Supplementary material

1 Introduction

This paper contains supplementary material to the main paper "*A birth-death process for feature allocation*".

2 Proofs

This section provides the proofs to theorems and equations provided in the paper. Before providing the actual proofs, we present some useful propositions that will help the reader understand the material that follows.

Poisson approximation to the Binomial distribution

Proposition 1. *(Poisson approximation) Suppose X_n is a random variable following the Binomial distribution with number of trials as n and success ratio as p_n, denoted by* Binomial(n, p_n)*, such that $n \to \infty$, $p_n \to 0$ and $np_n \to \lambda > 0$. Then, for $k = 0, 1, 2, \ldots,$*

$$X_n \to \text{Poisson}(\lambda), \qquad \textit{in distribution.}$$

In other words,

$$P(X_n = k) \to \frac{\lambda^k e^{-\lambda}}{k!}, \qquad \textit{as } n \to \infty.$$

Normal approximation to the Poisson distribution

Proposition 2. *(Normal approximation) Suppose X is a random variable following the Poisson distribution with mean λ such that $X \sim$ Poisson(λ). Then, for $\lambda \to \infty$,*

$$X \to \mathcal{N}\Big(\lambda, \sqrt{\lambda}\Big), \qquad \textit{in distribution.}$$

,where $\sqrt{\lambda}$ is the standard deviation.

Dirac approximation to Normal distribution

Proposition 3. *(Dirac approximation) Suppose X is a random variable following the Normal distribution with mean 1 and standard deviation $\sigma = \frac{1}{\lambda}$ such that $X \sim \mathcal{N}(1, \sigma)$. Then, for $\sigma \to 0$,*

$$X \to \delta(1), \qquad \textit{in distribution.}$$

In other words,

$$X \to 1, \qquad \textit{as } \sigma \to 0.$$

Proof of proposition (1) in the main paper:
Assume that $z(t)$ is a realization of the BDFP $(Z(t))$ over the finite interval $[0, T], T > 0$ and we write $(z(t))_{0 \le t \le T}$. With probability one the sample path $(z(t))_{0 \le t \le T}$ will only contain a finite number of jump events, each of which is either a birth or a death event. We write B and Q to denote the set of the features created or turned off by birth or death events respectively. We denote as $t_1, \ldots, t_J$ the times when the chain jumps, where $J = |B| + |Q|$.

The probability of observing a sample $(z(t))_{0 \le t \le T}$ can be written as the product of three factors; the probability of the initial state, the probability of each jump (event) and the probability of the interarrival times between the events. More specifically,

- Probability of the initial state at time $t = 0$:
  At time $t = 0$, the feature allocation $z(0)$ follows an IBP distribution.

$$\mathrm{P}(z(0)) = \frac{\alpha^{K_{z(0)+}}}{\prod_{h=1}^{\mathcal{H}_{z(0)}} K_h!} \exp\left(-\alpha H_N\right) \prod_{k=1}^{K_{z(0)+}} \frac{(N - m_k)!(m_k - 1)!}{N!} \tag{1}$$

  where $K_{z(0)+}$ and $\mathcal{H}_{z(0)}$ are the number of total and distinct nonzero features correspondingly in state $z(0)$ and m_k is the number of objects that possess feature k.

- Probability of all the interarrival times between events:
  The probability that the chain will not jump in the interval $[t_l, t_{l+1}]$ can be derived as follows, by dividing the time from t_l to t_{l+1} into k intervals of duration $\frac{t_{l+1} - t_l}{k}$, and letting $k \to \infty$:

$$\begin{aligned}
\mathrm{P}\begin{pmatrix}\text{no jump} \\ \text{in } [t_l, t_{l+1})\end{pmatrix} &= \lim_{k \to \infty} \prod_{i=0}^{k-1} \left(1 - q\left(t_l + \frac{i(t_{l+1} - t_l)}{k}\right)\frac{(t_{l+1} - t_l)}{k}\right) \\
&= \exp\left(\lim_{k \to \infty} \sum_{i=0}^{k-1} \log\left(1 - q\left(t_l + \frac{i(t_{l+1} - t_l)}{k}\right)\frac{(t_{l+1} - t_l)}{k}\right)\right) \\
&= \exp\left(-\int_{t_l}^{t_{l+1}} q(t)\mathrm{d}t\right) = \exp(-q_{z(t_l)}(t_{l+1} - t_l))
\end{aligned}$$

where we wrote $q(t) = q_{z(t_l)}$ and we also considered that the rate $q_{z(t_l)}$ is constant in the interval $t_{l+1} - t_l$. The interarrival times in $[0, T]$ are independent of each other, so the probability of all the interarrival times is a product as follows:

$$\begin{aligned}\prod_{l=0}^{|B|+|D|} \mathrm{P}\begin{pmatrix}\text{no jump}\\ \text{in } [t_l, t_{l+1})\end{pmatrix} &= \prod_{l=0}^{|B|+|Q|} \exp\Big(-\int_{t_l}^{t_{l+1}} q(\tau)\mathrm{d}\tau\Big)\\ &= \exp\Big(-\int_0^T q(t)\mathrm{d}t\Big)\\ &= \exp\Big(-\sum_{l=0}^{|B|+|Q|} (t_{l+1}-t_l) q_{z(t_l)}\Big) \qquad (2)\end{aligned}$$

where $t_0 = 0$ and $t_{|B|+|Q|+1} = T$. Note that the above product includes the probability of not transiting from the last jump time to T.

- Probability of all the events taking place during the interval $[0, T]$:
  The probability density that the chain jumps at time t_l to a new state is $q_{z(t_l-)z(t_l)}$, where t_l- denotes an infinitesimal time prior to t_l. The probability density of all the events taking place is:

$$\prod_{l=1}^{|B|+|Q|} q_{z(t_l-)z(t_l)} \qquad (3)$$

To compute the probability of the path we take the product of Equations (1),(2) and (3).

Proof of corollary (1) in the main paper:
The IBP is defined as the limit of the corresponding distribution over matrices with M columns as $M \to \infty$. The finite model, which gives the IBP in the limit as $M \to \infty$ (Griffiths & Ghahramani, 2011) is

$$\begin{aligned}\omega_k|\alpha &\sim \mathrm{Beta}\Big(\frac{\alpha}{M}\Big),\\ Z_{nk}|\omega_k &\sim \mathrm{Bernoulli}(\omega_k) \qquad (4)\end{aligned}$$

for $k = 1, \ldots, M$ and $n = 1, \ldots, N$.

The expected number of features at any time $t \in \mathbb{T}$ is

$$\mathbb{E}[N_f] = \int_{[0,1]\otimes\mathbb{X}} \nu_t(\mathrm{d}\omega\mathrm{d}x) = \int_{[0,1]} \rho(\mathrm{d}\omega) \int_{\mathbb{X}} \frac{\mu_0(\mathrm{d}x)}{D} = \frac{K}{D},\ K \to \infty \qquad (5)$$

We introduced $K \to \infty$ because $\int_{[0,1]} \rho(\mathrm{d}\omega) = \infty$. Since the expected number of features at any $t \in \mathbb{T}$ is given by $\frac{K}{D}$, the substitution of $M = \frac{K}{D}$ for $K \to \infty$ and $D = \frac{R}{\alpha}$, results in the generative model in Equation (12) in the main paper.

Proof of proposition (4) in the main paper:
For what follows, we use $N_B(\delta)$ to denote the number of birth events (actual appearance of a feature with at least one member in it) that is $N_B(\delta) = |\{\mathcal{B} = \{f_k\} : t_b^k \in \delta, \sum_{n=1}^{N} z_{nk}(t) \geq 1\}|$ and $N_F(\delta)$ to denote the number of feature events t_b at time interval δ, that is $N_F(\delta) = |\{\mathcal{F} = \{f_k\} : t_b^k \in \delta\}|$. Note here the difference between a birth and a feature event. Not all feature events are birth events.

- The number of features present in a time interval $[t, t+\delta]$ follows a Poisson distribution with intensity $\nu(\delta) = \int_{[0,1]} \rho(\mathrm{d}\omega) \int_{\mathbb{X}} \mu_0(\mathrm{d}x) \int_t^{t+\delta} \mathrm{d}t_b \int_0^{\infty} \mathrm{d}t_\omega = K\delta$ for $K \to \infty$, i.e. $N_f(\delta) \sim \text{Poisson}(K\delta)$. Since $K \to \infty$, use of Proposition (2) and Proposition (3) result in

$$N_f(\delta) \to K\delta, \text{ as } K \to \infty. \tag{6}$$

- The probability of activating a feature f_k given that it is created in the time interval δ is:

$$\begin{aligned} \pi = & \quad P(N_B(\delta) = 1 | N_F(\delta) = 1) \\ = & \int P(N_B(\delta) = 1, \omega_k | N_F(\delta) = 1) p(\omega_k) \mathrm{d}\omega_k \end{aligned} \tag{7}$$

where, based on Equation (12) in the main paper, $\omega_k \sim \text{Beta}\left(\frac{R}{K}, 1\right)$. Moreover, $P(N_B(\delta) = 1, \omega_k | N_F(\delta) = 1)$ is the probability that at time interval δ at least one object out of N grants membership of the feature given that only one feature is created at the interval δ, i.e

$$P(N_B(\delta) = 1, \omega_k | N_F(\delta) = 1) = P(n \geq 1) = 1 - P(n < 1) = 1 - (1 - \omega_k)^N$$

where we used the fact that $P(n \leq m)$ is given by the Binomial cumulative distribution function form $\mathcal{F}(m; N, \omega_k) = P(n \leq m) = \sum_{i=0}^{\lfloor m \rfloor} \binom{N}{i} \omega_k^i (1 - \omega_k)^{N-i}$. Here $\lfloor m \rfloor$ is the "floor" under m, i.e. the greatest integer less than or equal to m. Finally, Equation (7) becomes

$$\begin{aligned} \pi &= \int (1 - (1 - \omega_k)^N) \text{Beta}\left(\omega_k; \frac{R}{K}, 1\right) \mathrm{d}\omega_k \\ &= 1 - \int \frac{(1 - \omega_k^N)^N \omega_k^{\frac{R}{K} - 1}}{\mathrm{B}(\frac{R}{K}, 1)} d\omega_k \\ &= 1 - \frac{\mathrm{B}(\frac{R}{K}, N + 1)}{\mathrm{B}(\frac{R}{K}, 1)} \end{aligned} \tag{8}$$

where $\mathrm{B}(\cdot)$ is the Beta function.

In order to compute the birth rate at $[t, t+\delta]$ we need to compute the probability of having one only birth event at the time interval δ, i.e.

$$P(N_B(\delta) = 1) = \int P(N_B(\delta) = 1 | N_F(\delta)) P(N_F(\delta)) \mathrm{d}N_F. \tag{9}$$

The number of actual births given the number of features present follows a Binomial distribution, i.e. $N_B(\delta)|N_F(\delta) \sim \text{Binomial}(N_B(\delta); N_F(\delta), \pi)$. Since $N_F(\delta) \to K\delta \to \infty$ and $\pi \to 0$ as $K \to \infty$, following Proposition (1), this distribution can be approximated by

$$N_B(\delta)|N_F(\delta) \sim \text{Poisson}(N_B(\delta); \pi K\delta),\ K \to \infty \tag{10}$$

Based on Equation (10), Equation (9) now becomes

$$\begin{aligned} P(N_B(\delta) = 1) = & \int P(N_B(\delta) = 1|N_F(\delta))P(N_F(\delta))\mathrm{d}N_F \\ = & \ \text{Poisson}(1; \pi K\delta) \\ = & \ \pi K\delta e^{-\pi K\delta} \end{aligned} \tag{11}$$

We need to consider the case when $K \to \infty$ and we will focus on the product term $\pi K\delta$.

$$\begin{aligned} \lim_{K\to\infty} \pi K\delta \quad = & \quad \lim_{K\to\infty} \delta K\Big(1 - \frac{\mathrm{B}(\frac{R}{K}, N+1)}{\mathrm{B}(\frac{R}{K}, 1)}\Big) \\ \overset{\beta=\frac{1}{K}}{=} & \quad \lim_{\beta\to 0} \delta\Big(\frac{\mathrm{B}(\beta R, 1) - \mathrm{B}(\beta R, N+1)}{\beta \mathrm{B}(\beta R, 1)}\Big) \\ \overset{\text{L'Hopital}}{=} & \quad \lim_{\beta\to 0} \frac{R\mathrm{B}(\beta R, 1)(\psi(\beta R) - \psi(\beta R+1)) - R\mathrm{B}(\beta R, N+1)(\psi(\beta R) - \psi(\beta R+N+1))}{\beta R\mathrm{B}(\beta R, 1)(\psi(\beta R) - \psi(\beta R+1)) + \mathrm{B}(\beta R, 1)} \\ \overset{\beta\to 0}{=} & \quad \delta\frac{R(\psi(0) - \psi(1)) - R(\psi(0) - \psi(N+1))}{1} \\ = & \quad \delta R(\psi(N+1) - \psi(1)) = \delta R\sum_{n=1}^{N}\frac{1}{n} = \delta R H_N \end{aligned} \tag{12}$$

where we used the definition of the derivative of the Beta function $\frac{d\mathrm{B}(x,y)}{dx} = \mathrm{B}(x,y)(\psi(x) - \psi(x+y))$. Considering again the probability in Equation (11) and taking its limit as $K \to \infty$ and $\delta \to 0$ we have

$$P(N_B(\delta) = 1) = \delta R H_N e^{-\delta R H_N} \overset{\delta\to 0}{=} \delta R H_N \tag{13}$$

If we divide the result by δ, then the resulting rate is equal to the overall birth rate RH_N in BDFP. The probability of only one feature out of the N_F dying in the time interval δ is trivial $N_F(\delta)\frac{R}{\alpha}$ equal to the death rate in BDF process. We assumed that the probability of observing more than one events in $\delta \to 0$ is negligible.

Proof of proposition (5) in the main paper:
In the finite model, at any $t \in \mathbb{T}$, the constrained projection is computed as an integral over the space defined by the updated set of constraints $t - t_\omega < t_b < t$, $0 < t_\omega < \infty$ and $0 < t_b < T$ which can be summarised as $t - t_\omega < t_b < t$ and

$0 < t_\omega < t$. As such,

$$\begin{aligned}\nu_t(\mathrm{d}\omega\mathrm{d}x) &= \rho(\mathrm{d}\omega)\mu_0(\mathrm{d}x)\int_0^t\int_{t-t_\omega}^t g(\mathrm{d}t_b)\beta(\mathrm{d}t_\omega)\\ &= \rho(\mathrm{d}\omega)\mu_0(\mathrm{d}x)\Big(-te^{-Dt}+\frac{1-e^{-Dt}}{D}\Big)\end{aligned} \tag{14}$$

The exponential terms add a dependency on the index $t \in \mathbb{T}$. This is a result of constraining $t_b \in [0, T]$; the number of features present near the origin $t = 0$ is smaller than the number of features present later in time, since no features are allowed to be born in $t_b < 0$. This effect diminishes as $t \gg \frac{1}{D}$ since then $e^{-Dt} \to 0$ and $\nu_t(\mathrm{d}\omega\mathrm{d}x)$ coincides with the restricted projection in the infinite case, i.e. $\nu_t(\mathrm{d}\omega\mathrm{d}x) = \rho(\mathrm{d}\omega)\frac{\mu_0(\mathrm{d}x)}{D}$.

In the same fashion, in the finite model, the expected number of features present at any $t \in \mathbb{T}$ is $\int_0^1\int_{\mathbb{X}}\nu_t(\mathrm{d}\omega\mathrm{d}x) = K(-te^{-Dt}+\frac{1-e^{-Dt}}{D})$. Again, for $t \gg \frac{1}{D}$, the expected number of features at any t approaches $\frac{K}{D}$ and $K \to \infty$ in the infinite case. This is confirmed in Figure 1(a). It takes some time until the empirical mean number of features converges to the true expected value under the infinite model. In order to make sure that at any random time point in the range considered the process has IBP marginals, we have to make sure that the expected number of features at any time is equal to the true expected value, that is $\frac{K}{D} = \frac{K\alpha}{R}$. The discussion above ensures this considering the process at $t \gg \frac{1}{D}$, where the D is the death rate. In this case, we only allow data to live at the time range where IBP marginals are satisfied.

3 Posterior Simulation

The BDF process is a continuous-time Markov process and use of the forward-backward algorithm would facilitate inference. However, the exponentially large space of feature allocations makes inference hard. However, the equivalent BEP simplifies inference by allowing the application of a simple, efficient Gibbs sampling approach. For what follows, we present the posterior distributions for the linear-Gaussian likelihood model in Equation 15 and presented in Figure 3.

Likelihood term $p(\mathbf{Y}|\mathbf{S}, \mathbf{t}_b, \mathbf{t}_w, \mathbf{A}, \sigma_\epsilon)$ Given all the unknown parameters, the computation of the likelihood is straightforward. More precisely, given the values of $\mathbf{S}, \mathbf{t}_b$ and $\mathbf{t}_w$ the computation of the feature allocation matrices $\mathbf{Z}_t$ for $t = 1, \ldots, L$ is deterministic, as presented earlier, and as such, observing the data matrix $\mathbf{Y}_t$ is equal in distribution to observing matrix $\mathbf{Y}_t - \mathbf{Z}_t\mathbf{A}$ for which $y_{tnd} - \sum_{k=1}^{KT} z_{tnk}A_{kd} = \epsilon_{tnd} \sim \mathcal{N}(0, \sigma_\epsilon)$ holds. Consequently,

$$\begin{aligned}p(\mathbf{Y}|\mathbf{S}, \mathbf{t}_b, \mathbf{t}_w, \mathbf{A}, \sigma_\epsilon) &= \prod_{t=1}^{L} p(\mathbf{Y}_t|\mathbf{S}, \mathbf{t}_b, \mathbf{t}_w, \mathbf{A}, \sigma_\epsilon)\\ &= \prod_{t=1}^{L}\prod_{n}^{N}\prod_{d=1}^{D}\mathcal{N}\Big(y_{tnd} - \sum_{k=1}^{KT} z_{tnk}A_{kd}; 0, \sigma_\epsilon\Big)\end{aligned} \tag{15}$$

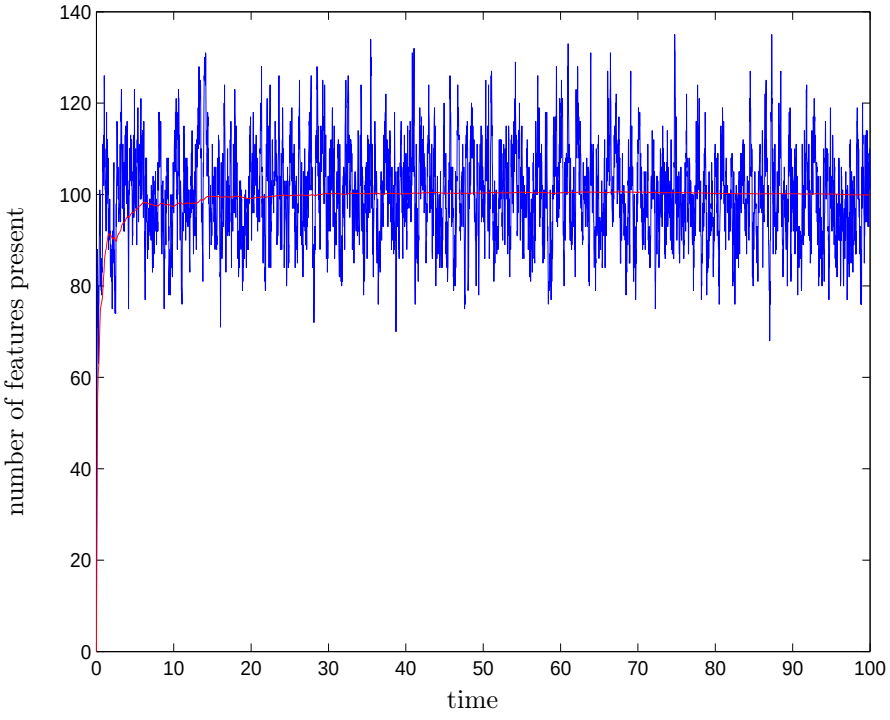


Figure 1: Traceplot and running mean plot of number of features present as a function of time for a sample from the finite BEP model. The plots were produced using a dataset of $N = 3$ and 5000 feature allocation matrices were drawn from the process over a period of $[0, 100]$. The number of features present is plotted over time. The values of hyperparameters chosen is $\alpha = 2, R = 20$ and $K = 1000$. Note the convergence to the mean $\mu = \frac{K\alpha}{R} = 100$ as indicated by the red traceplot; it takes some time until the number of features present converge to the expected mean.

Sample the parameters $\alpha, R, \mathbf{t}_b$ and $\mathbf{t}_w$. Due to lack of conjugacy, the posterior of these parameters has no closed form and exact computation is not feasible. For that reason, we used slice sampling (Neal, 2003). For completeness, we provide the conditional posteriors.

$$
\begin{aligned}
p(\alpha|t_w^k, R) \propto p(\mathbf{t}_w|\alpha, R)p(\alpha) &= \text{Gamma}(\alpha; \kappa_\alpha, \theta_\alpha) \prod_{k=1}^{KT} \text{Exponential}\left(t_w^k; \frac{R}{\alpha}\right) \\
p(R|\boldsymbol{\omega}, \mathbf{t}_w, \alpha) &\propto p(\boldsymbol{\omega}|R, \alpha)p(\mathbf{t}_w|R, \alpha)p(R) \\
&= \text{Gamma}(R; \kappa_R, \theta_R) \prod_{k}^{KT} \text{Beta}\left(\omega_k; \frac{R}{K}, 1\right) \text{Exponential}\left(t_w^k; \frac{R}{\alpha}\right) \\
p(t_b^k|\mathbf{Y}, \mathbf{S}, \mathbf{t}_b^{\neg k}, \mathbf{t}_w, \mathbf{A}, \sigma_\epsilon) &\propto p(\mathbf{Y}|\mathbf{S}, \mathbf{t}_b, \mathbf{t}_w, \mathbf{A}, \sigma_\epsilon)p(t_b^k) \\
&= \mathcal{U}(t_b^k; 0, T) \prod_{t=1}^{L} \prod_{n}^{N} \prod_{d=1}^{D} \mathcal{N}\left(y_{tnd} - \sum_k z_{tnk} A_{kd}; 0, \sigma_\epsilon\right) \\
p(t_w^k|\mathbf{Y}, R, \mathbf{S}, \mathbf{t}_w^{\neg k}, \mathbf{t}_b, \mathbf{A}, \alpha, \sigma_\epsilon) &\propto p(\mathbf{Y}|\mathbf{S}, \mathbf{t}_w, \mathbf{t}_b, \mathbf{A}, \sigma_\epsilon)p(t_w^k|R, \alpha) \\
&= \text{Exponential}\left(t_w^k; \frac{R}{\alpha}\right) \prod_{t=1}^{L} \prod_{n=1}^{N} \prod_{d=1}^{D} \mathcal{N}\left(y_{tnd} - \sum_k z_{tnk} A_{kd}; 0, \sigma_\epsilon\right)
\end{aligned}
$$

Sample the noise standard deviation σ_ϵ We put a Gamma prior over the precision, that is $\tau_\epsilon \sim \text{Gamma}(a_\epsilon, \beta_\epsilon)$, where $a_\epsilon, \beta_\epsilon$ are the shape and rate hyperparameters. The posterior over the precision is then,

$$
p(\tau_\epsilon|\mathbf{Y}, \mathbf{S}, \mathbf{t}_b, \mathbf{t}_w, \mathbf{A}) \propto p(\mathbf{Y}|\mathbf{S}, \mathbf{t}_b, \mathbf{t}_w, \mathbf{A}, \tau_\epsilon)p(\tau_\epsilon|a_\epsilon, \beta_\epsilon)
$$

The Gamma prior is conjugate to the Gaussian likelihood and thus, the conditional posterior is a Gamma distribution too. More precisely, the posterior has the form

$$
p(\tau_\epsilon|\mathbf{Y}, \mathbf{S}, \mathbf{t}_b, \mathbf{t}_w, \mathbf{A}) = \text{Gamma}\left(a_\epsilon + \frac{LND}{2}, \beta_\epsilon + \sum_{t=1}^{L} \sum_{n=1}^{N} \sum_{d=1}^{D} \frac{\epsilon_{tnd}^2}{2}\right)
$$

where $\epsilon_{tnd} = y_{tnd} - \sum_{k=1}^{KT} z_{tnk} A_{kd}$. Finally, we compute the standard deviation as $\sigma_\epsilon = \frac{1}{\sqrt{\tau_\epsilon}}$.

Sample the weights $\boldsymbol{\omega}$ Gibbs sampling is simple since the Beta prior $w_k \sim \text{Beta}\left(\frac{R}{K}, 1\right)$ is conjugate to the Binomial likelihood $p(\mathbf{S}|\boldsymbol{\omega}) = \prod_{k=1}^{KT} \text{Binomial}(m_k; N, \omega_k)$ resulting in the conditional posterior distribution $w_k \sim \text{Beta}\left(|m_k| + \frac{R}{K}, N - |m_k| + 1\right), i = 1, \ldots, KT$ where $|m_k|$ is the number of objects that have feature f_k in their potential, i.e. $S_{nk} = 1$.

$$
\begin{aligned}
p(\omega_k|\mathbf{S}) &\propto p(\mathbf{S}(:, k)|\omega_k)p(\omega_k) \\
&\propto \text{Binomial}(m_k; N, \omega_k)\text{Beta}\left(\frac{R}{K}, 1\right)
\end{aligned}
$$

$$\propto \quad \text{Beta}\left(\frac{R}{K} + m_k, 1 + N - m_k\right),$$

for $k = 1, \ldots, KT$.

Sample the feature potential matrix S The prior over the feature potential matrix $\mathbf{S}$ is a product of Bernoulli distributed parameters. More precisely,

$$p(\mathbf{S}|\boldsymbol{\omega}) = \prod_{n=1}^{N}\prod_{k=1}^{KT} p(S_{nk}|\boldsymbol{\omega}) = \prod_{n=1}^{N}\prod_{k=1}^{KT} \text{Bernoulli}(S_{nk};\omega_k)$$

The posterior over each matrix element S_{nk} is given by

$$p(S_{nk}|\mathbf{Y}, \omega, \mathbf{S}_{\neg nk}, \mathbf{t}_b, \mathbf{t}_w, \mathbf{A}, \sigma_\epsilon) \propto p(\mathbf{Y}|\mathbf{S}, \omega, \mathbf{t}_b, \mathbf{t}_w, \mathbf{A}, \sigma_\epsilon) p(S_{nk}|\boldsymbol{\omega})$$

We only need to consider $S_{nk} \in \{0, 1\}$, so we evaluate the right hand side for $S_{nk} = 0$ and $S_{nk} = 1$, normalize, and sample S_{nk} from the resulting Bernoulli posterior.

3.1 Getting it right

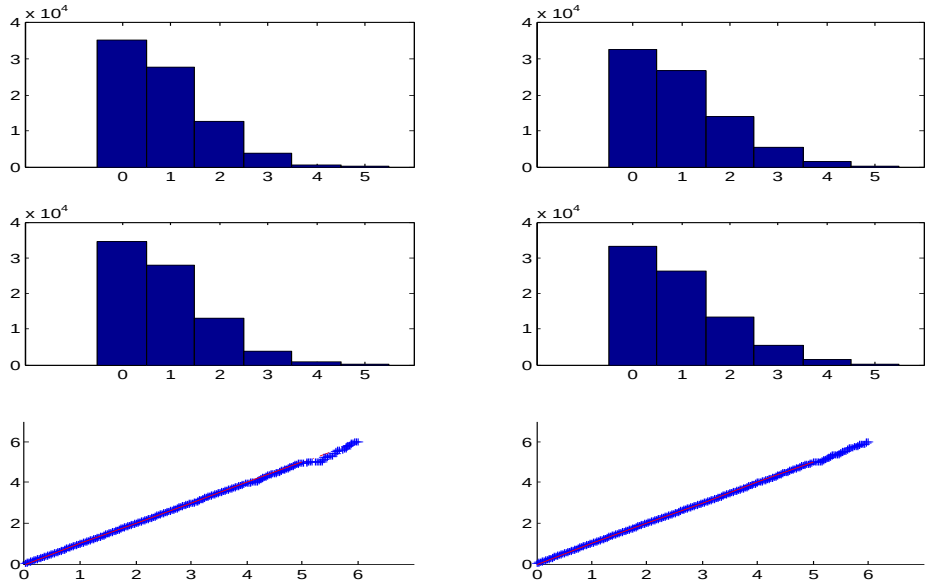
To validate our sampling alorithm for BEP we follow the joint distribution testing methodology of Geweke (2004). There are two ways to sample from the joint distribution, $P(Y, \theta)$ over parameters $\theta = \{R, \boldsymbol{\omega}, \boldsymbol{S}, \boldsymbol{t}_b, \boldsymbol{t}_w, \boldsymbol{A}, \alpha, \sigma_\epsilon\}$, and data, Y defined by a probabilistic model such as BEP. The first we will refer to as "marginal-conditional" sampling, shown in Algorithm 1. Both steps here are straightforward: sampling from the prior followed by sampling from the likelihood model. The second way, referred to as "successive-conditional" sampling is shown in Algorithm 2, where Q represents a single (or multiple) iteration(s) of our MCMC sampler. To validate our sampler we can then check, either informally or using hypothesis tests, whether the samples drawn from the joint $P(Y, \theta)$ in these two different ways appear to have come from the same distribution.

We apply this method to our sampler with just $N = 10$, $D = 2$ and $|\mathcal{F}| = 5$, two time points and all hyperparameters fixed as follows: For the shape and scale of α we set $\kappa_\alpha = 4, \theta_\alpha = 1$, for the shape and rate of ϵ we set $\alpha_\epsilon = 2, \beta_\epsilon = 2$, for the shape and scale of R we chose $\kappa_R = 4, \theta_R = 1$ and finally for the mean and standard deviation of A we set $\mu_A = 0, \sigma_A = 1$.

We draw 80K samples using both the marginal-conditional and successive-conditional procedures. We look at various characteristics of the samples, including the number of features at every time point, the mean of the factor loading matrix A.

The distribution of the number of features under the successive-conditional sampler matches that under the marginal-conditional sampler almost perfectly as shown in Figure 2. Both the histogram and the quantile-quantile plot show the similarity of the two distributions, with the straight line in the later indicating an almost perfect match. The deviation from a straight line in the upper

corner of the qq-plot is a result of there being fewer samples available to estimate these quantiles accurately. Under the successive-conditional sampler the average number of features is 0.85, 0.95 for the two locations while under the marginal-conditional is 0.83, 0.96 respectively with standard deviations 0.91, 1.01 and 0.92, 1.02 respectively: a hypothesis test did not reject the null hypothesis that the means of the two distributions are equal. While this cannot completely guarantee correctness of the algorithm and code, 80K samples is a large number for such a small model and thus provides strong evidence that our algorithm is correct. Figure 3 provides the same evidence.

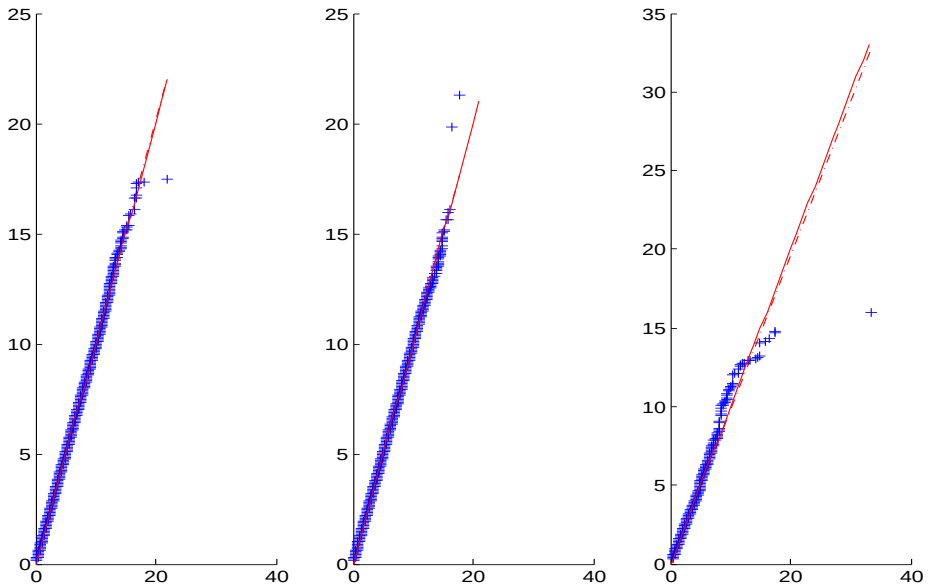


(a) First time location (b) Second time location

Figure 2: Comparing the distribution of the number of features under the marginal-conditional and successive conditional samplers. Figures in column (a) show the empirical distribution over the number of features under the marginal-conditional (first row) and successive conditional (second row) for the first time location. Respectively for the second time location in column (b). Figure in the last row show the qq-plots of the two empirical distributions for the two time locations. The agreement of the two distributions is evidence for the correctness of our MCMC sampler for the finite model.

4 Experiments

In the main paper, we presented results on real world dataset. To complete the analysis, we provide here further experiments on synthetic dataset.



(a) α (b) R (c) σ_ϵ

Figure 3: Comparing the distribution α, R and σ_ϵ under the marginal-conditional and successive conditional samplers. Figure shows the qq-plots of the two empirical distributions. The agreement of the two distributions is evidence for the correctness of our MCMC sampler for the finite model.

4.1 Synthetic dataset

We first explored the ability of our model to recover underlying structure using synthetic data. We generated observations $\mathbf{Y}_t$ from the BEP at 7 distinct time points, $t = 1, 2, \ldots, 7$. More specifically, we assumed $N = 20$ datapoints and 4 features with birth times, $\mathbf{t}_b = [0.1,\ 0.4,\ 0.8,\ 1.2]$ and life spans $\mathbf{t}_w = [0.8,\ 0.8,\ 0.8,\ 0.8]$. We also assumed a potential matrix S and derived the feature allocation matrices $\mathbf{Z}_t$ as determined by the BEP and shown in Figure 4. We used the four 6×6 images shown in Figure 4 (middle row, left) as features and collected them to construct the feature loading matrix $\mathbf{A}$ of size 4×36. Each row of this matrix corresponds to one of the 4 features. The synthetic data at each time point is then generated by superimposing the images using the linear Gaussian likelihood, i.e. $\mathbf{Y}_t = \mathbf{Z}_t\mathbf{A} + \epsilon$, where ϵ is the noise term which we take as Gaussian with standard deviation 0.5.

We ran both models, that is the BEP and a set of 7 independent IBP models (one at each time location). To evaluate predictive performance, we held out 10% of the data (elements in each Y_t). For inference, we ran the BEP sampler derived in Section 3 for 1000 MCMC iterations, which appeared sufficient for burnin. The total number of features is $KT = 12$, where we took $K = 6$ and $T = 2$. For the independent IBP model, we considered the same Gaussian likelihood

Algorithm 1: Marginal conditional

1: **for** $m = 1$ to M **do**
2: $\quad \theta^{(m)} \sim P(\theta)$
3: $\quad Y^{(m)} \sim P(Y|\theta^{(m)})$
4: **end for**

Algorithm 2: Successive conditional

1: $\theta^{(1)} \sim P(\theta)$
2: $Y^{(1)} \sim P(Y|\theta^{(1)})$
3: **for** $m = 2$ to M **do**
4: $\quad \theta^{(m)} \sim Q(\theta|\theta^{(m-1)}, Y^{(m-1)})$
5: $\quad Y^{(m)} \sim P(Y|\theta^{(m)})$
6: **end for**

	BEP	independent IBP
Train error	3.3934 ± 0.0714	$\mathbf{3.3428 \pm 0.0738}$
Test error	$\mathbf{3.4229 \pm 0.1771}$	4.7367 ± 0.4283
Train log likelihood	$-3,348 \pm 9.7123$	$\mathbf{-3,311 \pm 33.8831}$
Test log likelihood	$\mathbf{-381.7972 \pm 4.4620}$	-605.6935 ± 68.9909

Table 1: Results for synthetic Gaussian superimposition data

as for the BEP but with independent A and noise variance at each location. The quantitative results are presented in Table 4.1 where the likelihood and the error are averaged over the last 200 MCMC samples. Figure 4 shows the solutions found by the two models. The MCMC sample with the highest log probability under the posterior was used. The BEP successfully finds the true features and Z matrix, whereas the independent IBP model finds a solution where additional features are used in all of the 7 locations (see second and third row of Figure 4). The clean solution provided by BEP shows that leveraging the dependence among consecutive feature allocations greatly improves performance. Table 4.1 confirms this quantitatively: the BEP model performs considerably better both in terms of test error and likelihood. While the independent IBP looks good in terms of training error/likelihood, the big difference in the performance in terms of the test likelihood suggests this is overfitting.

Next, we explore the ability of BEP to recover latent features in a synthetic time series network data. We hand-constructed a set of six square binary matrices which encode the friendship links among $N = 20$ people evolving through time, as shown in Figure 5. People form groups which determine the links and non-links between them. As time passes, the partitioning of people changes: new friendship links are created while others break. The closer in time two snapshots are, the more similar we expect the related partitions to be. We ran BEP using the network likelihood model in Equation (16) for 2000 MCMC iterations keeping the last 400 samples for estimation and holding 10% of the data out for prediction. For comparison, we used independent LFRM models at each timepoint. The BEP model outperforms the independent LFRM in terms of both test error and likelihood (Table 4.1) while, analogously to the linear Gaussian setting, the independent LFRM seems to overfit yielding "better" values in

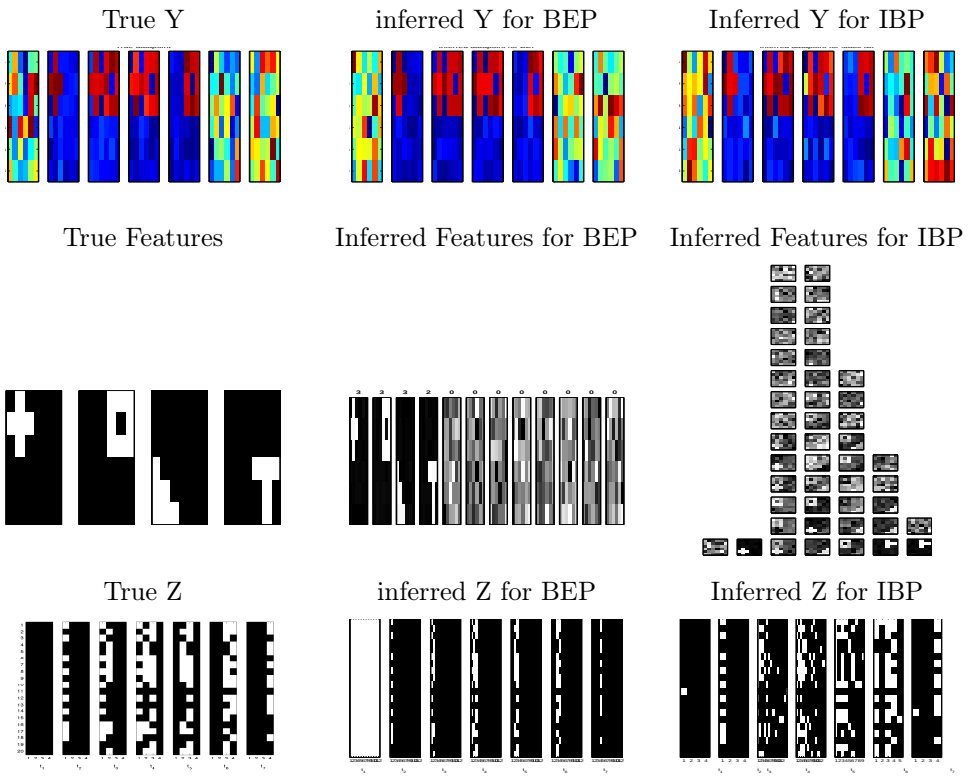


Figure 4: Linear Gaussian synthetic data experiment. **Top row:** the true and reconstructed observations for one datapoint ($n = 16$th) at the 7 different time locations. **Middle row:** the true and inferred features. The inferred features for BEP have been plotted along with a number on top indicating how often they are used, that is the total number of time locations at which they are active. Note that for the independent IBP model there are different factor loading matrices for each time location. **Bottom row:** true and inferred feature allocation matrices for the different time locations.

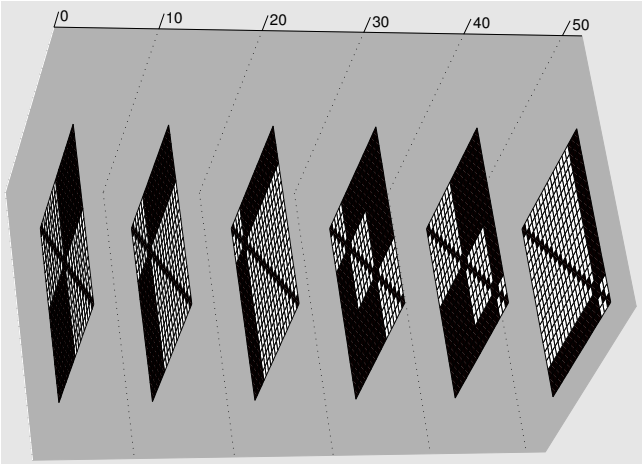


Figure 5: Synthetic network data.

	BEP	independent LFRM
Train error	$\mathbf{0.0688 \pm 0.0148}$	0.0968 ± 0.0286
Test error	$\mathbf{0.0745 \pm 0.0132}$	0.1211 ± 0.0194
Train log likelihood	-209.2373 ± 5.8391	$\mathbf{-203.8954 \pm 13.7807}$
Test log likelihood	$\mathbf{-9.2547 \pm 1.7358}$	-19.4331 ± 5.3122

Table 2: Results for synthetic link data

the train likelihood. Figures 6 and 7 offer a qualitative overview of the solution. Figure 6 shows the features found in the sample with the highest log probability under the posterior. Both show figures the slow evolution of the feature allocation dictated by the BEP; once objects are allocated to a feature, the feature membership remains the same until the feature dies. For instance, moving from time location 1 to 2 explains the data by keeping feature 17 alive (with the same 10 members) and introducing features 12 and 15. The LFRM is overly flexible since it assigns objects to features at each location independently. As such, it explains the observations at the second time location using two features created independently from the ones in the previous location.

4.2 ChIP-seq dataset

shows genomic annotations, from ChromHMM (Ernst et al., 2011) for the region we model in the main paper. The inferred solution for both the BEP and the independent IBP's are shown in Figure 9 and Figure 10 respectively. The BEP model allows for a smooth evolution of the latent feature allocation inferred as opposed to the independe IBP's where the latent structure is explained with rapid changing allocations and with considerable differences in the number of

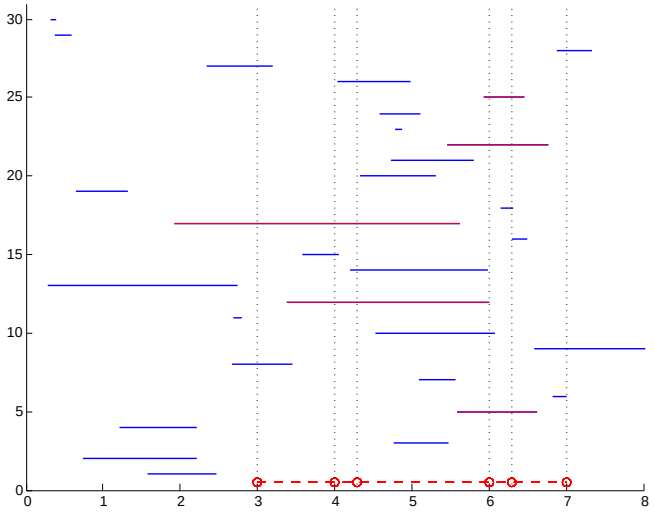


Figure 6: Inferred features using the BEP in he synthetic network data. Time is denoted on the x-axis along with the 6 time locations (dasehd red line), i.e [3 4 4.3 6 6.3 7]. The total number of features is $|\mathcal{F}| = 30$. Red colour is used to indicated the features that are alive for more than one location. The features that cross the vertical grey line at each time location are the ones present at that time.

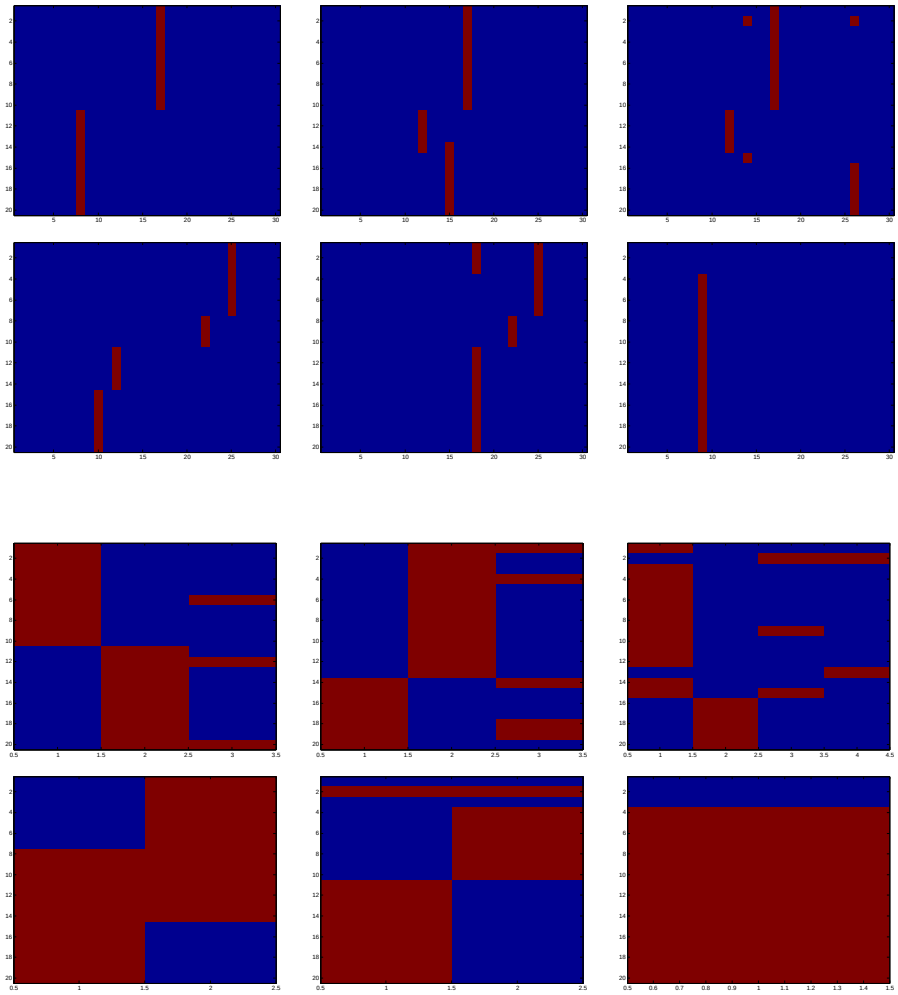


Figure 7: Inferred feature allocation matrices for the six locations (from left to right) in the synthetic link dataset. **First two rows:** Feature allocation matrices inferred by BEP. **Last two rows:** Feature allocation matrices inferred by independent LFRM.

features found at each location. As such, covariate dependence over the allocation is a better modelling approach.

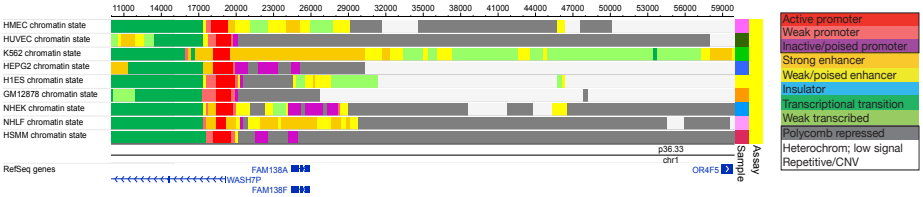


Figure 8: ChIP-seq data: Chromatin states for the genomic region we model. From the BEP reconstruction in Figure 5(b) we see the promoter region around 18kb-19kb with high H3K27ac, the transcribed region of the WASH7P gene from 8kb-18kb, and the repressive H3K29me3 and H3K9me3 marks further downstream, corresponding to polycomb repression and heterochromatin.

4.3 van de Bunt's dataset

In van de Bunt et al. (1999), 32 university freshman students in a given discipline at a Dutch university were surveyed at seven time points about who in their class they considered as friends. Initially, i.e. t_1, most of the students were unknown to each other. The first four time points are three weeks apart, whereas the last three time points are six weeks apart as showin in Figure 11.

References

Ernst, Jason, Kheradpour, Pouya, Mikkelsen, Tarjei S, Shoresh, Noam, Ward, Lucas D, Epstein, Charles B, Zhang, Xiaolan, Wang, Li, Issner, Robbyn, Coyne, Michael, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011. 14

Geweke, John. Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association*, 99:799–804, 2004. 9

Griffiths, Thomas L. and Ghahramani, Zoubin. The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, July 2011. 3

Neal, R M. Slice sampling. *The Annals of Statistics*, 31(3):705–741, 2003. 8

van de Bunt, Gerhard G, Van Duijn, Marijtje AJ, and Snijders, Tom AB. Friendship networks through time: An actor-oriented dynamic statistical network model. *Computational & Mathematical Organization Theory*, 5:167–192, 1999. 17

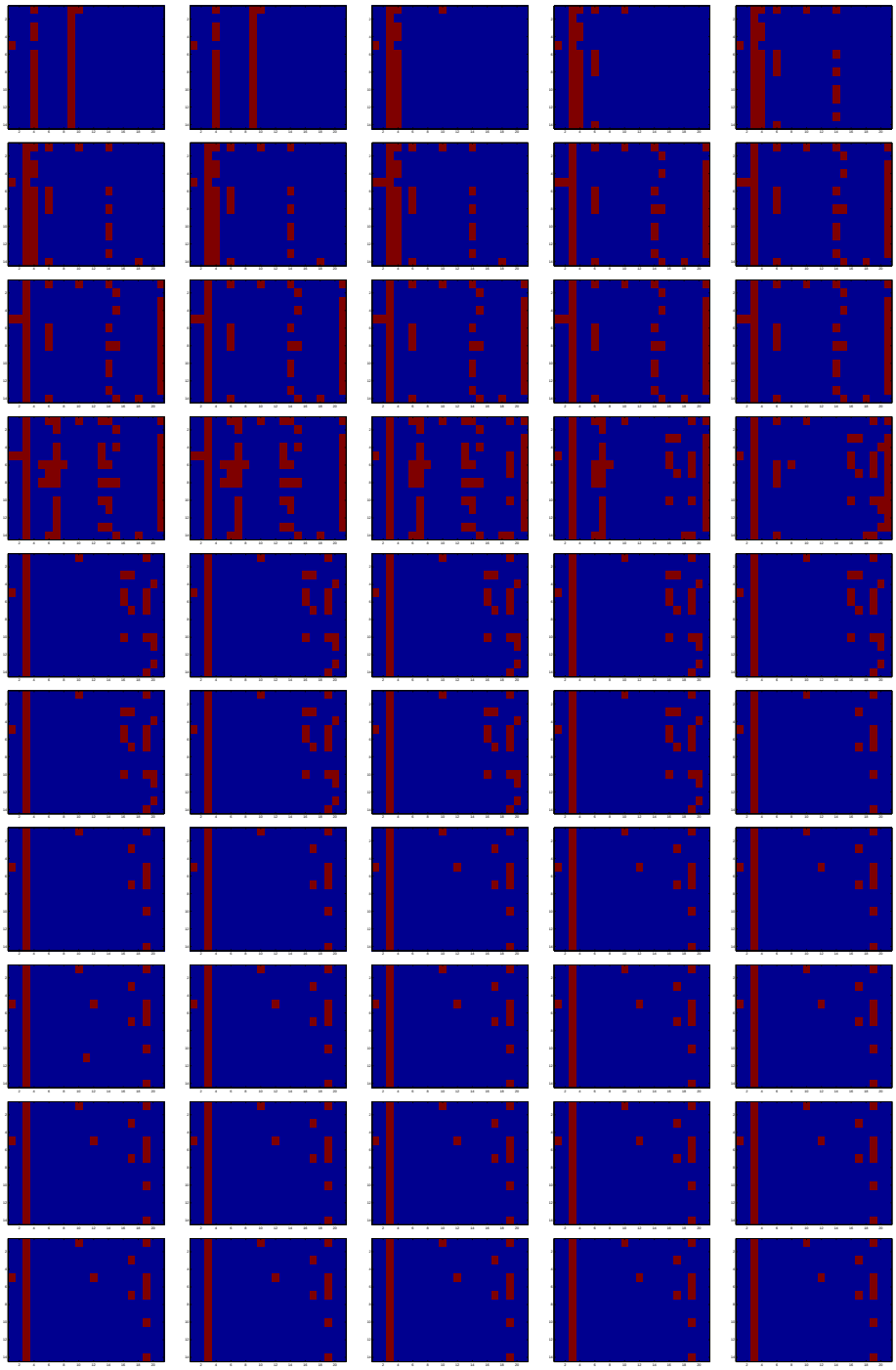


Figure 9: ChIP-seq dataset: Inferred feature allocation matrices for the BEP model. The allocation matrices for 50 locations (out of the 500) from left to right are shown. Each pair of adjacent matrices correspond to locations with 10 bins distance, i.e. 10^3 base pairs.

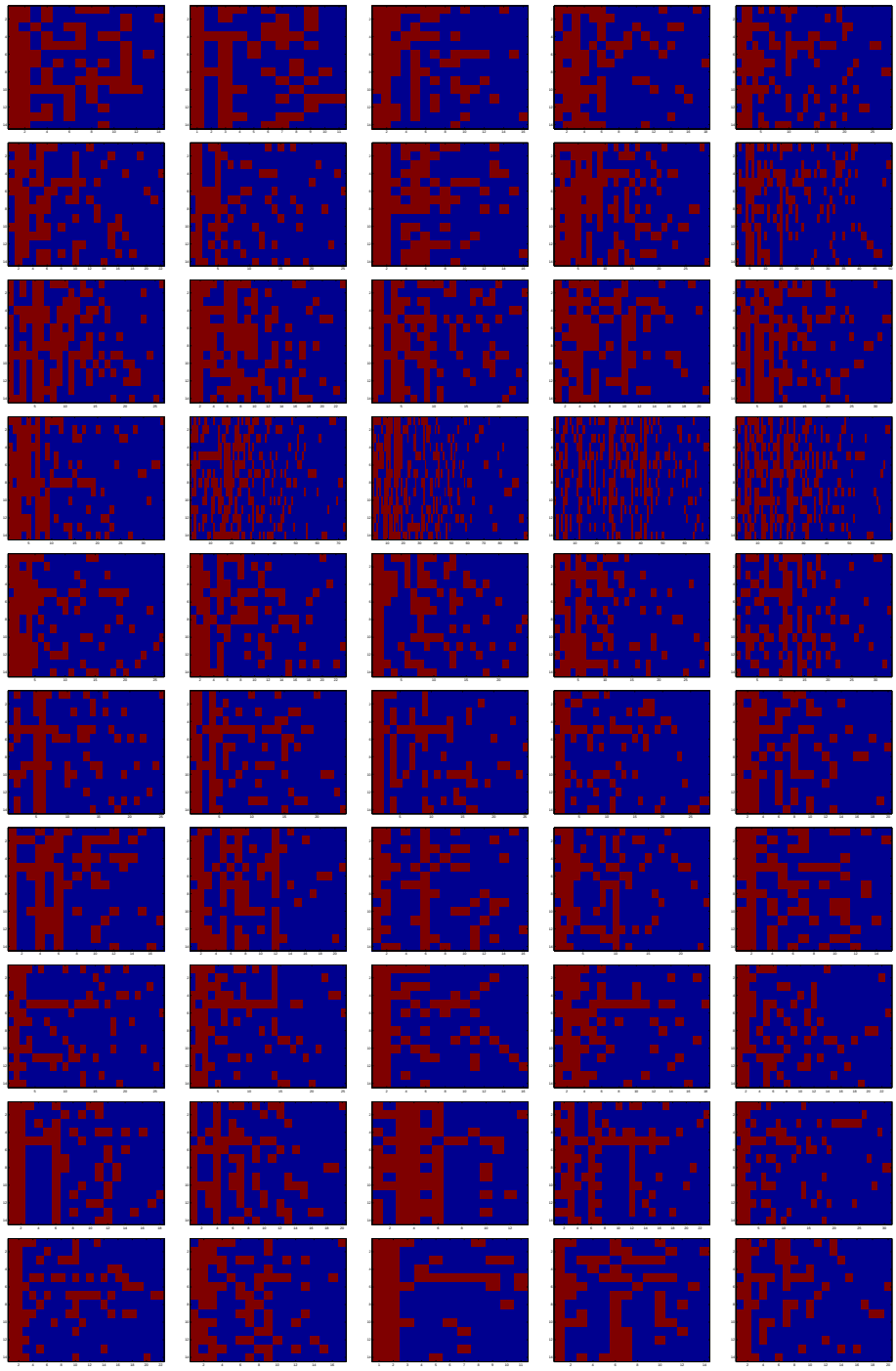


Figure 10: ChIP-seq dataset: Inferred feature allocation matrices for the independent IBP model. The allocation matrices for 50 locations (out of the 500) from left to right are shown. Each pair of adjacent matrices correspond to locations with 10 bins distance, i.e. 10^{19} base pairs.

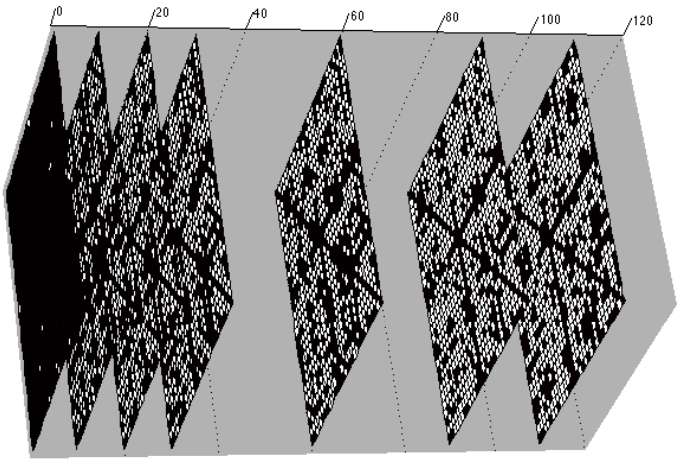


Figure 11: van de Bunt network data.