Utrecht University

Utrecht University

Bachelor thesis

On the consistency of the landmark Aalen-Johansen estimator

Olov Schavemaker

Supervised by
Dr. Cristian Spitoni

June 10, 2019

Contents

Chapter 1

Introduction

Our thesis concerns the consistent estimation of the transition probabilities of non-Markov multi-state models with right censoring. We will first explain what this means and why this is useful without diving into the mathematical details. Thereafter, we shall summarize this thesis' contents.

We begin by explaining what a multi-state model is. Suppose that an individual (not necessarily an organism) can be in several states. Not only that, but it can, over time, transition between those states. A multi-state model is a mathematical model of such a system. How about a concrete example?

Consider a fragile clockwork that can be in the following two states: working and broken. When I bought it, it was working. However, fragile as it is, it soon broke. I got it repaired, so now it is working again, but, with how fragile it is, it is only a matter of time before it breaks again. Unfortunately, the cycle of breaking and repairing does not last forever; every time the clockwork breaks, it becomes harder to repair, and, at some point, it will be broken beyond repair.

Although the clockwork can only be in two states, the situation is already quite complicated. The more it breaks, the more fragile it becomes, making it harder to repair. Consequently, the probability of transitioning from broken to working dwindles, until it eventually reaches 0. Important to note is that the transition probability depends on its past (how many times it was dropped).

If the transition probabilities are not dependent on the past (and thus solely dependent on the present), we call the multi-state model Markov. As this thesis will illustrate, Markov models are much easier to work with than their non-Markov brethren. However, a lot of practical situations cannot be adequately modelled by Markov models, because many things depend on their past (as illustrated by the clockwork example).

But this is not where the clockwork's story ends. I actually bought it as a birthday present for my friend. In spite of that, it is still just as fragile as before. That is to say, the clockwork will continue to break and get repaired unchangingly after the birthday party. The thing is, after the birthday party, I will no longer be able observe what state it is in; I gave it to my friend after all.

If, after some point in time, the state of an individual can no longer be observed, then we say that right censoring has occurred. The individual might still transition between states afterward, it is just that we will never know. In the clockwork example, right censoring occured the moment I gave my clockwork to my friend.

Admittedly, our introduction of right censoring in the clockwork's example was somewhat artificial. Examples where right censoring not only occurs naturally, but that are actually practically relevant, are, among others, clinical trials. Patients in clinical trials can be modelled as individuals in a multi-state model with states describing the efficacy or safety of the treatment. Examples of states might be "treatment is (not) working" and "(no) side effects". Since no clinical trial lasts forever, but some patients could still start showing side effects after the trial ends (transition from "no side effects" to "side effects"), right censoring naturally occurs.

While one could consider left and interval censoring as well, we shall not do so, because right censoring is the most common in practice [see 1, section III.2's epigraph].

Let us repeat that our thesis concerns the consistent estimation of the transition probabilities of non-Markov multi-state models with right censoring. We want consistent estimators, because consistency guarantees that, if we have enough individuals, our estimators will be close to the actual value they are trying to estimate.

Now that we have explained what the problem is, let us address how we will tackle this problem.

In Chapter 2, we will introduce conventions that will be used in the remainder of the thesis.

In Chapter 3, we will give a brief overview of measure theoretic probability. We shall introduce only those concepts that we need in the remainder of the thesis.

Chapter 4 introduces and explores the so-called product integral. The product integral will be used extensively to define estimators. After defining the product integral, we shall show its relation to the Volterra equation (which is the reason why it is so useful, as will become clear in Chapter 6).

Chapter 5 finally formally introduces the multi-state model. After an argosy of definitions, including those of the so-called cumulative transition hazard, which will play a key role in estimating the transition probabilities, and the transition probability, we will touch upon right censoring as well. We will show that, without right censoring, estimating the transition probabilities consistently would be very easy. We shall also formulate the "independent censoring assumption", whose importance will become apparent in the next chapter.

But what do we do when we do have right censoring?

In Chapter 6, we will estimate transition probabilities in the Markov case (the easier case). First, using the independent censoring assumption, we can derive a consistent estimator for the cumulative transition hazard, namely the Nelson-Aalen estimator. Thereafter, we shall show that the product integral of the Nelson-Aalen estimator, which is dubbed the Aalen-Johansen estimator, is a consistent estimator of the transition probabilities. We can get a consistent estimator of the transition probability by taking the product integral of a consistent estimator for the cumulative transition hazard because the transition probability and the transition hazard are related by a Volterra equation.

Unfortunately, in the non-Markov case, the Aalen-Johansen estimator is no longer consistent. As such, in Chapter 7, we will explore a modification of the Aalen-Johansen estimator, which does turn out to be consistent in the non-Markov case. The resulting estimator is called the landmark Aalen-Johansen estimator. It will be the product integral of a modification of the Nelson-Aalen estimator. This chapter's conclusion and the main result of this thesis is, as promised by the title, a proof of the consistency of the landmark Aalen-Johansen estimator.

Then there iss the final chapter, Chapter 8, which gives a glimpse of directions of further research.

Finally, there are five appendices that prove certain results that are used throughout the thesis. We recommend to at least check out Appendix E, in which a novel proof of the important Duhamel's equation (one of the the main ingredient in the proof of consistency of the Aalen-Johansen estimator) is given.

Chapter 2

Conventions

Throughout the thesis, the following conventions will be adhered.

- All matrices will be real square matrices.
- Calligraphic letters will denote σ-algebras, matrix-valued functions or partitions.
- "integrable" means
  - "P-integrable" for random variables
  - "Lebesgue integrable" for functions defined on a subset of $\mathbb{R}$
- Square brackets may denote the Iverson bracket.

Definition 2.0.1. Let p be a logical proposition. Then the Iverson bracket

$$[\mathrm{p}] = \begin{cases} 1 & \text{if p is true} \\ 0 & \text{if p is false} \end{cases}$$

- "$\restriction$" denotes the restriction of a function.
- "(a_n)" denotes a sequence.
- "$\lim_n$" means "$\lim_{n\to\infty}$".
- "$\equiv$" means "by definition equal to".
- "$\wp(A)$" denotes the power set of the set A.
- Bold letters will denote matrices or matrix-valued functions.
- If $\mathbf{A}$ is a matrix, then

  $$[\mathbf{A}]_{ij} = A_{ij}$$

  is the ij-th entry of $\mathbf{A}$ (this convention also holds for letters other than "A").

- "$\mathbf{I}$" denotes the identity matrix. If we want to specify that we are talking about the $K \times K$ identity matrix, we write $\mathbf{I}_K$.
- A "!" atop a binary relation symbol will mean that an explanation of the binary relation has been given precedingly or will be given shortly.
- "#A" denotes the cardinality of the set A.
- "$\lesssim$" means "less than and approximately equal to".
- "$\oplus$" denotes the direct sum of matrices. That is,

$$\mathbf{A} \oplus \mathbf{B} = \begin{Bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{Bmatrix}$$

  where $\mathbf{0}$ denotes the zero matrix. One can easily verify that $\oplus$ distributes over regular matrix addition and matrix multiplication.
- "ae" atop a binary relation symbol will mean that the binary relation holds almost everywhere.
- And finally,

$$\mathrm{diag}(a_1, \ldots, a_n) = \begin{Bmatrix} a_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_n \end{Bmatrix}$$

Chapter 3

Measure theoretic probability

In this chapter, we will introduce parts of probability theory from a measure theoretic point of view. The aim of this chapter is to curtly cover the mathematical background needed for later chapters. First, we will introduce basic probabilistic concepts such as random variables and their expectation and variance. Secondly, we will talk about convergence in probability. Thirdly, and finally, we introduce stochastic processes and filtrations.

Throughout this chapter and the rest of the thesis, we assume basic familiarity with measure theory. More concretely, we assume that the reader is familiar with the results in Chapter 1 and 2, section 4.1 and 4.2, and Chapter 5 of [2].

The following definitions are mostly taken from Chapter 10 of [2].

A probability space $(\Omega, \mathcal{A}, \mathrm{P})$ is a measure space such that $\mathrm{P}(\Omega) = 1$. If $\mathrm{A} \in \mathcal{A}$, then $\mathrm{P}(\mathrm{A})$ is the probability of the event A. For $\mathrm{B} \in \mathcal{A}$, the conditional probability of A given B is defined by

$$\mathrm{P}(\mathrm{A} \mid \mathrm{B}) = \frac{\mathrm{P}(\mathrm{A} \cap \mathrm{B})}{\mathrm{P}(\mathrm{B})}$$

provided $\mathrm{P}(\mathrm{B}) \neq 0$.

A real-valued random variable on $(\Omega, \mathcal{A}, \mathrm{P})$ is a measurable function from $(\Omega, \mathcal{A}, \mathrm{P})$ to $\mathbb{R}$. Let $(\mathrm{S}, \mathcal{B})$ be a measurable space. An $(\mathrm{S}, \mathcal{B})$-valued random variable is a measurable function from $(\Omega, \mathcal{A}, \mathrm{P})$ to $(\mathrm{S}, \mathcal{B})$.

Henceforth the underlying probability space of a random variable will always tacitly be assumed to be $(\Omega, \mathcal{A}, \mathrm{P})$.

If a real-valued random variable X is integrable, then its expected value

$$\mathrm{E}(\mathrm{X}) = \int \mathrm{X}\, \mathrm{dP}$$

exists and is finite. We call $\mathrm{E}(\mathrm{X}^2)$ the second moment of X. If X has finite second moment, then it has finite expectation because $|\mathrm{X}| \leq \mathrm{X}^2 + 1$. In this case, we define the variance of X by

$$\mathrm{var}(\mathrm{X}) = \mathrm{E}((\mathrm{X} - \mathrm{E}(\mathrm{X}))^2) = \mathrm{E}(\mathrm{X}^2) - (\mathrm{E}(\mathrm{X}))^2$$

which clearly exists and is finite as well.

Let $\mathscr{C}$ be a sub-σ-algebra of $\mathscr{A}$. If X is an integrable real-valued random variable, then the conditional expectation of X given $\mathscr{C}$ is defined by the following Radon-Nikodym derivative.

$$\mathrm{E}(\mathrm{X} \mid \mathscr{C}) = \frac{\mathrm{d}\mu}{\mathrm{d}(\mathrm{P} \restriction \mathscr{C})}$$

with

$$\mu : \mathscr{C} \ni \mathrm{C} \mapsto \int_{\mathrm{C}} \mathrm{X}\, \mathrm{dP} \in \mathbb{R}$$

Let us verify that the Radon-Nikodym derivative above is well defined. That entails verifying that

- $\mathrm{P} \restriction \mathscr{C}$ is a σ-finite positive measure on $(\Omega, \mathscr{C})$
- μ is a finite signed measure on $(\Omega, \mathscr{C})$
- μ is absolutely continuous with respect to $\mathrm{P} \restriction \mathscr{C}$.

Clearly $\mathrm{P} \restriction \mathscr{C}$ is a (σ-)finite positive measure on $(\Omega, \mathscr{C})$ because P is finite positive measure on $(\Omega, \mathscr{A})$. By the linearity of the integral and Lebesgue's dominated convergence theorem, μ is a signed measure on $(\Omega, \mathscr{C})$. Since X is integrable, μ is finite.

$$\begin{gathered}\text{Let } \mathrm{C} \in \mathscr{C}. \text{ If } (\mathrm{P} \restriction \mathscr{C})(\mathrm{C}) = 0, \text{ then } \mathrm{P}(\mathrm{C}) = 0, \\ \text{so } \mathrm{X}[\cdot \in \mathrm{C}] \text{ vanishes P-almost everywhere and hence } \mu(\mathrm{C}) = 0.\end{gathered}$$

That is, μ is absolutely continuous with respect to $\mathrm{P} \restriction \mathscr{C}$. As such, the Radon-Nikodym derivative of μ with respect to $\mathrm{P} \restriction \mathscr{C}$ is well defined.

We can now define conditional probabilities over sub-σ-algebras as follows: the conditional probability of A given $\mathscr{C}$ is given by

$$\mathrm{P}(\mathrm{A} \mid \mathscr{C}) = \mathrm{E}(\mathrm{X} \mid \mathscr{C})$$

with $\mathrm{X}(\omega) = [\omega \in \mathrm{A}]$.

The expected value and variance defined in this way satisfy all their usual properties that you might be used to from an introductory probability course.

Let us now focus our attention on convergence in probability.

Definition 3.0.1. Let $(S, \|\cdot\|)$ be a normed space and $X, (X_n)$ be S-valued random variables. Here $\mathcal{B}$ (the σ-algebra on S) is the σ-algebra generated by the topology induced by the metric induced by $\|\cdot\|$. Then X_n converges in probability to X if

$$\lim_n P\Big(\|X_n - X\| > \epsilon\Big) = \lim_n P(\{\omega \in \Omega : \|X_n(\omega) - X(\omega)\| > \epsilon\}) = 0$$

holds for all $\epsilon \geq 0$. We write $X_n \xrightarrow{p} X$.

Remark 3.0.2. The definition above only makes sense if

$$A \equiv \{\omega \in \Omega : \|X_n(\omega) - X(\omega)\| > \epsilon\}$$

is measurable. By the reverse triangle inequality, $\|\cdot\|$ is continuous and hence Borel measurable. Since the difference of measurable functions is measurable,

$$f(\cdot) \equiv \|(X_n - X)(\cdot)\|$$

is measurable as the composition of measurable functions, so

$$A = \{\omega \in \Omega : \|X_n(\omega) - X(\omega)\| > \epsilon\} = \{\omega \in \Omega : f(\omega) > \epsilon\} \in \mathcal{A}$$

and hence $P(A)$ is defined.

Definition 3.0.3. Let $\alpha \in S$ and X be the random variable defined by

$$X(\omega) = \alpha \text{ for all } \omega \in \Omega.$$

We say that X_n is a consistent estimator of α if $X_n \xrightarrow{p} X$. We write $X_n \xrightarrow{p} \alpha$.

Definition 3.0.4. We say that X_n converges in probability to ∞ if

$$\lim_n P\Big(\|X_n\| > K\Big) = \lim_n P(\{\omega \in \Omega : \|X_n(\omega)\| > K\}) = 1$$

holds for all $K \geq 0$. We write $X_n \xrightarrow{p} \infty$.

Remark 3.0.5. Since $f \equiv \|\cdot\| \circ X_n$ is measurable as the composition of measurable functions,

$$A \equiv \{\omega \in \Omega : \|X_n(\omega)\| > K\} = \{\omega \in \Omega : f(\omega) > K\} \in \mathcal{A}$$

so $P(A)$ is defined.

Having introduced convergence in probability, let us talk about stochastic processes and filtrations.

Definition 3.0.6. A stochastic process is a function

$$X : T \ni t \mapsto X(t)$$

such that the $X(t)$'s are random variables with the same domain and codomain. Say,

$$X(t) : (\Omega, \mathcal{A}, P) \to (S, \mathcal{B})$$

for all $t \in T$.

In practice, T is typically a set of integers (discrete time) or an interval of real numbers (continuous time). The set S is called the state space and, for $\omega \in \Omega$, the function

$$X(\cdot)(\omega) : T \to S$$

is called a sample path.

Example 3.0.7. Imagine flipping a fair coin thrice. This can be interpreted as a discrete stochastic process

$$X : \{1, 2, 3\} \ni \ell \mapsto X(\ell)$$

where $X(\ell)$ represents the ℓ-th coin flip. As such, $X(\ell)$ takes values in $\{H, T\}$ (the state space) and

$$P(X(\ell) = H) = P(X(\ell) = T) = 1/2$$

for all ℓ. An example of a sample path is

$$1 \mapsto T, 2 \mapsto T, 3 \mapsto T$$

(I got tails thrice in a row when I flipped a coin for this example...)

Definition 3.0.8. Let $(\Omega, \mathcal{A})$ be measurable space and T be a set of integers (discrete time) or an interval of real numbers (continuous time). A filtration on $(\Omega, \mathcal{A})$ is a function

$$\mathcal{F} : T \ni t \mapsto \mathcal{F}_t$$

where the $\mathcal{F}_t$'s are sub-σ-algebras of $\mathcal{A}$ that are increasing:

$$\text{if } s \leq t, \text{ then } \mathcal{F}_s \subseteq \mathcal{F}_t.$$

We typically considers filtrations in conjunction with stochastic processes. If we have a stochastic process X on T, then we consider filtrations $\mathscr{F}$ on T. The idea is that $\mathscr{F}_t$ contains information on X up to and including time t.

Example 3.0.9. Let us construct a filtration $\mathscr{F}$ on

$$(\Omega = \{\mathrm{H}, \mathrm{T}\}^3, \mathscr{A} = \wp(\Omega))$$

containing information on the stochastic process X from Example 3.0.7.

Let us start with $\mathscr{F}_1$. Up to and including the first flip, we only know whether the first flip resulted in heads or tails. If heads, then the final tally lies in

$$\mathrm{A_H} \equiv \{\mathrm{HTT}, \mathrm{HTH}, \mathrm{HHT}, \mathrm{HHH}\}$$

If tails, then the final tally lies in $\mathrm{A_T} \equiv \Omega \setminus \mathrm{A_H}$. Since $\mathscr{F}_1$ has to be a sub-σ-algebra of $\mathscr{A}$, let us define

$$\mathscr{F}_1 = \sigma\{\mathrm{A_H}, \mathrm{A_T}\} = \{\varnothing, \mathrm{A_H}, \mathrm{A_T}, \Omega\}$$

Up to and including the second flip, the possible tallies are HH, HT, TH and TT, so we know in which one of the following sets the final tally lies.

$$\begin{aligned}
\mathrm{A_{HH}} &\equiv \{\mathrm{HHT}, \mathrm{HHH}\} \\
\mathrm{A_{HT}} &\equiv \{\mathrm{HTT}, \mathrm{HTH}\} \\
\mathrm{A_{TH}} &\equiv \{\mathrm{THT}, \mathrm{THH}\} \\
\mathrm{A_{TT}} &\equiv \{\mathrm{TTT}, \mathrm{TTH}\}
\end{aligned}$$

As such, let us define

$$\mathscr{F}_2 = \sigma\{\mathrm{A_{HH}}, \mathrm{A_{HT}}, \mathrm{A_{TH}}, \mathrm{A_{TT}}\} = \{\varnothing, \mathrm{A_H}, \mathrm{A_T}, \mathrm{A_{HH}}, \mathrm{A_{HT}}, \mathrm{A_{TH}}, \mathrm{A_{TT}}, \Omega\}$$

And finally, up to and including the third flip, we know exactly what the final tally is, so

$$\mathscr{F}_3 \equiv \sigma\{\{\omega\} : \omega \in \Omega\} = \wp(\Omega) = \mathscr{A}$$

Now, $\mathscr{F}$ is indeed a filtration, because $\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \mathscr{F}_3$.

The previous examples were inspired by [3, Example 2.2.1].

Chapter 4

Product integral

In this chapter, we will introduce the concept of the product integral, which will appear prominently throughout the later chapters. First, we will define the product integral only for absolutely continuous functions. Secondly, we will show that it is related to a certain Volterra equation. Thirdly, we shall give a more general and more intuitive definition of the product integral, although we will not prove that it is well defined.

For brevity's sake, we will omit proofs of measurability of functions in this chapter. Nor shall we mention the use of such integral properties as monotonicity and the triangle inequality.

The author pieced together this section by consulting [4], [5], [6] and [1].

4.1 Definition

Before we can define the product integral, we need to give some auxiliary definitions.

Definition 4.1.1. A matrix-valued function $\mathbf{F}$ on $[a, b]$ is called absolutely continuous if there exists an integrable function $\mathbf{f}$ on $[a, b]$ such that

$$\mathbf{F}(x) = \mathbf{F}(a) + \int_a^x \mathbf{f}(t)\, dt$$

for all $x \in [a, b]$. One can easily verify that $\mathbf{f}$ is unique almost everywhere.

Let $\mathbf{F}$ be as in Definition 4.1.1 and define

$$\mathcal{I}_0(a, x; \mathbf{F}) = \mathbf{I}$$
$$\mathcal{I}_{n+1}(a, x; \mathbf{F}) = \int_a^x \mathcal{I}_n(a, t; \mathbf{F})\mathbf{f}(t)\, dt$$

In other words, for $n \geq 1$,

$$\mathcal{I}_n(a, x; \mathbf{F}) = \int_a^x \int_a^{x_1} \cdots \int_a^{x_{n-1}} \mathbf{f}(x_n) \cdots \mathbf{f}(x_1)\, dx_n \cdots dx_1$$

Since $\mathbf{f}$ is unique almost everywhere, $\mathcal{I}_n$ is well defined.

We are now ready to define the product integral.

Definition 4.1.2. Define the product integral as

$$\mathcal{P}(a, x; \mathbf{F}) = \sum_{n=0}^{\infty} \mathcal{I}_n(a, x; \mathbf{F})$$

To finish up this section, we will show that the series above is absolutely convergent and hence well defined. To do that, we introduce a specific matrix norm.

Let $\|\cdot\|$ be the following matrix norm.

$$\|\mathbf{A}\| = \max_i \sum_j |A_{ij}|$$

One easily verifies that it is submultiplicative.

Let $\mathbf{f}$ be a matrix-valued fucntion. If $\mathbf{f}$ is integrable, then so is $\|\mathbf{f}\|$ and

$$\| \int \mathbf{f}\| \leq \int \|\mathbf{f}\|$$

Indeed, $\mathbf{f}$ being integrable means that the f_{ij}'s are integrable, so the $|f_{ij}|$'s are integrable as well. Consequently,

$$\|\mathbf{f}\| = \max_i \sum_j |f_{ij}|$$

is integrable as the maximum of sums of integrable functions.

Furthermore,

$$\begin{aligned}\|\int \mathbf{f}\| &= \max_i \sum_j |\int f_{ij}| \\ &\leq \max_i \sum_j \int |f_{ij}| = \int \max_i \sum_j |f_{ij}| = \int \|\mathbf{f}\|\end{aligned}$$

which will be used extensively.

Let us return to our original goal of proving absolute convergence. They key argument is the following.

Lemma 4.1.3. Let

$$J = \int_a^x \|\mathbf{f}(t)\| \, dt$$

(note that $J < \infty$ because $\mathbf{f}$ is integrable). Then

$$\|\mathcal{I}_n(a, x, \mathbf{F})\| \leq J^n/n!$$

if we agree that $0^0 = 1$.

Proof. Clearly

$$\|\mathcal{I}_0(a, x, \mathbf{F})\| = \|\mathbf{I}\| = 1 = J^0/0!$$

If $n \geq 1$, then

$$\begin{aligned}\|\mathcal{I}_n(a, x; \mathbf{F})\| &= \left\| \int_a^x \int_a^{x_1} \cdots \int_a^{x_{n-1}} \mathbf{f}(x_n) \cdots \mathbf{f}(x_1) \, dx_n \cdots dx_1 \right\| \\ &\leq \int_a^x \int_a^{x_1} \cdots \int_a^{x_{n-1}} \|\mathbf{f}(x_1) \cdots \mathbf{f}(x_n)\| \, dx_n \cdots dx_1 \\ &\overset{!}{\leq} \int_a^x \int_a^{x_1} \cdots \int_a^{x_{n-1}} \|\mathbf{f}(x_1)\| \cdots \|\mathbf{f}(x_n)\| \, dx_n \cdots dx_1\end{aligned}$$

because $\|\cdot\|$ is submultiplicative.

If we let S_n denote the symmetric group on $\{1, \ldots, n\}$, then

$$n! \int_a^x \int_a^{x_1} \cdots \int_a^{x_{n-1}} \|\mathbf{f}(x_1)\| \cdots \|\mathbf{f}(x_n)\| \, dx_n \cdots dx_1 \stackrel{!}{=}$$
$$\sum_{\sigma \in S_n} \int_a^x \int_a^{x_{\sigma(1)}} \cdots \int_a^{x_{\sigma(n-1)}} \|\mathbf{f}(x_{\sigma(1)})\| \cdots \|\mathbf{f}(x_{\sigma(n)})\| \, dx_{\sigma(n)} \cdots dx_{\sigma(1)} =$$
$$\int_a^x \int_a^x \cdots \int_a^x \|\mathbf{f}(x_1)\| \cdots \|\mathbf{f}(x_n)\| \, dx_n \cdots dx_1 =$$
$$\left(\int_a^x \|\mathbf{f}(t)\| \, dt \right)^n$$

because the $\|\mathbf{f}(x_i)\|$'s commute. The result now readily follows. □

By the lemma above,

$$\|\mathcal{P}(a, x; \mathbf{F})\| \leq \sum_{n=0}^{\infty} \|\mathcal{I}_n(a, x; \mathbf{F})\| = \sum_{n=0}^{\infty} \frac{J^n}{n!} = e^J < \infty$$

establishing the absolute convergence of $\mathcal{P}(a, x; \mathbf{F})$.

4.2 Volterra equation

We will now show that the product integral $\mathcal{P}(a, x; \mathbf{F})$ is, in a sense, the unique solution to the Volterra equation

$$\mathbf{Z}(x) = \mathbf{I} + \int_a^x \mathbf{Z}(t)\mathbf{f}(t) \, dt$$

for all $x \in [a, b]$. Before we can give the precise statement, we need the following definition.

Definition 4.2.1. Let $I \subseteq \mathbb{R}$ be an interval. A function $f : I \to \mathbb{R}$ is said to have left limits if

$$f(x-) = \lim_{y \uparrow x} f(y)$$

exists for all $x \in I$ such that there exists a $y \in I$ such that $y < x$.

Similarly, f has right limits if

$$f(x+) = \lim_{y \downarrow x} f(y)$$

exists for all $x \in I$ such that there exists a $y \in I$ such that $x < y$.

A matrix-valued valued function is said to have left (right) limits if each of its components has left (right) limits.

Theorem 4.2.2. Not only is $\mathbf{Z}(x) = \mathcal{P}(a, x; \mathbf{F})$ a solution of

$$\mathbf{Z}(x) = \mathbf{I} + \int_a^x \mathbf{Z}(t)\mathbf{f}(t)\, dt$$

for all $x \in [a, b]$, but, if we require that $\mathbf{Z}$ has left and right limits, then $\mathbf{Z}(x) = \mathcal{P}(a, x; \mathbf{F})$ is the only solution.

Proof. First, we show that $\mathcal{P}(a, x; \mathbf{F})$ is a solution. That is,

$$\mathcal{P}(a, x; \mathbf{F}) - \mathbf{I} = \int_a^x \mathcal{P}(a, t; \mathbf{F})\mathbf{f}(t)\, dt$$

Indeed, by Fubini's theorem,

$$\begin{aligned}\mathcal{P}(a, x; \mathbf{F}) - \mathbf{I} &= \sum_{n=1}^{\infty} \mathcal{I}_n(a, x; \mathbf{F}) = \sum_{n=0}^{\infty} \int_a^x \mathcal{I}_n(a, t; \mathbf{F})\mathbf{f}(t)\, dt \\ &\overset{!}{=} \int_a^x \sum_{n=0}^{\infty} \mathcal{I}_n(a, t; \mathbf{F})\mathbf{f}(t)\, dt = \int_a^x \mathcal{P}(a, t; \mathbf{F})\mathbf{f}(t)\, dt\end{aligned}$$

We may use Fubini's theorem because, by Lemma 4.1.3,

$$\begin{aligned}\left\| \sum_{n=0}^{\infty} \int_a^x \mathcal{I}_n(a, t; \mathbf{F})\mathbf{f}(t)\, dt \right\| &\le \sum_{n=0}^{\infty} \int_a^x \|\mathcal{I}_n(a, t; \mathbf{F})\| \cdot \|\mathbf{f}(t)\|\, dt \\ &\overset{!}{\le} \sum_{n=0}^{\infty} \int_a^x \frac{1}{n!} \left(\int_a^t \|\mathbf{f}(u)\|\, du \right)^n \cdot \|\mathbf{f}(t)\|\, dt \\ &= \sum_{n=0}^{\infty} \int_0^J \frac{v^n}{n!}\, dv = \sum_{n=0}^{\infty} \frac{J^{n+1}}{(n+1)!} = e^J - 1 < \infty\end{aligned}$$

with

$$v = \int_a^t \|\mathbf{f}(u)\|\, du$$

Second, we show unicity. If $\mathbf{Z}$ has both left and right limits and satisfies

$$\mathbf{Z}(x) - \mathbf{I} = \int_a^x \mathbf{Z}(t)\mathbf{f}(t)\, dt$$

for all $x \in [a, b]$, then $\mathbf{Z}(x) = \mathcal{P}(a, x; \mathbf{F})$.

Indeed, if we define

$$\mathcal{P}^{(n)}(a, x; \mathbf{F}) = \sum_{k=0}^{n} \mathcal{I}_k(t)$$

then

$$\mathcal{P}^{(n+1)}(a, x; \mathbf{F}) = \mathbf{I} + \int_a^x \mathcal{P}^{(n)}(a, t; \mathbf{F})\mathbf{f}(t)\, dt$$

If we subtract the equation above from the Volterra equation with $\mathbf{Z}$, then we see that

$$\mathbf{Z}(x) - \mathcal{P}^{(n+1)}(a, x; \mathbf{F}) = \int_a^x (\mathbf{Z}(t) - \mathcal{P}^{(n)}(a, t; \mathbf{F}))\mathbf{f}(t)\, dt$$

Since $\mathbf{Z}$ has both left and right limits, by Theorem A.0.1, it is bounded on $[a, b]$. As such,

$$M \equiv \sup_{t\in[a,b]} \|\mathbf{Z}(t)\| < \infty$$

and hence

$$\|\mathbf{Z}(x) - \mathbf{I}\| \leq \int_a^x \|\mathbf{Z}(t)\| \cdot \|\mathbf{f}(t)\|\, dt \leq M \int_a^x \|\mathbf{f}(t)\|\, dt = M \cdot J$$

By induction on n, we will show that

$$\|\mathbf{Z}(x) - \mathcal{P}^{(n)}(a, x; \mathbf{F})\| \leq M\frac{J^{n+1}}{(n+1)!} = M\frac{1}{(n+1)!}\left(\int_a^x \|\mathbf{f}(t)\|\, dt\right)^{n+1}$$

We have just shown that it holds for $n = 0$. If it holds for n, then

$$\begin{aligned}
\|\mathbf{Z}(x) - \mathcal{P}^{(n+1)}(a, x; \mathbf{F})\| &\leq \int_a^x \|\mathbf{Z}(t) - \mathcal{P}^{(n)}(a, t; \mathbf{F})\| \cdot \|\mathbf{f}(t)\|\, dt \\
&\leq M \int_a^x \frac{1}{(n+1)!}\left(\int_a^t \|\mathbf{f}(u)\|\, du\right)^{n+1} \cdot \|\mathbf{f}(t)\|\, dt \\
&= M\frac{J^{n+2}}{(n+2)!}
\end{aligned}$$

so it holds for $n + 1$. As such, if $n \to \infty$, then

$$\mathbf{Z}(x) = \lim_n \mathcal{P}^{(n)}(a, x; \mathbf{F}) = \mathcal{P}(a, x; \mathbf{F})$$

proving unicity. □

4.3 General case

We will first give an alternative definition of the product integral and then we will generalize the alternative definition to a broader class of functions. As such, just like before, let $\mathbf{F}$ be a matrix-valued function on $[a, b]$ that is absolutely continuous.

Let $\mathcal{T} = \{t_i : i = 0, \ldots, n\}$ be a partition of $[a, b]$ (that is, $a = t_0 < t_1 < \cdots < t_n = b$) and $|\mathcal{T}| = \max_i(t_i - t_{i-1})$. If we then define

$$\begin{aligned}\prod_a^b (\mathbf{I} + d\mathbf{F}) &= \lim_{|\mathcal{T}|\to 0} \prod_{\mathcal{T}} (\mathbf{I} + \Delta\mathbf{F}) \\ &= \lim_{|\mathcal{T}|\to 0} \prod_{i=1}^{n} (\mathbf{I} + \mathbf{F}(t_i) - \mathbf{F}(t_{i-1}))\end{aligned}$$

(with the empty product equaling the identity matrix), then [by 1, Theorem II.6.4]

$$\prod_a^b (\mathbf{I} + d\mathbf{F}) = \mathcal{P}(a, b; \mathbf{F}) \tag{4.1}$$

For absolutely continuous $\mathbf{F}$, the limit above exists (because the product integral exists).

It was clearly the previous definition that inspired both the name and the notation of the product integral. To see this, we urge the reader to compare the definition above with the definition of the Riemann-Stieltjes integral.

The key observation to generalizing the product integral is that the limit above exists for more than just absolutely continuous $\mathbf{F}$'s. But, before we can formulate the definition of the generalized product integral, we need introduce the terms "cadlag" and "locally bounded variation".

Definition 4.3.1. A cadlag function is a right continuous function with left limits.

Remark 4.3.2. Cadlag is a french initialism of "continue à droite, limite à gauche" meaning "right continuous, left limits".

Definition 4.3.3. Let $\mathcal{T}$ be a partition of $[a, b]$. A matrix-valued function $\mathbf{F}$ on $[0, \infty)$ is said to be of locally bounded variation if

$$\sup_{\mathcal{T}} \sum_{\mathcal{T}} \|\Delta\mathbf{F}\| = \sup_{\mathcal{T}} \sum_{i=1}^{n} \|\mathbf{F}(t_i) - \mathbf{F}(t_{i-1})\| < \infty$$

for any $[a, b] \subseteq [0, \infty)$.

Theorem 4.3.4. Let $\mathcal{T}$ be a partition of $[a, b]$. If $\mathbf{F}$ is a matrix-valued function on $[0, \infty)$ that is cadlag and of locally bounded variation, then

$$\lim_{|\mathcal{T}|\to 0} \prod_{\mathcal{T}} (\mathbf{I} + \Delta\mathbf{F})$$

exists for any $[a, b] \subseteq [0, \infty)$.

The aforementioned theorem is basically a rewording of [1, Definition II.6.1].

Definition 4.3.5. Let $\mathcal{T}$ be a partition of $[a, b]$. We define the generalized product integral as

$$\prod_a^b (\mathbf{I} + d\mathbf{F}) = \lim_{|\mathcal{T}|\to 0} \prod_{\mathcal{T}} (\mathbf{I} + \Delta\mathbf{F})$$

for $\mathbf{F}$ as in Theorem 4.3.4.

Remark 4.3.6. Because the non-generalized product integral coincides with the generalized product integral, we will simply call them both just "product integral".

One can easily verify that if $\mathbf{F}$ is absolutely continuous, then it is (right) continuous and

$$\sum_{\mathcal{T}} \|\Delta\mathbf{F}\| \leq \int_a^b \|\mathbf{f}(t)\| \, dt < \infty$$

(because $\mathbf{f}$ is integrable) so it is of locally bounded variation.

Section 6.2 and 7.1 offer examples of the generalized product integral in action.

In case the reader is wondering why we did not give a proof of equation (4.1) or Theorem 4.3.4, it is because, in order to do so, we would have to do Lebesgue-Stieltjes integration, which is beyond the scope of this thesis.

Chapter 5

Multi-state models

In this chapter, we will give the formal definition of a multi-state model and other relevant definitions pertaining to multi-state models such as the (cumulative) transition hazard, the transition probability and the state occupation probability. We shall also touch base with right censoring and the independent censoring assumption.

5.1 Definitions

We will roughly follow [7, section 2].

Definition 5.1.1. A multi-state process is a continuous time stochastic process X on $[0, \infty)$ with a finite state space and cadlag sample paths.

We will typically, but not always (see the proof of Theorem 7.1.3), take the finite state space to be $\{1, \ldots, K\}$.

Remark 5.1.2. Fix ω. Since $X(\cdot)(\omega)$ is cadlag, it has left and right limits. By Theorem A.0.2, any function with left and right limits has but countably many discontinuities. Ergo, $X(\cdot)(\omega)$ is continuous almost everywhere.

What's more, X has only finitely many discontinuities (also called jumps) in finite time. Indeed, from the proof of Theorem A.0.2 we see that

$$\{t \in [0, n] : |X(t+)(\omega) - X(t-)(\omega)| > 1/2\}$$

is finite for all $n \in \mathbb{N}$. Now, since X takes values in $\{1, \ldots, K\}$, any discontinuity will be of size bigger than $1/2$, so X has only finitely many jumps in finite time.

Suppose we have n individuals, which we label with $1, \ldots, n$. We ascribe to each individual i a multi-state process $X_i^{\sim}$. That is what we call a multi-state model.

Henceforth we assume that the $X_i^{\sim}$'s are independent and identically distributed. Note that this assumption is a homogeneity assumption, in the sense that we are modelling a homogeneous population of identically distributed components. Furthermore, let $X^{\sim}$ be identically distributed to the $X_i^{\sim}$'s.

Definition 5.1.3. If $\mathbf{X}$ is a matrix-valued function on $[0,\infty)$ that has left limits, then

$$d\mathbf{X}(s) = \mathbf{X}(s) - \mathbf{X}(s-)$$

Let $i \in \{1,\dots,n\}$ and $j \neq k \in \{1,\dots,K\} \ni \ell, m$. Consider

- the counting process

  $$N_{ijk}^{\sim}(t) = \#\{u \leq t : X_i^{\sim}(u-) = j, X_i^{\sim}(u) = k\}$$

  which counts the number of direct transitions of subject i from state j to state k up to and including time t. By Remark 5.1.4, $N_{ijk}^{\sim}(t)$ is finite and hence well defined.

  Clearly

  $$N_{ijk}^{\sim}(t-) = \#\{u < t : X_i^{\sim}(u-) = j, X_i^{\sim}(u) = k\}$$

  so $N_{ijk}^{\sim}$ has left limits. As such,

  $$\begin{aligned} dN_{ijk}^{\sim}(t) &= N_{ijk}^{\sim}(t) - N_{ijk}^{\sim}(t-) \\ &= [X_i^{\sim}(t-) = j, X_i^{\sim}(t) = k] \in \{0,1\} \end{aligned}$$

  tells us whether individual i jumped from state j to state k at time t.

- the at-risk process

  $$Y_{ij}^{\sim}(t) = [X_i^{\sim}(t-) = j]$$

  which indicates whether individual i is in state j right before time t.

- the filtration

  $$\mathcal{F}_t^{\sim} = \sigma\{(N_{ijk}^{\sim}(u), Y_{ij}^{\sim}(u)) : u \leq t, i \in \{1,\dots,n\}, j \neq k \in \{1,\dots,K\}\}$$

- the transition hazard

$$\begin{aligned}\alpha_{jk}(t) &= \lim_{h\downarrow 0} \frac{P(X^{\sim}((t+h)-) = k \mid X^{\sim}(t-) = j)}{h} \\ &= \lim_{h\downarrow 0} \frac{P(X^{\sim}(t+h) = k \mid X^{\sim}(t-) = j)}{h}\end{aligned}$$

which we assume exists. We can then not-so-rigorously write

$$\begin{aligned}\alpha_{jk}(t)\,dt &= P(dN^{\sim}_{ijk}(t) = 1 \mid X^{\sim}_i(t-) = j) \\ &= E(dN^{\sim}_{ijk}(t) \mid X^{\sim}_i(t-) = j)\end{aligned}$$

because $X^{\sim}$ be identically distributed to the $X^{\sim}_i$'s. Let us finally define the $K \times K$ matrix $\boldsymbol{\alpha}(t)$ by

$$\begin{aligned}[\boldsymbol{\alpha}(t)]_{jk} &= \alpha_{jk}(t) \\ [\boldsymbol{\alpha}(t)]_{jj} &= -\sum_{k\neq j} \alpha_{jk}(t)\end{aligned}$$

- the cumulative transition hazard

$$A_{jk}(t) = \int_0^t \alpha_{jk}(u)\,du$$

In order for this to make sense, we assume that the α_{jk}'s are integrable over $[0,t]$. Also, define the $K \times K$ matrix $\mathbf{A}(t)$ analogously to $\boldsymbol{\alpha}(t)$.

- the transition probability

$$P_{\ell m}(s,t) = P(X^{\sim}(t) = m \mid X^{\sim}(s) = \ell)$$

We define the $K \times K$ matrix $\mathbf{P}(s,t)$ by

$$[\mathbf{P}(s,t)]_{\ell m} = P_{\ell m}(s,t)$$

Since $\sum_m P_{\ell m}(s,t) = 1$, it follows componentwise that

$$\boldsymbol{\alpha}(t) = \lim_{h\downarrow 0} \frac{\mathbf{P}(t-, t+h) - \mathbf{I}}{h}$$

- the state occupation probability

$$P_m(t) = P(X^{\sim}(t) = m)$$

Define the $1 \times K$ matrix (row vector) $\mathbf{P}(t)$ by

$$[\mathbf{P}(t)]_{1m} = P_m(t)$$

From the law of total probability, one can infer that

$$\mathbf{P}(t) = \pi(0)\mathbf{P}(0,t)$$

with $\pi(0)$ a $1 \times K$ matrix with

$$[\pi(0)]_{1k} = P(X^{\sim}(0) = k)$$

Finally, we call a multi-state model Markov if

$$P(X^{\sim}(t) = m \mid X^{\sim}(s) = \ell, \mathcal{F}^{\sim}_{s-}) = P(X^{\sim}(t) = m \mid X^{\sim}(s) = \ell)$$

If we consider time s to be the present, then the Markov property is intuitively saying that the probability that any individual, that is in state ℓ right now, will be in state m in the future does not dependent on where that individual was previously or how long they were there.

5.2 Right censoring

Let us begin by explaining what it means for a multi-state model to have right censoring.

Definition 5.2.1. A right-censoring time is a random variable $C : \Omega \to [0, \infty]$.

We say that a multi-state model has right censoring if, to every individual i, we have an associated right-censoring time C_i.

Suppose that the $X_i^{\sim}$'s and C_i's are independent and identically distributed. Let

$$\begin{aligned} X_i(t) &= X_i^{\sim}(t \wedge C_i) \\ N_{ijk}(t) &= \#\{u \leq t : X_i(u-) = j, X_i(u) = k, C_i \geq u\} \\ Y_{ij}(t) &= [X_i(t-) = j, C_i \geq t] \end{aligned}$$

which we call the censored multi-state, counting and at-risk processes. Intuitively, we can no longer observe what state individual i is in after time C_i.

Since

$$N_{ijk}(t) \leq N_{ijk}^{\sim}(t)$$

it is finite and hence well defined. Clearly it has left limits as well. As such,

$$\begin{aligned} dN_{ijk}(t) &= N_{ijk}(t) - N_{ijk}(t-) \\ &= [X_i(t-) = j, X_i(t) = k, C_i \geq t] \in \{0, 1\} \end{aligned}$$

tells us whether we observed that individual i jumped from state j to state k at time t.

We also want to highlight that Y_{ij} is "predictable" in a sense.

Theorem 5.2.2. $Y_{ij}(t) = Y_{ij}(t-)$. That is,

$$Y_{ij}(t) = 1 \Leftrightarrow Y_{ij}(t-) = 1, \text{ or equivalently, } Y_{ij}(u) = 1 \text{ for all } u \lesssim t.$$

Proof. We'll first show "$\Rightarrow$". Then we will show "$\Leftarrow$" by contraposition.

If $Y_{ij}(t) = 1$, then $X_i(t-) = j$ and $C_i \geq t$.

- If $X_i(t-) = j$, then $X_i((t-)-) = X_i(t-) = j$ as well.
- If $C_i \geq t$, then clearly $C_i \geq u$ for all $u < t$.

Thus, if $Y_{ij}(t) = 1$, then $Y_{ij}(t-) = 1$, or equivalently, $Y_{ij}(u) = 1$ for all $u \lesssim t$.

As for "$\Leftarrow$", suppose that $Y_{ij}(t) = 0$. Then $X_i(t-) \neq j$ or $C_i < t$.

- If $X_i(t-) \neq j$, then $X_i((t-)-) = X_i(t-) \neq j$ as well.
- If $C_i < t$, then $C_i = v$ for some $v < t$, so $C_i < u$ for all $v < u < t$.

Either case clearly establishes the contraposition. □

At this juncture, we would like to show you that right censoring makes the situation considerably more difficult. Indeed, without it ($C_i = \infty$ for all i), it follows readily from the weak law of large numbers (see Theorem B.0.1) that

$$\frac{1}{n}\sum_{i=1}^{n} [X_i(t) = m] = \frac{1}{n}\sum_{i=1}^{n} [X_i^{\sim}(t) = m] \xrightarrow{p} P_m(t)$$

because the $[X_i^{\sim}(t) = m]$'s are Bernoulli distributed with parameter

$$P(X_i^{\sim}(t) = m) = P(X^{\sim}(t) = m) = P_m(t)$$

Don't forget that the weak law of large numbers requires that the $[X_i^{\sim}(t) = m]$'s be independent and identically distributed, which readily follows from the fact that the the $X_i^{\sim}$'s are. So, without right censoring, estimating state occupation probabilities consistently is very easy.

In a similar way, one can show that without right censoring

$$\frac{1}{n}\sum_{i=1}^{n} [X_i(t) = m, X_i(s) = \ell] \xrightarrow{p} P(X^{\sim}(t) = m, X^{\sim}(s) = \ell)$$

so, by the continuous mapping theorem (Example D.0.3),

$$\begin{aligned}\frac{\sum_{i=1}^{n} [X_i(t) = m, X_i(s) = \ell]}{\sum_{i=1}^{n} [X_i(s) = \ell]} &\xrightarrow{p} \frac{P(X^{\sim}(t) = m, X^{\sim}(s) = \ell)}{P(X^{\sim}(s) = \ell)} \\ &= P(X^{\sim}(t) = m \mid X^{\sim}(s) = \ell) = P_{\ell m}(s,t)\end{aligned}$$

if $P(X^{\sim}(s) = \ell) > 0$. As such, without right censoring, the transition probabilities can be very easily consistently estimated as well.

Now, even with right censoring, the estimators above would clearly still be consistent if you disregard the censored individuals. That is, sum only over those i's such that $C_i = \infty$. However, we do not want to do that. Throwing away information reduces accuracy and we want to have as much accuracy as possible, even if it means having to do more (complicated) work.

Finally, we reveal the <u>independent censoring assumption</u>

$$P(dN_{ijk}(t) = 1 \mid Y_{ij}(t) = 1, \mathcal{F}_{t-}) = P(dN_{ijk}^{\sim}(t) = 1 \mid Y_{ij}^{\sim}(t) = 1, \mathcal{F}_{t-}^{\sim})$$

with

$$\mathcal{F}_t = \sigma\{(N_{ijk}(u), Y_{ij}(u)) : u \leq t, i \in \{1, \ldots, n\}, j \neq k \in \{1, \ldots, K\}\}$$

Intuitively, the independent censoring assumption enforces that it should be just as probable to register a jump with right censoring as it would be without right censoring.

Henceforth we shall assume that the independent censoring assumption is in effect.

Chapter 6

Markov case

We have seen that, without right censoring, estimating probabilities consistently is very easy. Since we don't want to throw away any data, we are going to have to do something more complicated in the case of right censoring.

In this chapter, we will have a look at the Markov case. The next chapter shall deal with the even more difficult non-Markov case.

First, we will derive that the transition probability can be written as the product integral of the cumulative transition hazard via a Volterra integral representation, which we will in turn derive from the so-called Chapman-Kolmogorov equations. Next, we will heuristically derive the Nelson-Aalen estimator from the independent censoring assumption. And finally, we shall touch upon the Aalen-Johansen estimator. Both the Nelson-Aalen and the Aalen-Johansen estimator turn out to be consistent.

6.1 Product integral representation

In order to derive the product integral representation we are after, we will first derive the so-called Chapman-Kolmogorov equations.

Theorem 6.1.1. If our multi-state model is Markov and $s \leq u \leq t$, then

$$\mathbf{P}(s,t) = \mathbf{P}(s,u)\mathbf{P}(u,t)$$

Proof. Because our multi-state model is Markov,

$$
\begin{aligned}
P_{hj}(s,t) &= P(X^{\sim}(t) = j \mid X^{\sim}(s) = h) \\
&= \sum_{\ell=1}^{K} P(X^{\sim}(t) = j, X^{\sim}(u) = \ell \mid X^{\sim}(s) = h) \\
&= \sum_{\ell=1}^{K} P(X^{\sim}(t) = j \mid X^{\sim}(u) = \ell, X^{\sim}(s) = h) P(X^{\sim}(u) = \ell \mid X^{\sim}(s) = h) \\
&\overset{!}{=} \sum_{\ell=1}^{K} P(X^{\sim}(t) = j \mid X^{\sim}(u) = \ell) P(X^{\sim}(u) = \ell \mid X^{\sim}(s) = h) \\
&= \sum_{\ell=1}^{K} P_{\ell j}(u,t) P_{h\ell}(s,u) = \sum_{\ell=1}^{K} P_{h\ell}(s,u) P_{\ell j}(u,t)
\end{aligned}
$$

The matrix equation now follows componentwise. □

In Appendix C, we present a proof of the following fact.

Theorem 6.1.2. If $s \leq t$, then

$$\mathbf{P}(s,t) = \mathbf{I} + \int_s^t \mathbf{P}(s,u)\boldsymbol{\alpha}(u)\, du$$

Since $X^{\sim}$ is cadlag, so is

$$t \mapsto [\mathbf{P}(s,t)]_{\ell m} = P_{\ell m}(s,t) = P(X^{\sim}(t) = m \mid X^{\sim}(s) = \ell)$$

As such, $t \mapsto \mathbf{P}(s,t)$ is cadlag and hence has left and right limits. Moreover, since

$$\mathbf{A}(t) = \int_0^t \boldsymbol{\alpha}(u)\, du$$

it follows that $\mathbf{A}$ is absolutely continuous on any $[s,t] \subseteq [0,\infty)$. As such, by Theorem 4.2.2, the following product integral representation now follows.

Corollary 6.1.3. If $s \leq t$, then

$$\mathbf{P}(s,t) = \mathcal{P}(s,t;\mathbf{A}) = \prod_s^t (\mathbf{I} + d\mathbf{A})$$

6.2 Nelson-Aalen estimator

In this section, we will heuristically show you where the Nelson-Aalen estimator comes from. The following argument is loosely based on [8, section 1.4 and 3.1.5] and [1, section II.1 and IV.1.1].

Note that

$$\begin{aligned} E(dN_{ijk}(t) \mid \mathcal{F}_{t-}) &= P(dN_{ijk}(t) = 1 \mid \mathcal{F}_{t-}) \\ &= P(dN_{ijk}(t) = 1, Y_{ij}(t) = 1 \mid \mathcal{F}_{t-}) \end{aligned}$$

because

$$dN_{ijk}(t) = 1 \Rightarrow X_i(t-) = j, C_i \geq t \Leftrightarrow Y_{ij}(t) = 1$$

(if we witness that individual i jumps from state j to state k at time t, then we must have known that they were at state j right before time t and that right censoring hasn't yet kicked in), so

$$\begin{aligned} E(dN_{ijk}(t) \mid \mathcal{F}_{t-}) = \\ P(dN_{ijk}(t) = 1, Y_{ij}(t) = 1 \mid \mathcal{F}_{t-}) = \\ P(Y_{ij}(t) = 1 \mid \mathcal{F}_{t-}) \cdot P(dN_{ijk}(t) = 1 \mid Y_{ij}(t) = 1, \mathcal{F}_{t-}) = \\ P(Y_{ij}(t-) = 1 \mid \mathcal{F}_{t-}) \cdot P(dN^{\sim}_{ijk}(t) = 1 \mid Y^{\sim}_{ij}(t) = 1, \mathcal{F}^{\sim}_{t-}) = \\ Y_{ij}(t-) \cdot P(dN^{\sim}_{ijk}(t) = 1 \mid X^{\sim}_i(t-) = j)\, dt = \\ Y_{ij}(t) \cdot \alpha_{jk}(t)\, dt \end{aligned}$$

because

- conditional probability (second equality)
- $Y_{ij}(t) = Y_{ij}(t-)$ (third and fifth equality)
- independent censoring assumption (third equality)
- our multi-state model is Markov (fourth equality)

Now define (for $j \neq k$)

$$\begin{aligned} \overline{N}_{jk}(t) &= \sum_{i=1}^{n} N_{ijk}(t) \\ \overline{Y}_j(t) &= \sum_{i=1}^{n} Y_{ij}(t) \end{aligned}$$

and the matrix $\overline{\mathbf{N}}(t)$ similarly to $\boldsymbol{\alpha}(t)$ and

$$\overline{\mathbf{Y}}_D(t) = \mathrm{diag}(\overline{Y}_1(t), \ldots, \overline{Y}_n(t))$$

Then, by summing over i, we see that

$$\mathrm{E}(\mathrm{d}\overline{\mathrm{N}}_{\mathrm{jk}}(\mathrm{t}) \mid \mathcal{F}_{\mathrm{t}-}) = \overline{\mathrm{Y}}_{\mathrm{j}}(\mathrm{t})\alpha_{\mathrm{jk}}(\mathrm{t})\ \mathrm{dt}$$

Although the differentials are already not so rigorous, now we are going to do something even more crude. Clearly $\mathrm{E}(\mathrm{d}\overline{\mathrm{N}}_{\mathrm{ijk}}(\mathrm{t}) \mid \mathcal{F}_{\mathrm{t}-})$ is impossible to obtain in practice, but, and here is the crux, $\mathrm{d}\overline{\mathrm{N}}_{\mathrm{ijk}}(\mathrm{t})$ is not. If we assume that the actual thing is close enough to its average (which sounds reasonable enough), then we get that

$$\mathrm{d}\overline{\mathrm{N}}_{\mathrm{jk}}(\mathrm{t}) \approx \overline{\mathrm{Y}}_{\mathrm{j}}(\mathrm{t})\alpha_{\mathrm{jk}}(\mathrm{t})\ \mathrm{dt}$$

At this juncture, we wish to divide by $\overline{\mathrm{Y}}_{\mathrm{j}}(\mathrm{t})$, which requires that $\overline{\mathrm{Y}}_{\mathrm{j}}(\mathrm{t}) > 0$. As such, we first multiply by

$$\mathrm{J}_{\mathrm{j}}(\mathrm{t}) \equiv \left[\overline{\mathrm{Y}}_{\mathrm{j}}(\mathrm{t}) > 0\right]$$

which yields that

$$\frac{\mathrm{J}_{\mathrm{j}}(\mathrm{t})}{\overline{\mathrm{Y}}_{\mathrm{j}}(\mathrm{t})}\ \mathrm{d}\overline{\mathrm{N}}_{\mathrm{jk}}(\mathrm{t}) \approx \mathrm{J}_{\mathrm{j}}(\mathrm{t})\alpha_{\mathrm{jk}}(\mathrm{t})\ \mathrm{dt}$$

if we agree that $0/0 = 0$. So,

$$\text{if it is very likely that } \overline{\mathrm{Y}}_{\mathrm{j}}(\mathrm{u}) > 0 \text{ for almost every } \mathrm{u} \in [0, \mathrm{t}],$$

then

$$\mathrm{A}^{\wedge}_{\mathrm{jk}}(\mathrm{t}) \equiv \text{“}\int_0^{\mathrm{t}} \frac{\mathrm{J}_{\mathrm{j}}(\mathrm{u})}{\overline{\mathrm{Y}}_{\mathrm{j}}(\mathrm{u})}\ \mathrm{d}\overline{\mathrm{N}}_{\mathrm{jk}}(\mathrm{u})\text{”} \approx \int_0^{\mathrm{t}} \mathrm{J}_{\mathrm{j}}(\mathrm{u})\alpha_{\mathrm{jk}}(\mathrm{u})\ \mathrm{du} \approx \int_0^{\mathrm{t}} \alpha_{\mathrm{jk}}(\mathrm{u})\ \mathrm{du} = \mathrm{A}_{\mathrm{jk}}(\mathrm{t})$$

where we have put quotation marks around the integral because we have not defined what it means to integrate with respect to $\mathrm{d}\overline{\mathrm{N}}_{\mathrm{jk}}(\mathrm{u})$. In order to properly do this, we would have to introduce stochastic integration, which is beyond the scope of this thesis.

Define $\mathbf{A}^{\wedge}(\mathrm{t})$ similarly to $\mathbf{A}(\mathrm{t})$ and that is the Nelson-Aalen estimator (we will give a different definition shortly).

Let (t_ℓ) be the increasing sequence of jump times of any individual from any state to any other state. Then t_ℓ is the ℓ-th jump time.

Let us justify that

$$\mathrm{A}^{\wedge}_{\mathrm{jk}}(\mathrm{t}) = \text{“}\int_0^{\mathrm{t}} \frac{\mathrm{J}_{\mathrm{j}}(\mathrm{u})}{\overline{\mathrm{Y}}_{\mathrm{j}}(\mathrm{u})}\ \mathrm{d}\overline{\mathrm{N}}_{\mathrm{jk}}(\mathrm{u})\text{”} = \sum_{\ell:\mathrm{t}_\ell \le \mathrm{t}} \frac{\mathrm{d}\overline{\mathrm{N}}_{\mathrm{jk}}(\mathrm{t}_\ell)}{\overline{\mathrm{Y}}_{\mathrm{j}}(\mathrm{t}_\ell)}$$

Since

$$d\overline{N}_{jk}(u) = \sum_{i=1}^{n} d\overline{N}_{ijk}(u)$$

it is equal to 0 if $u \neq t_\ell$ for some ℓ and $dN_{jk}(t_\ell)$ is the number of individuals that jumps from state j to state k at the ℓ-th jump time. Since there are only finitely many $t_\ell \leq t$, we are effectively integrating over a finite set and integrals over finite sets are finite sums.

Note that if $\overline{Y}_j(t_\ell) = 0$, then no individual was in state j right before the ℓ-th jump time, so $dN_{jk}(t_\ell) = 0$ and hence

$$\frac{d\overline{N}_{jk}(t_\ell)}{\overline{Y}_j(t_\ell)} = \frac{0}{0} = 0$$

is well defined. As such, the Nelson-Aalen estimator is well defined.

Our definition of the Nelson-Aalen estimator will be the matrix form of the sum above.

Definition 6.2.1. The Nelson-Aalen estimator is given by

$$\mathbf{A}^{\wedge}(t) = \sum_{\ell : t_\ell \leq t} \left(\overline{\mathbf{Y}}_D(t_\ell) \right)^{-1} d\overline{\mathbf{N}}(t_\ell)$$

where we agree that $0/0 = 0$.

And finally, as we have said earlier, the Nelson-Aalen estimator is consistent. A proof of the following theorem can be found in [1, section IV.1.2]. It is based on the so-called Lenglart's inequality [see 1, section II.5.2.1].

Theorem 6.2.2. Let $t \in [0, \infty)$. If

$$\overline{Y}_j^{(n)}(u) \xrightarrow{p} \infty \text{ for all } j$$

for almost every $u \in [0, t]$, then

$$\sup_{u \in [0,t]} \| \mathbf{A}_n^{\wedge}(u) - \mathbf{A}(u) \| \xrightarrow{p} 0$$

where we write $\overline{Y}_j^{(n)}$ & $\mathbf{A}_n^{\wedge}$ instead of $\overline{Y}_j$ & $\mathbf{A}^{\wedge}$ to remind you that they are dependent on the total number of individuals.

Upon further inspection, the assumption is plausible enough. Intuitively, it asserts that, if there are enough individuals, it is very likely that any state can have any number of individuals right before time u (for almost every $u \leq t$).

In particular, from Definition 3.0.4, one finds that

$$\lim_n P\left(\overline{Y}_j^{(n)}(u) > 0\right) = 1$$

for almost every $u \in [0,t]$, so, if there are enough individuals, then "it is very likely that $\overline{Y}_j(u) > 0$ for almost every $u \in [0,t]$".

Corollary 6.2.3. Let $t \in [0,\infty)$. Under the same condition as in Theorem 6.2.2,

$$\mathbf{A}_n^{\wedge}(t) \xrightarrow{p} \mathbf{A}(t)$$

6.3 Aalen-Johansen estimator

Since $\mathbf{P}(s,t) = \prod_s^t(\mathbf{I} + d\mathbf{A})$, let

$$\mathbf{P}^{\wedge}(s,t) = \prod_s^t (\mathbf{I} + d\mathbf{A}^{\wedge})$$

We call $\mathbf{P}^{\wedge}(s,t)$ the Aalen-Johansen estimator.

Let us justify that $\mathbf{A}^{\wedge}$ is in fact product integrable. Clearly it is cadlag and

$$\sum_{\mathcal{T}} \|\Delta\mathbf{A}^{\wedge}\| \leq \sum_{\ell : t_\ell \in [s,t]} \left\| \left(\overline{\mathbf{Y}}_D(t_\ell)\right)^{-1} d\overline{\mathbf{N}}(t_\ell) \right\| < \infty$$

for any partition $\mathcal{T}$ of $[s,t]$ (because it is a finite sum), so $\mathbf{A}^{\wedge}$ is of locally bounded variation as well.

From Definition 4.3.5, we can derive that

$$\mathbf{P}^{\wedge}(s,t) = \prod_s^t (\mathbf{I} + d\mathbf{A}^{\wedge}) = \lim_{|\mathcal{T}|\to 0} \prod_{\mathcal{T}} (\mathbf{I} + \Delta\mathbf{A}^{\wedge}) = \prod_{\ell : t_\ell \in [s,t]} (\mathbf{I} + d\mathbf{A}(t_\ell))$$

with $\mathcal{T}$ a partition of $[s,t]$. Note that the Aalen-Johansen estimator is thus a finite product. This observation is going to be important in the next chapter.

Just like the Nelson-Aalen estimator, the Aalen-Johansen estimator is consistent as well. A proof of the following theorem can be found in [1, section IV.4.2]. It is once again based on Lenglart's inequality, but this time the proof also uses the so-called Duhamel's equation.

In Appendix E, we lay out a novel proof of Duhamel's equation for the non-generalized product integral. It is (in our humble opinion) a very nice proof and, as such, we strongly encourage the reader to check it out.

Theorem 6.3.1. Let $s < v$. If

$$\overline{Y}_j^{(n)}(u) \xrightarrow{p} \infty \text{ for all } j$$

for almost every $u \in [s, v]$, then

$$\sup_{t \in [s,v]} \|\mathbf{P}_n^{\wedge}(s,t) - \mathbf{P}(s,t)\| \xrightarrow{p} 0$$

where we write $\overline{Y}_j^{(n)}$ & $\mathbf{P}_n^{\wedge}$ instead of $\overline{Y}_j$ & $\mathbf{P}^{\wedge}$ to remind you that they are dependent on the total number of individuals.

One way to prove the above theorem is to transfer the convergence in probability (consistency) of the Nelson-Aalen estimator,

$$\mathbf{A}^{\wedge} \xrightarrow{p} \mathbf{A}$$

to the Aalen-Johansen estimator

$$\mathbf{P}^{\wedge} = \prod (\mathbf{I} + d\mathbf{A}^{\wedge}) \xrightarrow{p} \prod (\mathbf{I} + d\mathbf{A}) = \mathbf{P}$$

This works because the the product integral is continuous and the continuous mapping theorem (Example D.0.2). Again, the details can be found in [1].

Corollary 6.3.2. Let $s \leq t$. Then, under the same condition as in Theorem 6.3.1,

$$\mathbf{P}_n^{\wedge}(s,t) \xrightarrow{p} \mathbf{P}(s,t)$$

Chapter 7

Non-Markov case

The Aalen-Johansen estimator is no longer consistent in the non-Markov case because we used the Markov assumption to derive the Chapman-Kolmogorov equations, which were used to derive a Volterra equation, from which we deduced that

$$\mathbf{P} = \prod (\mathbf{I} + d\mathbf{A})$$

As such, we must turn elsewhere. We introduce (as a product integral) and prove the consistency of the so-called landmark Aalen-Johansen estimator.

In order to define the landmark Aalen-Johansen estimator (henceforth simply called the LMAJ estimator), we first define

$$\overline{N}_{jk}^{(LM)}(t) = \sum_{i=1}^{n} N_{ijk}(t)[X_i(s) = \ell]$$
$$\overline{Y}_{j}^{(LM)}(t) = \sum_{i=1}^{n} Y_{ij}(t)[X_i(s) = \ell]$$

Define the matrices $\overline{\mathbf{N}}^{(LM)}$ & $\overline{\mathbf{Y}}_{D}^{(LM)}$ in the same way as $\overline{\mathbf{N}}$ & $\overline{\mathbf{Y}}_{D}$. If we let

$$\mathbf{A}^{\wedge}(t) = \sum_{\ell : t_\ell \leq t} \left(\overline{\mathbf{Y}}_{D}^{(LM)}(t_\ell) \right)^{-1} d\overline{\mathbf{N}}^{(LM)}(t_\ell)$$

then the LMAJ estimator is given by

$$P_{\ell m}^{LMAJ}(s,t) = \left[\prod_{s}^{t} (\mathbf{I} + d\mathbf{A}_{(LM)}^{\wedge}) \right]_{\ell m}$$

Note that the LMAJ estimator is basically the Aalen-Johansen estimator, but we only consider those individuals that were at state ℓ at time s.

7.1 Consistency proof

Before we dive into the consistency proof of the LMAJ estimator, we give two definitions and a lemma.

Define the $1 \times K$ matrix $\pi^{\wedge}(0)$ by

$$[\pi^{\wedge}(0)]_{1k} = \frac{1}{n} \sum_{i=1}^{n} [X_i(0) = k]$$

and define $\mathbf{P}^{\wedge}(t) = \pi^{\wedge}(0)\mathbf{P}^{\wedge}(0,t)$.

Lemma 7.1.1. Let $t \in [0, \infty)$. Under certain conditions [see 9, Theorem 5.3.1],

$$\sup_{u \in [0,t]} \|\mathbf{P}_n^{\wedge}(u) - \mathbf{P}(u)\| \xrightarrow{p} 0$$

where we write $\mathbf{P}_n^{\wedge}$ instead of $\mathbf{P}^{\wedge}$ to remind you that it is dependent on the total number of individuals.

Corollary 7.1.2. Let $t \in [0, \infty)$. Under the same conditions as Lemma 7.1.1,

$$\mathbf{P}_n^{\wedge}(t) \xrightarrow{p} \mathbf{P}(t)$$

With Corollary 7.1.2 under our belt, it is finally time for the consistency proof.

Theorem 7.1.3. Let us fix ℓ and s. Under the same conditions as Lemma 7.1.1 and the condition that $P(X^{\sim}(s) = \ell) > 0$,

$$P_{\ell m}^{LMAJ}(s,t) \xrightarrow{p} P_{\ell m}(s,t)$$

for all $s \leq t$.

Proof. The following proof is strongly based on [7, Appendix I].

Let us fix ℓ and s. Next, we define the multi-state process $X^{\star\sim}(t)$ with state space $\{\pm 1, \ldots, \pm K\}$ by

- $X^{\star\sim}(t) = +X^{\sim}(t)$ for $t < s$ and for $t \geq s$ if $X^{\sim}(s) = \ell$
- $X^{\star\sim}(t) = -X^{\sim}(t)$ for $t \geq s$ if $X^{\sim}(s) \neq \ell$

Note that we are, just like in the definition of the LMAJ estimator, conditioning on whether the process is in state ℓ at time s.

Note that, since $X^{\star\sim}(t)$ depends on the past through $X^{\star\sim}(s)$ for all $t > s$, the process $X^{\star\sim}$ is not Markov even if $X^{\sim}$ is.

Since the state $m \geq 1$ can only be reached if $X^{\sim}(s) = \ell$, it follows that, for $t \geq s$ and $m \geq 1$,

$$P^{\star}_{m}(t) \equiv P(X^{\star\sim}(t) = m) = P(X^{\sim}(t) = m, X^{\sim}(s) = \ell)$$

so

$$P_{\ell m}(s,t) = P(X^{\sim}(t) = m \mid X^{\sim}(s) = \ell) = \frac{P(X^{\sim}(t) = m, X^{\sim}(s) = \ell)}{P(X^{\sim}(s) = \ell)} = \frac{P^{\star}_{m}(t)}{P^{\star}_{\ell}(s)}$$

because $P^{\star}_{\ell}(s) = P(X^{\sim}(s) = \ell, X^{\sim}(s) = \ell) = P(X^{\sim}(s) = \ell)$.

By Corollary 7.1.2, the estimators $P^{\star\wedge}_{m}(t)$ & $P^{\star\wedge}_{\ell}(s)$ of $P^{\star}_{m}(t)$ & $P^{\star}_{\ell}(s)$ are consistent, so, by the continuous mapping theorem (Example D.0.3), their quotient consistently estimates $P_{\ell m}(s,t)$ if $P^{\star}_{\ell}(s) = P(X^{\sim}(s) = \ell) > 0$, which we assumed to be true.

Now, define $\pi^{\star\wedge}(0)$ & $\mathbf{A}^{\star\wedge}$ by plugging in $X^{\star}$ (instead of X) into the definition of $\pi^{\wedge}(0)$ & $\mathbf{A}^{\wedge}$. Then

$$\begin{aligned}
P^{\star\wedge}_{m}(t) &= \left[\pi^{\star\wedge}(0) \prod_{0}^{t} (\mathbf{I} + d\mathbf{A}^{\star\wedge})\right]_{1m} \\
&\overset{!}{=} \left[\pi^{\star\wedge}(0) \prod_{0}^{s} (\mathbf{I} + d\mathbf{A}^{\star\wedge}) \prod_{s}^{t} (\mathbf{I} + d\mathbf{A}^{\star\wedge})\right]_{1m} \\
&= \sum_{j} \left[\pi^{\star\wedge}(0) \prod_{0}^{s} (\mathbf{I} + d\mathbf{A}^{\star\wedge})\right]_{1j} \left[\prod_{s}^{t} (\mathbf{I} + d\mathbf{A}^{\star\wedge})\right]_{jm} \\
P^{\star\wedge}_{\ell}(s) &= \left[\pi^{\star\wedge}(0) \prod_{0}^{s} (\mathbf{I} + d\mathbf{A}^{\star\wedge})\right]_{1m}
\end{aligned}$$

because the product integral is finite product (see page 31).

If $j \neq \ell$, then

$$\left[\prod_{s}^{t} (\mathbf{I} + d\mathbf{A}^{\star\wedge})\right]_{jm} = 0$$

because, just before the first jump time after s, every individual is either at state ℓ or has been redirected to a negative state, from which the state $m \geq 1$ can never be reached.

Consequently,

$$\begin{aligned}\frac{\mathrm{P}_{\mathrm{m}}^{\star\wedge}(\mathrm{t})}{\mathrm{P}_{\ell}^{\star\wedge}(\mathrm{s})} &= \frac{[\pi^{\star\wedge}(0)\prod_{0}^{\mathrm{s}}(\mathbf{I}+\mathrm{d}\mathbf{A}^{\star\wedge})]_{1\ell}[\prod_{\mathrm{s}}^{\mathrm{t}}(\mathbf{I}+\mathrm{d}\mathbf{A}^{\star\wedge})]_{\ell\mathrm{m}}}{[\pi^{\star\wedge}(0)\prod_{0}^{\mathrm{s}}(\mathbf{I}+\mathrm{d}\mathbf{A}^{\star\wedge})]_{1\ell}} \\ &= \left[\prod_{\mathrm{s}}^{\mathrm{t}}(\mathbf{I}+\mathrm{d}\mathbf{A}^{\star\wedge})\right]_{\ell\mathrm{m}}\end{aligned}$$

which looks delightfully similar to the LMAJ estimator; the only difference being $\mathbf{A}^{\star\wedge}$ versus $\mathbf{A}^{\wedge}_{(\mathrm{LM})}$.

If $\mathrm{t} \geq \mathrm{s}$, then $\mathrm{X}^{\star\sim}$ only takes values in either the positive or the negative states, depending on whether $\mathrm{X}^{\sim}(\mathrm{s}) = \ell$ or $\mathrm{X}^{\sim}(\mathrm{s}) \neq \ell$, so there can be no jumps between states with different signs. As such, for $\mathrm{t} \geq \mathrm{s}$, the $2\mathrm{K} \times 2\mathrm{K}$ matrix $\mathbf{A}^{\star\wedge}(\mathrm{t})$ is a block diagonal matrix, consisting of two $\mathrm{K} \times \mathrm{K}$ blocks representing the positive states and the negative states:

$$\mathbf{A}^{\star\wedge}(\mathrm{t}) = \mathbf{A}_{+}^{\star\wedge}(\mathrm{t}) \oplus \mathbf{A}_{-}^{\star\wedge}(\mathrm{t})$$

Since $\oplus$ distributes over regular matrix addition and multiplication, $\mathbf{I}_{2\mathrm{K}} = \mathbf{I}_{\mathrm{K}} \oplus \mathbf{I}_{\mathrm{K}}$ and the product integral of $\mathbf{A}^{\star\wedge}$ is a finite product,

$$\prod_{\mathrm{s}}^{\mathrm{t}}(\mathbf{I}+\mathrm{d}\mathbf{A}^{\star\wedge}) = \prod_{\mathrm{s}}^{\mathrm{t}}(\mathbf{I}+\mathrm{d}\mathbf{A}_{+}^{\star\wedge}) \oplus \prod_{\mathrm{s}}^{\mathrm{t}}(\mathbf{I}+\mathrm{d}\mathbf{A}_{-}^{\star\wedge})$$

Since $\ell, \mathrm{m} \geq 1$, we hence find that

$$\left[\prod_{\mathrm{s}}^{\mathrm{t}}(\mathbf{I}+\mathrm{d}\mathbf{A}^{\star\wedge})\right]_{\ell\mathrm{m}} = \left[\prod_{\mathrm{s}}^{\mathrm{t}}(\mathbf{I}+\mathrm{d}\mathbf{A}_{+}^{\star\wedge})\right]_{\ell\mathrm{m}}$$

Let $j, k \geq 1$ and $j \neq k$. If $t \geq s$, then $dN^{\star}_{ijk}$ (with $N^{\star}_{ijk}$ the censored counting process) and the censored at-risk process and $Y^{\star}_{ij}$ used in

$$A^{\star\wedge}_{jk}(t) = \sum_{\ell : t_\ell \leq t} \frac{d\overline{N}_{jk}(t_\ell)}{\overline{Y}_j(t_\ell)}$$

are given by

$$\begin{aligned} dN^{\star}_{ijk}(t) &= dN_{ijk}(t) \cdot [X_i(s) = \ell] \\ Y^{\star}_{ij}(t) &= Y_{ij}(t) \cdot [X_i(s) = \ell] \end{aligned}$$

because, if $X_i(s) \neq \ell$, then $X^{\star}_i(t) < 0$ for $t \geq s$. For $t \geq s$, we hence find that

$$\begin{aligned} d\overline{N}^{\star}_{jk}(t) &= d\overline{N}^{(LM)}_{jk}(t) \\ \overline{Y}^{\star}_j(t) &= \overline{Y}^{(LM)}_j(t) \end{aligned}$$

so $\mathbf{A}^{\star\wedge}_{+}(t) = \mathbf{A}^{\wedge}_{(LM)}(t)$, which finally yields that

$$\begin{aligned} \frac{P^{\star\wedge}_m(t)}{P^{\star\wedge}_\ell(s)} &= \left[\prod_s^t (\mathbf{I} + d\mathbf{A}^{\star\wedge}) \right]_{\ell m} = \left[\prod_s^t (\mathbf{I} + d\mathbf{A}^{\star\wedge}_{+}) \right]_{\ell m} \\ &= \left[\prod_s^t (\mathbf{I} + d\mathbf{A}^{\wedge}_{(LM)}) \right]_{\ell m} = P^{LMAJ}_{\ell m}(s,t) \end{aligned}$$

is a consistent estimator for $P_{\ell m}(s,t)$. □

Chapter 8

Further research

Let us address the elephant in the room first. In light of the scope of the thesis, we have chosen not to include proofs of many key results (in their full generality). Notwithstanding, their inclusion would have been nice. In further research we could flesh out those proofs, which would make our thesis more self-contained.

Secondly, throughout the thesis we assumed that $X_i^{\sim}$'s are independent and identically distributed, which implies a homogeneous population. This is not the most realistic scenario. For example, you might be sicklier than your neighbor, meaning that the $X_i^{\sim}$'s would not be identically distributed. As such, investigating whether there is a consistent estimator of transition probabilities under weaker assumptions on the $X_i^{\sim}$'s would be a good direction for further research as well. If we could modify the consistency proof of the LMAJ estimator, that would be ideal.

Thirdly, up until now we have only considered non-Markov multi state models with right censoring, but, as mentioned priorly, there are more ways to censor, such as left and interval censoring. Having a consistent estimator for transition probabilities in non-Markov multi-state models with more general censoring would of course be nice. As such, investigating whether there is a consistent estimator of transition probabilities under more general censoring would also be a good direction for further research. Again, if we could modify the consistency proof of the LMAJ estimator, that would be ideal.

Finally, and most importantly, we have been estimating

$$P_{\ell m}(s,t) = P(X^{\sim}(t) = m \mid X^{\sim}(s) = \ell)$$

but really we want an estimator for

$$P_{\ell m}(s,t \mid \mathcal{F}_{s-}) = P(X^{\sim}(t) = m \mid X^{\sim}(s) = \ell, \mathcal{F}_{s-})$$

because, in a sense, $P_{\ell m}(s,t)$ averages over all possible $\mathcal{F}_{s-}$'s. Of course, $P_{\ell m}(s,t \mid \mathcal{F}_{s-})$ is much harder to estimate than $P_{\ell m}(s,t)$. Nevertheless, that does not mean we should not try. The final direction for further research we suggest is therefore: investigating whether $P_{\ell m}(s,t \mid \mathcal{F}_{s-})$ can be estimated consistently as well.

Appendix A

Left and right limits

Below you will find two results on functions with left and right limits that we have opted to leave out of the main text and instead showcase together here.

Theorem A.0.1. If $\mathbf{Z}$ is a matrix-valued function on $[a, b]$ that has left and right limits, then it is bounded.

Proof. We will prove this statement by contraposition.

Suppose that $\mathbf{Z}$ is unbounded. Then there exists (x_n) such that $x_n \to x \in [a, b]$ and $\|\mathbf{Z}(x_n)\| \to \infty$ (use Bolzano-Weierstraß if necessary). Since $x_n \to x$, it must have an increasing or decreasing subsequence (y_n). Either way, $\|\mathbf{Z}(y_n)\| \to \infty$. If $y_n \uparrow x$, then $\|\mathbf{Z}(x-)\| = \infty$, so the left limit $\mathbf{Z}(x-)$ does not exist. If $y_n \downarrow x$, then the right limit $\mathbf{Z}(x+)$ does not exist. □

The proof above is basically an adaptation of [10].

Theorem A.0.2. If $f : [0, \infty) \to \mathbb{R}$ has left and right limits, then it has only countably many discontinuities.

Proof. See [11]. The author feels very little for nearly verbatim copying the proof. □

Appendix B

Weak law of large numbers

Below we will present a proof of the weak law of large numbers, albeit not the most general one; we will assume finite variance even though it can be proved without that assumption.

The following proof has been adapted from [2, Theorem 10.2.1].

Theorem B.0.1. Let (X_n) be a sequence of independent and identically distributed real-valued random variables with finite second moment. Let

$$\begin{aligned} E(X_1) &= \mu \\ \operatorname{var}(X_1) &= \sigma^2 \end{aligned}$$

which are both finite. Define

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

Then $\overline{X}_n \xrightarrow{p} \mu$.

Proof. Because the X_i's are independent and identically distributed,

$$\begin{aligned} E(\overline{X}_n) &= \frac{E(X_1 + \cdots + X_n)}{n} = \frac{E(X_1) + \cdots + E(X_n)}{n} = \mu \\ \operatorname{var}(\overline{X}_n) &= \frac{\operatorname{var}(X_1 + \cdots + X_n)}{n^2} = \frac{\operatorname{var}(X_1) + \cdots + \operatorname{var}(X_n)}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

Let $\epsilon > 0$. Then

$$P\Big(|\overline{X}_n - \mu| > \epsilon\Big) = P\Big(|\overline{X}_n - \mu|^2 > \epsilon^2\Big) = P(\{\omega \in \Omega : f(\omega) > \epsilon^2\})$$

with

$$f(\cdot) = |\overline{X}_n(\cdot) - \mu|^2 = (\overline{X}_n(\cdot) - \mu)^2$$

Clearly $\overline{X}_n(\cdot) - \mu$ is measurable. Since $\mathbb{R} \ni x \mapsto x^2 \in \mathbb{R}$ is continuous and hence Borel measurable, f is measurable as the composition of measurable functions, so

$$\begin{aligned} P\Big(|\overline{X}_n - \mu| > \epsilon\Big) &= P(\{\omega \in \Omega : f(\omega) > \epsilon^2\}) \leq \frac{1}{\epsilon^2} \int f \, dP \\ &= \frac{E(f)}{\epsilon^2} = \frac{E\Big((\overline{X}_n - \mu)^2\Big)}{\epsilon^2} = \frac{E\Big(\Big(\overline{X}_n - E(\overline{X}_n)\Big)^2\Big)}{\epsilon^2} \\ &= \frac{\mathrm{var}(\overline{X}_n)}{\epsilon^2} = \frac{1}{\epsilon^2}\frac{\sigma^2}{n} \end{aligned}$$

It follows that

$$\lim_n P\Big(|\overline{X}_n - \mu| > \epsilon\Big) = 0$$

That is, $\overline{X}_n \xrightarrow{p} \mu$. □

Appendix C

Volterra equation for P

In this appendix, we will prove Theorem 6.2.1. That is to say, we will show that

$$\mathbf{P}(s,t) = \mathbf{I} + \int_s^t \mathbf{P}(s,u)\boldsymbol{\alpha}(u)\,du$$

for $s \leq t$. However, first we state (and prove) a preparatory lemma.

Lemma C.0.1. If $s \leq t$, then

$$\lim_{\delta \downarrow 0} \frac{1}{\delta} \int_t^{t+\delta} \mathbf{P}(s,u)\,du = \mathbf{P}(s,t)$$

Proof. Since $u \mapsto \mathbf{P}(s,u)$ is cadlag, it is right continuous, so

$$\begin{gathered}\text{if } \epsilon > 0, \text{ then } \exists \delta > 0 \text{ such that} \\ \text{if } t < u < t+\delta, \text{ then } \|\mathbf{P}(s,u) - \mathbf{P}(s,t)\| < \epsilon.\end{gathered}$$

As such,

$$\begin{aligned}\left\| \frac{1}{\delta} \int_t^{t+\delta} \mathbf{P}(s,u)\,du - \mathbf{P}(s,t) \right\| &= \left\| \frac{1}{\delta} \int_t^{t+\delta} \Big(\mathbf{P}(s,u) - \mathbf{P}(s,t) \Big)\,du \right\| \\ &\leq \frac{1}{\delta} \int_t^{t+\delta} \|\mathbf{P}(s,u) - \mathbf{P}(s,t)\|\,du \\ &\leq \frac{1}{\delta} \int_t^{t+\delta} \epsilon\,du = \epsilon\end{aligned}$$

The desired result now follows from the definition of the right-sided limit. □

Proof of Theorem 6.2.1. By Lebesgue's dominated convergence theorem,

$$\begin{aligned}
\mathbf{P}(s,t) - \mathbf{I} &= \mathbf{P}(s,t) - \mathbf{P}(s,s) \\
&= \lim_{h \downarrow 0} \frac{1}{h} \int_t^{t+h} \mathbf{P}(s,u)\, du - \lim_{h \downarrow 0} \frac{1}{h} \int_s^{s+h} \mathbf{P}(s,u)\, du \\
&= \lim_{h \downarrow 0} \frac{1}{h} \left(\int_t^{t+h} \mathbf{P}(s,u)\, du - \int_s^{s+h} \mathbf{P}(s,u)\, du \right) \\
&= \lim_{h \downarrow 0} \frac{1}{h} \left(\int_{s+h}^{t+h} \mathbf{P}(s,u)\, du - \int_s^t \mathbf{P}(s,u)\, du \right) \\
&= \lim_{h \downarrow 0} \frac{1}{h} \left(\int_s^t \mathbf{P}(s,u+h)\, du - \int_s^t \mathbf{P}(s,u)\, du \right) \\
&= \lim_{h \downarrow 0} \int_s^t \frac{\mathbf{P}(s,u+h) - \mathbf{P}(s,u)}{h}\, du \\
&= \lim_{h \downarrow 0} \int_s^t \mathbf{P}(s,u) \frac{\mathbf{P}(u,u+h) - \mathbf{I}}{h}\, du \\
&\overset{!}{=} \int_s^t \mathbf{P}(s,u) \boldsymbol{\alpha}(u)\, du
\end{aligned}$$

To justify the use of the dominated convergence theorem, we need to show that

$$\mathbf{P}(s,u) \frac{\mathbf{P}(u,u+h) - \mathbf{I}}{h} = \frac{\mathbf{P}(s,u+h) - \mathbf{P}(s,u)}{h} \text{ is measurable} \tag{C.1}$$

$$\lim_{h \downarrow 0} \mathbf{P}(s,u) \frac{\mathbf{P}(u,u+h) - \mathbf{I}}{h} \overset{\text{ae}}{=} \mathbf{P}(s,u) \boldsymbol{\alpha}(u) \tag{C.2}$$

$$\left| \left[\mathbf{P}(s,u) \frac{\mathbf{P}(u,u+h) - \mathbf{I}}{h} \right]_{jk} \right| \overset{\text{ae}}{\leq} g(u) \tag{C.3}$$

for $u \in [s,t]$ and some $[0,\infty]$-valued integrable function g. We can then use dominated convergence componentwise.

Since we are done once we have shown those three equations hold, let us finish up the proof by doing so.

1. Since $x \mapsto \mathbf{P}(s,x)$ is cadlag, it has left and right limits, so it is bounded on $[s,t]$ and continuous almost everywhere (see Appendix A). As such, by [2, Theorem 2.5.4], it's Riemann integrable and hence Lebesgue integrable (and thus measurable).

2. Clearly

$$\lim_{h \downarrow 0} \mathbf{P}(s,u) \frac{\mathbf{P}(u,u+h) - \mathbf{I}}{h} = \mathbf{P}(s,u) \lim_{h \downarrow 0} \frac{\mathbf{P}(u,u+h) - \mathbf{I}}{h}$$

Since $X^{\sim}$ is cadlag, so is

$$u \mapsto [\mathbf{P}(u,x)]_{\ell m} = P_{\ell m}(u,x) = P(X^{\sim}(x) = m \mid X^{\sim}(u) = \ell)$$

As such, $u \mapsto \mathbf{P}(u,x)$ is cadlag and hence has left and right limits. By Theorem A.0.2, it is continuous almost everywhere, so

$$\begin{aligned} \frac{\mathbf{P}(u,u+h)-\mathbf{I}}{h} &\overset{ae}{\equiv} \frac{\mathbf{P}(u-,u+h)-\mathbf{I}}{h} \Rightarrow \\ \lim_{h\downarrow 0} \frac{\mathbf{P}(u,u+h)-\mathbf{I}}{h} &\overset{ae}{\equiv} \lim_{h\downarrow 0} \frac{\mathbf{P}(u-,u+h)-\mathbf{I}}{h} = \boldsymbol{\alpha}(u) \end{aligned}$$

Equation (C.2) now follows.

3. As for the third one, since

$$\left| \left[\mathbf{P}(s,u) \frac{\mathbf{P}(u,u+h)-\mathbf{I}}{h} \right]_{jk} \right| \leq \left\| \mathbf{P}(s,u) \frac{\mathbf{P}(u,u+h)-\mathbf{I}}{h} \right\|$$

we need only find a $[0,\infty]$-valued integrable function g such that

$$\left\| \mathbf{P}(s,u) \frac{\mathbf{P}(u,u+h)-\mathbf{I}}{h} \right\| \overset{ae}{\leq} g(u)$$

Since $\mathbf{P}(s,u)$ is row stochastic,

$$\begin{aligned} \left\| \mathbf{P}(s,u) \frac{\mathbf{P}(u,u+h)-\mathbf{I}}{h} \right\| &\leq \|\mathbf{P}(s,u)\| \cdot \left\| \frac{\mathbf{P}(u,u+h)-\mathbf{I}}{h} \right\| \\ &= \left\| \frac{\mathbf{P}(u,u+h)-\mathbf{I}}{h} \right\| \\ &\overset{ae}{\equiv} \left\| \frac{\mathbf{P}(u-,u+h)-\mathbf{I}}{h} \right\| \\ &\overset{!}{\leq} \|\boldsymbol{\alpha}(u)\| + 1 = g(u) \end{aligned}$$

because, by right continuity, $\exists \delta > 0$ such that if $h < \delta$, then

$$\left\| \frac{\mathbf{P}(u-,u+h)-\mathbf{I}}{h} \right\| - \|\boldsymbol{\alpha}(u)\| \leq \left\| \frac{\mathbf{P}(u-,u+h)-\mathbf{I}}{h} - \boldsymbol{\alpha}(u) \right\| < 1$$

(reverse triangle inequality). Clearly $g \geq 0$. Since $\boldsymbol{\alpha}$ is integrable on $[s,t]$, so is $\|\boldsymbol{\alpha}\|$ and hence $g(\cdot) = \|\boldsymbol{\alpha}(\cdot)\| + 1$ (because $[s,t]$ is bounded).

We have shown that (C.1), (C.2) and (C.3) hold, which concludes the proof. □

Appendix D

Continuous mapping theorem

The following proof was greatly inspired by [12].

Theorem D.0.1. Let $(S, \|\cdot\|)$ be normed spaces and $(X_n), X$ be S-valued random variables. Suppose that $g : S \to S'$ (where $(S', \|\cdot\|')$ is another normed space) is a function whose set of discontinuities D_g satisfies $P(X \in D_g) = 0$.

(if g is continuous, then $D_g = \varnothing$, so vacuously $P(X \in D_g) = 0$ and hence the name "continuous mapping theorem")

If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$.

Proof. Let us fix $\epsilon > 0$. Define for all $\delta > 0$,

$$B_\delta = \{x \in S : x \notin D_g \text{ and } \exists y \in S \text{ such that } \|x - y\| < \delta \text{ and } \|g(x) - g(y)\|' > \epsilon\}$$

Because g is continuous on B_δ,

$$\lim_{\delta \downarrow 0} B_\delta \equiv \{x \in S : \lim_{\delta \downarrow 0} [x \in B_\delta] = 1\} = \varnothing$$

Suppose that $\|g(X_n) - g(X)\|' > \epsilon$. Then at least one of the following is true:

- either $X \in B_\delta$
- or $X \in D_g$
- or $\|X_n - X\| \geq \delta$.

In terms of probabilities,

$$P\Big(\|g(X_n) - g(X)\|' > \epsilon\Big) \leq P(X \in D_g) + P(X \in B_\delta) + P\Big(\|X_n - X\| \geq \delta\Big)$$

By assumption, $P(X \in D_g) = 0$.

Let $\eta > 0$. Since $\lim_{\delta \downarrow 0} B_\delta = \varnothing$, choose δ such that $P(X \in B_\delta) < \eta/2$.
Now, since $X_n \xrightarrow{p} X$, eventually

$$P\Big(\|X_n - X\| \geq \delta\Big) < \eta/2$$

That is, eventually

$$P\Big(\|g(X_n) - g(X)\|' > \epsilon\Big) < \eta$$

Therefore

$$\lim_n P\Big(\|g(X_n) - g(X)\|' > \epsilon\Big) = 0$$

so $g(X_n) \xrightarrow{p} g(X)$. □

Example D.0.2. Let $\alpha \in S$ and X be the random variable defined by

$$X(\omega) = \alpha \text{ for all } \omega \in \Omega.$$

Then, by the continuous mapping theorem, if g is continuous, then X_n is a consistent estimator of α, so $g(X_n)$ is a consistent estimator of $g(\alpha)$. Indeed,

$$X_n \xrightarrow{p} \alpha \Leftrightarrow X_n \xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X) \Leftrightarrow g(X_n) \xrightarrow{p} g(\alpha)$$

Example D.0.3. Let $\alpha, \beta \in \mathbb{R}$ and $(X_n), (Y_n)$ be real-valued random variables such that $X_n \xrightarrow{p} \alpha$ and $Y_n \xrightarrow{p} \beta$. We will show that if $\beta \neq 0$, then $X_n/Y_n \xrightarrow{p} \alpha/\beta$.

Note that $(X_n, Y_n) \xrightarrow{p} (\alpha, \beta)$. Indeed,

$$\begin{aligned} P(\|(X_n, Y_n) - (\alpha, \beta)\| > \epsilon) &= P(\|(X_n - \alpha, Y_n - \beta)\| > \epsilon) \\ &\leq P(|X_n - \alpha| + |Y_n - \beta| > \epsilon) \\ &\leq P(|X_n - \alpha| > \epsilon/2) + P(|Y_n - \beta| > \epsilon/2) \\ &= 0 + 0 = 0 \end{aligned}$$

because $\|(x, y)\| \leq |x| + |y|$ and

if $x + y > 2$, then at least one of the following is true: either $x > 1$ or $y > 1$.

Since $g(x, y) = x/y$ is a continuous function for $y \neq 0$, by the continuous mapping theorem, if $\beta \neq 0$, then $X_n/Y_n \xrightarrow{p} \alpha/\beta$.

Appendix E

Duhamel's equation

In this appendix, we present a novel proof of the so-called Duhamel's equation.

Theorem E.0.1. Let $\mathbf{F}$ and $\mathbf{G}$ be matrix-valued functions on $[a, b]$ that are absolutely continuous. Then

$$\mathcal{P}(a, b; \mathbf{F}) - \mathcal{P}(a, b; \mathbf{G}) = \int_a^b \mathcal{P}(a, x; \mathbf{F})\mathbf{h}(x)\mathcal{P}(x, b; \mathbf{G})\, dx$$

where $\mathbf{h} = \mathbf{f} - \mathbf{g}$.

The following lemma lies at the heart of the proof.

Lemma E.0.2. Let $n \geq m \geq 1$ and

$$c_{mn}(x) = \int_a^x \mathcal{I}_{n-m}(a, t; \mathbf{F})\mathbf{h}(t)\mathcal{I}_{m-1}(t, x; \mathbf{G})\, dt$$

Then

$$\int_a^b c_{mn}(x_1)\mathbf{g}(x_1)\, dx_1 = c_{(m+1)(n+1)}(b)$$

Proof. By Fubini's theorem,

$$\int_a^b c_{mn}(x_1)\mathbf{g}(x_1)\,dx_1 =$$

$$\int_a^b \left(\int_a^{x_1} \mathcal{F}_{n-m}(a, x_2; \mathbf{F})\mathbf{h}(x_2)\mathcal{F}_{m-1}(x_2, x_1; \mathbf{G})\,dx_2 \right)\mathbf{g}(x_1)\,dx_1 =$$

$$\int_a^b \int_a^{x_1} \mathcal{F}_{n-m}(a, x_2; \mathbf{F})\mathbf{h}(x_2)\mathcal{F}_{m-1}(x_2, x_1; \mathbf{G})\mathbf{g}(x_1)\,dx_2 dx_1 \stackrel{!}{=}$$

$$\int_a^b \int_{x_2}^b \mathcal{F}_{n-m}(a, x_2; \mathbf{F})\mathbf{h}(x_2)\mathcal{F}_{m-1}(x_2, x_1; \mathbf{G})\mathbf{g}(x_1)\,dx_1 dx_2 =$$

$$\int_a^b \mathcal{F}_{n-m}(a, x_2; \mathbf{F})\mathbf{h}(x_2)\left(\int_{x_2}^b \mathcal{F}_{m-1}(x_2, x_1; \mathbf{G})\mathbf{g}(x_1)\,dx_1 \right)dx_2 =$$

$$\int_a^b \mathcal{F}_{n-m}(a, x_2; \mathbf{F})\mathbf{h}(x_2)\mathcal{F}_m(x_2, b; \mathbf{G})\,dx_2 =$$

$$c_{(m+1)(n+1)}(b)$$

because, by definition,

$$\mathcal{F}_m(x_2, b; \mathbf{G}) = \int_{x_2}^b \mathcal{F}_{m-1}(x_2, x_1; \mathbf{G})\mathbf{g}(x_1)\,dx_1$$

We still need to justify our use of Fubini's theorem. By Lemma 4.1.3,

$$\begin{aligned}
\|\mathcal{F}_{n-m}(a, x_2; \mathbf{F})\| &\leq \frac{1}{(n-m)!}\left(\int_a^{x_2} \|\mathbf{f}(t)\|\,dt \right)^{n-m} \\
&\leq \frac{1}{(n-m)!}\left(\int_a^b \|\mathbf{f}(t)\|\,dt \right)^{n-m} \\
&= J^{n-m}/(n-m)!
\end{aligned}$$

with

$$J = \int_a^b \|\mathbf{f}(t)\|\,dt$$

Similarly,

$$\|\mathcal{F}_{m-1}(x_2, x_1; \mathbf{G})\| \leq K^{m-1}/(m-1)!$$

with

$$K = \int_a^b \|\mathbf{g}(t)\|\,dt$$

As such, we may use Fubini's theorem because

$$\left\| \int_a^b \int_a^{x_1} \mathcal{F}_{n-m}(a, x_2; \mathbf{F}) \mathbf{h}(x_2) \mathcal{F}_{m-1}(x_2, x_1; \mathbf{G}) \mathbf{g}(x_1)\, dx_2 dx_1 \right\| \leq$$

$$\int_a^b \int_a^{x_1} \|\mathcal{F}_{n-m}(a, x_2; \mathbf{F})\| \cdot \|\mathbf{h}(x_2)\| \cdot \|\mathcal{F}_{m-1}(x_2, x_1; \mathbf{G})\| \cdot \|\mathbf{g}(x_1)\|\, dx_2 dx_1 \leq$$

$$\int_a^b \int_a^{x_1} \frac{J^{n-m}}{(n-m)!} \cdot \|\mathbf{h}(x_2)\| \cdot \frac{K^{m-1}}{(m-1)!} \cdot \|\mathbf{g}(x_1)\|\, dx_2 dx_1 \leq$$

$$\frac{J^{n-m}}{(n-m)!} \frac{K^{m-1}}{(m-1)!} \int_a^b \int_a^{x_1} \|\mathbf{h}(x_2)\| \cdot \|\mathbf{g}(x_1)\|\, dx_2 dx_1 \leq$$

$$\frac{J^{n-m}}{(n-m)!} \frac{K^{m-1}}{(m-1)!} \int_a^b \int_a^{b} \|\mathbf{h}(x_2)\| \cdot \|\mathbf{g}(x_1)\|\, dx_2 dx_1 =$$

$$\frac{J^{n-m}}{(n-m)!} \frac{K^{m-1}}{(m-1)!} \int_a^b \|\mathbf{h}(x_2)\|\, dx_2 \int_a^b \|\mathbf{g}(x_1)\|\, dx_1 < \infty$$

because $\mathbf{h}$ (as the difference of integrable functions) and $\mathbf{g}$ are integrable. □

Lemma E.0.3. If $n \geq 1$, then

$$\mathcal{F}_n(a, b; \mathbf{F}) - \mathcal{F}_n(a, b; \mathbf{G}) = \sum_{m=1}^{n} c_{mn}(b)$$

Proof. We proceed by induction on n. Clearly

$$\begin{aligned} \mathcal{F}_1(a, b; \mathbf{F}) - \mathcal{F}_1(a, b; \mathbf{G}) &= \int_a^b \mathbf{f}(x_1)\, dx_1 - \int_a^b \mathbf{g}(x_1)\, dx_1 \\ &= \int_a^b \mathbf{h}(x_1)\, dx_1 = c_{11}(b) \end{aligned}$$

establishing the base case. Suppose that

$$\mathcal{F}_n(a, b; \mathbf{F}) - \mathcal{F}_n(a, b; \mathbf{G}) = \sum_{m=1}^{n} c_{mn}(b)$$

We will show that

$$\mathcal{F}_{n+1}(a, b; \mathbf{F}) - \mathcal{F}_{n+1}(a, b; \mathbf{G}) = \sum_{m=1}^{n+1} c_{m(n+1)}(b)$$

Since $\mathbf{f} = \mathbf{h} + \mathbf{g}$,

$$\begin{aligned}
&\mathcal{I}_{n+1}(a, b; \mathbf{F}) - \mathcal{I}_{n+1}(a, b, \mathbf{G}) = \\
&\int_a^b \int_a^{x_1} \cdots \int_a^{x_n} \mathbf{f}(x_{n+1}) \cdots \mathbf{f}(x_2)\mathbf{f}(x_1)\, dx_{n+1} \cdots dx_2 dx_1 - \\
&\int_a^b \int_a^{x_1} \cdots \int_a^{x_n} \mathbf{g}(x_{n+1}) \cdots \mathbf{g}(x_2)\mathbf{g}(x_1)\, dx_{n+1} \cdots dx_2 dx_1 = \\
&\int_a^b \int_a^{x_1} \cdots \int_a^{x_n} \mathbf{f}(x_{n+1}) \cdots \mathbf{f}(x_2)\mathbf{h}(x_1)\, dx_{n+1} \cdots dx_2 dx_1 + \\
&\int_a^b \int_a^{x_1} \cdots \int_a^{x_n} \mathbf{f}(x_{n+1}) \cdots \mathbf{f}(x_2)\mathbf{g}(x_1)\, dx_{n+1} \cdots dx_2 dx_1 - \\
&\int_a^b \int_a^{x_1} \cdots \int_a^{x_n} \mathbf{g}(x_{n+1}) \cdots \mathbf{g}(x_2)\mathbf{g}(x_1)\, dx_{n+1} \cdots dx_2 dx_1 = \\
&\int_a^b \mathcal{I}_n(a, x_1; \mathbf{F})\mathbf{h}(x_1)\, dx_1 + \int_a^b \Big(\mathcal{I}_n(a, x_1; \mathbf{F}) - \mathcal{I}_n(a, x_1; \mathbf{G})\Big)\mathbf{g}(x_1)\, dx_1 \overset{!}{=} \\
&c_{1(n+1)}(b) + \int_a^b \left(\sum_{m=1}^{n} c_{mn}(x_1)\right)\mathbf{g}(x_1)\, dx_1 = \\
&c_{1(n+1)}(b) + \sum_{m=1}^{n} \int_a^b c_{mn}(x_1)\mathbf{g}(x_1)\, dx_1
\end{aligned}$$

by the induction hypothesis. Applying Lemma E.0.2, the desired result readily follows. □

Proof of Theorem E.0.1. By Lemma E.0.3,

$$\begin{aligned}
\mathcal{P}(a, b; \mathbf{F}) - \mathcal{P}(a, b; \mathbf{G}) &= \sum_{n=0}^{\infty} \mathcal{I}_n(a, b; \mathbf{F}) - \sum_{n=0}^{\infty} \mathcal{I}_n(a, b; \mathbf{G}) \\
&\overset{!}{=} \sum_{n=1}^{\infty} \Big(\mathcal{I}_n(a, b; \mathbf{F}) - \mathcal{I}_n(a, b; \mathbf{G})\Big) \\
&= \sum_{n=1}^{\infty} \sum_{m=1}^{n} c_{mn}(b)
\end{aligned}$$

because $\mathcal{I}_0(a, b; \mathbf{F}) = \mathcal{I}_0(a, b; \mathbf{G}) = \mathbf{I}$. Using Fubini's theorem three more times,

$$
\begin{aligned}
\mathrm{P}(a,b;\mathbf{F})-\mathrm{P}(a,b;\mathbf{G}) &= \sum_{n=1}^{\infty}\sum_{m=1}^{n} c_{mn}(b) \overset{!}{=} \sum_{m=1}^{\infty}\sum_{n=m}^{\infty} c_{mn}(b) \\
&= \sum_{m=1}^{\infty}\sum_{n=m}^{\infty}\int_a^b \mathcal{F}_{n-m}(a,x;\mathbf{F})\mathbf{h}(x)\mathcal{F}_{m-1}(x,b;\mathbf{G})\,dx \\
&\overset{!}{=} \sum_{m=1}^{\infty}\int_a^b \sum_{n=m}^{\infty}\mathcal{F}_{n-m}(a,x;\mathbf{F})\mathbf{h}(x)\mathcal{F}_{m-1}(x,b;\mathbf{G})\,dx \\
&= \sum_{m=1}^{\infty}\int_a^b \Bigg(\sum_{n=m}^{\infty}\mathcal{F}_{n-m}(a,x;\mathbf{F})\Bigg)\mathbf{h}(x)\mathcal{F}_{m-1}(x,b;\mathbf{G})\,dx \\
&= \sum_{m=1}^{\infty}\int_a^b \mathcal{P}(a,x;\mathbf{F})\mathbf{h}(x)\mathcal{F}_{m-1}(x,b;\mathbf{G})\,dx \\
&\overset{!}{=} \int_a^b \sum_{m=1}^{\infty}\mathcal{P}(a,x;\mathbf{F})\mathbf{h}(x)\mathcal{F}_{m-1}(x,b;\mathbf{G})\,dx \\
&= \int_a^b \mathcal{P}(a,x;\mathbf{F})\mathbf{h}(x)\Bigg(\sum_{m=1}^{\infty}\mathcal{F}_{m-1}(x,b;\mathbf{G})\Bigg)\,dx \\
&= \int_a^b \mathcal{P}(a,x;\mathbf{F})\mathbf{h}(x)\mathcal{P}(x,b;\mathbf{G})\,dx
\end{aligned}
$$

One can easily justify these three uses of Fubini's theorem in a similar way as we have done in Lemma E.0.2. □

Bibliography

[1] Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. Statistical models based on counting processes. Springer Science & Business Media, 2012.

[2] Donald L Cohn. Measure theory. Springer.

[3] Ben Balkenende. "Brownian motion and Option pricing". B.S. thesis. 2018.

[4] Michael Baake and Ulrike Schlaegel. "The Peano-Baker series". In: Proceedings of the Steklov Institute of Mathematics 275.1 (2011), pp. 155–159.

[5] Antonín Slavík. Product integration, its history and applications.

[6] Richard D Gill and Soren Johansen. "A survey of product-integration with a view toward application in survival analysis". In: The annals of statistics 18.4 (1990), pp. 1501–1555.

[7] Hein Putter and Cristian Spitoni. "Non-parametric estimation of transition probabilities in non-Markov multi-state models: The landmark Aalen–Johansen estimator". In: Statistical methods in medical research 27.7 (2018), pp. 2081–2092.

[8] Odd Aalen, Ornulf Borgan, and Hakon Gjessing. Survival and event history analysis: a process point of view. Springer Science & Business Media, 2008.

[9] PWM Reijbroek. "An estimator for state occupation probabilities in non-Markov multistate models". MA thesis. 2017.

[10] user1551 (https://math.stackexchange.com/users/1551/user1551). How to prove that càdlàg (RCLL) functions on $[0, 1]$ are bounded? Mathematics Stack Exchange. URL: https://math.stackexchange.com/q/444879 (version: 2013-07-16). eprint: https://math.stackexchange.com/q/444879. URL: https://math.stackexchange.com/q/444879.

[11] Lord_Farin (https://math.stackexchange.com/users/43351/lord-farin). Cardinality of set of discontinuities of cadlag functions. Mathematics Stack Exchange. URL: https://math.stackexchange.com/q/502334 (version: 2016-12-17). eprint: https://math.stackexchange.com/q/502334. URL: https://math.stackexchange.com/q/502334.

[12] Wikipedia contributors.
Continuous mapping theorem — Wikipedia, The Free Encyclopedia.
https://en.wikipedia.org/w/index.php?title=Continuous_mapping_theorem&oldid=897508995. [Online; accessed 2-June-2019].
2019.