic_statistician/
Contents

1 Introduction

2 Dirichlet Processes - Nonparametric Clustering

3 Gaussian Processes - Nonparametric Regression

4 Case Study: The Relational Automatic Statistician System

5 Appendix
  - Conjugate Prior
  - Dirichlet Processes
  - Chinese Restaurant Processes
  - Beta Processes
  - Indian Buffet Processes
  - Nonparametric Hierarchical Models
Simple Example

Task: Toss a (potentially biased) coin $N$ times. Compute $\theta$, the probability of heads.

Suppose we observe: \{T, H, H, T\}. What do we think $\theta$ is? The maximum likelihood estimate is $\theta = 1/2$. Seems reasonable.

Now suppose we observe: \{H, H, H, H\}. What do we think $\theta$ is? The maximum likelihood estimate is $\theta = 1$. Seem reasonable?
Simple Example

Task: Toss a (potentially biased) coin $N$ times. Compute $\theta$, the probability of heads.

Suppose we observe: \{T, H, H, T\}. What do we think $\theta$ is? The maximum likelihood estimate is $\theta = 1/2$. Seems reasonable.

Now suppose we observe: \{H, H, H, H\}. What do we think $\theta$ is? The maximum likelihood estimate is $\theta = 1$. Seem reasonable?

Not really. Why?
Simple Example

When we observe \( \{H, H, H, H\} \), why does \( \theta = 1 \) seem unreasonable?

Prior knowledge! We believe coins generally have \( \theta \approx 1/2 \). How to encode this? By using a Beta prior on \( \theta \).
Simple Example

When we observe \{H, H, H, H\}, why does \( \theta = 1 \) seem unreasonable?

Prior knowledge! We believe coins generally have \( \theta \approx 1/2 \). How to encode this? By using a Beta prior on \( \theta \).
Bayesian Approach to Estimating $\theta$

Place a Beta$(a, b)$ prior on $\theta$. This prior has the form

$$p(\theta) \propto \theta^{a-1}(1 - \theta)^{b-1}.$$ 

What does this distribution look like?
Bayesian Approach to Estimating $\theta$

Place a Beta($a$, $b$) prior on $\theta$. This prior has the form

$$p(\theta) \propto \theta^{a-1}(1 - \theta)^{b-1}.$$

What does this distribution look like?
Bayesian Approach to Estimating $\theta$

After observing $X$, a sequence with $n$ heads and $m$ tails, the posterior on $\theta$ is:

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

$$\propto \theta^{a+n-1}(1-\theta)^{b+m-1}$$

$$\sim \text{Beta}(a+n, b+m).$$
Bayesian Approach to Estimating $\theta$

After observing $X$, a sequence with $n$ heads and $m$ tails, the posterior on $\theta$ is:

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$
$$ \propto \theta^{a+n-1}(1-\theta)^{b+m-1}$$
$$\sim \text{Beta}(a+n, b+m).$$

If $a = b = 1$ and we observe 5 heads and 2 tails, Beta(6, 3) looks like
The Dirichlet Distribution

We had

$$\pi \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$$

The Dirichlet density is defined as

$$p(\pi|\alpha) = \frac{\Gamma \left( \sum_{k=1}^{K} \alpha_k \right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \pi_1^{\alpha_1} \pi_2^{\alpha_2} \cdots \pi_K^{\alpha_K}$$

where $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$.

The expectations of $\pi$ are

$$E(\pi_i) = \frac{\alpha_i}{\sum_{i=1}^{K} \alpha_i}$$
A special case of the Dirichlet distribution is the Beta distribution when $K = 2$.

$$p(\pi | \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \pi^{\alpha_1-1}(1 - \pi)^{\alpha_2-1}$$
The Dirichlet Distribution

In three dimensions:

\[ p(\pi | \alpha_1, \alpha_2, \alpha_3) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \pi_1^{\alpha_1-1} \pi_2^{\alpha_2-1} (1 - \pi_1 - \pi_2)^{\alpha_3-1} \]

\[ \alpha = (2, 2, 2) \quad \alpha = (5, 5, 5) \quad (2, 2, 5) \]
Draws from the Dirichlet Distribution

\[ \alpha = (2, 2, 2) \]

\[ \alpha = (5, 5, 5) \]

\[ \alpha = (2, 2, 5) \]
The **Aggregation Property**: If

\[(\pi_1, \ldots, \pi_i, \pi_{i+1}, \ldots, \pi_K) \sim \text{Dir}(\alpha_1, \ldots, \alpha_i, \alpha_{i+1}, \ldots, \alpha_K)\]

then

\[(\pi_1, \ldots, \pi_i + \pi_{i+1}, \ldots, \pi_K) \sim \text{Dir}(\alpha_1, \ldots, \alpha_i + \alpha_{i+1}, \ldots, \alpha_K)\]

This is also valid for any aggregation:

\[\left(\pi_1 + \pi_2, \sum_{k=3}^{K} \pi_k\right) \sim \text{Beta}\left(\alpha_1 + \alpha_2, \sum_{k=3}^{K} \alpha_k\right)\]
Let $Z \sim \text{Multinomial}(\pi)$ and $\pi \sim \text{Dir}(\alpha)$.

Posterior:

$$p(\pi | z) \propto p(z | \pi)p(\pi)$$
$$= (\pi_1^{z_1} \cdots \pi_K^{z_K})(\pi_1^{\alpha_1-1} \cdots \pi_K^{\alpha_K-1})$$
$$= (\pi_1^{z_1+\alpha_1-1} \cdots \pi_K^{z_K+\alpha_K-1})$$

which is $\text{Dir}(\alpha + z)$. 
1 Introduction
2 Dirichlet Processes - Nonparametric Clustering
3 Gaussian Processes - Nonparametric Regression
4 Case Study: The Relational Automatic Statistician System

5 Appendix
- Conjugate Prior
- Dirichlet Processes
- Chinese Restaurant Processes
- Beta Processes
- Indian Buffet Processes
- Nonparametric Hierarchical Models
The Dirichlet Process Model

Parameters for the Dirichlet Process

- $\alpha$ - The concentration parameter.
- $G_0$ - The base measure. A prior distribution for the cluster specific parameters.

The Dirichlet Process (DP) is a distribution over distributions. We write

$$G \sim DP(\alpha, G_0)$$

to indicate $G$ is a distribution drawn from the DP.

It will become clearer in a bit what $\alpha$ and $G_0$ are.
The Dirichlet Process

**Definition:** Let $G_0$ be a probability measure on the measurable space $(\Omega, B)$ and $\alpha \in \mathbb{R}^+$. The *Dirichlet Process* $DP(\alpha, G_0)$ is the distribution on probability measures $G$ such that for any finite partition $(A_1, \ldots, A_m)$ of $\Omega$,

$$(G(A_1), \ldots, G(A_m)) \sim \text{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_m)).$$

(Ferguson, '73)
Mathematical Properties of the Dirichlet Process

Suppose we sample

- $G \sim DP(\alpha, G_0)$
- $\theta_1 \sim G$

What is the posterior distribution of $G$ given $\theta_1$?
Suppose we sample
- \( G \sim DP(\alpha, G_0) \)
- \( \theta_1 \sim G \)

What is the posterior distribution of \( G \) given \( \theta_1 \)?

\[
G|\theta_1 \sim DP \left( \alpha + 1, \frac{\alpha}{\alpha + 1} G_0 + \frac{1}{\alpha + 1} \delta_{\theta_1} \right)
\]

More generally

\[
G|\theta_1, \ldots, \theta_n \sim DP \left( \alpha + n, \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i} \right)
\]
Mathematical Properties of the Dirichlet Process

With probability 1, a sample $G \sim DP(\alpha, G_0)$ is of the form

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

(Sethuraman, '94)
The Stick-Breaking Process

- Define an infinite sequence of Beta random variables:

\[ \beta_k \sim \text{Beta}(1, \alpha) \quad k = 1, 2, \ldots \]

- And then define an infinite sequence of mixing proportions as:

\[ \pi_1 = \beta_1 \]

\[ \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \quad k = 2, 3, \ldots \]

- This can be viewed as breaking off portions of a stick:

\[ \frac{\beta_1}{\beta_1} \frac{\beta_2 (1-\beta_1)}{\beta_2 (1-\beta_1)} \ldots \]

- When \( \pi \) are drawn this way, we can write \( \pi \sim \text{GEM}(\alpha) \).
The Dirichlet Process Model

The Stick-Breaking Process

• We now have an explicit formula for each $\pi_k$:
  \[ \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \]

• We can also easily see that $\sum_{k=1}^{\infty} \pi_k = 1$ (wp1):

  \[
  1 - \sum_{k=1}^{K} \pi_k = 1 - \beta_1 - \beta_2 (1 - \beta_1) - \beta_3 (1 - \beta_1)(1 - \beta_2) - \cdots
  \]
  \[
  = (1 - \beta_1)(1 - \beta_2 - \beta_3 (1 - \beta_2) - \cdots)
  \]
  \[
  = \prod_{k=1}^{K} (1 - \beta_k)
  \]
  \[
  \to 0 \quad \text{(wp1 as } K \to \infty)\]

• So now $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ has a clean definition as a random measure
Contents

1 Introduction

2 Dirichlet Processes - Nonparametric Clustering

3 Gaussian Processes - Nonparametric Regression

4 Case Study: The Relational Automatic Statistician System

5 Appendix
  - Conjugate Prior
  - Dirichlet Processes
  - Chinese Restaurant Processes
  - Beta Processes
  - Indian Buffet Processes
  - Nonparametric Hierarchical Models
The 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
  - $m$th subsequent customer sits at a table drawn from the following distribution:

\[
\begin{align*}
P(\text{previously occupied table } i|\mathcal{F}_{m-1}) & \propto n_i \\
P(\text{the next unoccupied table}|\mathcal{F}_{m-1}) & \propto \alpha
\end{align*}
\]

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 Dirichlet Process Model

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 $(\phi_k)$ 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
The CRP Prior, Gaussian Likelihood, Conjugate Prior
The CRP and the DP

OK, so we’ve seen how the CRP relates to clustering. How does it relate to the DP?

Important fact: The CRP is exchangeable.

Remember De Finetti’s Theorem: If \((x_1, x_2, \ldots)\) are infinitely exchangeable, then

\[
\forall n \ p(x_1, \ldots, x_n) = \int \left( \prod_{i=1}^{n} p(x_i|G) \right) dP(G)
\]

for some random variable \(G\).
The CRP and the DP

OK, so we’ve seen how the CRP relates to clustering. How does it relate to the DP?

**Important fact:** The CRP is *exchangeable*.

Remember De Finetti’s Theorem: If \((x_1, x_2, \ldots)\) are *infinitely exchangeable*, then \(\forall n\)

\[
p(x_1, \ldots, x_n) = \int \left( \prod_{i=1}^{n} p(x_i|G) \right) dP(G)
\]

for some random variable \(G\).
The Dirichlet Process Model

The CRP and the DP

The Dirichlet Process is the De Finetti mixing distribution for the CRP.
The Dirichlet Process is the De Finetti mixing distribution for the CRP.

That means, when we integrate out $G$, we get the CRP.

$$p(\theta_1, \ldots, \theta_n) = \int \prod_{i=1}^{n} p(\theta_i | G) dP(G)$$
The Dirichlet Process Model

The CRP and the DP

The Dirichlet Process is the De Finetti mixing distribution for the CRP.

In English, this means that if the DP is the prior on $G$, then the CRP defines how points are assigned to clusters when we integrate out $G$. 
The Dirichlet Process Model

The DP, CRP, and Stick-Breaking Process Summary

\[ G \sim \text{DP}(\alpha, G_0) \]

The CRP describes the partitions of \( \theta \) when \( G \) is marginalized out.
Beta Processes

- Definition [Hjort, 90]: Let $H_0$ be a continuous probability measure $(\Omega, \mathcal{B})$ and $\alpha \in \mathbb{R}^+$. Then, Beta Process $BP(\alpha, H_0)$ is the distribution on probability measures $H$ such that for any (disjoint) finite partition $(A_1, \cdots, A_k)$ of $\Omega$ satisfies

$$H(A_i) \sim \text{Beta}(\alpha H_0(A_i), \alpha(1 - H_0(A_i)))$$

with $K \to \infty$ and $H_0(A_i) \to 0$ for $i = 1, \cdots, K$.

The beta process can be written in set function form,

$$H(w) = \sum_{k=1}^{\infty} \pi_k \delta_{w_k}(w)$$
Contents

1 Introduction

2 Dirichlet Processes - Nonparametric Clustering

3 Gaussian Processes - Nonparametric Regression

4 Case Study: The Relational Automatic Statistician System

5 Appendix

- Conjugate Prior
- Dirichlet Processes
- Chinese Restaurant Processes
- Beta Processes
- Indian Buffet Processes
- Nonparametric Hierarchical Models
Indian Buffet Processes

- Latent feature models
  - Clustering is not enough for mixed memberships
  - Grouping problem with overlapping clusters
  - Encode as binary matrix: Observation $n$ in cluster $k \iff x_{nk} = 1$
  - Alternatively: Item $n$ possesses feature $k \iff x_{nk} = 1$

- Indian buffet process (IBP)
  1. Customer 1 tries Poisson($\alpha$) dishes.
  2. Subsequent customer $n + 1$:
     - tries a previously tried dish $k$ with probability $\frac{n_k}{n+1}$
     - tries Poisson($\frac{\alpha}{n+1}$) new dishes.

- Properties
  - An exchangeable distribution over finite sets (of dishes).
  - Observation (=customer) $n$ in cluster (=dish) $k$ if customer ‘tries dish $k$’
Indian Buffet Process

- Alternative description

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

- Beta Process (BP)

Beta process is the de Finetti measure of the IBP.

Distribution on objects of the form

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

- 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.
- $\theta$ is a random measure
Binary matrices in left-order form

\[ \text{lof} \]
### Contents

1. Introduction
2. Dirichlet Processes - Nonparametric Clustering
3. Gaussian Processes - Nonparametric Regression
4. Case Study: The Relational Automatic Statistician System
5. Appendix
   - Conjugate Prior
   - Dirichlet Processes
   - Chinese Restaurant Processes
   - Beta Processes
   - Indian Buffet Processes
   - Nonparametric Hierarchical Models
Hierarchical Gaussian Processes

- Apply Bayesian representation recursively
  
  Split parameter $\Theta$: $\Theta \rightarrow \Psi$ and $\Theta|\Psi$

  $$
  p(data) = p(data|\Theta)p(\Theta) = p(data|\Theta)p(\Theta|\Psi)p(\Psi)
  $$

- Example: Hierarchical Gaussian process

  - Sample $\Psi \sim p(\Psi)$
    (e.g., large length-scale, mean 0)

  - Sample $\Theta|\Psi \sim p(\cdot|\Psi)$
    (e.g., smaller length scale, mean $\Psi$)

Decompose underlying pattern:
- Low-frequency component $\Psi$
- High-frequency component $\Theta$
Hierarchical Dirichlet Processes

- **Sampling scheme**
  - Sample $G_0 \sim DP(\gamma, H)$
  - Sample $G_1, G_2, \cdots \sim DP(\alpha, G_0)$
  - Sample $x_{ij} \sim G_j$ $G_1, G_2, \cdots$ have common ‘vocabulary’ of atoms.

- **Nonparametric Latent Dirichlet Allocation (LDA)**

\[
G_0 = \sum_{k=1}^{\infty} c_k \delta_{\theta_k^*} \quad G_j = \sum_{l=1}^{\infty} D_{ij} \delta_{\phi_{ij}}
\]

- $\theta_k = \text{finite probability (‘topic’)}$
- $c_k = \text{occurrence probability of topic } k$
- Document $j$ drawn from weighted combination of topics, with proportions $D_{ij}$ (‘admixture model’)

Jaesik Choi (UNIST)