Beta processes, stick-breaking, and power laws

T. Broderick, M. Jordan, J. Pitman

Presented by Jixiong Wang & J. Li

November 17, 2011

DP vs. BP

- Dirichlet Process
- Beta Process

DP vs. BP

- Dirichlet Process
- $G \sim DP(\alpha B_0)$:

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\psi_i}, \quad \sum_{i=1}^{\infty} \pi_i = 1$$

- Beta Process
- $G \sim BP(\theta, \gamma B_0)$:

$$G = \sum_{i=1}^{\infty} q_i \delta_{\psi_i}, \quad q_i \in (0, 1)$$

DP vs. BP

- Dirichlet Process
- $G \sim DP(\alpha B_0)$:

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\psi_i}, \quad \sum_{i=1}^{\infty} \pi_i = 1$$

- CRP - marginalize out π_i

- Beta Process
- $G \sim BP(\theta, \gamma B_0)$:

$$G = \sum_{i=1}^{\infty} q_i \delta_{\psi_i}, \quad q_i \in (0, 1)$$

- IBP - marginalize out q_i

DP vs. BP

- Dirichlet Process
- $G \sim DP(\alpha B_0)$:

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\psi_i}, \quad \sum_{i=1}^{\infty} \pi_i = 1$$

- CRP - marginalize out π_i
- Clustering framework

- Beta Process
- $G \sim BP(\theta, \gamma B_0)$:

$$G = \sum_{i=1}^{\infty} q_i \delta_{\psi_i}, \quad q_i \in (0,1)$$

- IBP - marginalize out q_i
- Featural framework

Poisson Point Process (PPP)

- PPP: A counting measure N such that $\forall A \in \mathcal{S}, N(A) \sim Pois(\mu(A))$

Poisson Point Process (PPP)

- PPP: A counting measure N such that $\forall A \in \mathcal{S}, N(A) \sim Pois(\mu(A))$

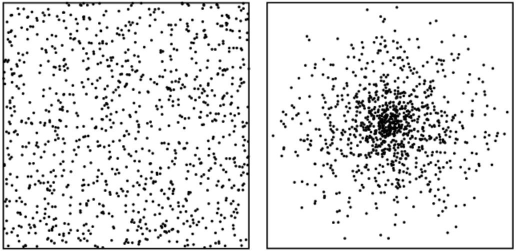


Figure: PPP realizations with different rate measure μ

Poisson Point Process (PPP)

- PPP: A counting measure N such that $\forall A \in \mathcal{S}, N(A) \sim Pois(\mu(A))$

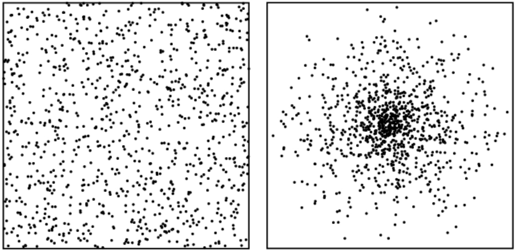


Figure: PPP realizations with different rate measure μ

- PPP is a completely random measure because for all disjoint subsets $A_1, ..., A_n \in \mathcal{S}$, $N(A_1), ..., N(A_n)$ are independent.
  - Note: DP is not a c.r.m..

Beta Process: $B \sim BP(\theta, \gamma B_0)$

BP is defined by a PPP that lives on $\Psi \times [0,1]$

- Rate measure: $\nu(d\psi, du) = \theta(\psi) u^{-1} (1-u)^{\theta(\psi)-1} du \gamma B_0(d\psi)$

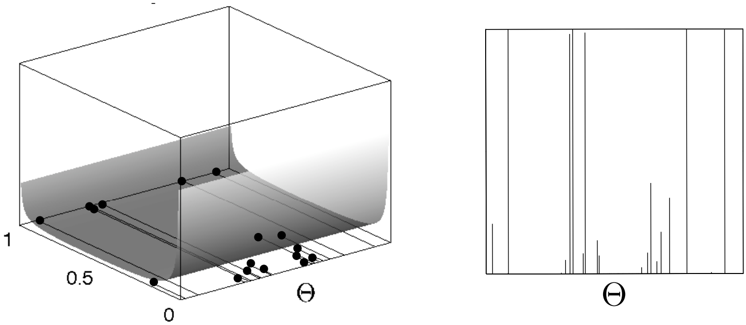
Beta Process: $B \sim BP(\theta, \gamma B_0)$

BP is defined by a PPP that lives on $\Psi \times [0,1]$

- Rate measure: $\nu(d\psi, du) = \theta(\psi) u^{-1} (1-u)^{\theta(\psi)-1} du \gamma B_0(d\psi)$
- To draw $B \sim BP(\theta, \gamma B_0)$
  - $\implies \Pi = \{(\psi_i, U_i)\}_i$
  - $\implies B = \sum_{i=1}^{\infty} U_i \delta_{\psi_i}$

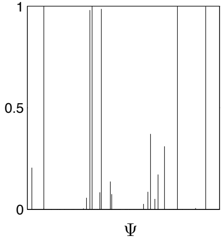
BP is defined by a PPP that lives on $\Psi \times [0, 1]$

- Rate measure: $\nu(d\psi, du) = \theta(\psi)u^{-1}(1-u)^{\theta(\psi)-1} du \gamma B_0(d\psi)$
- To draw $B \sim BP(\theta, \gamma B_0)$
  - $\implies \Pi = \{(\psi_i, U_i)\}_i$
  - $\implies B = \sum_{i=1}^{\infty} U_i \delta_{\psi_i}$



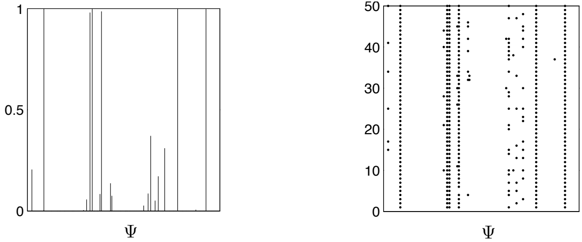
Bernoulli process & Binary feature matrix

- $Y \sim BeP(B)$: $Y = \sum_{i=1}^{\infty} b_i \delta_{\psi_i}$, where $b_i \sim Bern(U_i)$
- Draw $Y_1, \ldots, Y_N \sim BeP(B)$



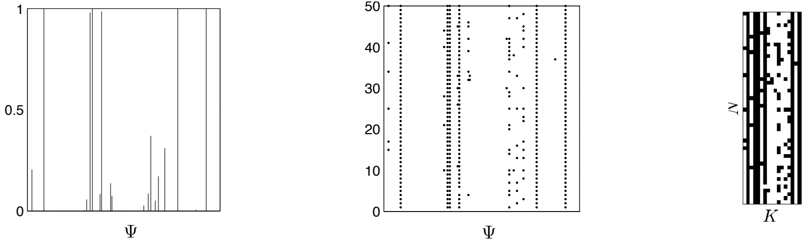
Bernoulli process & Binary feature matrix

- $Y \sim BeP(B)$: $Y = \sum_{i=1}^{\infty} b_i \delta_{\psi_i}$, where $b_i \sim Bern(U_i)$
- Draw $Y_1, \ldots, Y_N \sim BeP(B)$



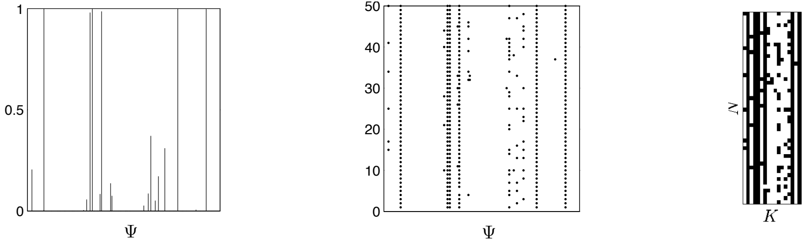
Bernoulli process & Binary feature matrix

- $Y \sim BeP(B)$: $Y = \sum_{i=1}^{\infty} b_i \delta_{\psi_i}$, where $b_i \sim Bern(U_i)$
- Draw $Y_1, ..., Y_N \sim BeP(B)$

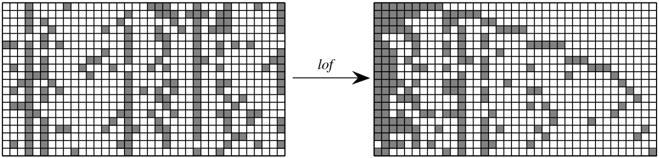


Bernoulli process & Binary feature matrix

- $Y \sim BeP(B)$: $Y = \sum_{i=1}^{\infty} b_i \delta_{\psi_i}$, where $b_i \sim Bern(U_i)$
- Draw $Y_1, ..., Y_N \sim BeP(B)$



- Form binary feature matrix $Z \sim$ BP-BeP(N, γ, θ)



[Ghahramani et al '06]

Stick-breaking construction of BP

$$
\begin{aligned}
B &= \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} V_{i,j}^{(i)} \prod_{l=1}^{i-1} (1 - V_{i,j}^{(l)}) \delta_{\psi_{i,j}} \\
C_i &\stackrel{iid}{\sim} \mathrm{Pois}(\gamma) \\
V_{i,j}^{(l)} &\stackrel{iid}{\sim} \mathrm{Beta}(1, \theta) \\
\psi_{i,j} &\stackrel{iid}{\sim} \frac{1}{\gamma} B_0.
\end{aligned}
$$

Stick-breaking construction of BP

$$
\begin{aligned}
B &= \sum_{i=1}^{\infty}\sum_{j=1}^{C_i} V_{i,j}^{(i)} \prod_{l=1}^{i-1}(1 - V_{i,j}^{(l)})\delta_{\psi_{i,j}} \\
C_i &\overset{iid}{\sim} \mathrm{Pois}(\gamma) \\
V_{i,j}^{(l)} &\overset{iid}{\sim} \mathrm{Beta}(1, \theta) \\
\psi_{i,j} &\overset{iid}{\sim} \frac{1}{\gamma} B_0.
\end{aligned}
$$

(Paisley et al 2010):

$$
\begin{aligned}
B = &\sum_{j=1}^{C_1} V_{1,j}^{(1)} \delta_{\psi_{1,j}} + \\
&\sum_{j=1}^{C_2} V_{2,j}^{(2)} (1 - V_{ij}^{(1)}) \delta_{\psi_{2,j}} + \\
&\sum_{j=1}^{C_3} V_{3,j}^{(3)} (1 - V_{3,j}^{(2,)})(1 - V_{3,j}^{(1)}) \delta_{\psi_{3,j}} + \ldots
\end{aligned}
$$

Stick-breaking construction of BP

(Paisley et al 2010):

$$
\begin{aligned}
B &= \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} V_{i,j}^{(i)} \prod_{l=1}^{i-1} (1 - V_{i,j}^{(l)}) \delta_{\psi_{i,j}} \\
C_i &\overset{iid}{\sim} \mathrm{Pois}(\gamma) \\
V_{i,j}^{(l)} &\overset{iid}{\sim} \mathrm{Beta}(1, \theta) \\
\psi_{i,j} &\overset{iid}{\sim} \frac{1}{\gamma} B_0.
\end{aligned}
$$

$$
\begin{aligned}
B = &\sum_{j=1}^{C_1} V_{1,j}^{(1)} \delta_{\psi_{1,j}} + \\
&\sum_{j=1}^{C_2} V_{2,j}^{(2)} (1 - V_{ij}^{(1)}) \delta_{\psi_{2,j}} + \\
&\sum_{j=1}^{C_3} V_{3,j}^{(3)} (1 - V_{3,j}^{(2,)})(1 - V_{3,j}^{(1)}) \delta_{\psi_{3,j}} + \ldots
\end{aligned}
$$

- Think of each i as a “round”
- It is “a multiple of stick-breaking DP”

Three parameter generalization

- 3 parameter stick-breaking ("a multiple of Pitman-Yor")

$$
\begin{aligned}
B &= \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} V_{i,j}^{(i)} \prod_{l=1}^{i-1} (1 - V_{i,j}^{(l)}) \delta_{\psi_{i,j}} \\
C_i &\stackrel{iid}{\sim} \text{Pois}(\gamma) \\
V_{i,j}^{(l)} &\stackrel{indep}{\sim} \text{Beta}(1 - \alpha, \theta + i\alpha) \\
\psi_{i,j} &\stackrel{iid}{\sim} \frac{1}{\gamma} B_0.
\end{aligned}
$$

Three parameter generalization

- 3 parameter stick-breaking ("a multiple of Pitman-Yor")

$$
\begin{aligned}
B &= \sum_{i=1}^{\infty}\sum_{j=1}^{C_i} V_{i,j}^{(i)} \prod_{l=1}^{i-1}(1 - V_{i,j}^{(l)})\delta_{\psi_{i,j}} \\
C_i &\overset{iid}{\sim} \mathrm{Pois}(\gamma) \\
V_{i,j}^{(l)} &\overset{indep}{\sim} \mathrm{Beta}(1-\alpha, \theta + i\alpha) \\
\psi_{i,j} &\overset{iid}{\sim} \frac{1}{\gamma} B_0.
\end{aligned}
$$

- 3 parameter $BP(\theta, \alpha, B_0)$. Rate measure:

$$
\begin{aligned}
\nu_{BP}(d\psi, du) =& B_o(d\psi) \times \mu_{BP}(du) \\
=& B_o(d\psi) \times \frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} u^{-1-\alpha}(1-u)^{\theta+\alpha-1} du
\end{aligned}
$$

Equivalence – An elegant proof

Proposition 1

B presented in the stick-breaking construction is equivalent to $B \sim BP(\theta, \alpha, B_0)$

Equivalence – An elegant proof

Proposition 1

B presented in the stick-breaking construction is equivalent to $B \sim BP(\theta, \alpha, B_0)$

Idea of proof:

- The stick-breaking representation is also a PPP, and induces rate measure ν

Equivalence – An elegant proof

Proposition 1

B presented in the stick-breaking construction is equivalent to $B \sim BP(\theta, \alpha, B_0)$

Idea of proof:

- The stick-breaking representation is also a PPP, and induces rate measure ν
- Therefore only need to show that $\nu = \nu_{BP}$

Power law behavior:

Power laws in clustering models:

- $K_{N,j} = \sum_{i=1}^{\infty} I(N_i = j)$
- $K_N = \sum_{i=1}^{\infty} I(N_i > 0)$
- Type 1: $K_N \sim cN^a, N \to \infty$
- Type 2: $K_{N,j} \sim \frac{a\Gamma(j-a)}{j!\Gamma(1-a)} cN^a, N \to \infty$

Power laws in featural models:

- Type 3: $P(k_n > M) \sim cM^{-a}$

Power law derivations: Type 1 and 2

Poissonization

Mean feature counts

Proposition 3

$K(t), K_j(t)$

$\Phi(t) = E[K(t)], \Phi_j(t) = E[K_j(t)]$

$K(N), K_j(N)$

$\Phi(N), \Phi_j(N)$

Lemma 4 & 5

$K_N, K_{N,j}$

$\Phi_N = E[K_N], \Phi_{N,j} = E[K_{N,j}]$

Proposition 6

$K(t)$ will be the number of such Poisson processes with points in the interval $[0, t]$

- $K(t) = \sum_i I|\Pi_i \bigcap [0, t]| > 0$

$K_j(t)$ will be the number of such Poisson processes with j points in the interval $[0, t]$

- $K_j(t) = \sum_i I|\Pi_i \bigcap [0, t]| = j$

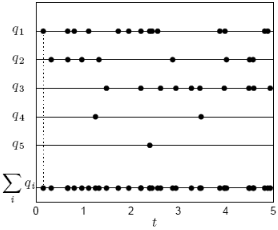


Figure 4: The first five sets of points, starting from the top of the figure, illustrate Poisson processes on the positive half-line in the range $t \in (0, 5)$ with respective rates $q_1, \ldots, q_5$. The bottom set of points illustrates the union of all points from the preceding Poisson point processes and is, therefore, itself a Poisson process with rate $\sum_i q_i$. In this example, we have for instance that $K(1) = 2$, $K(4) = 5$, and $K_2(4) = 1$.

Power law derivations

Theorem 2 (Part of Campbell's Theorem). *Let Π be a Poisson process on S with rate measure μ, and let $f : S \to \mathbb{R}$ be measurable. If $\int_S \min(|f(x)|, 1)\, \mu(dx) < \infty$, then*

$$\mathbb{E}\left[\sum_{X \in \Pi} f(X)\right] = \int_S f(x)\, \mu(dx). \tag{21}$$

$$\Phi(t) = \mathbb{E}[\sum_i (1 - e^{-tq_i})] = \int_0 (1 - e^{-tx})\, \nu(dx)$$

$$\Phi_N = \mathbb{E}[\sum_i (1 - (1 - q_i)^N)] = \int_0^1 (1 - (1 - x)^N)\, \nu(dx)$$

$$\Phi_j(t) = \mathbb{E}[\sum_i \frac{(tq_i)^j}{j!} e^{-tq_i}] = \frac{t^j}{j!} \int_0^1 x^j e^{-tx}\, \nu(dx)$$

$$\Phi_{N,j} = \binom{N}{j} \mathbb{E}[\sum_i q_i^j (1 - q_i)^{N-j}] = \binom{N}{j} \int_0^1 x^j (1 - x)^{N-j}\, \nu(dx).$$

Proposition 3. *Asymptotic behavior of the integral of ν of the following form*

$$\nu_1[0, x] := \int_0^x u\, \nu(du) \sim \frac{\alpha}{1 - \alpha} x^{1-\alpha} l(1/x), \quad x \to 0 \tag{27}$$

where l is a regularly varying function and $\alpha \in (0, 1)$ implies

$$\begin{aligned} \Phi(t) &\sim \Gamma(1 - \alpha) t^\alpha l(t), \quad t \to \infty \\ \Phi_j(t) &\sim \frac{\alpha \Gamma(j - \alpha)}{j!} t^\alpha l(t), \quad t \to \infty \quad (j > 1). \end{aligned}$$

Power law derivations

Lemma 4. *Let ν be σ-finite with $\int_0^\infty \nu(du) = \infty$ and $\int_0^\infty u\,\nu(du) < \infty$. Then the number of represented features has unbounded growth almost surely. The expected number of represented features has unbounded growth, and the expected number of features has sublinear growth. That is,*

$$K(t) \uparrow \infty \text{ a.s.}, \quad \Phi(t) \uparrow \infty, \quad \Phi(t) \ll t.$$

Lemma 5. *Suppose the $\{q_i\}$ are generated according to a Poisson process with rate measure as in Lemma 4. Then, for $N \to \infty$,*

$$|\Phi_N - \Phi(N)| < \frac{2}{N}\Phi_2(N) \to 0$$

$$|\Phi_{N,j} - \Phi_j(N)| < \frac{c_j}{N}\max\{\Phi_j(N), \Phi_{j+2}(N)\} \to 0.$$

for some constants c_j.

Proposition 6. *Suppose the $\{q_i\}$ are generated from a Poisson process with rate measure as in Lemma 4. For $N \to \infty$,*

$$K_N \overset{a.s.}{\sim} \Phi_N, \quad \sum_{k<j} K_{N,k} \overset{a.s.}{\sim} \sum_{k<j} \Phi_{N,k}.$$

Poissonization

Mean feature counts

Proposition 3

$$K(t), K_j(t) \qquad \Phi(t) = E[K(t)], \Phi_j(t) = E[K_j(t)]$$

$$K(N), K_j(N) \qquad \Phi(N), \Phi_j(N)$$

Lemma 4 & 5

$$K_N, K_{N,j} \qquad \Phi_N = E[K_N], \Phi_{N,j} = E[K_{N,j}]$$

Proposition 6

Power law derivations: Type 3

Let Z_i be a Bernoulli random variable with success probability q_i and such that all the Z_i are independent. Then $\mathbb{E}[\sum_i Z_i] = \sum_i q_i =: Q$. In this case, a Chernoff bound [Chernoff, 1952, Hagerup and Rub, 1990] tells us that, for any $\delta > 0$, we have

$$\mathbb{P}[\sum_i Z_i \geq (1+\delta)Q] \leq e^{\delta Q}(1+\delta)^{-(1+\delta)Q}.$$

When M is large enough such that $M > Q$, we can choose δ such that $(1+\delta)Q = M$. Then this inequality becomes

$$\mathbb{P}[\sum_i Z_i \geq M] \leq e^{M-Q} Q^M M^{-M} \quad \text{for } M > Q. \tag{31}$$

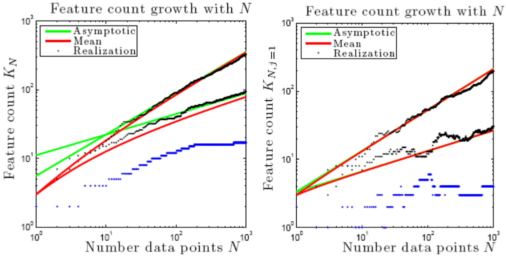
We see from Eq. (31) that the number of features $\sum_i Z_i$ that are expressed for a data point exhibits super-exponential tail decay and therefore cannot have a power law probability distribution when the sum of feature probabilities $\sum_i q_i$ is finite. For comparison, let $Z \sim \text{Pois}(Q)$. Then [Franceschetti et al., 2007]

$$\mathbb{P}[Z \geq M] \leq e^{M-Q} Q^M M^{-M} \quad \text{for } M > Q,$$

Simulation

- $\alpha = 0$ (classic), $\alpha = 0.3$ and $\alpha = 0.6$; $\gamma = 3$, $\theta = 1$.
- Generate 2000 random variables C_i and $\sum_{i=1}^{2000} C_i$ feature probabilities.
- With these probabilities, we generated $N = 1000$ data points, i.e., 1000 vectors of (2000) independent Bernoulli random variables.

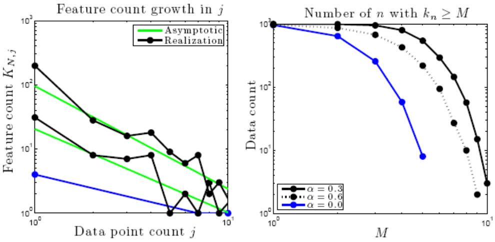
Simulation: Type 1 & 2



$$\phi_N = \mathbb{E}[K_N] = \mathbb{E}\left[\sum_{n=1}^{N} \text{Pois}\left(\gamma \frac{\theta}{n+\theta}\right)\right] = \sum_{n=1}^{N} \gamma \frac{\theta}{n+\theta} \sim \gamma\theta \log(N).$$

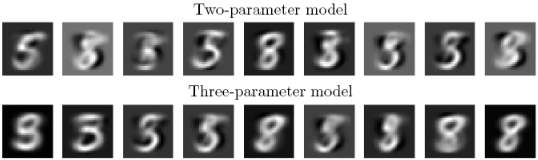
$$\begin{aligned}\Phi_{N,1} = \mathbb{E}[K_{N,1}] &= \binom{N}{1} \int_0^1 x^1 (1-x)^{N-1} \cdot \theta x^{-1} (1-x)^{\theta-1} \, dx \\ &= N\theta \cdot \frac{\Gamma(1)\Gamma(N-1+\theta)}{\Gamma(N+\theta)} = \theta \frac{N}{N-1+\theta} \sim \theta,\end{aligned}$$

Simulation: Type 3

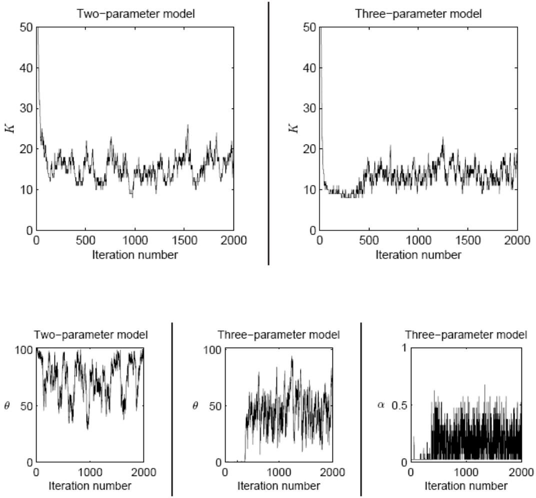


Experimental results

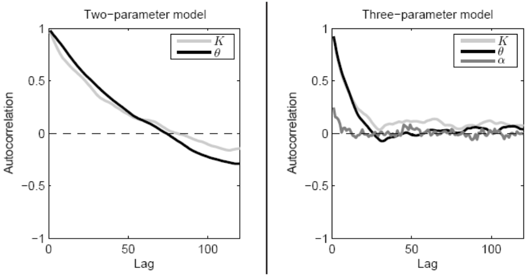
- Beta process coupled with a discrete factor analysis model.
- Handwritten digit: 28x28 pixels projected into 50 dimensions with PCA.



Experimental results



Experimental results



Conclusions

- (BP, stick-breaking, IBP) – (DP, stick-breaking, CRP)
- Three-parameter generalization of BP – Pitman-Yor generalization of DP
- Type 1 & 2 power laws follow from the three-parameter model.
- Type 3: an open problem to discover new class of stochastic process.