Phase-type distributions

Fernando Zepeda

TU Berlin - Lufthansa Systems

September 27, 2016

Outline

Introduction

Some data do not follow necessarily the form of any standard distribution (unconventional forms):

- Probability distributions may posses any shape
- Rugged and multimodal data distributions

The **phase-type (PH)** family of distributions have some useful properties in fitting this kind of data

- They are dense in the field of all positive-valued distributions $[0, \infty)$
- Thus, they can be used to approximate any positive-valued distribution Bladt. et al. (2003)
- Markovian properties make them analytically flexible

Motivation

Goal: Estimate

- Buying probabilities (and ticketing probabilities)
- Inter-arrival time of bookings

by fitting PH distributions to them given booking historical data

Markov Chains

Let **S** denote a countable set of states, and let $\{X(t)\}_{t\geq 0}^{\infty}$ be a stochastic process with state space **S**

- Let $n \in \mathbb{N}$ be the size of **S**

$\{X(t)\}_{t\geq 0}^{\infty}$ is a continuous-time Markov chain **(CTMC)**, if it is characterized by the *Markov property*:

$$\boxed{\begin{array}{c} Prob(X(t_{k+1}) = x_{k+1} | X(t_k) = x_k, \ldots, X(t_0) = x_0) = \\ Prob(X(t_{k+1}) = x_{k+1} | X(t_k) = x_k) \end{array}}$$

for any $t_{k+1} \geq t_k \geq \ldots \geq t_1 \geq t_0 \geq 0$ and $x_l \in \mathbf{S}$

The transition between states x are defined by the values $p_t(i,j)$ which are contained in a matrix with transition probabilities $\mathbf{P_t}$

The state probabilities at time t are denoted by

$$p_t(j) = Prob(X(t) = j), j \in \mathbf{S} \text{ with } \sum_j p_t(j) = 1$$

$$\pi(0) = [p_0(1), p_0(2), \ldots, p_0(n)]\text{: initial probability vector}$$

Random times between states transitions $V_1, V_2, \ldots, V_n$ are exponentially distributed random variables with parameter $\lambda(i)$ for state i

V_i: holding time in state i with distribution function:

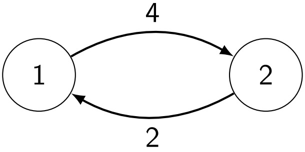
$$Prob(V_i \leq t) = 1 - e^{-\lambda(i)t}, t \geq 0$$

The probabilistic behavior if the CTMC can be summarize in terms of its infinitesimal generator: a $n \times n$ matrix $\mathbf{Q}$ with components:

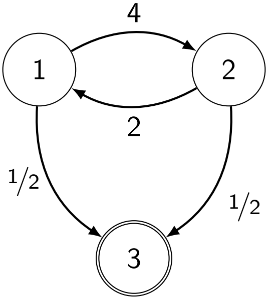
$$\mathbf{Q(i,j)} = \begin{cases} -\lambda(i) & \text{if } i = j, \\ \lambda(i,j) & \text{if } i \neq j. \end{cases}$$

- Since $\lambda(i) \geq 0$, it follows $\mathbf{Q}(i,i) \leq 0$: *diagonal elements are non-positive*
- Transition to j feasible in state i then $\mathbf{Q}(i,j) > 0$, otherwise $\mathbf{Q}(i,j) = 0$: *non-diagonal elements must be non-negative*
- $\sum_j \mathbf{Q}(i,j) = 0$

Absorbing Markov Chains



(a) CTMC with transient states only



(b) CTMC with one absorbing state

Figure 1: Different Markov chains and their transition representation

Absorbing Markov Chains

Some definitions

A state j is reachable from another i when we have

$$p_t(i,j) = Prob(X(t+u) = j | X(u) = i) > 0$$

for some t

Def.
States i and j can communicate with each other if i is reachable from j and viceversa.

Let $\mathbf{C} \subset \mathbf{S}$. If all states in $\mathbf{C}$ communicate, we call it communicating set

Def.
A subset $\mathbf{C}$ of the state space $\mathbf{S}$ is said to be closed if $\mathbf{P}(i,j) = 0$ for any $i \in \mathbf{C}, j \notin \mathbf{C}$.

If $\mathbf{C}$ consists of a single state, say i, then i is said to be an absorbing state.

It holds for i that $\mathbf{P}(i, i) = 1$: *a process can never leave a closed set after entering it*

Def.

A state $i \in \mathbf{S}$ is transient, if the probability of returning to i after leaving it is less than 1.

PH distributions and absorbing Markov chains

If every state in a Markov chain is either absorbing or transient, then the Markov chain is called *absorbing Markov chain*

Particularly interesting case: absorbing Markov chains with one single absorbing state

Assumption: state space $\mathbf{S}$ of the continuous time absorbing Markov process $\{X(t)\}_{t \geq 0}^{\infty}$ is finite and contains the set of **transient** states $\mathbf{S}_T = \{1, \ldots, n\}$ and a single absorbing state $n + 1$

$$\mathbf{Q} = \begin{pmatrix} \mathbf{D_0} & \mathbf{d_1} \\ \mathbf{0} & 0 \end{pmatrix}$$

is the generator matrix, where:

- $\mathbf{D_0}$: describes transitions between transient states
- $\mathbf{d_1}$: transition from transient states to the absorbing state
- row vector $\mathbf{0}$: since no transition from the absorbing state to the transient states

$$
\mathbf{Q} = \begin{array}{cc} n \times n & n \times 1 \\ \left(\begin{array}{cc} \boxed{\mathbf{D_0}} & \boxed{\mathbf{d_1}} \\ \boxed{\mathbf{0}} & \boxed{0} \end{array} \right) & \\ 1 \times n & 1 \times 1 \end{array}
$$

is the generator matrix, where:

- $\mathbf{D_0}$: describes transitions between transient states
- $\mathbf{d_1}$: transition from transient states to the absorbing state
- row vector $\mathbf{0}$: since no transition from the absorbing state to the transient states

Since the states are transient, implies:

$$\lim_{t \to \infty} Prob(X(t) < n + 1) = 0,$$

i.e. the absorption occurs with probability 1.

$\mathbf{D_0}$ is invertible and matrix $(-\mathbf{D_0})^{-1}$ is the fundamental matrix of the absorbing CTMC

Def. Phase-type distribution (PHD)

A PHD is defined as the distribution of the lifetime X or the time to absorption from the set of transient states $\mathbf{S_T}$ of an absorbing continuous Markov process $\{X(t)\}_{t \geq 0}^{\infty}$

Phase Type Distributions (PHD)

The Markov process starts in an arbitrary state from $\mathbf{S} = \mathbf{S_T} \cup \mathbf{S_A}$, so the vector

$$\pi = [\pi(1), \ldots, \pi(n)]$$

represents the inital probabilites for transient states and $\pi(n+1)$ gives the probability for a direct start in the absorbing (assumption: $\pi(n+1) = 0$)

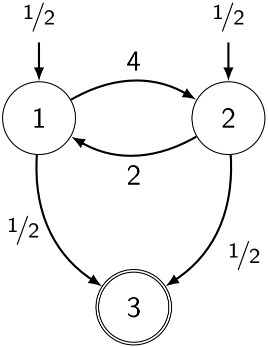
Phase Type Distributions (PHD)

Simple example

$$\mathbf{S} = \mathbf{S_T} \cup \mathbf{S_A} = \{1, 2\} \cup \{3\}$$

$$\mathbf{Q} = \left(\begin{array}{cc|c} -4.5 & 4 & 0.5 \\ 2 & -2.5 & 0.5 \\ \hline 0 & 0 & 0 \end{array}\right)$$

$$\pi = [\frac{1}{2}, \frac{1}{2}]$$



PHD Representation

We can say now that the random variable X describing the time until absorption is of phase-type with the vector-matrix tuple representation:

$$(\pi, \mathbf{D_0})$$

since the vectors $\mathbf{d_1}$ and the value $\pi(n+1)$ are implicitly given by the matrix $\mathbf{D_0}$

The size of the PHD is the size of the sub-generator matrix $\mathbf{D_0}$, n

Analytical Properties of PHD

$$F(x) = \mathbb{1} - \pi e^{\mathbf{D_0} x}\mathbb{1}, \text{ for } x \geq 0 \text{ (c.d.f.)}$$

$$f(x) = \pi e^{\mathbf{D_0} x}\mathbf{d_1}, \text{ for } x \geq 0 \text{ (p.d.f.)}$$

$$\mu_i = E[X^i] = i!\pi(-\mathbf{D_0^{-1}})^i\mathbb{1}, \text{ PHD } i\text{th moment}$$

Note 1: $\mathbf{M} = -\mathbf{D_0^{-1}}$ is the moment matrix

$$\rightarrow \mu_i = E[X^i] = i!\pi\mathbf{M}^i\mathbb{1}$$

Note 2: $-\mathbf{D_0^{-1}}(i,j) = \mathbf{M}(i,j)$ is the expected total time spent in phase j before absorption, given that initial phase is i

PH Classes

Based on the structure of the underlying Markov chain several classes of PH distributions can be distinguished

- Structure of PH representation often has an impact on its application
- Some structures allow more efficient solutions

Most important distinction:

1. Acyclic phase-type distributions (APH)
2. General phase-type distributions

PH Classes

Based on the structure of the underlying Markov chain several classes of PH distributions can be distinguished

- Structure of PH representation often has an impact on its application
- Some structures allow more efficient solutions

Most important distinction:

1. Acyclic phase-type distributions (APH)

2. General phase-type distributions

PH Classes

Based on the structure of the underlying Markov chain several classes of PH distributions can be distinguished

- Structure of PH representation often has an impact on its application
- Some structures allow more efficient solutions

Most important distinction:

1. Acyclic phase-type distributions (APH)
   - Hyper-Erlang distributions
   - Hyper-Exponential distributions
2. General phase-type distributions

Hyper-Erlang distributions (HEr PHD)

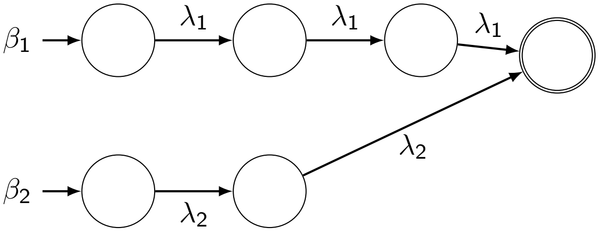


Figure 2: Hyper-Erlang distribution with two branches
$(\beta, m, \mathbf{b}, \lambda) = ([\beta_1, \beta_2], 2, [3, 2], [\lambda_1, \lambda_2])$

Size of the Hyper-Erlang distribution: $n = \mathbf{b}\mathbb{1}$

Special case: $(\beta, m, \mathbf{b}, \lambda) = ([\beta_1 = 1], 1, \mathbf{b}, \lambda_1) \rightarrow$ Erlang distr.

Hyper-Exponential distributions (HEx PHD)

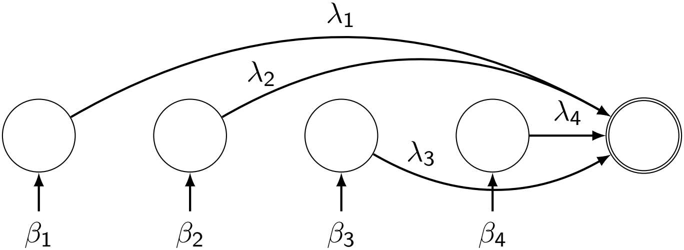


Figure 3: Hyper-Exponential distribution with four branches $(\beta, n, \lambda) = ([\beta_1, \beta_2, \beta_3, \beta_4], 4, [\lambda_1, \lambda_2, \lambda_3, \lambda_4])$

Special case: $(\beta, n, \lambda) = ([\beta = 1], 1, \lambda_1) \rightarrow$ Exponential distr.

Note: HEx are subclass of HEr when $\mathbf{b} = \mathbb{1}$ and $\mathbf{m} = n$

Canonical Representation for APH distributions

CF-1

Def.

The Canonical Form 1 (CF-1 form) is a bi-diagonal representation (β, Λ), where β initial probabilities and rates λ_i in Λ are in increasing order: $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$

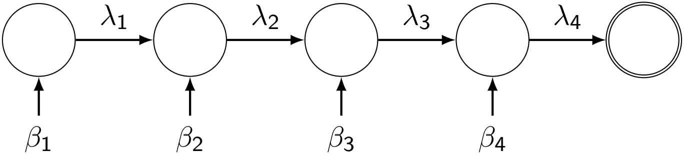


Figure 4: APH distribution in CF-1 form

Canonical Representation for APH distributions

Every acyclic phase type distribution (APH) with a Markovian representation of size n has an unique CF-1 representation of the same size.

- A general APH has $n-1$ initial probabilities and n^2 entries in the sub-generator matrix $\mathbf{D_0}$, i.e. the number of parameters is n^2+n-1
- In the CF-1 form $\mathbf{D_0}$ is an upper bi-diagonal matrix with $\lambda_{i,i} = -\lambda_{i+1,i}$, therefore the CF-1 form has $2n-1$ parameters

PH-Fitting Tools

Two general classes of algorithms

- Analytical: direct computation of parameters
  - Moment Matching: μ_1, μ_2, μ_3
- Statistical: maximum likelihood method (iterative process)
  - G-FIT: Hyper-Erlang distributions
  - PhFit: Acyclic PHD CF-1 form

Summary

- PHD can be used to approximate any positive-valued distribution
- They represent the distribution of the time to absorption of an absorbing continuous time Markov process $\{X(t)\}_{t\geq 0}^{\infty}$
- They are defined by the tuple $(\pi, \mathbf{D_0})$
- Two big families: Hyper-Erlang and Hyper-Exponential distributions
- Acyclic phase-type distributions deserve an special attention
- They can be reduced in a canonical form in order to estimate their parameters $(2n - 1)$
- Fitting techniques

References

Bucholtz, Kriege & Felko (2014). Input Modeling with Phase-Type Distributions and Markov Models: Theory and Applications

Cumani, A (1982). On the canonical representation of homogeneous Markov processes modelling failure-time distributions microelectronics and reliability 22. 582-602.

Reinicke, P & Horváth G. (2012). Phase-type distributions for realistic modeling in discrete-event simulation. *In Proceedings of the 5th International ICST Conference on Simulation Tools and Techniques* (pp.283-290). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).