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

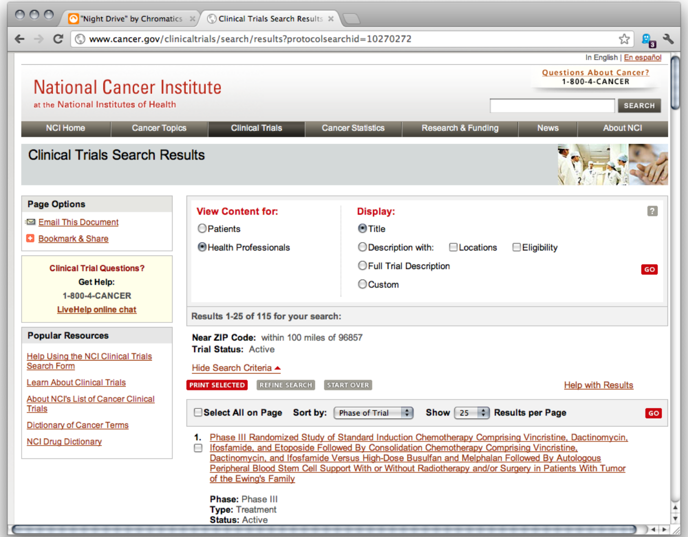
Link to completely random measures

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:

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

We are given right censored observations:

$$X_i \text{ lifetime,} \tag{2}$$
$$T_i \text{ time of last observation,} \tag{3}$$
$$d_i \text{ censoring indicator,} \tag{4}$$
$$c_i \text{ time of censoring,} \tag{5}$$
$$T_i = \min\{X_i, c_i\}, \tag{6}$$
$$d_i = \mathbb{I}\{X_i \leq c_i\}. \tag{7}$$

Survival analysis [Cox, 1972]

We want to estimate the hazard rate:

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

We are given right censored observations:

$$\begin{align} &X_i \text{ lifetime}, \tag{2}\\ &T_i \text{ time of last observation}, \tag{3}\\ &d_i \text{ censoring indicator}, \tag{4}\\ &c_i \text{ time of censoring}, \tag{5}\\ &T_i = \min\{X_i, c_i\}, \tag{6}\\ &d_i = \mathbb{I}\{X_i \leq c_i\}. \tag{7}\end{align}$$

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:

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

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

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:

$$
\begin{aligned}
p(h(t) = u | T_i, d_i) &\propto \Pr(T_i, d_i | h(t) = u) p(h(t) = u), && (11) \\
&= u^{\#i: T_i \in [t, t+\delta) \text{ and } d_i = 1} (1-u)^{\#i: T_i \geq t+\delta} \alpha_t(u) && (12) \\
&= u^{dN(t)} (1-u)^{Y(t) - dN(t)} \alpha_t(u). && (13)
\end{aligned}
$$

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

$$
\begin{aligned}
h(t) &\sim \mathrm{Beta}(c(t)\mu_\delta(t), c(t)(1 - \mu_\delta(t))), && (14) \\
h(t) | T_i, d_i &\sim \mathrm{Beta}(c(t)\mu_\delta(t) + dN(t), c(t)\mu_\delta(t) + Y(t) - dN(t)), && (15)
\end{aligned}
$$

$\mu_\delta(t) = \mu[t, t+\delta)$ 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:

$$\begin{aligned} p(h(t) = u | T_i, d_i) &\propto \Pr(T_i, d_i | h(t) = u) p(h(t) = u), && (11) \\ &= u^{\# i : T_i \in [t, t+\delta) \text{ and } d_i = 1} (1-u)^{\# i : T_i \geq t+\delta} \alpha_t(u) && (12) \\ &= u^{dN(t)} (1-u)^{Y(t) - dN(t)} \alpha_t(u). && (13) \end{aligned}$$

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

$$\begin{aligned} h(t) &\sim \text{Beta}(c(t)\mu_\delta(t), c(t)(1-\mu_\delta(t))), && (14) \\ h(t) | T_i, d_i &\sim \text{Beta}(c(t)\mu_\delta(t) + dN(t), c(t)\mu_\delta(t) + Y(t) - dN(t)), && (15) \end{aligned}$$

$\mu_\delta(t) = \mu[t, t+\delta)$ 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:

$$p(h(t)=u|T_i,d_i) \propto \Pr(T_i,d_i|h(t)=u)p(h(t)=u), \tag{11}$$
$$= u^{\#i:T_i\in[t,t+\delta)\text{ and } d_i=1}(1-u)^{\#i:T_i\geq t+\delta}\alpha_t(u) \tag{12}$$
$$= u^{dN(t)}(1-u)^{Y(t)-dN(t)}\alpha_t(u). \tag{13}$$

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

$$h(t) \sim \mathrm{Beta}(c(t)\mu_\delta(t), c(t)(1-\mu_\delta(t))), \tag{14}$$
$$h(t)|T_i,d_i \sim \mathrm{Beta}(c(t)\mu_\delta(t)+dN(t), c(t)\mu_\delta(t)+Y(t)-dN(t)), \tag{15}$$

$\mu_\delta(t)=\mu[t,t+\delta)$ 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:

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

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

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

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

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

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

$\mu_\delta(t)=\mu[t,t+\delta)$ 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:

$$F(t) \asymp 1 - \prod_{k=0}^{\lfloor t/\delta \rfloor} (1 - h(k\delta)). \tag{16}$$

$$\asymp 1 - \exp(-\underbrace{\sum_{k=0}^{\lfloor t/\delta \rfloor} h(k\delta)}_{\text{limit is } A(t)}) \tag{17}$$

Theorem

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

$$A(t) = \lim_{\delta \to 0^+} \sum_{k=0}^{\lfloor t/\delta \rfloor} h(k\delta). \tag{18}$$

Continuous hazard rates

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

$$F(t) \asymp 1 - \prod_{k=0}^{\lfloor t/\delta \rfloor} (1 - h(k\delta)). \tag{16}$$

$$\asymp 1 - \exp(-\underbrace{\sum_{k=0}^{\lfloor t/\delta \rfloor} h(k\delta))}_{\text{limit is } A(t)} \tag{17}$$

Theorem

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

$$A(t) = \lim_{\delta \to 0^+} \sum_{k=0}^{\lfloor t/\delta \rfloor} h(k\delta). \tag{18}$$

Properties of the cumulative hazard

Corollary

1. $A(0) = 0$,
2. $A(t_i) - A(t_{i-1})$ *are independent for all* $0 \leq t_1 < t_2 < \dots$,
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 $B_1, \dots, B_n$ are disjoint then $A(B_1), \dots, A(B_n)$ are independent.

Properties of the cumulative hazard

Corollary

1. $A(0) = 0$,
2. $A(t_i) - A(t_{i-1})$ *are independent for all* $0 \leq t_1 < t_2 < \dots$,
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 $B_1, \dots, B_n$ are disjoint then $A(B_1), \dots, A(B_n)$ are independent.

Representation as a CRM

By the Lévy–Khinchine representation theorem (from lecture 2), there exists a measure $\lambda(du, ds)$ such that for all functions $f(s)$ on $\mathbb{R}_{\geq 0}$:

$$\mathbb{E}\left[\exp\left(-\int_0^\infty f(s)A(ds)\right)\right] = \exp\left(-\int_0^\infty \int_0^1 1 - e^{-uf(s)}\lambda(du, ds)\right), \tag{19}$$

$$\lambda(du, ds) = c(s)u^{-1}(1-u)^{c(s)-1}\mu(ds). \tag{20}$$

Write $A \sim \mathrm{BP}(c, \mu)$ in this case.

Representation as a CRM

By the Lévy–Khinchine representation theorem (from lecture 2), there exists a measure $\lambda(du, ds)$ such that for all functions $f(s)$ on $\mathbb{R}_{\geq 0}$:

$$\mathbb{E}\left[\exp\left(-\int_0^\infty f(s)A(ds)\right)\right] = \exp\left(-\int_0^\infty \int_0^1 1 - e^{-uf(s)}\lambda(du, ds)\right), \tag{19}$$

$$\lambda(du, ds) = c(s)u^{-1}(1-u)^{c(s)-1}\mu(ds). \tag{20}$$

Write $A \sim \mathrm{BP}(c, \mu)$ in this case.

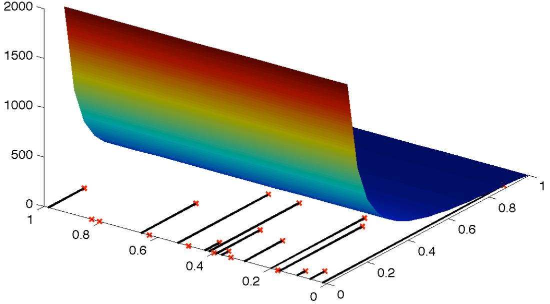
Link to completely random measures

Corollary

A beta process $A \sim \mathrm{BP}(c, \mu)$ *is a completely random measure* s.t.*:*

$$A = \sum_{k=1}^{\infty} w_k \delta_{s_k}, \tag{21}$$

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.

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

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,
- $w_1, \ldots, w_k \sim \mathcal{N}(1, 1)$.

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

Latent feature models

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

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

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,
- $w_1, \ldots, w_K \sim \mathcal{N}(1, 1)$.

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

Latent feature models

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

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

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,
- $w_1, \ldots, w_K \sim \mathcal{N}(1, 1)$.

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

Latent feature models

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

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

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,
- $w_1, \ldots, w_k \sim \mathcal{N}(1, 1)$.

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

Prior for features

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

$$p(\pi_k | z_1, \ldots, z_n) \propto p(z_1, \ldots, z_n | \pi_k) p(\pi_k), \tag{24}$$

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

This is the same situation as for the hazard function, suggesting a beta prior for π_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:

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

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

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

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

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:

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

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

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

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

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:

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

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

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

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

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(α) dishes.
- Customer $i > 1$ picks a previously chosen dish with probability m_k/i and Poisson(α/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(α) dishes.
- Customer $i > 1$ picks a previously chosen dish with probability m_k/i and Poisson(α/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(α) dishes.
- Customer $i > 1$ picks a previously chosen dish with probability m_k/i and Poisson(α/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:

$$\alpha^K \exp\left(-\alpha \sum_{i=1}^{n} 1/i\right) \prod_{i=1}^{K} \frac{(m_k - 1)!(n - m_k)!}{n!} h(\theta_k^*). \tag{29}$$

Applications to machine learning:

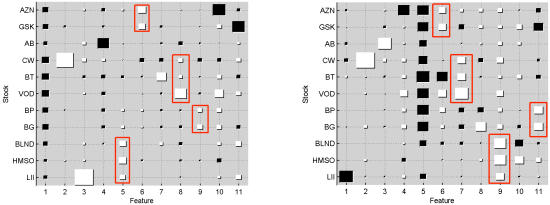
- Elimination by aspects choice model [Görür et al., 2006],
- Infinite ICA [Knowles and Ghahramani, 2007, Doshi et al., 2009].
- Latent feature relational model [Miller et al., 2009].
- Word frequency models [Teh and Görür, 2009].

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.

$$Y_i = G(Z_i \odot X_i) + E, \tag{30}$$

- $Z \sim \mathrm{IBP}(c, \mu)$,



(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.

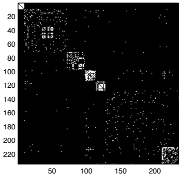
Figure 16: Application to financial data set.

Applications: latent feature relational model [Miller et al., 2009]

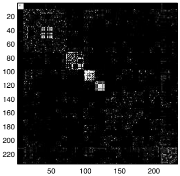
Prior for directed graphs. Each vertex has a latent binary feature vector z_i. Probability of an edge between vertices is an inner product of the feature vectors passed through a sigmoid.

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

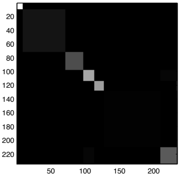
	Countries single	Countries global	Alyawarra single	Alyawarra global
LFRM w/ IRM	0.8521 ± 0.0035	$\mathbf{0.8772} \pm 0.0075$	0.9346 ± 0.0013	$\mathbf{0.9183} \pm 0.0108$
LFRM rand	$\mathbf{0.8529} \pm 0.0037$	0.7067 ± 0.0534	$\mathbf{0.9443} \pm 0.0018$	0.7127 ± 0.030
IRM	0.8423 ± 0.0034	0.8500 ± 0.0033	0.9310 ± 0.0023	0.8943 ± 0.0300
MMSB	0.8212 ± 0.0032	0.8643 ± 0.0077	0.9005 ± 0.0022	0.9143 ± 0.0097



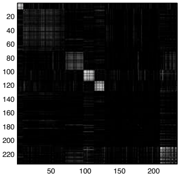
(a) True relations



(b) Feature predictions

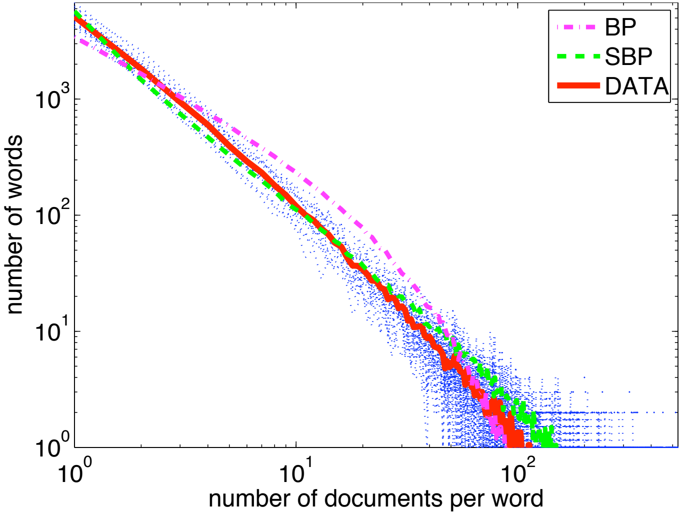


(c) IRM predictions



(d) MMSB predictions

Language modelling [Teh and Görür, 2009].



Beta process conditionals[Thibaux and Jordan, 2007]

Let $A = \sum w_k \delta_{sk}$ be a beta process with base measure μ. 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.

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

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

$$z_{ik} \sim \text{Bernoulli}(w_k). \tag{33}$$

Beta process conditionals[Thibaux and Jordan, 2007]

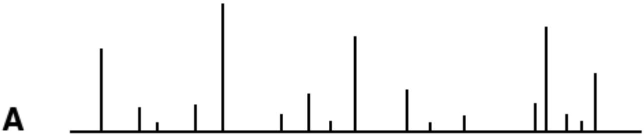
Let $A = \sum w_k \delta_{sk}$ be a beta process with base measure μ. If $\mu[0, \infty) = \alpha$, then $\mathbb{E}[\sum w_k] = \alpha < \infty$. This means, if we sample from Bernoulli distributions with weight w_k at each of the atoms of A, we will get a finite number of 1s.

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

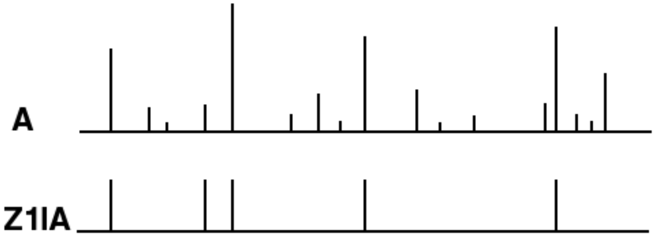
$$Z_i = \sum_{k=1}^{\infty} z_{ik} \delta_{sk}, \tag{32}$$

$$z_{ik} \sim \text{Bernoulli}(w_k). \tag{33}$$

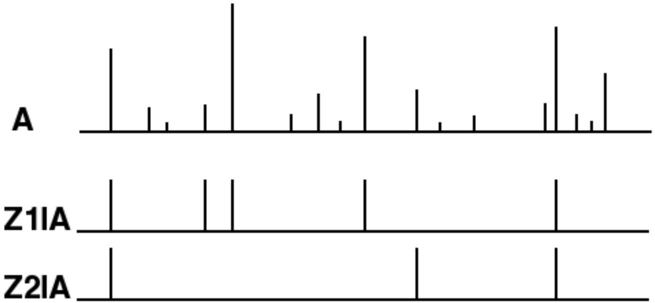
Beta process conditionals



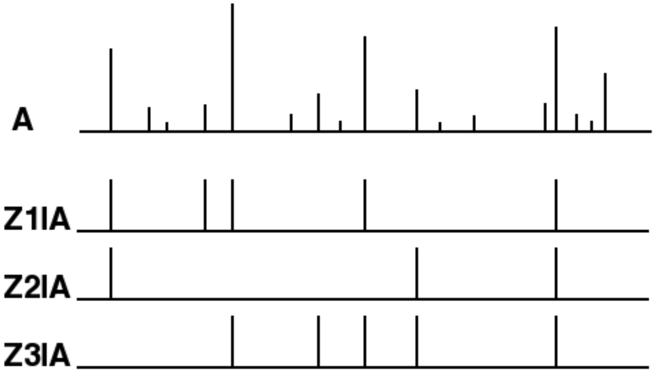
Beta process conditionals



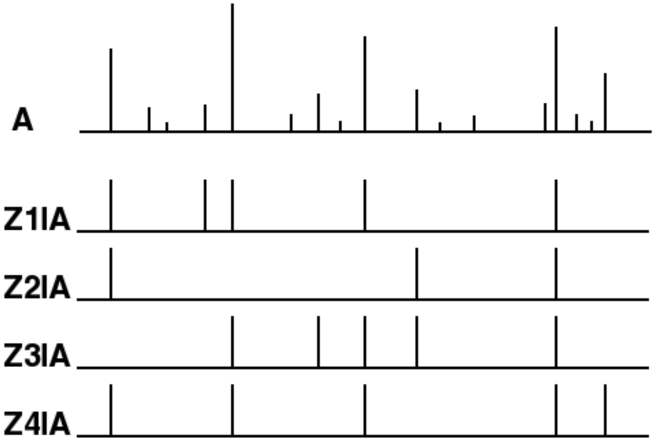
Beta process conditionals



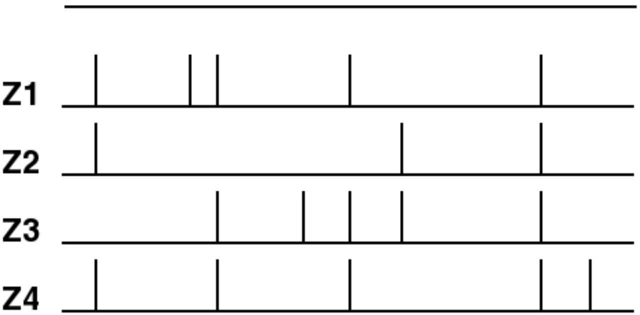
Beta process conditionals



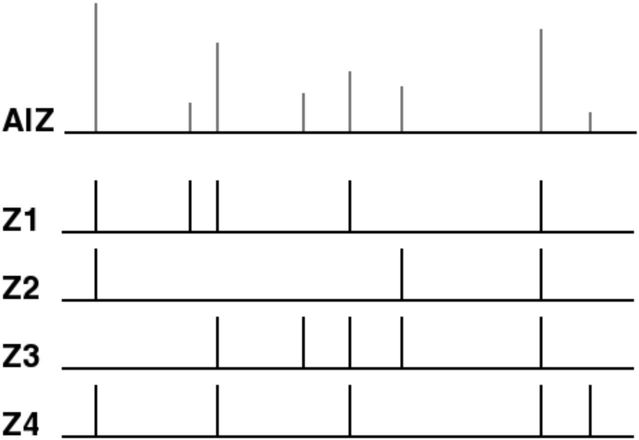
Beta process conditionals



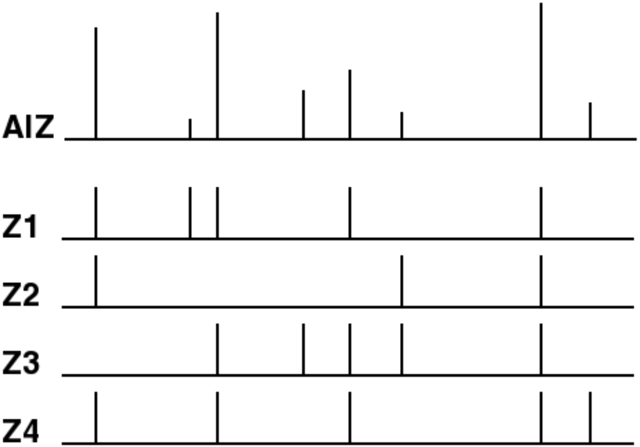
Beta process conditionals



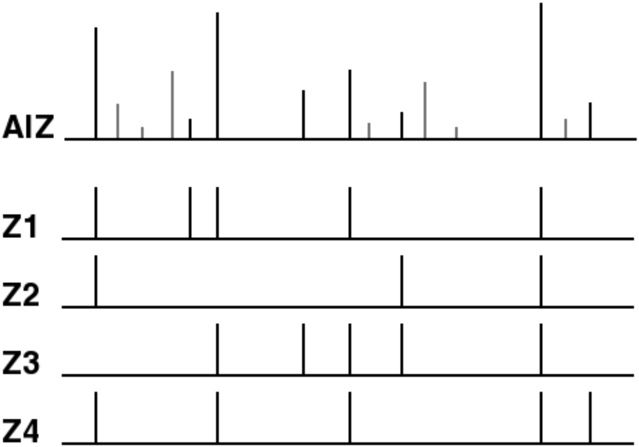
Beta process conditionals



Beta process conditionals



Beta process conditionals



Beta process conditionals[Thibaux and Jordan, 2007]

$$A = \sum_{k=1}^{\infty} w_k \delta_{sk}, \tag{34}$$

$$Z_i = \sum_{k=1}^{\infty} z_{ik} \delta_{sk}, \tag{35}$$

$$z_{ik} \sim \text{Bernoulli}(w_k), i = 1, \ldots, n. \tag{36}$$

Then,

$$A|Z_1, \ldots, Z_n = \sum_{k=1}^{K} F_{nk} \delta_{s_k^*} + \sum_{k=1}^{\infty} w_k^n \delta_{sk}. \tag{37}$$

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

$$F_{nk} \sim \text{Beta}(m_k, n - m_k + c), \tag{38}$$

$$\tag{39}$$

And (w_k^n, s_k) 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]

$$A = \sum_{k=1}^{\infty} w_k \delta_{sk}, \tag{34}$$

$$Z_i = \sum_{k=1}^{\infty} z_{ik} \delta_{sk}, \tag{35}$$

$$z_{ik} \sim \text{Bernoulli}(w_k), i = 1, \ldots, n. \tag{36}$$

Then,

$$A|Z_1, \ldots, Z_n = \sum_{k=1}^{K} F_{nk} \delta_{s_k^*} + \sum_{k=1}^{\infty} w_k^n \delta_{sk}. \tag{37}$$

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

$$F_{nk} \sim \text{Beta}(m_k, n - m_k + c), \tag{38}$$

$$\tag{39}$$

And (w_k^n, s_k) 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:

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

$$z_k^* \sim \text{Bernoulli}\left(\frac{m_k}{n+1}\right), z_k^n = \text{Bernoulli}(w_k^n). \tag{41}$$

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

$$\sum_{k=1}^{\infty} z_k^n = \int_0^\infty \int_0^1 cu^{-1}(1-u)^{n+c-1} du\mu(ds), \tag{42}$$

$$= \frac{c}{c+n}\mu[0,\infty). \tag{43}$$

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:

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

$$z_k^* \sim \text{Bernoulli}\left(\frac{m_k}{n+1}\right), z_k^n = \text{Bernoulli}(w_k^n). \tag{41}$$

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

$$\sum_{k=1}^{\infty} z_k^n = \int_0^{\infty} \int_0^1 cu^{-1}(1-u)^{n+c-1} du \mu(ds), \tag{42}$$

$$= \frac{c}{c+n} \mu[0, \infty). \tag{43}$$

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 I

- Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 187(34).
- Doshi, F., Miller, K. T., Van Gael, J., and Teh, Y. W. (2009). Variational inference for the Indian buffet process. In *JMLR Workshop and Conference Proceedings: AISTATS 2009*, volume 5, pages 137–144.
- Görür, D., Jäkel, F., and Rasmussen, C. E. (2006). A choice model with infinitely many latent features. In *Proceedings of the International Conference on Machine Learning*, volume 23.
- Griffiths, T. L. and Ghahramani, Z. (2005). Infinite latent feature models and the indian Buffet process. Technical Report 001, Gatsby Computational Neuroscience Unit, UCL.
- Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, 18(3):1259–1294.
- Knowles, D. and Ghahramani, Z. (2007). Infinite sparse factor analysis and infinite independent components analysis. In *International Conference on Independent Component Analysis and Signal Separation*, volume 7 of *Lecture Notes in Computer Science*. Springer.
- Miller, K., Griffiths, T., and Jordan, M. (2009). Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems*, volume 22.
- Teh, Y. W. and Görür, D. (2009). Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems*, volume 22, pages 1838–1846.
- Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 11, pages 564–571.