The beta process: survival analysis, latent feature models, and the Indian buffet process

Lloyd Elliott

Acknowledgements: Yee Whye Teh, Vinayak Rao

March 30, 2012
Outline

Conjugate priors for survival analysis

Link to completely random measures

Indian buffet process

Applications to machine learning
Survival analysis [Cox, 1972]
Let $X \geq 0$ be the lifetime of a process with cdf $F(t)$. 
Survival analysis [Cox, 1972]

We want to estimate the hazard rate:

\[ h(t) = \lim_{\delta \to 0^+} \delta^{-1} \Pr(X \leq t + \delta | X > t). \]  

(1)

We are given right censored observations:

\( X_i \) lifetime,  \hspace{1cm} (2)
\( T_i \) time of last observation,  \hspace{1cm} (3)
\( d_i \) censoring indicator,  \hspace{1cm} (4)
\( c_i \) time of censoring,  \hspace{1cm} (5)
\( T_i = \min\{X_i, c_i\} \)  \hspace{1cm} (6)
\( d_i = \mathbb{I}\{X_i \leq c_i\} \).  \hspace{1cm} (7)
Survival analysis [Cox, 1972]

We want to estimate the hazard rate:

\[ h(t) = \lim_{\delta \to 0^+} \delta^{-1} \Pr(X \leq t + \delta | X > t). \]  \hspace{1cm} (1)

We are given right censored observations:

- \( X_i \) lifetime, \hspace{1cm} (2)
- \( T_i \) time of last observation, \hspace{1cm} (3)
- \( d_i \) censoring indicator, \hspace{1cm} (4)
- \( c_i \) time of censoring, \hspace{1cm} (5)
- \( T_i = \min\{X_i, c_i\} \), \hspace{1cm} (6)
- \( d_i = \mathbb{I}\{X_i \leq c_i\} \). \hspace{1cm} (7)
Discrete approximation [Hjort, 1990]

First, we will look at the sets \([t, t + \delta)\) for \(t = 0, \delta, 2\delta, \ldots\)

\[ h(t) = \Pr(X \in [t, t + \delta)| T \geq t). \]  

(8)

Define the counting process \(N(t)\) and the number at risk \(Y(t)\) as follows:

\[
dN(t) = \sum_{i=1}^{n} \mathbb{I}\{T_i \in [t, t + \delta) \text{ and } d_i = 1\},
\]

(9)

\[
Y(t) = \sum_{i=1}^{n} \mathbb{I}\{T_i \geq t\}.
\]

(10)
Discrete approximation [Hjort, 1990]

First, we will look at the sets \([t, t + \delta)\) for \(t = 0, \delta, 2\delta, \ldots\)

\[ h(t) = \Pr(X \in [t, t + \delta) | T \geq t). \tag{8} \]

Define the counting process \(N(t)\) and the number at risk \(Y(t)\) as follows:

\[ dN(t) = \sum_{i=1}^{n} \mathbb{I}\{T_i \in [t, t + \delta) \text{ and } d_i = 1\}, \tag{9} \]

\[ Y(t) = \sum_{i=1}^{n} \mathbb{I}\{T_i \geq t\}. \tag{10} \]
Hazard rates

We assume that hazard rates $h(t)$ are independent r.v.'s in $[0, 1]$. Suppose that **a priori** $h(t)$ is distributed as $\alpha_t(u)$.

**Theorem**

The posterior density of $h(t)$ after observing $(T_i, d_i)_{i=1}^n$ is:

$$p(h(t) = u|T_i, d_i) \propto \Pr(T_i, d_i|h(t) = u)p(h(t) = u),$$

$$= u\#i:T_i \in [t, t+\delta) \text{ and } d_i = 1(1 - u)\#i:T_i \geq t + \delta \alpha_t(u)$$

$$= udN(t)(1 - u)^{Y(t) - dN(t)}\alpha_t(u).$$

This suggests that we should place a beta prior on $h(t)$:

$$h(t) \sim \text{Beta}(c(t)\mu_\delta(t), c(t)(1 - \mu_\delta(t))),$$

$$h(t)|T_i, d_i \sim \text{Beta}(c(t)\mu_\delta(t) + dN(t), c(t)\mu_\delta(t) + Y(t) - dN(t)),$$

$\mu_\delta(t) = \mu[t, t + \delta)$ is a mean measure and $c(t) \geq 0$ is a concentration.
Hazard rates

We assume that hazard rates $h(t)$ are independent r.v.'s in $[0, 1]$. Suppose that \textit{a priori} $h(t)$ is distributed as $\alpha_t(u)$.

**Theorem**

The posterior density of $h(t)$ after observing $(T_i, d_i)_{i=1}^n$ is:

$$p(h(t) = u | T_i, d_i) \propto \Pr(T_i, d_i | h(t) = u) p(h(t) = u),$$  
(11)

$$= u \#i: T_i \in [t, t+\delta) \text{ and } d_i = 1 \cdot (1 - u) \#i: T_i \geq t + \delta \alpha_t(u)$$  
(12)

$$= u^{dN(t)} (1 - u)^{Y(t) - dN(t)} \alpha_t(u).$$  
(13)

This suggests that we should place a beta prior on $h(t)$:

$$h(t) \sim \text{Beta}(c(t)\mu_\delta(t), c(t)(1 - \mu_\delta(t))),$$  
(14)

$$h(t) | T_i, d_i \sim \text{Beta}(c(t)\mu_\delta(t) + dN(t), c(t)\mu_\delta(t) + Y(t) - dN(t)).$$  
(15)

$\mu_\delta(t) = \mu[t, t + \delta]$ is a mean measure and $c(t) \geq 0$ is a concentration.
Hazard rates

We assume that hazard rates $h(t)$ are independent r.v.'s in $[0, 1]$. Suppose that a priori $h(t)$ is distributed as $\alpha_t(u)$.

**Theorem**

The posterior density of $h(t)$ after observing $(T_i, d_i)_{i=1}^n$ is:

$$p(h(t) = u | T_i, d_i) \propto \Pr(T_i, d_i | h(t) = u) p(h(t) = u),$$

$$= u^{\#i: T_i \in [t, t+\delta) \text{ and } d_i = 1} (1 - u)^{\#i: T_i \geq t + \delta} \alpha_t(u)$$

$$= u^{dN(t)} (1 - u)^{Y(t) - dN(t)} \alpha_t(u).$$

This suggests that we should place a beta prior on $h(t)$:

$$h(t) \sim \text{Beta}(c(t)\mu_\delta(t), c(t)(1 - \mu_\delta(t))),$$

$$h(t) | T_i, d_i \sim \text{Beta}(c(t)\mu_\delta(t) + dN(t), c(t)\mu_\delta(t) + Y(t) - dN(t)),$$

$\mu_\delta(t) = \mu[t, t + \delta]$ is a mean measure and $c(t) \geq 0$ is a concentration.
Hazard rates

We assume that hazard rates $h(t)$ are independent r.v.’s in $[0, 1]$. Suppose that a priori $h(t)$ is distributed as $\alpha_t(u)$.

Theorem

**The posterior density of $h(t)$ after observing $(T_i, d_i)_{i=1}^n$ is:**

$$p(h(t) = u | T_i, d_i) \propto \Pr(T_i, d_i | h(t) = u)p(h(t) = u),$$

$$= u^\#i: T_i \in [t, t+\delta) \text{ and } d_i = 1 (1 - u)^\#i: T_i \geq t + \delta \alpha_t(u)$$

$$= u^{dN(t)} (1 - u)^{Y(t) - dN(t)} \alpha_t(u).$$

This suggests that we should place a beta prior on $h(t)$:

$$h(t) \sim \text{Beta}(c(t)\mu_\delta(t), c(t)(1 - \mu_\delta(t))),$$

$$h(t) | T_i, d_i \sim \text{Beta}(c(t)\mu_\delta(t) + dN(t), c(t)\mu_\delta(t) + Y(t) - dN(t)),$$

$\mu_\delta(t) = \mu[t, t + \delta]$ is a mean measure and $c(t) \geq 0$ is a concentration.
Continuous hazard rates

We can write the cdf of the lifetime $X$ in terms of the hazard rate:

$$ F(t) \approx 1 - \prod_{k=0}^{[t/\delta]} (1 - h(k\delta)) . $$  \hspace{1cm} (16)

$$ \approx 1 - \exp(- \sum_{k=0}^{[t/\delta]} h(k\delta)) $$  \hspace{1cm} (17)

Theorem

Let $\mu$ be a measure and let $c(t) \geq 0$ be piecewise continuous. The cumulative hazard exists & is called a beta process:

$$ A(t) = \lim_{\delta \to 0^+} \sum_{k=0}^{[t/\delta]} h(k\delta) . $$  \hspace{1cm} (18)
Continuous hazard rates

We can write the cdf of the lifetime $X$ in terms of the hazard rate:

$$F(t) \approx 1 - \prod_{k=0}^{[t/\delta]} (1 - h(k\delta)). \quad (16)$$

$$\approx 1 - \exp(- \sum_{k=0}^{[t/\delta]} h(k\delta)) \quad (17)$$

limit is $A(t)$

Theorem

Let $\mu$ be a measure and let $c(t) \geq 0$ be piecewise continuous. The cumulative hazard exists & is called a beta process:

$$A(t) = \lim_{\delta \to 0^+} \sum_{k=0}^{[t/\delta]} h(k\delta). \quad (18)$$
Properties of the cumulative hazard

Corollary

1. $A(0) = 0$,
2. $A(t_i) - A(t_{i-1})$ are independent for all $0 \leq t_1 < t_2 < \ldots$,
3. $A(t)$ is right continuous,

The beta process $A$ can be seen as a measure on $\mathbb{R}_{\geq 0}$ by defining $A(t_0, t_1] = A(t_1) - A(t_0)$. By the above corollary, $A$ is a completely random measure (CRM): if $B_1, \ldots, B_n$ are disjoint then $A(B_1), \ldots, A(B_n)$ are independent.
Properties of the cumulative hazard

Corollary

1. \( A(0) = 0 \),
2. \( A(t_i) - A(t_{i-1}) \) are independent for all \( 0 \leq t_1 < t_2 < \ldots \),
3. \( A(t) \) is right continuous,

The beta process \( A \) can be seen as a measure on \( \mathbb{R}_{\geq 0} \) by defining \( A(t_0, t_1] = A(t_1) - A(t_0) \). By the above corollary, \( A \) is a completely random measure (CRM): if \( B_1, \ldots, B_n \) are disjoint then \( A(B_1), \ldots, A(B_n) \) are independent.
By the Lévy–Khinchine representation theorem (from lecture 2), there exists a measure $\lambda(du, ds)$ such that for all functions $f(s)$ on $\mathbb{R}_{\geq 0}$:

$$\mathbb{E} \left[ \exp \left( - \int_0^\infty f(s) A(ds) \right) \right] = \exp \left( - \int_0^\infty \int_0^1 1 - e^{-uf(s)} \lambda(du, ds) \right),$$

(19)

$$\lambda(du, ds) = c(s)u^{-1}(1 - u)^{c(s) - 1}\mu(ds).$$

(20)

Write $A \sim \text{BP}(c, \mu)$ in this case.
By the Lévy–Khinchine representation theorem (from lecture 2), there exists a measure $\lambda(du, ds)$ such that for all functions $f(s)$ on $\mathbb{R}_{\geq 0}$:

$$E \left[ \exp \left( - \int_0^\infty f(s)A(ds) \right) \right] = \exp \left( \int_0^\infty \int_0^1 1 - e^{-uf(s)} \lambda(du, ds) \right),$$  

(19)

$$\lambda(du, ds) = c(s)u^{-1}(1 - u)^{c(s)-1}\mu(ds).$$  

(20)

Write $A \sim \text{BP}(c, \mu)$ in this case.
A beta process $A \sim BP(c, \mu)$ is a completely random measure s.t.:

$$A = \sum_{k=1}^{\infty} w_k \delta_{s_k},$$

where $(w_k, s_k)_{k=1}^{\infty}$ is a Poisson process on $[0, 1] \times \mathbb{R}_{\geq 0}$ with rate

$$\lambda(du, ds) = cu^{-1}(1 - u)^{c-1} \mu(ds).$$
Latent feature models
Suppose $s_1, \ldots, s_K$ are features, and $z_{ik}$ indicates if data item $i$ has feature $k$.

$$z_{ik} = \begin{cases} 
1 & \text{if data item } i \text{ has feature } k, \\
0 & \text{otherwise}.
\end{cases} \quad (22)$$

This is a popular situation in Bayesian statistics, for example the elimination by aspects choice model [Görür et al., 2006]. Subjects are asked ‘with whom they would prefer to spend an hour of conversation’ given pairs from 9 celebrities (Rumelhart and Greeno 1971).

1. Celebrities have features $z_i$,
2. Subjects form preferences based on the features.

Generative process:

- A binary feature matrix $Z$ is selected,
- $w_1, \ldots, w_k \sim \mathcal{N}(1, 1)$.

$$\Pr(i \text{ beats } j) \propto \sum_{k=1}^{K} w_k Z_i(s_k)(1 - Z_j(s_k)), \quad (23)$$
Latent feature models

Suppose $s_1, \ldots, s_K$ are features, and $z_{ik}$ indicates if data item $i$ has feature $k$.

$$z_{ik} = \begin{cases} 1 & \text{if data item } i \text{ has feature } k, \\ 0 & \text{otherwise.} \end{cases} \quad (22)$$

This is a popular situation in Bayesian statistics, for example the elimination by aspects choice model [Görür et al., 2006]. Subjects are asked ‘with whom they would prefer to spend an hour of conversation’ given pairs from 9 celebrities (Rumelhart and Greeno 1971).

1. Celebredies have features $z_i$,
2. Subjects form preferences based on the features.

Generative process:

- A binary feature matrix $Z$ is selected,
- $w_1, \ldots, w_k \sim \mathcal{N}(1, 1)$.

$$\Pr(i \text{ beats } j) \propto \sum_{k=1}^{K} w_k z_i(s_k)(1 - z_j(s_k)), \quad (23)$$
Latent feature models

Suppose \( s_1, \ldots, s_K \) are features, and \( z_{ik} \) indicates if data item \( i \) has feature \( k \).

\[
Z_{ik} = \begin{cases} 
1 & \text{if data item } i \text{ has feature } k, \\
0 & \text{otherwise.} 
\end{cases}
\]  

(22)

This is a popular situation in Bayesian statistics, for example the elimination by aspects choice model [Görür et al., 2006]. Subjects are asked ‘with whom they would prefer to spend an hour of conversation’ given pairs from 9 celebrities (Rumelhart and Greeno 1971).

1. Celebredies have features \( z_i \),
2. Subjects form preferences based on the features.

Generative process:

- A binary feature matrix \( Z \) is selected,
- \( w_1, \ldots, w_k \sim \mathcal{N}(1, 1) \).

\[
\Pr(i \text{ beats } j) \propto \sum_{k=1}^{K} w_k Z_i(s_k) (1 - Z_j(s_k)),
\]  

(23)
Latent feature models
Suppose \( s_1, \ldots, s_K \) are features, and \( z_{ik} \) indicates if data item \( i \) has feature \( k \).

\[
z_{ik} = \begin{cases} 
1 & \text{if data item } i \text{ has feature } k, \\
0 & \text{otherwise.}
\end{cases}
\]

This is a popular situation in Bayesian statistics, for example the elimination by aspects choice model [Görür et al., 2006]. Subjects are asked ‘with whom they would prefer to spend an hour of conversation’ given pairs from 9 celebrities (Rumelhart and Greeno 1971).

1. Celebredies have features \( z_i \),
2. Subjects form preferences based on the features.

Generative process:

- A binary feature matrix \( Z \) is selected,
- \( w_1, \ldots, w_k \sim \mathcal{N}(1, 1) \).

\[
\Pr(i \text{ beats } j) \propto \sum_{k=1}^{K} w_k z_i(s_k)(1 - z_j(s_k)), \tag{23}
\]
Prior for features

Let $\pi_k$ be the prior probability of having feature $s_k$. If we assume the $\pi_k$ are independent r.v.s, the posterior densities are:

$$p(\pi_k|z_1, \ldots, z_n) \propto p(z_1, \ldots, z_n|\pi_k)p(\pi_k), \quad (24)$$

$$= \pi_k^{m_k}(1 - \pi_k)^{n-m_k}p(\pi_k). \quad (25)$$

This is the same situation as for the hazard function, suggesting a beta prior for $\pi_k$. 

Latent feature models
[Griffiths and Ghahramani, 2005]

Assume the prior probability of having feature $s_k$ is $\pi_k \sim \text{Beta}(\alpha/K, 1)$. The marginal probability of $Z$ is:

$$\Pr(Z) = \prod_{k=1}^{K} \int_0^1 \prod_{i=1}^{n} \Pr(z_{ik} = 1 | \pi_k) \rho(\pi_k) d\pi_k, \quad (26)$$

$$= \prod_{k=1}^{K} \frac{\alpha/K \Gamma(m_k + \alpha/K) \Gamma(n - m_k + 1)}{\Gamma(n + 1 + \alpha/K)}. \quad (27)$$

As $K \to \infty$, the expected number of nonzero columns of $Z$ is finite.

$$\lim_{K \to \infty} \Pr([Z]) = \alpha^{K^+} \exp \left(-\alpha \sum_{i=1}^{n} 1/i \right) \prod_{k=1}^{K^+} \frac{(n - m_k)! (m_k - 1)!}{n!}. \quad (28)$$

Here, $K^+$ is the number of nonzero columns.
Latent feature models
[Griffiths and Ghahramani, 2005]

Assume the prior probability of having feature $s_k$ is $\pi_k \sim \text{Beta}(\alpha/K, 1)$. The marginal probability of $Z$ is:

$$
\Pr(Z) = \prod_{k=1}^{K} \int_0^1 \prod_{i=1}^{n} \Pr(z_{ik} = 1|\pi_k) p(\pi_k) d\pi_k,
$$

(26)

$$
= \prod_{k=1}^{K} \frac{\alpha/K}{\Gamma(n+1+\alpha/K)} \frac{\Gamma(m_k + \alpha/K)\Gamma(n - m_k + 1)}{\Gamma(n + 1 + \alpha/K)}.
$$

(27)

As $K \to \infty$, the expected number of nonzero columns of $Z$ is finite.

$$
\lim_{K \to \infty} \Pr([Z]) = \alpha^{K+} \exp \left( -\alpha \sum_{i=1}^{n} 1/i \right) \prod_{k=1}^{K+} \frac{(n - m_k)! (m_k - 1)!}{n!}.
$$

(28)

Here, $K^+$ is the number of nonzero columns.
Latent feature models
[Griffiths and Ghahramani, 2005]

Assume the prior probability of having feature $s_k$ is $\pi_k \sim \text{Beta}(\alpha/K, 1)$. The marginal probability of $Z$ is:

$$ \Pr(Z) = \prod_{k=1}^{K} \int_{0}^{1} \prod_{i=1}^{n} \Pr(z_{ik} = 1 | \pi_k) p(\pi_k) d\pi_k, $$

$$ = \prod_{k=1}^{K} \frac{\alpha/K \Gamma(m_k + \alpha/K) \Gamma(n - m_k + 1)}{\Gamma(n + 1 + \alpha/K)}.$$

(26)

As $K \to \infty$, the expected number of nonzero columns of $Z$ is finite.

$$ \lim_{K \to \infty} \Pr([Z]) = \alpha^{K^+} \exp \left( -\alpha \sum_{i=1}^{n} 1/i \right) \prod_{k=1}^{K^+} \frac{(n - m_k)! (m_k - 1)!}{n!}.$$

(28)

Here, $K^+$ is the number of nonzero columns.
The Indian buffet process
[Griffiths and Ghahramani, 2005]
n customers enter an Indian buffet in sequence.

- Customer 1 chooses Poisson(\(\alpha\)) dishes.
- Customer \(i > 1\) picks a previously chosen dish with probability \(m_k/i\) and Poisson(\(\alpha/i\)) new dishes. (\(m_k\) is the \# of customers who have already chosen dish \(k\).)

The IBP is exchangeable and it induces a prior on binary matrices with \(n\) rows and an arbitrary number of columns.

- Row \(i\), column \(k\) indicates if customer \(i\) chose dish \(k\).
- Columns are labelled with draws \(s_k\).
- Posterior probability is:

\[
\alpha^K \exp \left( -\alpha \sum_{i=1}^{n} 1/i \right) \prod_{i=1}^{K} \frac{(m_k - 1)! (n - m_k)!}{n!} h(\theta_k^*). \tag{29}
\]
The Indian buffet process
[Griffiths and Ghahramani, 2005]

$n$ customers enter an Indian buffet in sequence.

- Customer 1 chooses $\text{Poisson}(\alpha)$ dishes.
- Customer $i > 1$ picks a previously chosen dish with probability $m_k/i$ and $\text{Poisson}(\alpha/i)$ new dishes. ($m_k$ is the \# of customers who have already chosen dish $k$.)

The IBP is exchangeable and it induces a prior on binary matrices with $n$ rows and an arbitrary number of columns.

- Row $i$, column $k$ indicates if customer $i$ chose dish $k$.
- Columns are labelled with draws $s_k$.
- Posterior probability is:

$$
\alpha^K \exp \left( -\alpha \sum_{i=1}^{n} 1/i \right) \prod_{i=1}^{K} \frac{(m_k - 1)! (n - m_k)!}{n!} h(\theta_k^*). \tag{29}
$$
The Indian buffet process
[Griffiths and Ghahramani, 2005]

$n$ customers enter an Indian buffet in sequence.

- Customer 1 chooses Poisson($\alpha$) dishes.
- Customer $i > 1$ picks a previously chosen dish with probability $m_k/i$ and Poisson($\alpha/i$) new dishes. ($m_k$ is the \# of customers who have already chosen dish $k$.)

The IBP is exchangeable and it induces a prior on binary matrices with $n$ rows and an arbitrary number of columns.

- Row $i$, column $k$ indicates if customer $i$ chose dish $k$.
- Columns are labelled with draws $s_k$.
- Posterior probability is:

$$\alpha^K \exp \left( -\alpha \sum_{i=1}^{n} 1/i \right) \prod_{i=1}^{K} \frac{(m_k - 1)!(n - m_k)!}{n!} h(\theta_k^*).$$

(29)
Applications to machine learning:

- Elimination by aspects choice model [Görür et al., 2006],
- Infinite ICA [Knowles and Ghahramani, 2007, Doshi et al., 2009].
- Latent feature relational model [Miller et al., 2009].
- Word frequency models [Teh and Görür, 2009].
Applications: infinite ICA
[Knowles and Ghahramani, 2007, Doshi et al., 2009]

Given signals $Y_i$. Assume latent sources $X$ are selected by a binary feature matrix, and then mixed by $G$.

$$Y_i = G(Z_i \odot X_i) + E,$$

1. $Z \sim \text{IBP}(c, \mu)$,

(a) Hinton diagram of the average mixing matrix, $G$, for iICA$_2$ applied to the financial dataset.

(b) Hinton diagram of the mixing matrix for FastICA (pow3) applied to the financial dataset.

Figure 16: Application to financial data set.
Applications: latent feature relational model
[Miller et al., 2009]

Prior for directed graphs. Each vertex has a latent binary feature vector $z_i$. Probability of an edge between vertices is an inner product of the feature vectors passed through a sigmoid.

- $Z \sim \text{IBP}(\alpha)$,
- $\Pr(e_{ij} = 1) = \text{sigmoid}(z_i^T B z_j^T)$.

<table>
<thead>
<tr>
<th></th>
<th>Countries single</th>
<th>Countries global</th>
<th>Alyawarra single</th>
<th>Alyawarra global</th>
</tr>
</thead>
<tbody>
<tr>
<td>LFRM w/ IRM</td>
<td>0.8521 ± 0.0035</td>
<td>0.8772 ± 0.0075</td>
<td>0.9346 ± 0.0013</td>
<td>0.9183 ± 0.0108</td>
</tr>
<tr>
<td>LFRM rand</td>
<td>0.8529 ± 0.0037</td>
<td>0.7067 ± 0.0534</td>
<td>0.9443 ± 0.0018</td>
<td>0.7127 ± 0.030</td>
</tr>
<tr>
<td>IRM</td>
<td>0.8423 ± 0.0034</td>
<td>0.8500 ± 0.0033</td>
<td>0.9310 ± 0.0023</td>
<td>0.8943 ± 0.0300</td>
</tr>
<tr>
<td>MMSB</td>
<td>0.8212 ± 0.0032</td>
<td>0.8643 ± 0.0077</td>
<td>0.9005 ± 0.0022</td>
<td>0.9143 ± 0.0097</td>
</tr>
</tbody>
</table>

Figure 2: Predictions for all algorithms on the NIPS coauthorship dataset. In (a), a white entry is shown in Figure 1(e). Though this matrix is seen in Figure 2(a). We again learned models for the latent feature relational model, the IRM and the MMSB, and to demonstrate the effectiveness of our full algorithm. For the Alyawarra dataset, we had no known covariates. For the countries dataset, the coauthorship relationship is symmetric, we learned a full, symmetric weight matrix allowing partial membership is not as restrictive. However, on this dataset, the IRM outperformed weights for each of the relations. This highlights the importance of proper initialization. To demonstrate that the covariates are helping, but that even without them, our model does well, we ran the

Acknowledgments

We showed empirically that the nonparametric latent feature model performs well at link prediction of what can be inferred by this model. As a consequence, our model is strictly more expressive than a traditional class-based approach. The success of this model can be traced to its richer representations, which make it able to capture subtle patterns of interaction much better than class-based models.

Many local optima in the feature space, further motivating the need for good initialization.

The importance of proper initialization also does well, beating the IRM initialization on both datasets. However, the random initialization also performs well, beating the IRM initialization on both datasets. However, the random feature matrix from the IRM and reported the AUC on the held-out data. Using five restarts for each method, the LFRM w/ IRM performed best with an AUC of 0.9509, the LFRM rand was next with an AUC of 0.9281, and the IRM performed last with an AUC of 0.9136.

In (c), we show predictions for the IRM. Notice that all members of classes interact similarly. For visualization, we have ordered the authors by their class assignments. In (d), we show predictions for the MMSB. Notice that all members of classes interact similarly. For visualization, we have ordered the authors by their class assignments.
Language modelling [Teh and Görür, 2009].

Figure 2: Power-law properties of the 20newsgroups dataset. The faint dashed lines are the distributions of words in the documents in each class, the solid curve is the mean of these lines. The dashed lines are the means of the word distributions generated by the ML parameters for the beta process (pink) and the SBP (green).

Table 1: Classification performance of SBP and beta process (BP). The $j$th column (denoted $1:j$) shows the cumulative rank $j$ classification accuracy of the test documents. The three numbers after the models are the percentages of training, validation and test sets respectively.

<table>
<thead>
<tr>
<th></th>
<th>BP</th>
<th>SBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>85.5(±0.6)</td>
<td>85.5(±0.4)</td>
</tr>
<tr>
<td>valid</td>
<td>91.6(±0.3)</td>
<td>91.9(±0.4)</td>
</tr>
<tr>
<td>test</td>
<td>94.2(±0.2)</td>
<td>94.4(±0.2)</td>
</tr>
</tbody>
</table>

Statistics of word occurrences; see Figure 2). We also plotted the characteristics of data simulated from the models using the estimated ML parameters. The SBP has a much better fit than the beta process to the power-law properties of the corpora.

In the second experiment we tested the two models on categorizing test documents into one of the 20 newsgroups. Since this is a discriminative task, we optimized the parameters in both models to maximize the cumulative ranked classification performance. The rank $j$ classification performance is defined to be the percentage of documents where the true label is among the top $j$ predicted classes (as determined by the IBP conditional probabilities of the documents under each of the 20 newsgroup classes). As the cost function is not differentiable, we did a grid search over the parameter space, using $20$ values of $\alpha, c$ and $\sigma$ each, and found the parameters maximizing the objective function on a validation set separate from the test set. To see the effect of sample size on model performance we tried splitting the documents in each newsgroup into 20% training, 20% validation and 60% test sets, and into 60% training, 20% validation and 20% test sets. We repeated the experiment five times with different random splits of the dataset. The ranked classification rates are shown in Table 1. Figure 3 shows that the SBP model has generally higher classification performances than the beta process.

5 Discussion

We have introduced a novel stochastic process called the stable-beta process. The stable-beta process is a generalization of the beta process, and can be used in nonparametric Bayesian featural models with an unbounded number of features. As opposed to the beta process, the stable-beta process has a number of appealing power-law properties. We developed both an Indian buffet process and a stick-breaking construction for the stable-beta process and applied it to modeling word occurrences in document corpora. We expect the stable-beta process to find uses modeling a range of natural phenomena with power-law properties.
Let $A = \sum w_k \delta_{sk}$ be a beta process with base measure $\mu$. If $\mu[0, \infty) = \alpha$, then $E[\sum w_k] = \alpha < \infty$. This means, if we sample from Bernoulli distributions with weight $w_k$ at each of the atoms of $A$, we will get a finite number of 1s.

$$A = \sum_{k=1}^{\infty} w_k \delta_{sk}, \quad (31)$$

$$Z_i = \sum_{k=1}^{\infty} z_{ik} \delta_{sk}, \quad (32)$$

$$z_{ik} \sim \text{Bernoulli}(w_k). \quad (33)$$
Beta process conditionals [Thibaux and Jordan, 2007]

Let $A = \sum w_k \delta_{sk}$ be a beta process with base measure $\mu$. If $\mu[0, \infty) = \alpha$, then $\mathbb{E}[\sum w_k] = \alpha < \infty$. This means, if we sample from Bernoulli distributions with weight $w_k$ at each of the atoms of $A$, we will get a finite number of 1s.

$$A = \sum_{k=1}^{\infty} w_k \delta_{sk}, \quad (31)$$

$$Z_i = \sum_{k=1}^{\infty} z_{ik} \delta_{sk}, \quad (32)$$

$$z_{ik} \sim \text{Bernoulli}(w_k). \quad (33)$$
Beta process conditionals
Beta process conditionals

![](image)

A

Z1IA
Beta process conditionals

A
Z1IA
Z2IA
Beta process conditionals

A
Z1IA
Z2IA
Z3IA
Beta process conditionals
Beta process conditionals
Beta process conditionals

AIZ

Z1

Z2

Z3

Z4
Beta process conditionals

\[ AIZ \]

\[ Z1 \]

\[ Z2 \]

\[ Z3 \]

\[ Z4 \]
Beta process conditionals
Beta process conditionals [Thibaux and Jordan, 2007]

\[ A = \sum_{k=1}^{\infty} w_k \delta_{s_k}, \]  
\[ Z_i = \sum_{k=1}^{\infty} z_{ik} \delta_{s_k}, \]  
\[ z_{ik} \sim \text{Bernoulli}(w_k), i = 1, \ldots, n. \]

Then,

\[ A|Z_1, \ldots, Z_n = \sum_{k=1}^{K} F_{nk} \delta_{s_k^*} + \sum_{k=1}^{\infty} w^n_k \delta_{s_k}. \]

where \( (s_k^*) = \{ s_k : \exists i \text{ s.t. } z_{ik} = 1 \} \) and

\[ F_{nk} \sim \text{Beta}(m_k, n - m_k + c), \]  
\[ (w^n_k, s_k) \text{ are drawn from a Poisson process with rate} \]
\[ cu^{-1}(1 - u)^{n+c-1} du \mu(ds). \]
Beta process conditionals [Thibaux and Jordan, 2007]

\[ A = \sum_{k=1}^{\infty} w_k \delta_{s_k}, \quad (34) \]

\[ Z_i = \sum_{k=1}^{\infty} z_{ik} \delta_{s_k}, \quad (35) \]

\[ z_{ik} \sim \text{Bernoulli}(w_k), \quad i = 1, \ldots, n. \quad (36) \]

Then,

\[ A|Z_1, \ldots, Z_n = \sum_{k=1}^{K} F_{nk} \delta_{s^*_k} + \sum_{k=1}^{\infty} w^n_k \delta_{s_k}. \quad (37) \]

where \((s^*_k) = \{s_k : \exists i \text{ s.t. } z_{ik} = 1\}\) and

\[ F_{nk} \sim \text{Beta}(m_k, n - m_k + c), \quad (38) \]

\[ (w^n_k, s_k) \text{ are drawn from a Poisson process with rate } \]

\[ cu^{-1}(1 - u)^{n+c-1} du \mu(ds). \]
Beta process conditionals [Thibaux and Jordan, 2007]

Furthermore, the conditional distribution of $Z_{n+1}$ with $A$ marginalized can be found as follows:

$$Z_{n+1} = \sum_{k=1}^{K} z_k^* \delta_{s_k^*} + \sum_{k=1}^{\infty} z_k^n \delta_{s_k},$$  \hspace{1cm} (40)$$

$$z_k^* \sim \text{Bernoulli} \left( \frac{m_k}{n+1} \right), z_k^n = \text{Bernoulli}(w^n_k).$$  \hspace{1cm} (41)$$

And as before $(w^n_k, s_k)$ are drawn from a Poisson process with rate $cu^{-1}(1 - u)^{n+c-1} du \mu(ds)$. So:

$$\sum_{k=1}^{\infty} z_k^n = \int_{0}^{\infty} \int_{0}^{1} cu^{-1}(1 - u)^{n+c-1} du \mu(ds),$$  \hspace{1cm} (42)$$

$$= \frac{c}{c + n} \mu[0, \infty).$$  \hspace{1cm} (43)$$

This is the link to the IBP.
Beta process conditionals [Thibaux and Jordan, 2007]

Furthermore, the conditional distribution of $Z_{n+1}$ with $A$ marginalized can be found as follows:

$$Z_{n+1} = \sum_{k=1}^{K} z^*_k \delta^*_s + \sum_{k=1}^{\infty} z^n_k \delta_s,$$

$$z^*_k \sim \text{Bernoulli}\left(\frac{m_k}{n+1}\right), \quad z^n_k = \text{Bernoulli}(w^n_k).$$

And as before $(w^n_k, s_k)$ are drawn from a Poisson process with rate $cu^{-1}(1 - u)^{n+c-1}du\mu(ds)$. So:

$$\sum_{k=1}^{\infty} z^n_k = \int_0^\infty \int_0^1 cu^{-1}(1 - u)^{n+c-1}du\mu(ds),$$

$$= \frac{c}{c + n}\mu[0, \infty).$$

This is the link to the IBP.
Outline

Conjugate priors for survival analysis

Link to completely random measures

Indian buffet process

Applications to machine learning
References


