Learning with Labeled and Unlabeled Multi-Step Transition Data for Recovering Markov Chain from Incomplete Transition Data

Masahiro Kohjima , **Takeshi Kurashima** and **Hiroyuki Toda**
NTT Service Evolution Laboratories, NTT Corporation, Japan
{masahiro.kohjima.ev, takeshi.kurashima.uf, hiroyuki.toda.xb}@hco.ntt.co.jp

Abstract

Due to the difficulty of comprehensive data collection, created by factors such as privacy protection and sensor device limitations, we often need to analyze incomplete transition data where some information is missing from the ideal (complete) transition data. In this paper, we propose a new method that can estimate, in a unified manner, Markov chain parameters from incomplete transition data that consist of hidden transition data (data from which visited state information is partially hidden) and dropped transition data (data from which some state visits are dropped). A key to developing the method is regarding the hidden and dropped transition data as labeled and unlabeled multi-step transition data, where the labels represent the number of steps required for each transition. This allows us to describe the generative process of multi-step transition data, and thus develop a new probabilistic model. We confirm the effectiveness of the proposal by experiments on synthetic and real data.

1 Introduction

The Markov chain (MC) has been applied to various dynamic systems such as queuing systems, marketing and finance [Neuts, 1981; Pfeifer and Carraway, 2000; Frydman and Schuermann, 2008]. Due to the development and widespread use of sensor devices, MC variants are now used for analyzing the dynamics of urban cities such as traffic, and people flows [Crisostomi *et al.*, 2011; Fan *et al.*, 2015; Iwata *et al.*, 2017].

Since the transition probabilities of MC are unknown in practice, we need to estimate them from observed transition data. In the ideal case, for example, if the information of every visit of each person to each state (location) is available for analyzing people flow in a city [1], the transition probability can be estimated directly from the number of transitions between the states [Billingsley, 1961]. However, comprehensive data collection is rare because of, for example, constraints imposed by privacy protection, coverage area deficiencies, and limited sensor device precision. Thus in practice the actual data to be analyzed differs from the ideal *complete* transition data; the transition data available is *incomplete* as some information is missing.

We found that it is essential to be able to analyze two common types of incomplete transition data: *hidden* transition data where visited state information is partially hidden and *dropped* transition data where visits to states are partially dropped. See Figure 1. The difference between the two is whether we are aware of the existence of "missing states" or not in the sequence of transitions.

An example of dropped transition data is the transition data of subscribers provided by mobile phone companies. The data contain the transition between states (locations) only where the user stays for more than a certain time to ensure privacy protection and to save memory cost. Therefore, as shown in Fig. 1, although the (true) complete transition consists of two steps, $a \rightarrow b \rightarrow c$, the visit of state b is dropped and only the transition $a \rightarrow c$ is recorded so the user appears to have visited only state a and c. An example of hidden transition data is the data collected by ourselves using e.g., GPS logger or mobile phone apps. Since the data often contains missing values, the current state (location) information may not be recorded in some logs although the log exists. For example, if the visited state information of state b is missing from $a \rightarrow b \rightarrow c$, only the transition $a \rightarrow ? \rightarrow c$ is recorded.

In this paper, we propose a new method that can, in a unified manner, estimate MC parameters from an incomplete transition data set that includes dropped and/or hidden transition data. A key to developing the method is to regard hidden and dropped transition data as labeled and unlabeled multi-step transition data, where the label represents the number of steps required for each transition. See the bottom of Fig. 1. The example of the hidden transition data, $a \rightarrow ? \rightarrow c$, and that of dropped transition data, $a \rightarrow c$, can be seen as the labeled multi-step transition $a \rightarrow c$ where the number of steps is 2 and the unlabeled multi-step transition $a \rightarrow c$ where the number of steps is unknown, respectively. This yields the following description of the generative process of multi-step transition data: the labels indicating the number of steps, k, is first determined and the next state is determined by multi-step transition probability, which, according to MC theory, is

[1] Hereinafter we consider that people's movements are represented by the transition between discretized areas or meshed cells similar to [Fan *et al.*, 2015] for simplicity.

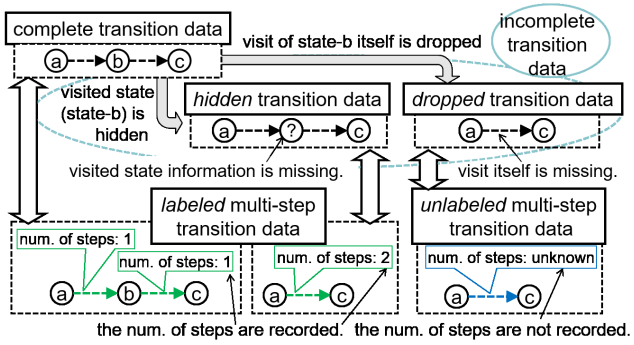


Figure 1: Examples and relation between complete and incomplete transition data and labeled and unlabeled multi-step transition data.

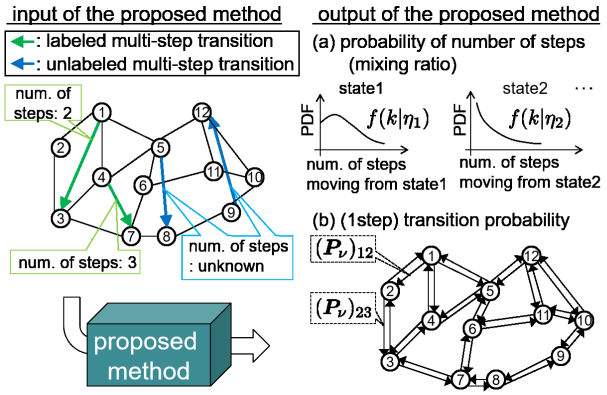


Figure 2: Problem formulation. Proposed method estimates (a) probability of generated number steps and (b) 1-step transition probability from labeled and/or unlabeled multi-step transition data.

given by raising the transition probability to the power of k. Figure 2 shows the input and output of the proposed method. By estimating MC parameters, we can recover the (1-step) transition probability that underlies multi-step transitions. It is also shown that the proposed method can be seen as a semi-supervised learning method of (possibly infinite) mixture of exponentiated transition probabilities.

We also develop a Majorization Minimization (MM) algorithm [Hunter and Lange, 2004; De Leeuw, 1994] that can guarantee that the objective function improves with each iteration. Note that our formulation also allows us to use complete transition data since it can be represented by labeled multi-step transition data whose labels are all 1. Moreover, our formulation, model, and algorithms are not limited to the case of analyzing people flow and indeed is valid in various scenarios where the available data are represented by labeled and unlabeled multi-step transition data.

The contributions of this paper are summarized below:

- We consider the problem of recovering MC parameters from labeled and/or unlabeled multi-step transition data.
- We develop a new model that can handle labeled and/or unlabeled multi-step transition data; it can be seen as a semi-supervised learning method of (possibly infinite) mixture of exponentiated transition probability.
- We develop an MM algorithm and give proof of its convergence. A method of handling infinite mixtures with theoretical support is also provided.
- We confirm the effectiveness of the proposal by experiments on synthetic and real data sets.

The rest of this paper is organized as follows. §2 details related works and §3 provides a definition and theory of the Markov chain. The proposed method is presented in §4. §5 details the experiments conducted and §6 concludes the paper.

2 Related Works

Recently, new problem formulations for estimating MC parameters have been investigated. For example, [Morimura *et al.*, 2013; Kumar *et al.*, 2015] tackle the problem of recovering MC from a steady state distribution. [Iwata *et al.*, 2017; Akagi *et al.*, 2018] tackle the estimation problem by using snapshots of population data. However, the problem of MC recovery from multi-step transition data, especially unlabeled multi-step transition data made from dropped transition data, has yet to be studied. Moreover, although handling just hidden transition data is simple since the EM algorithm can be adopted by treating missing state information as latent variables, no unified approach that can handle both dropped and hidden transition data exists. Hidden Markov Model (HMM) cannot be applied to our problem since the time-step of (true) complete transition data and that of dropped transition data do not have one to one correspondence and an infinitely large number of steps may be dropped. This study provides a unified approach to handling the two types of data as well as complete transition data.

Our problem formulation can be seen as *semi-supervised learning* since the relation between the labeled and unlabeled multi-step data are similar to the way labeled and unlabeled data are treated in semi-supervised clustering [Ghahramani and Jordan, 1994; Nigam *et al.*, 2000; Basu *et al.*, 2002]. However, we emphasize our motivation comes from recovering MC from incomplete transition data.

Our model for MC recovery from multi-step transition data can be seen as a mixture model. However, our model differs from the "standard" mixture of Markov chains that have different component distributions [Goodman, 1961; Frydman, 1984; Gupta *et al.*, 2016]; the k-th component distribution of our model is the transition probability raised to the power of k. This property allows us to approximate the component with large k by a steady state distribution under a mild assumption; it enables us to construct a new type of infinite mixture model that differs from Dirichlet process mixture models [Ferguson, 1973; Antoniak, 1974; Neal, 2000; Blei and Jordan, 2006; Teh, 2010].

3 Preliminaries

Let $\mathcal{X} = \{1, 2, \cdots, |\mathcal{X}|\}$ be a state space. A discrete time Markov chain (MC) on $\mathcal{X}$ is a stochastic process $\{X_t; t =$

	support	η	$h(k)$	$T(k)$	$A(\eta)$	μ
Categorical	$k = \{0 \text{ or } 1, \cdots, K\}$	$\{\log(\lambda_i/\lambda_K)\}_i$	1	$\{I(k{=}i)\}_i$	$\log(\sum_k e^{\eta_k})$	$\{e^{\eta_i}/\sum_k e^{\eta_k}\}_i$
Geometric	$k = \{1, 2, \cdots\}$	$\log(1-\lambda)$	1	k	$\eta - \log\{1 - e^{\eta}\}$	$1/\{1 - e^{\eta}\}$
Poisson	$k = \{0, 1, \cdots\}$	$\log(\lambda)$	$1/k!$	k	$\exp(\eta)$	$\exp(\eta)$
ZTP	$k = \{1, 2, \cdots\}$	$\log(\lambda)$	$1/k!$	k	$\log(e^{e^{\eta}} - 1)$	$e^{e^{\eta}+\eta}/\{e^{e^{\eta}} - 1\}$

Table 1: Examples of exponential family

$0, 1, 2, \cdots\}$ that satisfies the following Markov property:

$$Pr(X_{t+1} = x_{t+1} | X_\ell = x_\ell; \ell = 0, \cdots, t)$$
$$= Pr(X_{t+1} = x_{t+1} | X_t = x_t) \quad (\forall x_\ell \in \mathcal{X}, \forall t \in \mathbb{Z}_{\geq 0}).$$

MC $\mathcal{M}$ is thus defined by $\mathcal{M} := \{\mathcal{X}, \mathcal{P}\}$, where $\mathcal{P} : \mathcal{X} \times \mathcal{X} \to [0, 1]$ is the transition probability, $\mathcal{P}(x_{next}|x) = Pr(X_{t+1} = x_{next} | X_t = x)$. We also employ adjacency information of $\Gamma = \{\Gamma_i\}_{i \in \mathcal{X}}$, where Γ_i is the set of reachable states from state i by one-step transitions.

Later we will use the following theoretical results.

Theorem 1. (e.g. Theorem (2.1) [Durrett, 1999]) *k-step transition probability is given by $\boldsymbol{P}$ to the power of k, $\boldsymbol{P}^k$.*

For example, 2-step and 3-step transition probabilities are given by $\boldsymbol{P}^2$ and $\boldsymbol{P}^3$, respectively. Note that 0-step transition probability is self-transition since a matrix to the power of 0 is an identity matrix. If MC $\mathcal{M}$ is *irreducible* and *aperiodic*, the following theorem holds.

Theorem 2. (e.g. Theorem (4.5) [Durrett, 1999]) *Let $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^{|\mathcal{X}|}$ be the steady state probability of Markov chain $\mathcal{M}$. Then, $(\boldsymbol{P}^k)_{ij} \xrightarrow[k \to \infty]{} \pi_j$.*

This states that each row of $\boldsymbol{P}^k$ converges to the steady state probability at the limit.

4 Proposed Method

This section describes the proposed method for estimating Markov chain parameters from multi-step transition data. Figure 2 shows our problem formulation.

4.1 Data

We assume that incomplete transition data are available in the form of unlabeled and labeled multi-step transition data, as shown in Fig 1. We denote the unlabeled multi-step transition data made from dropped transition data (where the number of steps is unknown) as $\mathcal{D}_{um} = \{N_{ij}\}_{ij \in \mathcal{X}}$ where N_{ij} denotes the number of transitions from state i to state j. We also denote the labeled multi-step transition data made from hidden transition data as $\mathcal{D}_{\ell m} = \{M_{ijk}\}_{ij \in \mathcal{X}, k \in \mathbb{Z}_{\geq 0}}$ where M_{ijk} denotes the number of transitions from state i to state j by k-step transitions. If complete transition data is available, we consider that the data are represented by part of the labeled multi-step transition data, $\{M_{ij1}\}_{ij}$. We denote the set of the two data types as $\mathcal{D} = \mathcal{D}_{um} \cup \mathcal{D}_{\ell m}$ and refer to it as multi-step transition data. We also use adjacency information Γ as an optional input. In the case of, for example, analysis of people flows in a city, Γ can be derived from maps or street network information. If such adjacency information is not available, we consider Γ_i to be a set of all states, $\mathcal{X}$. Although Γ is optional, its use is recommended if the labeled multi-step data are unavailable or limited since Γ strongly helps in estimating the number of steps present in the unlabeled multi-step data.

4.2 Model Component

Here we show the model components used by the proposed method.

Label Probability. We decide to use the exponential family for modeling generated labels that indicate the number of steps of each transition since it can express various types of distributions. The density of exponential family f is given by

$$f(k|\eta) := h(k) \exp\big\{\eta \cdot T(k) - A(\eta)\big\}, \tag{1}$$

where η is the *natural parameter*, $T(k)$ is *sufficient statistics* and $A(\eta)$ is the log-normalizer. Examples of distributions belonging to the exponential family are categorical distribution (Cat), geometric distribution (Geo), Poisson distribution (Poi), and zero-truncated Poisson distribution (ZTP):

$$\mathrm{Cat}(k|\lambda) = \textstyle\prod_i \lambda_i^{I(k=i)},\ \mathrm{Geo}(k|\lambda) = (1-\lambda)^{k-1}\lambda,$$
$$\mathrm{Poi}(k|\lambda) = \lambda^k \exp(-\lambda)/k!,\ \mathrm{ZTP}(k|\lambda) = \lambda^k/\{(e^\lambda - 1)k!\}.$$

By assigning specific values to $T(x)$ and $A(\eta)$, the above distributions are represented by Eq. (1) (See Table 1). We denote the cumulative density function of f as F, where $F(C|\eta) := \int_{-\infty}^{C} f(k|\eta)dk$. In a later section, we use *mean-value parameter* μ, which is defined as $\mu := \mathbb{E}_{f(k|\eta)}[T(k)] = \partial A(\eta)/\partial \eta$. Note that natural parameter η and mean-value parameter μ have a one-to-one correspondence [Amari and Nagaoka, 2007]. Converting μ into η may require numerical computation; we use Newton method for ZTP. We will also use the property that the Hessian of the log-normalizer is positive-definite, since it corresponds to the variance.

Transition Probability. We use parameter $\boldsymbol{\nu}$ to model the (1-step) transition probability of a MC. To emphasize parameter dependency, we denote the model of the transition probability as $\boldsymbol{P}_\nu$; the (i, j)-th element of matrix $\boldsymbol{P}_\nu$ represents the (1-step) transition probability from state i to j, i.e., $Pr(X_{t+1} = j | X_t = i, \nu) = (\boldsymbol{P}_\nu)_{ij}$. We denote a Markov chain constructed using the model with parameter ν as $M_\nu = \{\mathcal{X}, P_\nu\}$. Examples of the model are the following tabular model and log-linear model.

Model 1. (Tabular Model) *Let us define ν as $\nu = \{\{p_{ij}\}_{j \in \Gamma_i}\}_{i \in \mathcal{X}}$ where $\sum_j p_{ij} = 1$ for all i. The tabular model can be defined as $(\boldsymbol{P}_\nu)_{ij} = p_{ij}$ if $j \in \Gamma_i$, and $(\boldsymbol{P}_\nu)_{ij} = 0$ otherwise.*

Model 2. (Log-Linear Model) *Let us define* ν *as* $\nu = \{\boldsymbol{v}, \boldsymbol{w}\}$. *Then,* $(\boldsymbol{P}_\nu)_{ij} = \frac{\exp\{v_{ij}+\boldsymbol{\phi}(i,j)^T\boldsymbol{w}\}}{\sum_{j'\in\Gamma_i}\exp\{v_{ij'}+\boldsymbol{\phi}(i,j')^T\boldsymbol{w}\}}$ *if* $j \in \Gamma_i$ *and* $(\boldsymbol{P}_\nu)_{ij} = 0$ *otherwise, where* $\boldsymbol{\phi}(i,j)$ *is a feature vector, such as the (inverse) distance between states* i *and* j [2].

4.3 Generative Process

The generative process of multi-step transition data can be described using the model component explained in the previous section. We consider that parameters $\theta = \{\boldsymbol{\eta}, \boldsymbol{\nu}\}$ are determined following prior distribution $P(\theta) = P(\boldsymbol{\eta})P(\boldsymbol{\nu})$. $P(\boldsymbol{\eta})$ is a conjugate prior of the exponential family,

$$P(\boldsymbol{\eta}) = \prod\nolimits_{i=1}^{|\mathcal{X}|} \mathcal{Z}(\xi_0, \mu_0)\exp\big\{\eta_i \cdot \xi_0 - \mu_0 A(\eta_i)\big\},$$

where the normalizer $\mathcal{Z}(\xi_0, \mu_0)$ is determined by the chosen distribution (See Table 1). As the prior of ν, we can use any distribution. For the case of the log-linear model, a Gaussian distribution is frequently used:

$$P(\boldsymbol{\nu}) = \mathcal{N}(\boldsymbol{\nu}|\mathbf{0}, \alpha_0^{-1}\boldsymbol{I}_{|\boldsymbol{\nu}|}) \propto \exp(-\alpha_0\|\boldsymbol{\nu}\|^2/2).$$

Note that ξ_0, μ_0 and α_0 are hyperparameters[3].

The stochastic process describing a multi-step transition is determined in the following manner. We denote the process as $\{X'_s; s = 0, 1, 2, \cdots\}$ where s are the time steps of data generation [4]. (i) The initial state, X'_0, is determined by some initial state probability. At every data generation time step s, (ii) the number of steps K_s when moving from current state $X'_s = i$ is determined following $P(K_s = k|X'_s = i)$ defined using exponential family f,

$$P(K_s = k|X'_s = i) = f(k|\eta_i). \tag{2}$$

Then, (iii) next state visited by k step transition is determined following $P(X'_{s+1}|X'_s = i, K_s = k)$ which is defined as, from Theorem 1, $\boldsymbol{P}_\nu$ to the power of k,

$$P(X'_{s+1} = j|X'_s = i, K_s = k) = (\boldsymbol{P}_\nu^k)_{ij}. \tag{3}$$

Note that k is a latent variable for unlabeled data and an observed variable for labeled data. The following equations are derived from Eq. (2)(3):

$$P(X'_{s+1} = j, K_s = k|X'_s = i) = f(k|\eta_i)(\boldsymbol{P}_\nu^k)_{ij} \tag{4}$$

$$P(X'_{s+1} = j|X'_s = i) = \sum\nolimits_k f(k|\eta_i)(\boldsymbol{P}_\nu^k)_{ij}. \tag{5}$$

We can interpret this model as being a mixture model because the transition probability defined by Eq. (5) where k is marginalized out is given by the weighted sum of exponentiated transition probability. However, unlike the "standard" mixture model of the Markov chain e.g. [Goodman, 1961; Frydman, 1984; Gupta *et al.*, 2016], the mixing ratio of our model is given by an exponential family and the k-th component is the transition probability to the power of k. Moreover, when we adopt an exponential family that has unbounded support such as a Poisson distribution, the model becomes an *infinite* mixture model; a way of handling infinite summation is detailed later.

From Eq. (4)(5), the log-likelihood of the labeled and unlabeled multi-step transition data are given by

$$\log P(\mathcal{D}_{\ell m}|\theta) = \sum\nolimits_{i,j,k} M_{ijk}\log\{f(k|\eta_i)(\boldsymbol{P}_\nu^k)_{ij}\},$$

$$\log P(\mathcal{D}_{um}|\theta) = \sum\nolimits_{i,j} N_{ij}\log\big\{\sum\nolimits_k f(k|\eta_i)(\boldsymbol{P}_\nu^k)_{ij}\big\}.$$

The logarithm of joint probability of data $\mathcal{D}$ and parameter θ is then given by

$$\log P(\mathcal{D}, \theta) = \log P(\mathcal{D}_{\ell m}|\theta) + \log P(\mathcal{D}_{um}|\theta) + \log P(\theta).$$

Parameter θ is estimated by optimizing this objective.

$$\hat{\theta} = \arg\min_\theta \mathcal{L}(\theta), \quad \mathcal{L}(\theta) := -\log P(\mathcal{D}, \theta). \tag{6}$$

Note that we can interpret term $-\log P(\theta)$ as the regularization term; choosing the other prior distribution yields the other type of regularization.

4.4 Majorization Minimization Algorithm

For minimizing objective function Eq. (6), we use a majorization-minimization (MM) algorithm [Hunter and Lange, 2004; De Leeuw, 1994] that can guarantee that the objective function is improved in each iteration.

The MM scheme indirectly minimizes objective $\mathcal{L}$ by using the majorizing function $\mathcal{G}$ which is defined as

$$\mathcal{G}(\theta, \boldsymbol{Z}) = -\sum\nolimits_{i,j,k}(M_{ijk} + N_{ij}z_{ijk})\log\{f(k|\eta_i)(\boldsymbol{P}_\nu^k)_{ij}\} + \sum\nolimits_{i,j,k} N_{ij}z_{ijk}\log z_{ijk} - \log P(\theta)$$

where $\boldsymbol{Z} = \{z_{ijk}\}$ is an auxiliary variable satisfying $z_{ijk} \geq 0$ $(\forall(i,j,k))$, $\sum_k z_{ijk} = 1$ $(\forall(i,j))$. Majorizing function $\mathcal{G}$ has the following two properties:

$$1.\ \mathcal{L}(\theta) \leq \mathcal{G}(\theta, \boldsymbol{Z}), \quad 2.\ \mathcal{L}(\theta) = \min_{\boldsymbol{Z}} \mathcal{G}(\theta, \boldsymbol{Z}). \tag{7}$$

Note that the equality holds if and only if

$$z_{ijk} = f(k|\eta_i)(\boldsymbol{P}_\nu^k)_{ij}\big/\big\{\sum\nolimits_{k'} f(k'|\eta_i)(\boldsymbol{P}_\nu^{k'})_{ij}\big\}. \tag{8}$$

MM uses the following 2 step procedures to minimize majorizing function $\mathcal{G}$:

1. Minimize $\mathcal{G}$ w.r.t. θ, 2. Minimize $\mathcal{G}$ w.r.t. $\boldsymbol{Z}$.

Parameter θ is updated in the following manner.

Update of $\boldsymbol{\eta}$. Let ∇_θ be the partial derivative operator w.r.t. θ. Setting the partial derivative w.r.t. η_i equal to zero, $\nabla_{\eta_i}\mathcal{G}(\theta, \boldsymbol{Z}) = 0$, yields

$$\mu_i = \frac{\xi_0 + \sum_{j,k}(M_{ijk} + N_{ij}z_{ijk})T(k)}{\mu_0 + \sum_{j,k} M_{ijk} + \sum_j N_{ij}}, \tag{9}$$

where μ_i is the mean value parameter of $f(k|\eta_i)$. Thus, parameter η_i is obtained by converting μ_i.

[2] If no such information is available, the term related to the feature and parameters $\boldsymbol{w}$ can be dropped from the model. Note that the tabular model is a special case of the log-linear model since the model without feature vector and parameter $\boldsymbol{w}$ is equivalent to the tabular model if we define $p_{ij} = \exp\{v_{ij}\}/\sum_{j'\in\Gamma_i}\exp\{v_{ij'}\}$.

[3] We set $\xi_0 = \mu_0 = \alpha_0 = 1.0$ in the experiments.

[4] Time step s does not correspond to t, the time step of the underlying Markov chain, since the number of (1-step) transitions conducted up to step s' is given by $\sum_{s=0}^{s'} K_s$.

Algorithm 1 Majorization Minimization Algorithm for Markov Chain Recovery from Multi-step Transition Data

Input: $\mathcal{D}$: multi-step transition data, Γ: adjacency information (optional), α_0, ξ_0, μ_0: hyperparameters
Output: $\hat{\boldsymbol{\theta}} = \{\hat{\eta}, \hat{\lambda}\}$: estimated parameters
1: Initialize $\boldsymbol{\theta}$.
2: **repeat**
3: Update $\boldsymbol{Z}$ following Eq. (8) or (12).
4: Update $\boldsymbol{\nu}$ by numerical optimization.
5: Update $\boldsymbol{\eta}$ following Eq. (9) or (13).
6: **until** A stopping condition is met

Update of $\boldsymbol{\nu}$. To update $\boldsymbol{\nu}$ we need some numerical optimization techniques such as gradient descent. Gradient descent based algorithms update the parameter at optimization step ℓ as follows, $\nu_{\ell+1} \leftarrow \nu_\ell - \gamma_\ell \boldsymbol{G}_\ell^{-1} \nabla_\nu \mathcal{G}(\theta_\ell, \boldsymbol{Z})$, where γ_ℓ is the learning rate, $\boldsymbol{G}_\ell$ is the identity matrix for gradient descent, and Hessian is used as the Newton method. ∇_ν is the partial derivative operator w.r.t. ν. The partial derivative can be computed as

$$\nabla_\nu \mathcal{G}(\theta_\ell, \boldsymbol{Z}) = -\sum\nolimits_{i,j,k} (M_{ijk} + N_{ij} z_{ijk}) \frac{\nabla_\nu (\boldsymbol{P}_\nu^k)_{ij}}{(\boldsymbol{P}_\nu^k)_{ij}} + \alpha_0 \nu. \tag{10}$$

The experiments described later use the L-BFGS method [Liu and Nocedal, 1989]. The optimization process is summarized in Alg. 1.

We provide here the convergence property of the proposed MM algorithm. To ensure the generality of the analysis to cover the use of various numerical optimization techniques for updating ν, we make the following assumption, similar to the generalized EM algorithm [Dempster *et al.*, 1977].

Assumption 1. *Let ν_{old} and ν_{new} be parameter ν before and after the update by some (numerical) optimization, respectively. Then, $\mathcal{G}((\eta, \nu_{new}), \boldsymbol{Z}) \leq \mathcal{G}((\eta, \nu_{old}), \boldsymbol{Z})$.*

Under assumption 1, the following theorem holds.

Theorem 3. *The proposed MM algorithm makes objective $\mathcal{L}(\theta)$ monotonically decreasing. The value is invariant if and only if θ is at a stationary point.*

Proof. (Theorem 3) Let us denote the parameters and the auxiliary variables that satisfy $\mathcal{L}(\theta) = \mathcal{G}(\theta, \boldsymbol{Z})$ as $\theta_{old} = (\eta_{old}, \nu_{old})$ and $\boldsymbol{Z}_{old}$, respectively. We also denote η after the first step of the optimization given by Eq. (9) as η_{new} and $\boldsymbol{Z}$ after the second step given by Eq. (8) as $\boldsymbol{Z}_{new}$. Since $A(\eta)$ is convex, $\mathcal{G}$ is also convex w.r.t. η; $\mathcal{G}((\eta_{new}, \nu), \boldsymbol{Z}) \leq \mathcal{G}((\eta, \nu), \boldsymbol{Z})$ holds $(\forall \eta)$. From the property of $\boldsymbol{Z}$ shown in Eq. (7), $\mathcal{G}(\theta, \boldsymbol{Z}_{new}) \leq \mathcal{G}(\theta, \boldsymbol{Z})$ holds $(\forall \boldsymbol{Z})$. Using Assumption 1, $\mathcal{L}(\theta_{old}) = \mathcal{G}(\theta_{old}, \boldsymbol{Z}_{old}) \geq \mathcal{G}((\eta_{new}, \nu_{old}), \boldsymbol{Z}_{old}) \geq \mathcal{G}((\eta_{new}, \nu_{new}), \boldsymbol{Z}_{old}) \geq \mathcal{G}((\eta_{new}, \nu_{new}), \boldsymbol{Z}_{new}) = \mathcal{L}(\theta_{new})$. □

4.5 Handling Infinite Mixtures

This section shows how to handle infinite mixtures. The key is the use of approximation for computing the infinite summation in $P(X'_{s+1}|X'_s)$ (Eq. (5)) by adopting the following mild assumption.

Assumption 2. *Markov chain $M_{\boldsymbol{\nu}}$ is irreducible and aperiodic for any parameter ν.*

This assumption allows us to use Theorem 2. By using (sufficiently large) truncation level K_{tr}, we can approximate $P(X'_{s+1}|X'_s)$ by $\tilde{P}(X'_{s+1}|X'_s)$ as follows:

$$\begin{aligned} &\tilde{P}(X'_{s+1} = j | X'_s = i) \\ &\quad = \sum\nolimits_{k=1}^{K_{tr}} f(k|\eta_i) (\boldsymbol{P}_\nu^k)_{ij} + \{1 - F(K_{tr}|\eta_i)\} (\boldsymbol{\pi}_\nu)_j, \end{aligned} \tag{11}$$

where $\boldsymbol{\pi}_\nu$ is the steady state probability of Markov chain $M_{\boldsymbol{\nu}}$. Thus infinite mixture models are handled by using $K = K_{tr} + 1$ components, where the final component has mixing ratio $1 - F(K_{tr}|\eta_i)$ and steady state probability $\boldsymbol{\pi}_\nu$. Under Assumption 2, the following theorem holds.

Theorem 4. *Given that Assumption 2 holds, there exist constants $\alpha \in (0, 1)$ and $C > 0$ such that $\max_{i \in \mathcal{X}} \|\tilde{P}(X'_{s+1}|X'_s = i) - P(X'_{s+1}|X'_s = i)\|_{TV} \leq C\alpha^{K_{tr}}$, where $\|\cdot\|_{TV}$ is the total variation distance* [5] .

Then, the infinite mixture model of Eq. (5) is well approximated by the approximated model of Eq. (11) with sufficiently large K_{tr}. Proof is shown by using the following theorem.

Theorem 5. [Levin and Peres, 2017] *If P is irreducible and aperiodic, with stationary distribution π, there exist constants $\beta \in (0, 1)$ and $D > 0$ such that $\max_{x \in \mathcal{X}} \|P^k(\cdot|x) - \pi\|_{TV} \leq D\beta^k$.*

Proof. (Theorem 4) Since f is a discrete distribution, there exists a finite value $f_{\max} := \max_{k,i} f(k|\eta_i)$.

$$\begin{aligned} &\|\tilde{P}(x_{s+1}|x_s = i) - P(x_{s+1}|x_s = i)\|_{TV} \\ &= \frac{1}{2} \sum\nolimits_j \sum\nolimits_{k'=K_{tr}+1}^{\infty} f(k'|\eta_i) |(\boldsymbol{P}_\nu^{k'})_{ij} - (\boldsymbol{\pi}_\nu)_j| \\ &\leq \frac{1}{2} \sum\nolimits_j \sum\nolimits_{k'=K_{tr}+1}^{\infty} f_{\max} |(\boldsymbol{P}_\nu^{k'})_{ij} - (\boldsymbol{\pi}_\nu)_j| \\ &\leq \sum\nolimits_{k'=K_{tr}+1}^{\infty} f_{\max} D\beta^{k'} = (f_{\max} D\beta^{K_{tr}+1})/(1-\beta). \end{aligned}$$

Setting $C = f_{\max} D\beta/(1-\beta)$ and $\alpha = \beta$ completes the proof. □

The algorithm for (approximate) infinite mixtures is derived in an analogous manner to the finite model.

Update of $\boldsymbol{Z}$. Similar to Eq. (8),

$$z_{ijk} = \begin{cases} \frac{f(k|\eta_i)(\boldsymbol{P}_\nu^k)_{ij}}{\sum_{k'=1}^{K_{tr}} f(k'|\eta_i)(\boldsymbol{P}_\nu^{k'})_{ij} + \{1 - F(K_{tr}|\eta_i)\}(\boldsymbol{\pi}_\nu)_j} & (\text{if } k \leq K_{tr}) \\ \frac{\{1 - F(K_{tr}|\eta_i)\}(\boldsymbol{\pi}_\nu)_j}{\sum_{k'=1}^{K_{tr}} f(k'|\eta_i)(\boldsymbol{P}_\nu^{k'})_{ij} + \{1 - F(K_{tr}|\eta_i)\}(\boldsymbol{\pi}_\nu)_j} & (\text{otherwise}) \end{cases}. \tag{12}$$

[5] The total variation (TV) distance between probability distribution μ and ν on $\mathcal{X}$ is formally defined as $\|\mu - \nu\|_{TV} = \max_{\mathcal{A} \subset \mathcal{X}} |\mu(\mathcal{A}) - \nu(\mathcal{A})|$. From Proposition 4.2 in [Levin and Peres, 2017], TV is also computed as $\|\mu - \nu\|_{TV} = 1/2 \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)|$.

Update of $\boldsymbol{\eta}$. Similar to Eq. (9),

$$\mu_i = \Big[\xi_0 + \sum\nolimits_j \Big\{\sum\nolimits_{k=1}^{K_{tr}} (M_{ijk} + N_{ij} z_{ijk}) T(k) + (M_{ijK} + N_{ij} z_{ijK}) \mathbb{E}_{f_{tr}(x|\eta_i, K_{tr}, \infty)}[T(x)]\Big\}\Big] \Big/ \Big[\mu_0 + \sum\nolimits_{j,k} M_{ijk} + \sum\nolimits_j N_{ij}\Big]. \quad (13)$$

where $f_{tr}(x|\eta, a, b)$ is the *truncated* distribution whose support lies in the range $(a, b]$ [6]and $\mathbb{E}_{f_{tr}}$ is the expectation of the distribution

Update of $\boldsymbol{\nu}$. Similar to Eq. (10),

$$\nabla_\nu \mathcal{G}(\theta_\ell, \boldsymbol{Z}) = -\sum\nolimits_{i,j} \Big\{\sum\nolimits_{k=1}^{K_{tr}} (M_{ijk} + N_{ij} z_{ijk}) \frac{\nabla_\nu (\boldsymbol{P}_\nu^k)_{ij}}{(\boldsymbol{P}_\nu^k)_{ij}} + (M_{ijK} + N_{ij} z_{ijK}) \frac{\nabla_\nu (\boldsymbol{\pi}_\nu)_j}{(\boldsymbol{\pi}_\nu)_j}\Big\} + \alpha_0 \nu.$$

The partial derivative of steady state $\boldsymbol{\pi}_\nu$ can be computed as [7] $\nabla_\nu \boldsymbol{\pi}_\nu = \boldsymbol{\pi}_\nu \nabla_\nu \boldsymbol{P}_\nu (\boldsymbol{I} - \boldsymbol{P}_\nu + \mathbf{1}^T \mathbf{1})^{-1}$, where $\mathbf{1}$ is a row vector whose elements are all ones.

5 Experiment

This section confirms that the proposed method well handles both unlabeled and labeled multi-step transition data. To evaluate how well the proposed method can recover 1-step transition probability, we use test data given in the form of complete transition data (1-step transition data), see Fig. 1.

5.1 Setting

Synthetic data. In the synthetic data experiment, we set the number of states to 15 and randomly generated chain edges following [Morimura *et al.*, 2009]. The true transition probability was set using the log-linear model where parameters ν^* and feature ϕ were generated using a standard normal distribution. We also added symmetric Dirichlet noise with parameter of 0.3 to the transition probability by taking the weight sum with $\beta = 0.1$ for the noise term and $1 - \beta = 0.9$ for the transition probability. The true label probability was set using a Poisson distribution whose parameters $\{\lambda_i^*\}$ were generated by a gamma distribution with shape $= 2.0$, scale $= 1.0$ [8]. We generated training and test data sets by generating episodes (sequences of states from initial state) with 20 steps in common. The number of episodes used in generating the test data were 1000. We prepared 5 sets of training and test data.

Real data. In the real data experiment, we used car probe data provided by NAVITIME JAPAN Co, Ltd. The dataset is a collection of GPS trajectories of individuals who used a car navigation application in the greater Tokyo area, Japan. We divided the region using an approximately 5km $\times$ 5km grid mesh; the total number of mesh cells was approximately 150. We used the data recorded during the period between 2015.4.13 to 2015.4.17 (5 working days in total) in the morning (6:00 am to 10:59 am). The number of unique users per day was, on average, approximately 8000. The data were made by converting the GPS trajectories into sequences of visited mesh cells (states). We excluded the states that appeared fewer than 20 times per day on average and used only episodes containing more than 2 steps. Here we did not use feature ϕ. Since the data are complete transition data, training data were made by randomly extracting the visited state following the generative process using ZTP, whose parameters are set via a gamma distribution analogous to the synthetic data. Training and test data were made from the logs of one day and that of the next day, respectively. We ran 4 trials using 4 sets of training and test data.

Evaluation Measure. As the evaluation metric, we used the negative test log likelihood computed using the test data given in the form of complete transition data. The negative test log likelihood is defined as $(1/\mathcal{T}_{test}) \sum_{i,j \in \mathcal{X}} -\tilde{n}_{ij} \log \hat{p}_{ij}$, where $\mathcal{T}_{test}$ is the number of total transitions and $\tilde{n}_{ij}$ is the number of transitions from state i to state j in the test. $\hat{p}_{ij}$ is the estimated 1-step transition probability. The results of proposed method using log-linear model for transition probability and Poisson and ZTP for label probability with K=10 are reported [9]. A lower value indicates 1-step transition probability is precisely recovered.

Baseline Methods. Since our method is the first method that can estimate MC parameters from multi-step transition data, we decided to compare the proposed method with the simple baseline methods that use only labeled multi-step transition data. Note that existing methods including HMM and Bayesian nets are not designed to recover 1-step transition probability. Using the log-linear model similar to the proposed method, the parameter of the baseline, $\hat{\nu}_{\text{base}} = \arg\min_\nu \big\{-\sum_{i,j,k} M_{ijk} \log\{(\boldsymbol{P}_\nu^k)_{ij}\} - \log P(\nu)\big\}$, was obtained by numerical optimization. A comparison with this baseline also confirms the usefulness of the proposed method that can handle unlabeled data.

5.2 Results

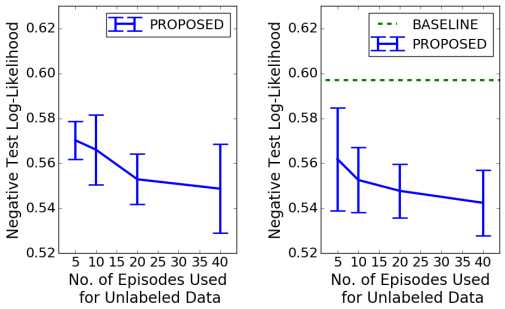
Quantitative Result. Figure 3 shows the results of the synthetic data experiment. We can confirm that the performance of the proposed method improves as the amount of unlabeled multi-step transition data increases for both cases wherein the labeled multi-step transition data is and is not used; this verifies that the proposed method well handles unlabeled multi-step transition data. From Fig. 3b, it is also confirmed that the proposed method outperforms the baseline when challenged

[6]The truncated distribution is often used in survival analysis and is formally defined as $f_{tr}(x|\eta, a, b) = \{F(b|\eta) - F(a|\eta)\}^{-1} f(x|\eta)$ if x in $(a, b]$ and $f_{tr}(x|\eta, a, b) = 0$ otherwise.

[7]From $\boldsymbol{\pi}_\nu = \mathbf{1}(\boldsymbol{I} - \boldsymbol{P}_\nu + \mathbf{1}^T \mathbf{1})^{-1}$ as shown in Proposition (2.14.1) [Resnick, 2002].

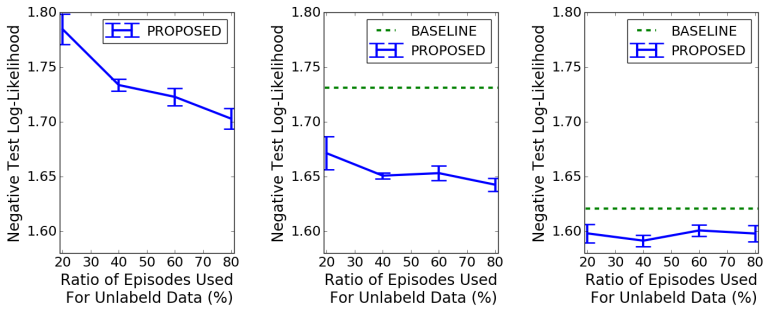
[8]λ_i has mean and variance of 2.0. The 95th percentile of Poisson distribution with $\lambda_i = 4.0$ lies between 7 and 8.

[9]In this experiment, "sensitivity of the truncated level K" described later, shows that stable performance is demonstrated when K is set to 10 or more.



(a) w/o Labeled Data (b) w/ Labeled Data

Figure 3: Results of synthetic data experiment in which the number of episodes used for unlabeled multi-step transition data was varied. The performances (a) without and (b) with labeled multi-step transition data made from 3 episodes are shown. Lower values are better.



(a) w/o Labeled Data (b) w/ Labeled Data (5%) (c) w/ Labeled Data (10%)

Figure 4: Results of real data experiment in which the ratio of episodes used for unlabeled multi-step transition data was varied. The performances (a) without and (b)(c) with labeled multi-step transition data are shown. The labeled multi-step transition data were made from (b) 5% and (c) 10% of the episodes. Lower values are better.

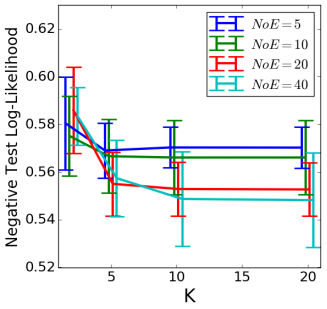
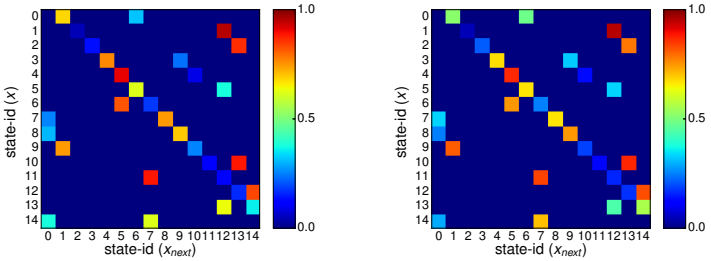


Figure 5: Result of varying truncation level K in the proposed method. NoE is the number of episodes used for unlabeled data.



(a) True (b) Proposed

Figure 6: (a) True and estimated transition probability yielded by (b) Proposed method in synthetic data experiment where the number of episodes used for unlabeled multi-step transition data is 40.

with labeled multi-step transition data. This verifies the usefulness of our method in its ability to handle both labeled and unlabeled data.

Figure 4 shows that similar results were obtained in the real data experiment. Moreover, by comparing the performance of the proposed method with that of the baseline in Fig. 4bc, the degree of improvement is large when the amount of labeled multi-step data is small (5%); this implies that the proposed method is more effective when the amount of labeled data is limited. We can also confirm that the performance of the proposed method improves as the amount of labeled data increases. These results confirm the effectiveness of our method for recovering 1 step transition probability from multi-step transition data.

Sensitivity of the truncation level K. Figure 5 shows the results of the synthetic data experiment when the truncation level (or the number of components) of the proposed method, K, was varied. It shows that the performance improves as the value of K increases, and converges when K is approximately 10. This demonstrate that the proposed method provides stable performance when K is set to 10 or more.

Qualitative Result. Figure 6 illustrates the true and estimated 1-step transition probability from the synthetic data experiment. It confirms that the estimated probability by the proposed method is closer to the true probability. This also implies that the proposed method can accurately estimate MC parameters.

6 Conclusion

This paper proposed a new model and algorithm for estimating MC parameters from labeled and unlabeled multi-step transition data for recovering 1-step transition probability from incomplete transition data. We proved the convergence of the algorithm and introduced an approximation scheme for handling infinite mixture models that has strong theoretical support. We also evaluated the effectiveness of the proposed method using synthetic and real data. Remaining future work is to theoretically analyze the performance of the proposed method. Developing Bayesian algorithms is also a promising direction.

References

[Akagi *et al.*, 2018] Yasunori Akagi, Takuya Nishimura, Takeshi Kurashima, and Hiroyuki Toda. A fast and accurate method for estimating people flow from spatiotemporal population data. In *International Joint Conferences on Artificial Intelligence*, pages 3293–3300, 2018.

[Amari and Nagaoka, 2007] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

[Antoniak, 1974] Charles E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974.

[Basu *et al.*, 2002] Sugato Basu, Arindam Banerjee, and Raymond Mooney. Semi-supervised clustering by seeding. In *In Proceedings of 19th International Conference on Machine Learning*, 2002.

[Billingsley, 1961] Patrick Billingsley. *Statistical inference for Markov processes*. University of Chicago Press, 1961.

[Blei and Jordan, 2006] David M Blei and Michael I Jordan. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[Crisostomi *et al.*, 2011] Emanuele Crisostomi, Stephen Kirkland, and Robert Shorten. A google-like model of road network dynamics and its application to regulation and control. *International Journal of Control*, 84(3):633–651, 2011.

[De Leeuw, 1994] Jan De Leeuw. Block-relaxation algorithms in statistics. In *Information systems and data analysis*, pages 308–324. Springer, 1994.

[Dempster *et al.*, 1977] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[Durrett, 1999] Richard Durrett. *Essentials of stochastic processes*, volume 1. Springer, 1999.

[Fan *et al.*, 2015] Zipei Fan, Xuan Song, Ryosuke Shibasaki, and Ryutaro Adachi. Citymomentum: an online approach for crowd behavior prediction at a citywide level. In *International Joint Conference on Pervasive and Ubiquitous Computing*, pages 559–569, 2015.

[Ferguson, 1973] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[Frydman and Schuermann, 2008] Halina Frydman and Til Schuermann. Credit rating dynamics and markov mixture models. *Journal of Banking & Finance*, 32(6):1062–1075, 2008.

[Frydman, 1984] Halina Frydman. Maximum likelihood estimation in the mover-stayer model. *Journal of the American Statistical Association*, 79(387):632–638, 1984.

[Ghahramani and Jordan, 1994] Zoubin Ghahramani and Michael I Jordan. Supervised learning from incomplete data via an em approach. In *Advances in neural information processing systems*, pages 120–127, 1994.

[Goodman, 1961] Leo A Goodman. Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*, 56(296):841–868, 1961.

[Gupta *et al.*, 2016] Rishi Gupta, Ravi Kumar, and Sergei Vassilvitskii. On mixtures of markov chains. In *Advances in neural information processing systems*, pages 3441–3449, 2016.

[Hunter and Lange, 2004] David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

[Iwata *et al.*, 2017] Tomoharu Iwata, Hitoshi Shimizu, Futoshi Naya, and Naonori Ueda. Estimating people flow from spatiotemporal population data via collective graphical mixture models. *ACM Transactions on Spatial Algorithms and Systems*, 3(1):2, 2017.

[Kumar *et al.*, 2015] Ravi Kumar, Andrew Tomkins, Sergei Vassilvitskii, and Erik Vee. Inverting a steady-state. In *International Conference on Web Search and Data Mining*, pages 359–368, 2015.

[Levin and Peres, 2017] David A Levin and Yuval Peres. *Markov chains and mixing times*. American Mathematical Soc., 2017.

[Liu and Nocedal, 1989] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[Morimura *et al.*, 2009] Tetsuro Morimura, Eiji Uchibe, Junichiro Yoshimoto, and Kenji Doya. A generalized natural actor-critic algorithm. In *Advances in Neural Information Processing Systems*, pages 1312–1320, 2009.

[Morimura *et al.*, 2013] Tetsuro Morimura, Takayuki Osogami, and Tsuyoshi Idé. Solving inverse problem of markov chain with partial observations. In *Advances in Neural Information Processing Systems*, pages 1655–1663, 2013.

[Neal, 2000] Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[Neuts, 1981] Marcel F Neuts. *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Johns Hopkins University Press, 1981.

[Nigam *et al.*, 2000] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134, 2000.

[Pfeifer and Carraway, 2000] Phillip E Pfeifer and Robert L Carraway. Modeling customer relationships as markov chains. *Journal of interactive marketing*, 14(2):43–55, 2000.

[Resnick, 2002] Sidney I Resnick. *Adventures in stochastic processes*. Springer Science & Business Media, 2002.

[Teh, 2010] Yee Whye Teh. Dirichlet process. *Encyclopedia of machine learning*, pages 280–287, 2010.