Estimating Non-Homogeneous Intensity Matrices in Continuous Time Multi-State Markov Models

Gerald Lebovic

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Dalla Lana School of Public Health
University of Toronto

©Copyright 2011 Gerald Lebovic
Estimating Non-Homogeneous Intensity Matrices in Continuous Time Multi-State Markov Models

Gerald Lebovic

Doctor of Philosophy
Dalla Lana School of Public Health
University of Toronto
April 2011

Abstract

Multi-State-Markov (MSM) models can be used to characterize the behaviour of categorical outcomes measured repeatedly over time. Kalbfleisch and Lawless (1985) and Gentleman et al. (1994) examine the MSM model under the assumption of time-homogeneous transition intensities. In the context of non-homogeneous intensities, current methods use piecewise constant approximations which are less than ideal. We propose a local likelihood method, based on Tibshirani and Hastie (1987) and Loader (1996), to estimate the transition intensities as continuous functions of time. In particular the local EM algorithm suggested by Betensky et al. (1999) is employed to estimate the in-homogeneous intensities in the presence of missing data.

A simulation comparing the piecewise constant method with the local EM method is examined using two different sets of underlying intensities. In addition, model assessment tools such as bandwidth selection, grid size selection, and bootstrapped percentile intervals are examined. Lastly, the method is applied to an HIV data set to examine the intensities with regard to depression scores. Although computationally intensive, it appears that this method is viable for estimating non-homogeneous intensities and outperforms existing methods.
Acknowledgements

There are numerous people, both colleagues and family alike that deserve many thanks for helping me produce this thesis. I would first and foremost like to thank my committee, Patrick Brown, George Tomlinson and James Stafford for all their guidance without which this thesis would not be possible. Patrick’s sharp skills and his open door policy, George’s thorough and logical thinking and direction, and Jamie’s experience and motivation were all instrumental in completing this thesis. I also would like to express gratitude to my internal and external examiners, Michael Escobar and Thierry Duchesne for their helpful comments and to Steve Fan and Lennon Li for many long and helpful discussions. I would also like to thank Steve for providing this template. I would like to acknowledge NSERC and the University of Toronto for granting me the financial assistance to complete my Ph.D. work. In addition, I express my gratitude to Dr. Liviana Calzavara of the Social and Behavioural Health Science Unit for permitting me to use the Polaris data.

Numerous family members have also been vital in my progress. I would like to thank my parents for their care, support and belief in me as well as my siblings, extended family and friends for their encouragement. Lastly, to my dear wife, Hadassah, through thick and thin you believed in me and had the graceful patience to listen to me discuss my work when no one else wanted to! I thank you for your devotion, motivation, support and confidence in me during this time. This accomplishment is for both of us to share and savor equally together. We have finally done it!
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.1</td>
<td>Local EM in MSM Models</td>
<td>35</td>
</tr>
<tr>
<td>3.1.2</td>
<td>Details of the E-Step</td>
<td>41</td>
</tr>
<tr>
<td>3.2</td>
<td>Simulation</td>
<td>46</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Simulation Setup</td>
<td>46</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Results: Time Varying Intensities</td>
<td>49</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Results: Piecewise Constant Intensities</td>
<td>53</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Additional Simulations</td>
<td>57</td>
</tr>
<tr>
<td>3.3</td>
<td>Remarks</td>
<td>57</td>
</tr>
<tr>
<td>4</td>
<td>Bandwidth Selection and Variance Estimation</td>
<td>64</td>
</tr>
<tr>
<td>4.1</td>
<td>Bandwidth Selection</td>
<td>64</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Cross Validation and Local EM</td>
<td>64</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Bandwidth Selection with Simulated Data</td>
<td>66</td>
</tr>
<tr>
<td>4.2</td>
<td>Bootstrap and Confidence Intervals</td>
<td>72</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Percentile Bootstrap</td>
<td>73</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Confidence Intervals with Simulated Data</td>
<td>74</td>
</tr>
<tr>
<td>4.3</td>
<td>Remarks</td>
<td>78</td>
</tr>
<tr>
<td>5</td>
<td>Example: Polaris Study</td>
<td>80</td>
</tr>
<tr>
<td>5.1</td>
<td>Background</td>
<td>80</td>
</tr>
<tr>
<td>5.2</td>
<td>Methodology</td>
<td>81</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Local EM Method</td>
<td>81</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Permutation Test</td>
<td>82</td>
</tr>
<tr>
<td>5.3</td>
<td>Results</td>
<td>83</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Local EM</td>
<td>83</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Permutation Test</td>
<td>88</td>
</tr>
<tr>
<td>5.4</td>
<td>Remarks</td>
<td>96</td>
</tr>
</tbody>
</table>
# List of Figures

1.1 Two individuals’ paths of transitions over an interval \((0, t)\). The \(x\) indicates the fixed time points at which the subjects are observed. ........................................ 3

1.2 MSM model: transitions between various states .............................. 4

2.1 Intensity estimates for both constant and piecewise constant intensities ....................................................... 25

3.1 Various scenarios for computing \(\tilde{\gamma}_{ij}\) ...................................................... 39

3.2 Underlying true intensities used in the time varying simulation ............... 47

3.3 Time Varying intensity estimates \(\delta = 1\) .................................................. 50

3.4 Time Varying intensity estimates \(\delta = 1/3\) ............................................. 51

3.5 Time varying intensity estimates; various grid sizes .......................... 52

3.6 Piecewise constant intensity estimates \(\delta = 1\) ..................................... 54

3.7 Piecewise constant intensity estimates \(\delta = 1/3\) ................................. 55

3.8 Piecewise constant intensity estimates for a fixed bandwidth of 2 ............. 56

3.9 Constant intensity estimates for various bandwidths .......................... 58

3.10 Intensity estimates for independent data .................................................. 59

4.1 Comparison of the ISE and CV methods for the time varying simulated dataset 69

4.2 Comparison of the ISE and CV methods for the PWC simulated dataset ........ 71

4.3 Bootstrap percentile intervals using a bandwidth of 2 ............................ 75

4.4 Bootstrap percentile intervals using a bandwidth of 3 ............................ 76
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Intensity estimates for HIV positive subjects</td>
<td>84</td>
</tr>
<tr>
<td>5.2</td>
<td>Intensity estimates for HIV negative subjects</td>
<td>85</td>
</tr>
<tr>
<td>5.3</td>
<td>Intensity estimates corresponding to the optimal bandwidth</td>
<td>87</td>
</tr>
<tr>
<td>5.4</td>
<td>Intensity estimates corresponding to a bandwidth of 9 months</td>
<td>89</td>
</tr>
<tr>
<td>5.5</td>
<td>Histograms of Sum of Squares (times 3 - 12 months)</td>
<td>91</td>
</tr>
<tr>
<td>5.6</td>
<td>Percentile plots times 3 -12 months</td>
<td>92</td>
</tr>
<tr>
<td>5.7</td>
<td>Histograms of Sum of Squares (times 13 - 60 months)</td>
<td>94</td>
</tr>
<tr>
<td>5.8</td>
<td>Percentile plots times 13 -60 months</td>
<td>95</td>
</tr>
<tr>
<td>5.9</td>
<td>Histograms of Sum of Squares (times 3 - 60 months)</td>
<td>97</td>
</tr>
<tr>
<td>5.10</td>
<td>Percentile plots times 3 -60 months</td>
<td>98</td>
</tr>
<tr>
<td>5.11</td>
<td>Marginal probabilities for HIV positive and negative groups</td>
<td>99</td>
</tr>
<tr>
<td>5.12</td>
<td>Conditional probabilities for a HIV positive and negative individual</td>
<td>100</td>
</tr>
<tr>
<td>5.13</td>
<td>Examining one individual’s observed time points in relation to the grid</td>
<td>101</td>
</tr>
<tr>
<td>A.1</td>
<td>Additional simulation $\delta = 1$</td>
<td>118</td>
</tr>
<tr>
<td>A.2</td>
<td>Additional simulation $\delta = 1/3$</td>
<td>119</td>
</tr>
</tbody>
</table>
List of Tables

4.1 ISE values and CV scores from the time varying simulated dataset . . . . . 68
4.2 Comparison of the ISE and CV methods for the PWC simulated data . . . . 70
4.3 Comparison of the ISE using 4 different values for $\delta$, and 3 different bandwidths 71
4.4 Effect of increasing the sample size on bandwidth selection . . . . . . . . . . 72

5.1 Cross validation for both HIV positive and negative participants . . . . . . . 86
Chapter 1

Introduction

In many longitudinal studies measuring various aspects of disease, a categorical outcome is often of interest. Frequently, the categorical outcome has an inherent order. For example, subjects with chronic diseases may be asked to rate their quality of life as being poor, fair, or good. These types of data require models that treat the outcome as ordinal and standard logistic or Poisson regression methods are not valid. Several methods have been proposed and we will briefly discuss these methods.

1.1 Ordinal Models

The most straightforward model for these types of data is the polytomous logistic model discussed by Hosmer and Lemeshow [20]. This model essentially extends the logistic model by comparing each response to a prespecified response. If there are a total of $K$ states then the model computes $K - 1$ separate logits, each comparing a particular state to the baseline. A numerical procedure is used to find the parameter estimates. The most frequently used ordinal model is the proportional odds or cumulative logit model discussed in McCullagh [35]. This model assumes that the categories are contiguous intervals on a continuous scale. It uses cumulative logits to model the data. The model has $K - 1$ cutpoints or distinct
intercepts, yet only one shape, or slope parameter, $\beta$ per predictor. The benefit of this model is primarily when collapsing categories as $\beta$ does not change, only the intercept values will vary. When a subset of explanatory variables do not adhere to the proportional odds assumptions, a partial proportional odds model can be used as suggested by Peterson and Harrell Jr [39]. Other models such as the continuation ratio model proposed by Feinberg [14], adjacent category model discussed by Agresti [2], and the stereotype model suggested by Anderson [5] are also designed to analyze ordinal data. A thorough review of these models, with applications, can be found in Agresti [2], Ananth and Kleinbaum [3] and Armstrong and Sloan [6]. These models have been extended to the generalized estimating equation (GEE) framework by Lipsitz et al. [30] and Miller et al. [36] and to generalized linear mixed model (GLMM) methodology by Liu and Hedeker [31] and Raman and Hedeker [41]. A review of ordinal methods applicable to longitudinal data is discussed by Hines and Hines [19] and Agresti [1]. Transition models represent a distinct approach to longitudinal ordinal data.

The transition model primarily differs from the other methods because the correlation of observations in time can be directly attributed to past observations and the past observations are therefore treated as additional predictor variables. These methods have been applied by Korn and Whittemore [28] and Zeger and Qaqish [48] as well as others. For example, in a quality of life study, one might expect a subject’s reported quality of life today to depend on previously reported quality of life scores. Let $Y(t)$ be the observed state, at time $t$. Then in essence, a transition model states that the current observation $Y(t)$ may depend on the previous observations $Y(t-1), Y(t-2), \ldots, Y(0)$. The Markov property simplifies this assumption by stating that the current observation only depends on the observation immediately prior to the current observation. Or, more formally,

$$P(Y(t) = i_t | Y(t-1) = i_{t-1}, Y(t-2) = i_{t-2} \ldots Y(0) = i_0) = P(Y(t) = i_t | Y(t-1) = i_{t-1})$$

for all states $i_t, i_{t-1}, \ldots, i_0$. These models have been extended to binary and count data and
Figure 1.1: Two individuals’ paths of transitions over an interval $(0, t)$. The x indicates the fixed time points at which the subjects are observed.

A thorough treatment can be found in Diggle [11].

In our setting of longitudinal ordinal data, we are interested in determining the probability of transitioning between various states. Figure 1.1 illustrates paths taken through three states for two participants observed at 4 time points, $t_1$ to $t_4$. Participants are observed at fixed time points and may transition between states at any point along the time interval. Between time $t-1$ and time $t$, the probability of transitioning from state $i$ to state $j$ is given by $p_{ij}$. These probabilities can be arranged in a square matrix, denoted as $P$, where the $(i, j)^{th}$ entry is $p_{ij}$. Suppose subjects are free to move among a total of $n$ possible states then Figure 1.2 indicates the possible transitions between any of these states. In the special case where the outcome is ordinal, then transitions can only take place between adjacent states, as seen in Figure 1.2 with the curved (red) lines removed. Referring back to the quality of life example, the transition matrix can be used to determine the probability that subjects may transition among various states over a period of time. In particular the probability of
transitioning from state $i$ to state $j$ in $n$ transitions can be found as

$$p^{(n)}_{ij} = \sum_{s=1}^{\infty} p_{is}p^{(n-1)}_{sj}$$

where $p^{(0)}_{ij}$ is 1 when $i = j$ and 0 otherwise.

The matrix $P$ is governed by an underlying intensity matrix $Q$ which can be defined in a similar fashion to $P$. $Q$ is a square matrix where the $(i, j)^{th}$ entry, $q_{ij}$, is the instantaneous probability of transitioning from state $i$ to state $j$ in a small time interval $\Delta t$. A more formal definition is provided in Chapter 2. It is this matrix that we are primarily interested in.

Multi state Markov (MSM) models examined by Kalbfleisch and Lawless [23] as well as Gentleman et al. [15], employ the Markov property to estimate both $Q$ and $P$. The most prevalent method examined is to assume that the underlying intensities are constant with respect to time or piecewise constant with respect to time. When intensities are assumed to be constant with respect to time then regardless of the length of time on study the same
underlying $Q$ is assumed. However, realistically, these rates are often non-homogeneous as they will depend on time. This is especially true in the field of medicine where advances in therapy change frequently with time and can have a significant impact on the outcome. The piecewise constant approach provides a rather simplistic solution to this problem by assuming the intensities are constant over particular intervals or piecewise constant. Though this may improve the estimate of $Q$, it still falls short of the true value, especially when $Q$ is thought to vary to an important degree with respect to time.

Another approach to estimating intensities in the context of MSM is provided by Andersen [4]. A solution employing the product integral is used to relate the transition matrix to the intensity matrix when the intensity is non-homogeneous with respect to time. A non-parametric method is used to estimate the transition probabilities based on non-parametric estimates of the cumulative intensity. This method will be discussed in more detail in section 2.3. Our approach will introduce a method that can estimate non-homogeneous intensities $Q(t)$ as a function of time. It will not rely on the more complicated and obscure form of the product integral [16] and will estimate the actual intensities rather than the cumulative intensities.

The next chapter, chapter 2, will introduce the existing methods and illustrate them with an example. It will also introduce local likelihood-based methods and the local EM algorithm which will lay the foundation for chapter 3. Chapter 3 introduces our method in the context of the local likelihood and presents the use of the method through analysis of two simulated datasets. Chapter 4 provides tools for two important aspects of model assessment, namely bandwidth selection and confidence interval construction. These tools are applied to the simulated data from chapter 3. Chapter 5 applies the methods developed in the thesis to the analysis of a real dataset consisting of longitudinal data on HIV participants. The last chapter discusses some concluding remarks as well as some ideas for further research.
Chapter 2

Literature Review

2.1 Relationship Between Transition Probabilities and Transition Intensities

Multi-state Markov models are frequently used with panel data that are correlated in time. These models have been used in the literature [25, 26] to examine balanced, equally spaced data. The foundation underlying these models is the Markov property, in particular, any presently observed outcome is only dependent on the previous observation and none of the observations prior to that. This property applies to both discrete and continuous data as we will now describe in more detail. References expanding on the material in the following section can be found in Taylor and Karlin [45], Ross [43] Grimmett and Stirzaker [17].

2.1.1 Markov Process

A first order Markov Process, $Y(t)$, is a stochastic process in which future knowledge about the process is provided only by the current state and is not altered with the additional knowledge of past states. More formally, in the continuous time setting, a Markov process,
$Y(t), t > 0$, can be defined as,

$$P(Y(t_{k+1}) = y_{k+1}|Y(t_k) = y_k, \ldots, Y(t_0) = y_0) = P(Y(t_{k+1}) = y_{k+1}|Y(t_k) = y_k)$$

for all states $y_{k+1}, \ldots, y_0$ and all time points $t_{k+1}, t_k, \ldots, t_0$, where $0 \leq t_0, \ldots, t_k, t_{k+1} \leq T$, $T$ the maximum time. For the remainder of this thesis the above probability will be denoted as $P_{y_k, y_{k+1}}(t_k, t_{k+1})$.

When the state space is a finite or a countable set and the time index consists of $(0,1,2, \ldots)$ then the Markov process is said to be a discrete time process. Again, in this setting, only the most recent observation is pertinent and

$$P(Y(t_k) = y_k|Y(t_k - 1) = y_{k-1}, \ldots, Y(t_0) = y_0) = P(Y(t_k) = y_k|Y(t_k - 1) = y_{k-1})$$

for discrete times $t_k, t_0$. If (the process is time homogeneous and) the time interval $t_k - t_{k-1}$ is equally spaced for all $k$ then $P_{y_{k-1}, y_k}(t_{k-1}, t_k)$ can simply be written as $P_{y_{k-1}, y_k}$ which is the one step transition probability from $y_{k-1}$ to $y_k$. A matrix, $P$, containing all possible one step transition probabilities is called the transition matrix. To compute the $n^{th}$-step transition, the $P$ matrix is simply raised to the $n^{th}$ power. If transition probabilities are non-homogeneous and vary with respect to time, then the appropriate transition matrices will need to be multiplied to obtain the $n^{th}$ step transition.

Computing the probability transition matrix in the continuous setting is more complex than in the discrete setting. To construct the matrix $P$ analogous to the one in discrete time, underlying intensities of the process are required. We will now examine how the intensities of the process are related to the transition probabilities.
2.1.2 Intensities

The intensity between two states, \(i\) and \(j\), can be thought of as the rate of change of the probability \(P_{ij}\) in a very small time interval, \(\Delta t\). More formally, we define the intensity, \(q_{ij}(t)\) from state \(i\) to state \(j\) at time \(t\) as:

\[
q_{ij}(t) = \lim_{\Delta t \to 0} \frac{P_{ij}(t,t+\Delta t)}{\Delta t} \quad \text{for } i \neq j
\]

for any given time \(\{t : 0 < t < T\}\) and interval length \(\Delta t > 0\).

Analogous to the probability transition matrix, an intensity matrix \(Q\) can be constructed. It contains all possible intensities between the various states. For example, an outcome containing \(K\) states would have the following intensity matrix, \(Q\).

\[
Q(t) = \begin{bmatrix}
q_{11}(t) & q_{12}(t) & \ldots & q_{1K}(t) \\
q_{21}(t) & q_{22}(t) & \ldots & q_{2K}(t) \\
\vdots & \vdots & \ddots & \vdots \\
q_{K1}(t) & q_{K2}(t) & \ldots & q_{KK}(t)
\end{bmatrix}
\]

For all intensity matrices there are two constraints placed on the row entries, \(q_{ij}(t)\).

1. The off diagonals must be non-negative i.e. \(q_{ij}(t) \geq 0\) for \(i \neq j\).

2. The rows must sum to zero i.e. \(\sum_{j=1}^{K} q_{ij}(t) = 0\).

For example, suppose there are \(K = 3\) possible states of interest, then under the above constraints, \(Q(t)\) would have form:

\[
Q(t) = \begin{bmatrix}
-q_{12}(t) - q_{13}(t) & q_{12}(t) & q_{13}(t) \\
q_{21}(t) & -q_{21}(t) - q_{23}(t) & q_{23}(t) \\
q_{31}(t) & q_{32}(t) & -q_{31}(t) - q_{32}(t)
\end{bmatrix}
\]
The diagonals of the matrix are simply the “rate” at which people do not transition and remain in their state, while the off diagonals are the rates at which subjects transition into other states. In most cases the intensity matrix is used, via matrix exponentials, to determine the transition matrix. When intensities are time homogeneous these results are well established as we will now discuss.

2.1.3 Time-Homogeneous Intensities

When intensities are treated as being time homogeneous then the dependency on time can be removed. In this setting it can be shown that $P$ has the closed form solution:

$$P = \exp(Qt) = \sum_{r=0}^{\infty} \frac{Q^r t^r}{r!} \quad (2.2)$$

This closed form solution, is only valid with time homogeneous intensities as it is based on certain assumptions. We will now examine the assumptions as well as the derivation of equation (2.2), discussed extensively in Ross [43] and Grimmett and Stirzaker [17], to understand the problematic issues for non-homogeneous intensities.

The matrix $P$ is time dependent and when $Q$ is assumed to be time-homogeneous we have

$$P_{ij}(s,t) = P_{ij}(t-s,0) = P_{ij}(t-s) \quad \text{for } 0 \leq s \leq t$$

To emphasize the dependence of $P$ on time, the transition matrix should be denoted as $P(t)$. However, when intensities are time homogeneous the dependence of $P$ on time will be omitted when its meaning is unambiguous.

**Theorem 1** The family $P(t): t \geq 0$ is a stochastic semigroup. That is, it satisfies the following conditions:

1. $P(0) = I_{K \times K}$ where $K$ is the number of states.
2. \( \mathbf{P}(t) \) is stochastic (i.e. \( \mathbf{P}(t) \) has non-negative entries and rows that sum up to 1.)

3. Chapman-Kolmogorov equations apply; \( \mathbf{P}(s + t) = \mathbf{P}(s)\mathbf{P}(t) \) if \( s, t \geq 0 \).

In addition, if \( \lim_{t \to 0} \mathbf{P}(t) = I_{K \times K} \) then the semigroup is standard. Therefore, in a very small interval, \( \Delta t \):

a. No transitions may occur with probability \( P_{ii}(\Delta t) + o(h) \), where \( o(h) \) is the probability of transitioning in and out of state \( i \) in the interval \( \Delta t \).

b. There may be a transition to a new state with probability \( P_{ij}(\Delta t) + o(h) \), where \( o(h) \) is the probability of 2 or more transitions during the small interval \( \Delta t \).

Therefore, in a small interval, \( \Delta t \), \( P_{ij}(\Delta t) \) is linear. Now, using this result, we have the following Lemma.

**Lemma 2.1.1**

a. The off diagonals of \( \mathbf{Q} \), or the probability of transitioning, are:

\[
\lim_{\Delta t \to 0} \frac{P_{ij}(\Delta t)}{\Delta t} = q_{ij} \quad \text{for } i \neq j
\]

b. The diagonals, or the probability of not transitioning, can be found as:

\[
\lim_{\Delta t \to 0} \frac{1 - P_{ii}(\Delta t)}{\Delta t} = \sum_{i \neq j} q_{ij} = -q_{ii}
\]
Combining \(a\) and \(b\) and using matrix notation yields:

\[
\lim_{\Delta t \to 0} \frac{1}{\Delta t} (P(\Delta t) - I) = Q
\]

Essentially in a short interval, \(\Delta t\), the probability of transitioning from \(i\) to \(j\) is \(P_{ij}(\Delta t) \simeq q_{ij}\Delta t\) for \(i \neq j\). Also, applying the constraints when \(i = j\) leads to \(1 - P_{ii}(\Delta t) \simeq (\sum_{i \neq j} q_{ij})\Delta t = -q_{ii}\Delta t\), or rearranging, \(P_{ii}(\Delta t) \simeq 1 + q_{ii}\Delta t\). These properties relating \(Q\) to \(P\) only hold if the semigroup is stochastic as well as standard.

**Chapman-Kolmogorov Equations**

Thus far we have shown how \(Q\) can be determined from \(P\). However, \(P\) can also be found from \(Q\) by using the Chapman-Kolmogorov equations. These equations state for any two states \(i\) and \(j\):

\[
P_{ij}(t + \Delta t) = \sum_{l=0}^{\infty} P_{il}(t)P_{lj}(\Delta t)
\]

(2.3)

The proof of (2.3) is rather straightforward and can be found in Ross [43].

The Kolmogorov equations are used to derive the relationship between the transition probability matrix, \(P\) and the intensity matrix, \(Q\). The Kolmogorov backward equation has the form:

\[
P_{ij}(t + \Delta t) = \sum_{l=0}^{\infty} P_{il}(t)P_{lj}(\Delta t)
\]

\[
P_{ij}(t + \Delta t) - P_{ij}(t) = \sum_{l=0}^{\infty} P_{il}(t)P_{lj}(\Delta t) - P_{ij}(t)
\]

Which, using Lemma 2.1.1, dividing both sides by \(\Delta t\), and taking the limit as \(\Delta t \to 0\) we obtain the differential equation,

\[
P'_t = QP_t
\]

(2.4)
Similarly, to compute the Kolmogorov forward equation we have:

\[ P_{ij}(\Delta t + t) - P_{ij}(t) = \sum_{l=0}^{\infty} P_{il}(\Delta t)P_{lj}(t) - P_{ij}(t) \]

and again using the same methods as the backward equation this leads to the differential equation,

\[ P'_{t} = P_{t}Q \quad (2.5) \]

Equations (2.4) and (2.5) can be solved and when \( Q \) is time homogeneous the solution is:

\[ P = \exp(Qt) \]

More details on the derivation of the Kolmogorov forward and backward differential equations can be found in Grimmett and Stirzaker [17].

In section 2.2 we will show how to use this result to obtain the MLE of \( Q \).

### 2.1.4 Ordinal Models

In an ordinary MSM setting, with the exception of the constraints mentioned previously, there are typically no restrictions on the transitions as they are permitted between any two states. However, in many studies the outcome is measured on an ordinal scale and that requires some restrictions in the construction of the intensity matrix. In this scenario, subjects must transition to an adjacent state, even if only very briefly, before moving to another, non-adjacent, state. In the quality of life example, it is highly probable that a subject transitioning from a state of poor to good, passed through, even if ever so briefly, the state of fair. Therefore, the intensity matrix will be tridiagonal, as intensity values are only possible between adjacent states, all other entries will be 0. The off diagonal entries are non-negative and represent the rate at which subjects transition to other states. The
diagonal elements may be viewed as a constraint and represent the rate of remaining in the same state. Only the off-diagonals need to be estimated since the diagonals are constrained so the rows sum to zero. The $Q$ matrix, in the ordinal setting with $K$ states can be constructed as

$$Q = \begin{bmatrix} -q_{12} & q_{12} & 0 & 0 & \ldots & 0 \\ q_{21} & -(q_{21} + q_{23}) & q_{23} & 0 & \ldots & 0 \\ 0 & q_{32} & -(q_{32} + q_{34}) & q_{34} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & q_{K-1,K} & -q_{K-1,K} \end{bmatrix}$$

(2.6)

Although imposing the ordinal structure on the outcome limits the number of parameters, it is clear that as $K$ increases the number of parameters requiring estimation also increases. In general, for $K$ ordinal states it is necessary to have $2(K - 2) + 2$ parameters. The following sections will illustrate methods for obtaining the intensity estimates while assuming there are $K = 3$ possible outcomes.

### 2.2 Maximum Likelihood Estimation for Time-Homogeneous Intensities

In the time homogeneous setting, methods to estimate the intensity matrix have been discussed in Kalbfleisch and Lawless [23] and [15]. Suppose we have an ordinal outcome with $K = 3$ states. The intensity matrix $Q$ would have the form

$$Q = \begin{bmatrix} -q_{12} & q_{12} & 0 \\ q_{21} & -(q_{21} + q_{23}) & q_{23} \\ 0 & q_{32} & -q_{32} \end{bmatrix}$$

(2.7)
and estimating the four intensities $q_{12}, q_{21}, q_{23}, q_{32}$ would result in the entire intensity matrix being known. Let $\boldsymbol{\theta} = (q_{11}, q_{21}, q_{23}, q_{32})$ denote the vector of intensities. In the general setting, when the outcome is not necessarily ordinal, $\boldsymbol{\theta}$ would be a $c \times 1$ vector with at most $K(K-1)$ entries. Our objective is to maximize the likelihood to obtain estimates of $\boldsymbol{\theta}$.

### 2.2.1 Construction and Estimation of the Likelihood

The likelihood function for both the balanced and unbalanced cases, as found in Kalbfleisch and Lawless [23] and Gentleman et al. [15] respectively, can be constructed in a straightforward manner. Referring back to Figure 1.1, suppose $N$ individuals observed at time points $t_{ij}, i = 1, \ldots, N; j = 1, \ldots, J_i$, can transition between states at any time along an interval. Let $S_i(t_{ij})$ denote the state occupied by individual $i$ at time $t_{ij}$. In addition, with $k = 1, \ldots, K$ ordinal states, and $\boldsymbol{\theta} = (q_{11}, q_{21}, q_{23}, \ldots, q_{KK-1})$ then the unconditional likelihood for all $N$ individuals is:

$$L = \prod_{i=1}^{N} P_{S_i(t_{io})}(0; \boldsymbol{\theta}) \prod_{j=1}^{J_i} P_{S_i(t_{ij-1}), S_i(t_{ij})}(t_{ij-1}, t_{ij}; \boldsymbol{\theta})$$  \hspace{1cm} (2.8)

and, conditional on the initial state, the likelihood is simply:

$$\prod_{i=1}^{N} \prod_{j=1}^{J_i} P_{S_i(t_{ij-1}), S_i(t_{ij})}(t_{ij-1}, t_{ij}; \boldsymbol{\theta})$$  \hspace{1cm} (2.9)

In order to obtain $\mathbf{P}$ we use the $\mathbf{Q}$ matrix. To use equation (2.2) the matrix exponentials are required. Though there are numerous ways to compute the matrix exponential [37], matrix decomposition methods were used as they were readily implemented in R and as suggested by Kalbfleisch and Lawless [23] in most cases eigenvalue decomposition is a valid and straightforward method. When eigenvalue decomposition fails, other methods such as Jordan canonical decomposition may be used. Using the $K$ distinct eigenvalues, $d_1, \ldots, d_K$
we have
\[ P(t; \theta) = A \text{diag}(e^{d_1 t}, \ldots, e^{d_K t}) A^{-1} \]

where \( A \) is a matrix containing the \( K \) eigenvectors and the dependence of \( A \) and \( d_k \) on \( \theta \) is suppressed. In matrix notation, if we let \( D = \text{diag}(d_1, \ldots, d_K) \) then we have:

\[ P = A e^D A^{-1} \]

In general, the solution of \( P_{ij} \) in terms of \( \theta \) is a lengthy and complicated expression. In Appendix A.1 we provide the solution for \( P_{ij} \) in terms of \( \theta \) when the outcome consists of \( K = 3 \) ordinal states.

Obtaining the MLE of \( \theta \) is accomplished most efficiently by having both the first and second derivatives of the likelihood function. For simplicity, we first re-write equation (2.9) using slightly different notation for computational purposes.

\[ L = \prod_{i=1}^{N} \prod_{j=1}^{J_i} \prod_{u,v=1}^{K} P_{u,v}^{\gamma_{ijuv}}(t_{ij-1}, t_{ij}; \theta) \]  

(2.10)

Where \( K \) is the number of states, \( u, v = 1, \ldots, K \) and:

\[ \gamma_{ijuv} = \begin{cases} 
1, & \text{if } S_i(t_{ij-1}) = u \text{ and } S_i(t_{ij}) = v \\
0, & \text{otherwise}
\end{cases} \]
From equation (2.10) we can derive:

\[
\ell = \sum_{i=1}^{N} \sum_{j=1}^{J_i} \sum_{u,v=1}^{K} \gamma_{ijuv} \log P_{uv}(t_{ij-1}, t_{ij}; \theta) \tag{2.11}
\]

\[
\frac{\partial \ell}{\partial \theta_w} = \sum_{i=1}^{N} \sum_{j=1}^{J_i} \sum_{u,v=1}^{K} \gamma_{ijuv} \frac{\partial}{\partial \theta_w} P_{uv}(t_{ij-1}, t_{ij}; \theta) \frac{\partial}{\partial \theta_v} P_{uv}(t_{ij-1}, t_{ij}; \theta) \tag{2.12}
\]

\[
\frac{\partial^2 \ell}{\partial \theta_w \partial \theta_z} = \sum_{i=1}^{N} \sum_{j=1}^{J_i} \sum_{u,v=1}^{K} \gamma_{ijuv} \left\{ \frac{\partial^2}{\partial \theta_w \partial \theta_z} P_{uv}(t_{ij-1}, t_{ij}; \theta) - \frac{\partial}{\partial \theta_w} P_{uv}(t_{ij-1}, t_{ij}; \theta) \frac{\partial}{\partial \theta_z} P_{uv}(t_{ij-1}, t_{ij}; \theta) \right\} P_{uv}^2(t_{ij-1}, t_{ij}; \theta) \tag{2.13}
\]

Making use of the first and second derivatives of the log-likelihood, the Newton-Raphson algorithm can be used to iteratively solve for \( \theta \). The Newton-Raphson algorithm is used both in the time homogeneous setting as well as in the time varying approach in the following chapter. Therefore we will briefly outline the algorithm as can be found in Knight [27].

**Newton Raphson Algorithm**

The algorithm can be used to solve a system of non-linear equations, in our case the likelihood equations. To obtain the MLE of \( \theta \), the score function, equation (2.12), can be set to zero and then solved to obtain the estimates. The algorithm, based on Taylor series expansions about a point \( x \), will numerically solve the equation for an unknown \( x_0 \). In the context of the likelihood, starting values for \( \theta \) are chosen and the \( p^{th} \) estimate is updated as:

\[
\hat{\theta}^{(p+1)} = \hat{\theta}^{(p)} + S(\hat{\theta}^{(p)}) H^{-1}(\hat{\theta}^{(p)})
\]

where

\[
S(\theta) = \left( \frac{\partial \ell}{\partial \theta_1}, \ldots, \frac{\partial \ell}{\partial \theta_c} \right)
\]
and $H(\theta)$ is a $c \times c$ matrix where the $w, z$ entry is $\frac{\partial^2 \ell}{\partial w \partial z}$.

The iterative procedure continues until a convergence criterion, $|\hat{\theta}^{(p+1)} - \hat{\theta}^{(p)}| < C$ is met.

Due to the complex nature of the second derivative, Kalbfleisch and Lawless [23] suggest using a quasi-Newton procedure for obtaining $\hat{\theta}$. For this procedure it is simply the $E[H(\theta)]$ that is substituted in place of $H(\theta)$ and it can be expressed as

$$
E \left[ - \frac{\partial^2 l}{\partial \theta_w \partial \theta_z} \right] = \sum_{i=1}^{N} \sum_{j=1}^{J_i} \sum_{u,v=1}^{K} \gamma_{iju} \frac{\partial^2}{\partial \theta_w} P_{uv}(t_{ij-1}, t_{ij}; \theta) \frac{\partial}{\partial \theta_z} P_{uv}(t_{ij-1}, t_{ij}; \theta)
$$

(2.14)

where,

$$
\gamma_{iju} = \begin{cases} 
1, & \text{if } S_i(t_{ij-1}) = u \\
0, & \text{otherwise}
\end{cases}
$$

Details on the derivation can be found in Appendix A.2.

We use the expectation of the second derivative in place of the second derivative in the quasi-Newton algorithm. The overwhelming advantage is to avoid direct computation of (2.13). Although direct computation of $S(\theta)$ is also complicated, an indirect approach, using eigenvalue decomposition can be used to find the derivatives. For more details see Jennrich and Bright [22] and Kalbfleisch and Lawless [23].

In the special case where individuals are observed on a balanced grid, then $J_i = J$ for all $N$ individuals and equation (2.9) has the simpler form:

$$
\prod_{i=1}^{N} \prod_{j=1}^{J} P_{S_i(t_{ij-1}), S_i(t_{ij})}(t_{ij-1}, t_{ij}; \theta)
$$

Or, analogous to equation (2.10)

$$
\prod_{j=1}^{J} \prod_{u,v=1}^{K} P_{uv-1, uv}(t_{j-1}, t_j; \theta)
$$
Similarly the form of the log-likelihood, score function, and second derivative are:

\[
\ell = \sum_{j=1}^{J} \sum_{u,v=1}^{K} n_{juv} \log P_{uv}(t_{j-1}, t_j; \theta)
\]

\[
\frac{\partial \ell}{\partial \theta_w} = \sum_{j=1}^{J} \sum_{u,v=1}^{K} n_{juv} \frac{\partial}{\partial \theta_w} P_{uv}(t_{j-1}, t_j; \theta)
\]

\[
\frac{\partial^2 \ell}{\partial \theta_w \partial \theta_z} = \sum_{j=1}^{J} \sum_{u,v=1}^{K} n_{juv} \left\{ \frac{\partial^2}{\partial \theta_w \partial \theta_z} P_{uv}(t_{j-1}, t_j; \theta) - \frac{\partial}{\partial \theta_w} P_{uv}(t_{j-1}, t_j; \theta) \frac{\partial}{\partial \theta_z} P_{uv}(t_{j-1}, t_j; \theta) \right\}
\]

and the expectation of the second derivative is expressed as

\[
E \left[ -\frac{\partial^2 \ell}{\partial \theta_w \partial \theta_z} \right] = \sum_{j=1}^{J} \sum_{u,v=1}^{K} N_j(t_{j-1}) \frac{\partial}{\partial \theta_w} P_{uv}(t_{j-1}, t_j; \theta) \frac{\partial}{\partial \theta_z} P_{uv}(t_{j-1}, t_j; \theta) \frac{P_{uv}(t_{j-1}, t_j; \theta)}{P_{uv}(t_{j-1}, t_j; \theta)}
\]

where \( n_{juv} \) is the number of people in both state \( u \) at time \( j - 1 \) and state \( v \) at time \( j \) and \( N_j(t_{j-1}) = \sum_{v=1}^{K} n_{juv} \) is the number of people in state \( u \) at time \( j - 1 \).

Computationally, estimating the intensity and transition matrices with balanced data is a lot faster. This is because the intervals, \( t_{j-1} - t_j \), are identical for all participants and need not be computed individually for each subject. This is reflected in the likelihood by having one less summation term.

### 2.2.2 Example - Polaris Cohort Study

Using a portion of data from the Polaris Cohort Study [8], we will illustrate the above methods to find the ML estimates of \( \theta \). The Polaris HIV Seroconversion Study is a social and behavioural study designed to capture many aspects of physical and emotional health related to HIV. Enrollment into the study began in June 1998 and all interviews were completed by April 2007. Subjects were interviewed every 6 months and were asked a variety of both qualitative and quantitative questions. The study recruited recent HIV seroconverters as well as HIV negative controls, all individuals from Ontario, Canada. HIV seroconverters had a documented HIV positive test in the last 2 years and HIV negative
controls had a documented HIV negative test in the same time frame. Although a total of 459 subjects were enrolled in the study, for the purposes of this example a subset of 70 individuals were selected for analysis. This sample excluded participants with only one interview, any missing observations, females, heterosexual males, HIV negative subjects, as well as HIV negative subjects who seroconverted while on the study. The maximum time on study for the 70 participants was 7.5 years.

One objective of the study was to examine subjects’ attitudes towards the class of HIV medications known as Highly Active Antiretroviral Treatment (HAART). More specifically, the question of interest was “What effect will HAART have on quality of life?”. Although there was a choice of five ordinal responses, we collapsed it into the following three categories (a) make it worse (state 1), (b) no impact or too soon to tell (state 2), (c) improve or better it (state 3). The Q matrix was set up as in equation (2.7) and the likelihood for the data was constructed as in equation (2.10). In this example we want to obtain estimates for $\theta = (q_{12}, q_{21}, q_{23}, q_{32})$, where, for example, $q_{12}$ is the intensity associated with a jump from state 1 to state 2 and time $t$ is measured in years.

The msm package [21] in R was used to fit the model, obtaining MLEs of $\theta$. The msm function requires initial estimates for $\theta$ and then uses the method described above to obtain maximum likelihood estimates. In order to compute initial estimates, $\hat{\theta}_0$, the method described in Kalbfleisch and Lawless [23] was used by setting up a transition table and having the initial estimates approximated as:

$$P_{uv}(t_{ij-1}, t_{ij}) \approx \exp[q_{uv}(t_{ij} - t_{ij-1})] \Rightarrow q_{uv} \approx \log[P_{uv}(t_{ij-1}, t_{ij})/(t_{ij} - t_{ij-1})]$$

The idea being to use one interval and assume either people have one transition or none and, assuming time-homogeneous intensities, we can find reasonable starting values for the intensities. In our example, since participants were followed-up approximately every
6 months, we used \( t_{ij} - t_{ij-1} = 0.5 \) years to obtain initial estimates. Using the above method we have \( \theta_0 = (1.62, 4.795, 1.386, 0.33) \). Using the methods described in the previous section we obtain \( \hat{\theta} = (1.19, 0.28, 1.20, 0.19) \) or in terms of the intensity matrix:

\[
Q = \begin{bmatrix}
-1.19 & 1.19 & 0 \\
0.28 & -1.49 & 1.20 \\
0 & 0.19 & -0.19
\end{bmatrix}
\]

Figure 2.1 plots these time homogeneous intensities and we can see that \( q_{12} \) and \( q_{23} \) are relatively low while \( q_{21} \) and \( q_{32} \) are significantly higher. This can be interpreted as individuals being very unlikely to transition from higher, (more optimistic), states to lower, (less optimistic), states but they are more likely to transition from lower states to higher ones. Individuals who believe that HAART will either improve quality of life or at least keep it the same are not likely to change and believe that it will worsen their quality of life. On the other hand those who started out believing HAART will either worsen their quality of life or have no impact on it are more likely to change their minds and respectively think it will either have no impact on it or improve it.

We can also obtain the standard errors of \( \hat{\theta} = SE(\theta) = (0.375, 0.117, 0.212, 0.041) \) from \( H(\theta) \). Furthermore, we can obtain the probability matrix \( P(t) \) based on the \( Q \) matrix. To do this we use equation (2.2) where the particular form for the probabilities are the equations found in Appendix A.1. If we use \( t = 1/2 \) year we have:

\[
P(1/2) = \begin{bmatrix}
0.353 & 0.347 & 0.300 \\
0.083 & 0.313 & 0.604 \\
0.011 & 0.094 & 0.895
\end{bmatrix}
\]

Note since we are assuming time-homogeneous intensities we can use equation (2.2) for any time interval length.
Thus, we can see from the data, after 6 months subjects who initially believed the medication would improve quality of life tended to stay in that state with probability of about 0.9. Those who started in state 2, and were unsure about the impact of HAART tended to either move to state 3 or remain where they were. Lastly, those who started in state 1, and felt HAART would worsen their quality of life, had about an equal chance of transitioning to any of the other states or remaining in the current one. Examining the transition matrix over a five year interval we have:

\[
P(5) = \begin{bmatrix}
0.041 & 0.141 & 0.818 \\
0.034 & 0.133 & 0.833 \\
0.030 & 0.129 & 0.841
\end{bmatrix}
\]

Now, it seems no matter which state a person started in, 5 years later there was more than an 80% chance that individuals would have transitioned into, or remained in state 3, and felt HAART would improve their quality of life. There was less than a 5% chance that after 5 years individuals would remain or transition into state 1, and feel HAART would negatively affect their quality of life.

One may notice that as the time interval lengthens the transition probabilities tend to converge to particular values, irrespective of the starting states. This idea, known as the stationary probability distribution, can be thought of as the probability of finding the process in state \( j \) regardless of the initial state. By computing \( P^n \) as \( n \to \infty \) the stationary distribution can be found. In particular, using eigenvalue decomposition we can rewrite \( P \) as \( P = ALA^{-1} \) where \( L = \text{diag} \ell_1, \ldots, \ell_K \) \( \ell_k \) are distinct eigenvalues and \( A \) is a \( k \times k \) matrix of right eigenvectors corresponding to \( \ell_k \). Then \( P^n = AL^nA^{-1} \) which is easier to compute. A more comprehensive discussion can be found in Taylor and Karlin [45] and Ross [43].
2.3 Non-homogeneous Transition Intensities

Often, it is too simplistic to assume that the underlying intensities are constant over time. For example, referring back to the HIV example in section 2.2.2, it is more realistic to assume that as science and medicine evolve, the rate at which people change their opinion about HAART will likely change. They are likely to quickly accept that newer medications will improve their quality of life. Therefore, we need to employ methods that can account for intensities that are a function of time. Define $\theta$ as a time dependent vector of intensities, $\theta = \{\theta(t) | 0 < t < T\}$, that we are interested in estimating where $T$ is the maximum observed time. We shall refer to $\theta(t)$ in the context of non-homogeneous intensities. In our example $\theta(t) = [q_{12}(t), q_{21}(t), q_{23}(t), q_{32}(t)]$, each entry is a function of time.

One method suggested by Andersen [4] employs the product integral, denoted as $\Pi$, to relate the transition probability matrix to the cumulative intensities. The product integral can be defined as

$$Z(t) = \Pi_{s \in [0,t]}(I + X(ds)) = \lim_{\max|t_i-t_{i-1}| \to 0} \prod (I + X(t_i) - X(t_{i-1}))$$

where $X(t)$ is a $K \times K$ matrix valued function of time $t$ and $I$ is the identity matrix. In the MSM context, let $A$ be a $K \times K$ matrix where the entries $A_{uv}, u \neq v$ are the integrated intensity function for transitions from state $u$ to $v$. Then the relationship between $P$ and $A$ can be found as

$$P(s,t) = \Pi_{(s,t]}(I + dA(u))$$

The method uses the Nelson-Aalen estimator as an estimate of the cumulative intensities. This method, requires the use of the cumulative intensities and furthermore seems difficult to implement easily in the context of ordinal MSM models. For more detail the reader is referred to Andersen [4].

A more common and easily implemented method to estimate $\theta(t)$ is to use a piecewise
constant (PWC) approach which we will now discuss briefly.

### 2.3.1 Piecewise Constant Models

The idea, a simple and straightforward one, is to divide up the entire time interval, into $R$ contiguous, disjoint segments, $\tau_1, \ldots, \tau_R$, where $\tau_r$ is the interval from time $t_{r-1}$ to $t_r$ and $\tau_r < \tau_{r'}$ if $r < r'$. We treat each segment as a distinct time homogeneous process and apply the methods discussed above. The segments need not be the same size. This idea, suggested by Faddy [13], and more explicitly by Kay [25] is an extension of the time homogeneous case. Suppose we have segments, then $Q$ and $\theta$ all depend on time as:

\[
Q(t) = Q_r \quad t \in \tau_r \quad (2.15)
\]
\[
\theta(t) = \theta_r \quad t \in \tau_r \quad (2.16)
\]

where $Q_r$ and $\theta_r$ can be found using the method discussed above. Computing $P(0, t_j)$, for a $t_j$ in segment $\tau_r$, entails multiplying all of the transition matrices across the various intervals as seen below:

\[
P(0, t_j) = \prod_{b=1}^{r-1} P(b)(\tau_b) P(\tau_{r-1}, t_j) \quad (2.17)
\]

where $P(b)$ is the transition matrix using $Q_b$ for the $b^{th}$ segment, $\tau_b$.

In a case where subjects are observed on an equally spaced grid and segments are divided up along these time points, then $P_{uv}(0, t_j)$ would simply be the $(u, v)^{th}$ element of the matrix in equation (2.17). However, when data are not equally spaced, (or, for that matter, they may be equally spaced but segments are not created at the breakpoints), then observations would be considered missing at the breakpoints. To resolve this, Kay [25] suggested a method to account for all possible pathways between the last observed state in segment $b - 1$ and the first observed state in segment $b$.

For example, suppose a breakpoint is created at a point $t'$ between two time points $t_j$
and $t_k$. Then the likelihood contribution from interval $t_j, t_k$ for individual $i$ can be found as:

$$L_i = \sum_{l=1}^{K} P_{ul}^{(1)}(t_j, t') P_{lv}^{(2)}(t', t_k)$$

(2.18)

for states $u,v$. A likelihood ratio test can be used to determine whether the piecewise constant model is a better fit than the constant model. In the following section we will extend the example from section 2.2.2 to the piecewise constant framework. The likelihood ratio test will be used to determine whether there is a difference between the constant and piecewise constant approaches with regards to the HIV data.

### 2.3.2 Example: Polaris Cohort Study (Continued)

Using the HIV data described in section 2.2.2 we will construct a piecewise constant model to examine whether the intensities may change over time. We chose cutpoints at times 2 and 5 years. Participants who do not have observations at times 2 or 5 will contribute to the likelihood using equation (2.18). In particular, with 3 segments this necessitates the computation of all possible pathways from the last observed time in segment 1 to the first observed time in segment 2 and from the last observed time in segment 2 to the first observed time in segment 3. Again, the msm package was used to fit these models.

Initial parameter estimates were obtained in the same manner as the time homogeneous case in section 2.2.2. We require estimates of $Q(t)$, which can be broken into the following three distinct intensity matrices,

$$Q(t) = \begin{cases} 
Q_1, & \text{if } 0 \leq t < 2 \\
Q_2, & \text{if } 2 \leq t < 5 \\
Q_3, & \text{if } 5 \leq t < 7.5 
\end{cases}$$
Figure 2.1: Intensity estimates for both constant and piecewise constant intensities; (—) \( q_{12} \); (—) \( q_{21} \); (—) \( q_{23} \); (—) \( q_{32} \).

Applying the \texttt{msm} function to each of the three intervals separately we found the estimated quantities \( \hat{Q}_1, \hat{Q}_2, \hat{Q}_3 \) as

\[
\hat{Q}_1 = \begin{bmatrix}
-0.53 & 0.53 & 0 \\
0.10 & -2.51 & 2.42 \\
0 & 0.15 & -0.15
\end{bmatrix}
\quad \hat{Q}_2 = \begin{bmatrix}
-1.91 & 1.91 & 0 \\
0.37 & -1.60 & 1.22 \\
0 & 0.23 & -0.23
\end{bmatrix}
\quad \hat{Q}_3 = \begin{bmatrix}
-1.04 & 1.04 & 0 \\
0.30 & -1.33 & 1.03 \\
0 & 0.15 & -0.15
\end{bmatrix}
\]

Figure 2.1 compares the PWC estimated intensities with the previous time homogeneous estimates obtained in section 2.2.2. On the right side, looking at the PWC intensities we can see that there is not much variation in the piecewise estimates for intensities \( q_{32} \) and \( q_{21} \) and they are close to the estimates obtained from the constant method. This implies that participants starting in either state 2 or 3 were unlikely to transition to lower states over the course of the study as the rate of transition did not vary much. However, the other two intensities, \( q_{23} \) and \( q_{12} \), appear to vary over time and are significantly different than the time homogeneous intensities. For those starting in state 2 they had a high rate of transition...
up to state 3 during the first couple of years and a moderate rate of transition during the following three years. For subjects starting in state 1, they initially had a relatively low rate of transition up to state 2 and during years 2 to 5 of follow up they had a significantly high rate of transition from state 1 to 2. During the last 2.5 years of study both of these groups had moderate rates of transition out of states 1 and 2 but these rates were lower than the earlier years of the study. Though subjects do eventually transition from the lower rated states to higher ones, those in the lowest rated state, (make it worse), appear to take more time to do so, with the bulk of the transitions occurring after the second year on study. The likelihood ratio test below will test whether there is a significant difference between the piecewise and constant estimates.

In addition to the intensity matrices, we obtained estimates of $\hat{P}(0, 1/2)$ and $\hat{P}(0, 5)$ which were:

$$
\hat{P}(0, 1/2) = \begin{bmatrix}
0.599 & 0.145 & 0.256 \\
0.027 & 0.125 & 0.848 \\
0.003 & 0.052 & 0.945
\end{bmatrix}, \quad \hat{P}(0, 5) = \begin{bmatrix}
0.087 & 0.071 & 0.842 \\
0.013 & 0.058 & 0.929 \\
0.010 & 0.057 & 0.933
\end{bmatrix}
$$

Comparing the current probability transition matrices to the ones from the time-homogeneous example, we can see after six months the probability of being in state 3 is slightly higher if the person originated in states 3 or 2 and is slightly lower (0.256 vs. 0.30) if the person originated in state 1. At the end of year 5, using the piecewise approach, we can see the probability of being in state 3, regardless of originating state, is higher than when the constant method was used, with those who originated in state 3 or 2 having a more than 90% chance of occupying those states after 5 years. Examining Figure 2.1(b), a plausible explanation can be the relatively large intensity value for $q_{23}$ early on in the study. This increases the probability of transitioning into state 3 as compared to later on in the study. Another interesting observation can be seen by examining $q_{12}$. Unlike $q_{23}$, during the first
couple of years on study the intensity is low and then jumps higher during years 2 through 5. This seems to imply that there is a delayed effect for those transitioning from the lowest rated state to a higher one, while those transitioning from state 2 up to state 3 seem to do so at the beginning of study and stabilize over time.

To determine whether the piecewise model is a better fit to the data than the straightforward constant approach, we employ a likelihood ratio test. Under the null hypothesis we assume \( q_{ijr} = q_{ij} \) for all intervals \( \tau_r, r = 1, 2, 3 \). The alternative hypothesis is that at least two of the \( q_{ijr} \) are not equal. Under the null hypothesis we have a total of 4 parameters to estimate while under the alternative we have 12 parameters to estimate, implying 8 degrees of freedom. We have:

\[
-2L_0 = 377.5204 \quad \text{and} \quad -2L_a = 371.3702.
\]

This leads to:

\[
X^2 = -2(L_0 - L_a) = 6.1502 \quad \text{has a chi-square distribution on 8 df}
\]

Therefore, we do not reject \( H_0 \) (\( p = 0.63 \)) at the \( \alpha = 0.05 \) level, and we conclude that there is no improvement in going to the PWC model. One must note that the PWC approach is far from certain to estimate time varying intensities, as its estimates are conditional on the breakpoints. If other breakpoints had been selected, it is possible that the result of the likelihood ratio test may show the piecewise model being a significantly better fit than the constant model.

### 2.3.3 Local Likelihood and EM Algorithm

The difficulty in the estimation of time varying intensities arises because the assumptions mentioned in section 2.1.3 are no longer valid. The existing known relationship between the intensity matrix \( Q(t) \) and transition matrix \( P(t) \) proposed by Andersen [4], discussed in 2.3 is quite complex, difficult to implement and uses the cumulative intensity function. As discussed, a more simplistic approach using piecewise constant intensities attempts to
address this issue, yet this method is less than ideal. It is highly unlikely that intensities are constant over segments of time and experience discontinuities between segments. Therefore, a different approach that permits the intensities to vary over time will be proposed to resolve the issue. It will assume no specific functional form of the intensities and it will make use of the EM algorithm in the context of the local likelihood. We will briefly review both of these topics before proceeding to the next chapter where they are implemented to estimate non-homogeneous intensities.

**EM Algorithm**

The EM (Expectation-Maximization) algorithm [10] is a commonly used algorithm especially in the presence of missing data. Suppose we have incomplete data $X$ and complete data $Y$. $X$ is the observed data and there exists a many-to-one mapping from the unobserved data $Y$ to $X$. Let $X$ have joint density function $f_X(x; \theta)$ and $Y$ have joint density function $f_Y(y; \theta)$ and we have known functions $g_i$ such that $X_i = g_i(Y)$. The idea is to find the joint density of $X$ by integrating over the joint density of $Y$ as:

$$f_X(x; \theta) = \int_D f_Y(y; \theta)dy$$

where $D$ is the subset of $y$ such that $x_i = g_i(y)$. In the presence of discrete data the integral is replaced with a summation.

The complexity arises due to the fact that we can only estimate $\theta$ using the observed, incomplete data. To avoid this problem the EM algorithm is employed. The idea is to use the observed data to “fill in” the missing values of the complete data. In the E Step we condition on the incomplete data and find the expectation of the likelihood function which at the $p^{th}$ iteration is:

$$\ell^{(p)}(\theta) = E[\log f_Y(y; \theta)|X = x; \hat{\theta}^{(p)}]$$
The M Step of the algorithm provides an updated estimate of $\theta$. This new update, $\hat{\theta}^{(p+1)}$, is found by maximizing the expected likelihood function found in the E Step. Typically, numerical procedures such as Fisher Scoring or Newton-Raphson are used to obtain these estimates. These two steps are iterated until a prespecified convergence criterion is met. Our approach to estimating non-homogeneous intensity matrices will employ the EM algorithm and will be discussed in more detail in the following chapter.

Local Likelihood

Local likelihood as discussed in Tibshirani and Hastie [46] and Loader [33], extends the smoothing techniques found in scatterplot smoothing to the likelihood setting. More detail can be found in those references. Suppose we have $n$ independent pairs of data $(x_i, y_i)$, $i = 1, \ldots, n$ from an independent variable $X$ and a response variable $Y$. If we define $\mu(x_i)$ as a smooth yet unknown function, and $\epsilon_i$ as an error term then we assume a model of

$$Y_i = \mu(x_i) + \epsilon_i$$

The $\epsilon_i$s are assumed to be independent and identically distributed random variables with mean zero and finite variance. Locally, about a point $x_0$, the smooth function, $\mu(\cdot)$, is assumed to be well approximated by a polynomial of degree $p$,

$$\mu(x) = a_0 + a_1(x_0 - x) + \frac{1}{2}a_2(x_0 - x)^2 + \ldots + \frac{1}{n}a_n(x_0 - x)^p \quad x \in \mathcal{N}(x_0) \quad (2.19)$$

where $\mathcal{N}(x_0)$ denotes a neighbourhood of $x_0$ and weighted least squares, as outlined below, can be used to find $\hat{\mu}(x_0)$.

As discussed in Cleveland [9], there are four items that must be specified for local scatterplot smoothing. We must determine the degree of the polynomial, $p$, a convergence criterion,
C, a neighbourhood size, h, and a weight function, K. The first item, choosing the degree of the polynomial is a choice that is made based on the bias-variance trade off. While selecting a higher order polynomial will result in less bias, the large number of coefficients will increase the variability of the estimate. In addition, practical considerations, such as computational time and ability, typically limit the polynomial to degree two or a quadratic function, which will often suffice. The second item, the convergence criterion, is typically specified by the user and will not be discussed here. The last two items, the neighbourhood size and weight function will now be discussed in more detail.

Selecting the neighbourhood size will determine the smoothness of the estimate $\hat{\mu}$. For a point $x_0$, define $h(x_0)$ as a bandwidth of size h and a span or window of size $[x_0 - h(x_0), x_0 + h(x_0)]$. The bandwidth determines the size of the window, and only observations falling within this pre-specified window are used. Choosing a small bandwidth may result in too few observations falling within the window leading to a noisy and highly variable fit. Alternatively, selecting a large bandwidth can result in significant bias as it can over smooth the underlying function and lead to poor estimates locally. Various methods, as we will discuss in later chapters, can be used to aid in bandwidth selection to help with the bias-variance trade-off. The weight function, or kernel, influences the appearance of the estimate and can also minimally affect smoothness of the estimate, though not to the same degree as the bandwidth. Silverman [44] provides a thorough examination of this topic though in general the kernel must satisfy the following property: $\int_{-\infty}^{+\infty} K(u)du = 1$. Other, often desirable, properties of the kernel are: (1) it is symmetric about a point $x_0$, $K(x_0 - x) = K(x_0 + x)$ and (2) the weights are non-negative, with those observations outside the window getting assigned weights of zero and typically, closer observations to $x_0$ are assigned larger weights than those further away. Several examples of commonly used kernels are uniform, Gaussian and Epanechnikov. The coefficients of the local regression
model can then be found by minimizing,

\[ \sum_{i=1}^{n} K_h(x_i - x)(Y_i - a_0 - a_1(x_i - x) - \frac{1}{2}a_2(x_i - x)^2 - \ldots - \frac{1}{n}a_n(x_i - x)^p)^2 \]

where

\[ K_h(x_i - x) = K \left( \frac{x_i - x}{h(x)} \right) \cdot \frac{1}{h} \]

Local likelihood extends local regression to the likelihood setting. Again, suppose there are \( n \) independent realizations of two random variables \( X \) and \( Y \) with density function

\[ Y_i|X = x_i \sim f(Y, \theta_i) \]

where \( \theta \) is a function of the covariates \( x_i \), \( \theta_i = \theta(x_i) \). Then the global likelihood can be constructed as:

\[ L(\theta) = \prod_{i=1}^{n} f(y_i, \theta_i) \]
\[ \ell(\theta) = \sum_{i=1}^{n} \log f(y_i, \theta_i) \]

Unlike an ordinary approach which would assume a parametric model and use maximum likelihood estimation to estimate \( \theta \), the local likelihood approach only assumes that \( \theta \) is a smooth function of \( x \). Using a polynomial of degree \( p \) to estimate \( \theta \), the local likelihood can be written as,

\[ \ell_x(a) = \sum_{i=1}^{n} K_h(x_i - x) \log[f(Y_i, a_0 + a_1(x_i - x) + \frac{1}{2}a_2(x_i - x)^2 + \ldots + \frac{1}{n}a_n(x_i - x)^p)] \] (2.20)

In particular, setting \( p = 0 \) yields a local constant estimate, or \( \hat{\theta}(x) = \hat{a}_0 \), and using a polynomial of degree 1 yields a local linear estimate.
Local EM

To enable us to estimate non-homogeneous intensities, we will use a local EM method introduced by Betensky et al. [7] which uses the EM algorithm in the context of the local likelihood. The basic idea is to use the two step EM algorithm in the local setting. The E-step enables us to find the complete data, while the M-step locally maximizes the likelihood to obtain local parameter estimates. We will briefly outline the algorithm in the general sense and in the following chapter we will illustrate how it applies in the MSM framework.

Let $G$ be the set of grid points and

- Choose an initial value for $\theta = \hat{\theta}^{(0)}(x)$. This value will usually be chosen using an educated guess based on the particular scenario.

- For each $x \in G$ compute the E Step

$$\ell_x(\theta) = E[\log K_h(x_i - x)f_Y(y; \theta)|X = x; \hat{\theta}^{(p)}]$$

- For each $x \in G$, the M step, which provides an updated value for $\theta$, can be found by calculating

$$\hat{\theta}^{(p+1)}(x) = \arg \max_{\theta} \ell_x(\theta)$$

- The E and M steps are repeated until convergence has been achieved.

It is important to realize that we have only provided a general framework for the local EM, depending on the density function the E and M steps may be very simple or difficult to compute. In the following chapter the local EM will be introduced to MSM models with an ordinal outcome. We will illustrate how both the E and M steps can be computed and
illustrate an example of the algorithm by analyzing a simulated dataset. To do this, R code was developed to implement the local EM in the MSM context.
Chapter 3

Non-Homogeneous Intensities

3.1 Local EM Method

The previous chapter discussed methods as they apply to intensities that are constant or piecewise constant with respect to time. In this chapter we turn our attention to developing a method that will enable us to estimate time varying intensities. This method will be developed using the local EM algorithm discussed at the end of the last chapter. Regardless of whether the collected data is evenly spaced and balanced or whether it is unequally spaced and unbalanced, the method will enable the non-homogeneous intensities to be estimated. The general case, involving unbalanced, unequally spaced data will be presented and the balanced, equally spaced case can be derived as a special case. In addition, this new approach will allow us to handle missing data. Missing data can arise simply because some subjects are purposely observed less frequently than others or simply because patients have missed follow-up appointments. Unlike the standard existing algorithms, our method will explicitly account for the missing data of either kind. In order to provide the details of the local EM algorithm in the MSM context we will first introduce (and review) some notation that will be used for the remainder of the thesis.
Notation

$k = 1, \ldots, K$ is the occupied state

Let $i = 1, \ldots, N$ be individuals

$t_{ij} = \text{time of } j^{th} \text{ observation for individual } i, j = 1, \ldots, J_i$ indexes observed times

$J = \max_i J_i$

$t_i^- (x) = t_{ir}$ such that $r = \max_j (t_{ij} \leq x)$

$t_i^+ (x) = t_{ir}$ such that $r = \min_j (t_{ij} \geq x)$

$Y = \{Y_i(t_{ij})|i = 1, \ldots, N, j = 1, \ldots, J_i\}$

$Y_i(t) =$ an indicator vector of length $K$ with entries $Y_{ik}(t)$ where

$Y_{ik}(t) = \begin{cases} 1, & \text{if individual } i \text{ is in state } k \text{ at time } t \\ 0 & \text{otherwise} \end{cases}$

$S_i(t_{ij}) = k | Y_{ik}(t) = 1$

3.1.1 Local EM in MSM Models

To construct the local EM in the context of MSM models, let us first imagine that all data is available. That is, suppose there are $N$ individuals that are continuously observed over the interval $(0, T)$ so that all transitions are observed. Furthermore, recall that $\theta$ is of length $c$ and contains $q_{Y_i(t), Y_i(t+\Delta t)}$ such that $Y_i(t) \neq Y_i(t + \Delta t)$. Then, using the likelihood of a point
process the global likelihood and log likelihood can be written as:

\[ L(\theta|Y) = \prod_{i=1}^{N} P(\text{No transitions where there aren’t any}) \times \prod_{i=1}^{N} P(\text{transitions at exact times}) \]

\[ L(\theta|Y) = \prod_{i=1}^{N} \left[ \exp \left( -\int_{0}^{T} \sum_{v \neq S_i(s)} q_{S_i(s)v}(s) ds \right) \prod_{j=1}^{J_i} q_{S_i(t_{ij-1})S_i(t_{ij})} \right] \]

\[ \log L(\theta|Y) = \sum_{i=1}^{N} \left[ -\int_{0}^{T} \sum_{v \neq S_i(s)} q_{S_i(s)v}(s) ds + \sum_{j=1}^{J_i} \log q_{S_i(t_{ij-1})S_i(t_{ij})} \right] \]

\[ \ell(\theta|Y) = \sum_{i=1}^{N} \left[ -\int_{0}^{T} \sum_{v \neq S_i(s)} q_{S_i(s)v}(s) ds + \sum_{j=1}^{J_i} \log q_{S_i(t_{ij-1})S_i(t_{ij})} \right] \]

\[
(3.1)
\]

where \( q_{S_i(s)v} \) is the intensity of transitioning out of state \( S_i(s) \) and into state \( v \). In addition, \( q_{S_i(t_{ij-1})S_i(t_{ij})} \) is the intensity associated with the known transition \( S_i(t_{ij-1}) \) to \( S_i(t_{ij}) \).

In the context of the local likelihood we can place a kernel into equation (3.1) as:

\[ \ell_t(\theta) = K_h(t_{ij} - t) \cdot \ell(\theta|Y) \]

\[ = \sum_{i=1}^{N} -\int_{0}^{T} \sum_{v \neq S_i(s)} K_h(t_{ij} - t) q_{S_i(s)v}(s) ds + \sum_{i=1}^{N} \sum_{j=1}^{J_i} K_h(t_{ij} - t) \log q_{S_i(t_{ij-1})S_i(t_{ij})} \]

\[
(3.2)
\]

where \( \log q_{S_i(t_{ij-1})S_i(t_{ij})} \) can be approximated locally using a polynomial as discussed in section 2.3.3.

**Construction of Local EM**

A discrete approximation of (3.2) can be constructed which can then be used in realistic scenarios where not all the data is observed. A grid of size \( G \) points can be constructed on which the likelihood will be evaluated at these discrete time points. Define \( \delta \) as a parameter that controls the coarseness of the grid. The total number of points on the grid \( G \) is \( J/\delta \).

When \( \delta = 1 \) there are \( J \) points and a value of less than one indicates a finer grid for \( G \), while
3 Non-Homogeneous Intensities

a value greater than 1 indicates a coarser grid. Furthermore, define \( g_j = \delta \cdot j, \quad j = 1, \ldots, G \) as the \( j^{th} \) time point along the grid. Then, when all transitions are observed, we can rewrite equation (3.1) as:

\[
\ell(\theta|Y) = \sum_{i=1}^{N} \log P\{Y_i(s)|s \in (0, T_i)\}
\]  

(3.3)

where \( T_i \) is the maximum time for individual \( i \), though typically \( T_i = T \) for all \( i \). Equation (3.3) is (3.1) rewritten using the transition probabilities which are functions of \( \theta \). To evaluate the likelihood we can use a discrete approximation on a grid of size \( \delta \) by:

\[
\ell(\theta|Y) \approx \bar{\ell}(\theta|Y) = \sum_{i=1}^{N} \log P\{Y_i(s)|s \in (g_1, \ldots, g_J)\}
\]  

(3.4)

where \( \bar{\ell} \) is the likelihood for the reduced data. Recalling from section 2.2.1 that:

\[
\gamma_{iuv} = \begin{cases} 
1, & \text{if } S_i(t_{ij} - 1) = u \text{ and } S_i(t_{ij}) = v \\
0, & \text{otherwise}
\end{cases}
\]

and using the format of (2.11) the local likelihood corresponding to (3.4) can be written as:

\[
\bar{\ell}_g(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{G} \sum_{u,v=1}^{K} K_h(g_j - t) \gamma_{iuv} \log P_{uv}(g_{j-1}, g_j|\theta)
\]  

(3.5)

Being that this is a discrete approximation we are losing some data. Therefore, decreasing \( \delta \) will lead to a better approximation, which will be discussed in more detail in chapter 4.

In reality, all of a subject’s transitions over an extended period of time cannot be observed. Individuals on study are typically scheduled to come in at prespecified times. Typically participants are not observed aside from these specified times and often, for various reasons, the data is missing at those prespecified time points and is frequently irregularly spaced. Therefore, the discrete approximation suggested here will be used in this scenario. The local EM method will allow us to estimate the underlying time varying intensities in
Suppose we have complete data $X$ and incomplete or observed data $Z$, then general idea for the EM algorithm, as outlined in section 2.3.3 is:

E-Step: Use the observed data, $Z$, together with an estimated intensity vector $\hat{\theta}(\cdot)$ to find $E[X|Z]$ where $\hat{X}$ is the data that is wanted and $\hat{\theta}(\cdot)$ is the estimated intensity vector over the entire time interval. That is, find:

$$X^{(p)} = E\left[X|Z; \hat{\theta}^{(p)}(\cdot)\right]$$

M-Step: Using $X^{(p)}$, find an updated estimate for $\hat{\theta}^{(p+1)}(\cdot)$

In our context we require the expected number of transitions between states for unobserved time points. That is, at missing time points $\gamma_{ijuv}$ is unknown. Therefore, we construct a $K \times K$ matrix, $\tilde{\gamma}_{ij}$ with $u,v$ entry $\tilde{\gamma}_{ijuv}$ which are the expected number of transitions from $u$ to $v$. This can be better understood by examining Figure 3.1 which summarizes the 3 unique scenarios that arise when constructing, $\tilde{\gamma}_{ij}$. For a unique subject we may have any of the following: (1) exactly one missing observation between 2 known observations, $t_2$, or a missing observation adjacent to a known observation, $t_4$ to $t_5$ (2) no missing observations between any two observational times, times $t_3$ to $t_4$ (3) two or more consecutive missing observations, times $t_5$ and $t_6$. These three scenarios correspond to the following three matrices for $\tilde{\gamma}_{ij}$

\[
\begin{align*}
\text{(1)} & \quad \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
\text{(2)} & \quad \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
\text{(3)} & \quad \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}
\end{align*}
\]
Figure 3.1: Illustration of the 3 scenarios for computing $\tilde{\gamma}_{ij}$. Times 1 to 2, 2 to 3, 4 to 5 and 6 to 7 correspond to scenario (1). Times 3 to 4, 7 to 8, and 8 to 9 are known, scenario (2). Time 5 to 6 corresponds to scenario (3) and requires the computation of the joint probability.
Based on this idea, in our framework and using a grid of $G$ points, the general equation for the $p^{th}$ iteration of local EM algorithm can be written, as:

$$
\hat{\theta}^{(p+1)}(t) = \arg \max_\theta \mathbb{E} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{G} \sum_{u,v=1}^{K} K_h(g_j - t) \gamma_{ijuv} \log P_{u,v}(g_{j-1}, g_j; \theta) \bigg| Y; \theta^{(p)}(\cdot) \right\}
$$

(3.6)

which can be divided into the E-step

$$
\tilde{\gamma}^{(p)}_{ijuv} = \mathbb{E} \left[ \gamma_{ijuv} \big| Y; \theta^{(p)}(\cdot) \right]
$$

(3.7)

and letting $\tilde{\gamma}_{ijuv} = \mathbb{E} \left[ \gamma_{ijuv}; \theta^{(p)}(\cdot), Y \right]$ be the missing data, the M-step can be used to obtain $\hat{\theta}(t)$ by maximizing:

$$
\hat{\theta}^{(p+1)}(t) = \arg \max_\theta \sum_{i=1}^{N} \sum_{j=1}^{G} \sum_{u,v=1}^{K} K_h(g_j - t) [\tilde{\gamma}^{(p)}_{ijuv} \log (P_{u,v}(g_{j-1}, g_j; \theta))]}
$$

(3.8)

at the $(p + 1)^{th}$ iteration.

The E-step, details of which will be provided in section 3.1.2, enables us to provide the expected number of transitions for the unobserved states. Unlike many other EM implementations, in the context of MSM models computing the expected observations in the E-step is not sufficient to produce the iterations in (3.8). Rather the expected probability of pairs of successive states, in other words the probability of undergoing each possible transition, is required. Since the transition probabilities, $P_{uv}$, governed by $\theta$, are known in the E-step, they can be taken outside the expectation and the log-likelihood is linear in $\gamma_{ijuv}$. This permits the expectation to be taken as shown in equation (3.7). In the M-step, equation (3.8), the $\tilde{\gamma}_{ijuv}$ will be inserted and computed based on the appropriate scenario. Notice that when a subject’s transition between two adjacent grid points is observed then $\tilde{\gamma}_{ijuv} = \gamma_{ijuv}$ as in equation (2.11). Successive applications of the M-step for all values of $g$ produces
an estimate for the complete $\hat{\theta}(\cdot)$. The M-Step is dependent on the kernel selected and bandwidth size. Different choices will provide different values for $\hat{\theta}(t)$. Though we have described the local EM algorithm we have glossed over many of the details of the E-step which we will now discuss.

### 3.1.2 Details of the E-Step

The purpose of the E-step is to enable us to deal with the missing observations, whether they may be missing by design or for other reasons. In particular, we need to find $\tilde{\gamma}_{ijuv}$ to enable the computation of (3.8). This can be done by finding:

$$
\tilde{\gamma}_{ij} = E[Y_i(g_j)Y_i(g_j)'|Y]
= E[Y_i(g_j)Y_i(g_j)'|Y_i(t_i^- (g_j)), Y_i(t_i^+ (g_j))]
$$

(3.9)

where (3.9) is true by the Markov property. In order to facilitate the computation of (3.9), we will first find the expected transitions in a scenario where one of the states is known, scenario (1) above. Without loss of generality let us assume that $S_i(g_{j-1})$ is observed and $S_i(g_j)$ is missing. The method is similar for the reverse scenario. Since we observe $S_i(g_{j-1})$ we only need to find:

$$
E[(Y_i(g_j))|Y_i(g_{j-1}), Y_i(t_i^+ (g_j))]
$$

as the other $K - 1$ rows of $\tilde{\gamma}_{ij}$ will simply contain entries of 0. We will first demonstrate the E-step in this scenario with a simple example before providing the general formula.

Suppose a longitudinal study has 3 time points, $t_1, t_2, t_3$ and the outcome of interest has $K = 3$ possible states. An individual is observed at times $t_1$ and $t_3$ but their data is missing at time $t_2$. That is, we need to find $Y_i(t_2)$. Let $Y_i(t_1)$ be the indicator vector corresponding
to the state occupied at \( t_1 \) and similarly \( Y_i(t_3) \) for \( t_3 \). To find \( Y_i(t_2) \) we need:

\[
P(Y_i(t_2) = (Y_{i1}(t_2) \ Y_{i2}(t_2) \ Y_{i3}(t_2))' | Y_i(t_1), Y_i(t_3))
\]

Therefore we need to find the individual elements, \( Y_{ik}(t_2) \), \( k = 1 \ldots K \). These individual elements are mutually exclusive Bernoulli random variables. Therefore in our example with \( K = 3 \) we have

\[
E[Y_i(t_2)|Y_i(t_1), Y_i(t_3)] = P[Y_i(t_2)|Y_i(t_1), Y_i(t_3)]
\]

\[
= P[Y_{ik}(t_2) = 1|Y_i(t_1), Y_i(t_3)] \quad (k = 1, 2, 3)
\]

Although this can be solved by examining each entry of the matrix one at a time, it is much more efficient to use matrix notation, where the probability can be computed as

\[
P(Y_i(t_2)|Y_i(t_1), Y_i(t_3)) = \frac{\text{diag}(P(t_2, t_3)Y_i(t_3)Y_i'(t_1)P(t_1, t_2))}{Y_i'(t_1)P(t_1, t_2)P(t_2, t_3)Y_i(t_3)}
\]

In the more general setting, for any subject at a given grid point \( g_j \) we can find \( Y_i(g_j) \).

When the subject is observed at both the first and last time points, this can be found as:

\[
E[(Y_i(g_j))|Y_i(g_{j-1}), Y_i(t_i^+(g_j))] = \frac{\text{diag}(P(g_j, t_{i}^+(g_j))Y_i(t_{i}^+(g_j))Y_i'(g_{j-1})P(g_{j-1}, g_j))}{Y_i'(g_{j-1})P(g_{j-1}, g_j)P(g_j, t_{i}^+(g_j))Y_i(t_{i}^+(g_j))}
\]

(3.10)

where \( Y_i(t_{i}^+(g_j)) \) is the nearest observed state following \( g_j \) and \( P(g_{j-1}, g_j) \) and \( P(g_j, t_{i}^+(g_j)) \) are the transition probability matrices that go from \( g_{j-1} \) to \( g_j \) and \( g_j \) to \( t_{i}^+(g_j) \) respectively.

An extension of this scenario is when a subject is missing the last observation. The dependence on \( Y_i(t_{i}^+(g_j)) \) is no longer possible. In this case the time of the last observed state, \( t_i^-(g_j) \), and the corresponding state indicator vector, \( Y_i(t_i^-(g_j)) \), are located and observations missing prior to time \( t_i^-(g_j) \) can be found using equation (3.10). For observations
3 Non-Homogeneous Intensities

missing after time $t_i^-(g_J)$ their expected probabilities can be found by:

$$E[Y_i(g_j)|Y_i(t_i^-(g_J))] = Y_i(t_i^-(g_J))P(t_i^-(g_J), g_j) \quad g_j = t_i^-(g_J), \ldots, g_J$$ (3.11)

Details of equations (3.10) and (3.11) can be found in appendix A.3.

When data from the first observation time is missing we elected to assign probabilities to the states based on the distribution of the observable data at the initial time point. For example, with $K = 3$ states and $N = 100$ subjects of which 25 are missing the first observation we can assign them probabilities based on the remaining 75 observations. If the proportions based on the 75 observations are 0.2, 0.4, 0.4 for states 1, 2 and 3 respectively, then all 25 missing observations would be assigned those weights. When very few observations are available at the first time point we may elect to move the starting point to the second or third time point where more observations are available. (Since this Markov process is not time-reversible, a stationary distribution cannot be determined.)

In the presence of two consecutive missing observations along the grid, computing the expectation is slightly more complex. In this scenario the number of individuals transitioning between these unobserved states is unknown. Unlike a missing state that is adjacent to an observed state where equations (3.10) and (3.11) can be used, in the presence of two or more consecutive grid points with unobserved states these equations can no longer be used. To understand this, consider an example of 100 individuals that transition between states over time. When consecutive states are unobserved, the subjects’ transitions remain unknown. Therefore, we must compute the joint probability for any two consecutive unobserved states, $\tilde{\gamma}_{ij}$, which will be used in the construction of the likelihood when the data is missing. Conditioning on $Y_i(t_i^-(g_{j-1}))$ and $Y_i(t_i^+(g_j))$ the expected transitions can be found as

$$\tilde{\gamma}_{ij} = E[Y_i(g_{j-1})Y_i(g_j)|Y_i(t_i^-(g_{j-1})), Y_i(t_i^+(g_j))] =$$

$$\frac{\text{diag}([Y_i(t_i^-(g_{j-1}))]P(t_i^-(g_{j-1}), g_{j-1}))P(g_{j-1}, g_j)\text{diag}(P(g_j, t_i^+(g_j))[Y_i(t_i^+(g_j))])}{Y_i(t_i^-(g_{j-1})) \cdot P(t_i^-(g_{j-1}), t_i^+(g_j)) \cdot Y_i(t_i^+(g_j))}$$ (3.12)
Note that with two consecutive missing observations along the grid $t_i^- (g_{j-1})$ and $t_i^- (g_j)$ are identical.

Equation (3.12) is used for two or more consecutive missing states when states both prior to and following the missing data are available (the derivation for (3.12) can also be found in Appendix A.3). In other situations, equations (3.10) and (3.11) will suffice. We will now illustrate a brief example using each of the three scenarios found in Figure 3.1.

Example:

Referring to Figure 3.1 we can see that there are $K = 3$ states, a total of $J = 9$ time points, with observations missing at times $t_2, t_5$ and $t_6$.

**Time $t_2$:** The expected probabilities can be found using (3.10) with $Y_i(t_i^- (g_{j-1})) = Y_i(1) = (1,0,0)'$, $Y_i(t_i^+ (g_j)) = Y_i(3) = (0,1,0)'$ and $g_j = 2$:

$$E[Y_i(2)|Y_i(1) = (1,0,0)', Y_i(3) = (0,1,0)'] = \frac{\text{diag}(Y_i'(4)P(5,6)Y_i(7)P(5,6))}{Y_i'(4)P(4,7)Y_i(7)}$$

This yields matrix (1) above.

**Time $t_3$ to $t_4$:** This is simply an observed transition from state 2 to state 2 and matrix (2) above can be used.

**Time $t_5$ and $t_6$:** With two consecutive missing observations we need to use (3.12) to construct the form of matrix (3) above for the transition between these two time points

$$E[P(5,6)|Y_i(4) = (0,1,0)', Y_i(7) = (0,0,1)'] = \frac{\text{diag}(Y_i'(4)P(5,6))P(5,6)\text{diag}(Y_i'(7)P(5,6))}{Y_i'(4)P(4,7)Y_i(7)}$$

With all the details of the EM algorithm accounted for we now illustrate the implementation of the algorithm.
Implementation of the local EM

The local EM algorithm can now be readily implemented as follows:

1. Choose a kernel, $K$, bandwidth, $h$, grid size $\delta$, and convergence criterion $C$.

2. Choose reasonable starting values, typically the estimates of time homogeneous intensities, $\theta^{(0)}(t) = \theta$, as discussed in Kalbfleisch and Lawless [23].

3. $\tilde{\gamma}_{ij}^{(0)} = E(\gamma_{ij} | Y, \theta^{(0)}(t))$.

4. Compute $\theta^{(1)}(t) = \arg \max_{\theta} \Pr(\tilde{\gamma}_{ij}^{(0)}, \theta^{(0)}(t))$. Repeat for $t \in (0, G)$.

5. Update: $\tilde{\gamma}_{ij}^{(1)} = E(\gamma_{ij} | Y, \theta^{(1)}(t))$.

6. Repeat until $|\theta^{(p)}(t) - \theta^{(p+1)}(t)| < C$.

In this paragraph we discuss a few general remarks about the implementation. As discussed in section 2.3.3 the choice of bandwidth is the most crucial. The size of the bandwidth will always be in units of the original grid. Suppose we are using a bandwidth of 3 and setting $\delta$ to 1, then for a given point, $t$, on the grid the neighbourhood extends 3 units to both the right and left of $t$. If we decrease $\delta$ to 1/3, then the neighbourhood will extend from $t - 3$ units to $t + 3$ units on the original grid or from $t - 9$ to $t + 9$ units on the new grid. In addition, when implementing the E-Step, one must take care to use the transition matrix for the appropriate time interval. Although the E-Step uses the complete intensity estimate, $\theta^{(p)}(\cdot)$, it uses time varying transition matrices and one must ensure that the probabilities reflect this. Furthermore, we have made no mention about the computation of the M-Step at the endpoints. Generally locally constant estimates tend to produce unreliable results at the endpoints as mentioned in Tibshirani and Hastie [46], Loader [32]. At the endpoints, the kernel covers less data points and therefore this resulted in increased variability. The kernel function was multiplied by a constant to enable it to integrate to 1. For example, at the actual endpoints themselves there is only half as much date in $\mathcal{N}(x_0)$ and therefore the weights of all the data points in the neighbourhood were increased by a factor of 2. Though this may not be ideal, this approach did give us reasonable results and did not lead to convergence problems. The upcoming section will now illustrate the local EM method through the use of simulated datasets.
3.2 Simulation

In this section we examine two separate simulated datasets that were constructed to illustrate the local EM method and compare it to the piecewise constant method. The first simulated dataset contained underlying realistic time varying intensities. The idea of this was to see how the method performed in a realistic setting. The second simulated dataset examined underlying PWC intensities. This was done to see how the local EM performed in an unlikely setting where the intensities were PWC. Although not as realistic, it is necessary to see whether the method can adapt adequately to intensity functions with abrupt changes. Both datasets will now be presented in more detail.

3.2.1 Simulation Setup

The approach for simulating data was similar for both datasets with the exception of the underlying intensities. There were a total of \( N = 300 \) individuals, each with 20 follow-up time points and an outcome of \( K = 3 \) ordinal states. Of the 6000 observations 30%, or 1800, of them were randomly removed to illustrate the algorithm in the context of missing data. For both simulations the intensity matrix had the form

\[
Q(t) = \begin{bmatrix}
-q_{11}(t) & q_{11}(t) & 0 \\
q_{21}(t) & -q_{21}(t) - q_{23}(t) & q_{23}(t) \\
0 & q_{33}(t) & -q_{33}(t)
\end{bmatrix}
\]

and our parameters to be estimated are \( \theta(t) = [q_{11}(t), q_{21}(t), q_{23}(t), q_{33}(t)|0 < t \leq 20] \).

For the time varying scenario the underlying intensities used were:
as shown in Figure 3.2. Realistic scenarios were chosen for each of the intensities. A linear intensity can be reasonable when transitioning between states where the rate of change can increase linearly with time. Both concave and convex intensities represent rates that are initially decreasing (increasing), plateau for a period of time, and then increase (decrease) with time. Lastly, an S-shaped intensity is ideal when rates stay constant in the beginning and end of an interval but increase sometime during the middle of the interval.

For the piecewise constant intensities the underlying intensities used were:

\[ q_{12}(t) = 0.2 + 0.02t \quad 0 < t \leq 20 \]
\[ q_{21}(t) = 0.4(t - 10)^2/100 + 0.2 \quad 0 < t \leq 20 \]
\[ q_{23}(t) = -0.4(t - 10)^2/100 + 0.6 \quad 0 < t \leq 20 \]
\[ q_{32}(t) = 0.2 + (1/(1+\exp(-t-10)/2.5))/1.5 \quad 0 < t \leq 20 \]
These intensities were chosen to illustrate a unique example and are not thought to be very realistic. Due to the ordinal nature of the outcome, in both simulated datasets $\theta(t)$ is of length 4. Only from state 2 can individuals transition to 1 of 2 states (states 1 or 3). For individuals occupying states 1 or 3, individuals must transition through state 2 before proceeding to another state.

Using Monte Carlo methods, a non-homogeneous Poisson process was used to simulate the data from the above intensities. We used the approach outlined by Lewis and Shedler [29] that simulates a non-homogeneous Poisson process by thinning and it is briefly outlined here. We use the notation $\lambda$ in place of $q_{uv}$. Let $\lambda(t_i)$ be the intensity evaluated at time $t_i$ and $\lambda \cdot 20$ be an upper bound placed on the intensity and used to find the maximum mean number of transitions that occurred during the interval $(0, 20)$. That is $\lambda \cdot 20 = max_s \lambda(s)$. A Poisson random variable was used to determine the number of transitions and if greater than zero a uniform random variable was used to determine transition times. The process was then thinned using a binomial random variable with probability $\frac{\lambda(t_i)}{\lambda \cdot 20}$ and if accepted, the minimum transition time was used as the transition time. A multinomial probability was used to determine which state the individual transitions into using the appropriate intensities at the transition time. Once the observed transition times and states were simulated, a grid of size 1 unit was constructed with pre chosen observed time points. To be
realistic, the simulated transition times which in reality would not be observed, were not included. Rather, only states at observed time points were recorded. When a transition took place between observed time points \( t_{ij-1} \) and \( t_{ij} \) then the state at time \( t_{ij} \) reflected the new state. With time varying intensities, the R package \texttt{msm} can no longer be used to estimate the intensities. Instead, R code found in Appendix A.5, was used to implement the local EM algorithm. User defined functions were employed and the \texttt{optim} function was used to solve equation (3.8). A lower bound of 0 was placed on the intensities due to the constraints discussed in section 2.1.2. No convergence problems were found even with intensity estimates close to 0. Nevertheless, should this problem arise, one may wish to reparameterize the intensities using a log transformation as suggested by Kalbfleisch and Lawless [23].

### 3.2.2 Results: Time Varying Intensities

Figures 3.3 and 3.4 below illustrate the performance of the local EM algorithm. Various bandwidth sizes and grid sizes were chosen. The wide black line represents the true intensity values. Both the time varying and piecewise constant intensities are displayed in the plots to contrast the two approaches. For the time varying approach, an Epanechnikov kernel [44] was selected with bandwidth size \( h \) with:

\[
K_h(u) = \frac{3}{4h} (1 - \left(\frac{u}{h}\right)^2) I(|u| \leq h)
\]

The support of the kernel was \( 2h \). Examining Figure 3.3, the smallest bandwidth, \( h = 1 \), (red line), produces some very rough and unreliable results, including sharp peaks and troughs. A bandwidth of 1 provides weights of zero to adjacent time points and therefore should be considered a running mean or moving average smoother [18] with no neighbouring time points. Increasing the bandwidth size yields smoother estimates (blue line), though as the bandwidth increases (green line) it tends to underestimate peaks and overestimate troughs. Similar, though more muted, findings can be seen in Figure 3.4. The piecewise constant method used the \texttt{msm} package to obtain estimates. Comparing the two approaches, it indicates that the local EM method provides estimates that
are ‘closer’ to the true intensities than the PWC estimates. It also appears that this approach is better than the constant method. Chapter 4 will examine methods to compare the constant, piecewise constant and local EM approaches. In addition to varying the bandwidth size, varying $\delta$, the grid size, results in different estimates for $\theta(t)$. Figure 3.5 demonstrates the effect of varying the coarseness of the grid while maintaining the same bandwidth. As can be seen, creating a finer grid (green line), results in smoother estimates for $\theta(t)$. Aside from varying the bandwidth size and grid size, we also used several different initial starting points for $\hat{\theta}^{(0)}(t)$ that all resulted in the same estimates $\hat{\theta}(t)$ when using the local EM algorithm. A uniform kernel was also used and results were very similar to the Epanechnikov kernel. Section 3.3 discusses these items in more detail.

Figure 3.3: Time Varying intensity estimates for various bandwidths with a fixed grid of $\delta = 1$: true $\theta$ (---); $\theta$ PWC (- - -); BW 1 (--); BW3 (---); BW8 (---).
Figure 3.4: Time Varying intensities for various bandwidths with a fixed grid of $\delta = 1/3$: true $\theta$ (—); $\theta$ PWC ( - - -); BW 1 (—); BW3 (—); BW8 (—).
Figure 3.5: Time varying intensity estimates for a fixed bandwidth of 2: true $\theta$ (—); $\delta = 1$ (—); $\delta = 1/2$ (—); $\delta = 1/3$ (—).
3.2.3 Results: Piecewise Constant Intensities

As mentioned earlier, the primary goal of the simulated dataset with PWC intensities was to demonstrate the local EM method when intensities have discontinuities and abrupt jumps. In this case we would expect the piecewise constant method to estimate the intensities better than the local EM method. Again an Epanechnikov kernel was used with various bandwidth and grid sizes. Figures 3.6 and 3.7 below examine the effect of changing the bandwidth size for various grid levels. In addition, Figure 3.8 looks at the effect of changing the grid size while holding the bandwidth constant. First looking at the local EM method it is obvious that a small bandwidth of size 1 (red line) seems to reasonably approximate the underlying intensities and certainly outperforms the larger bandwidths of size 3 and 8 (blue and green lines). This is probably because the smaller bandwidths are better suited to estimate the intensities at their jump points while larger intensities are more likely to smooth these discontinuities. The piecewise constant intensities appear to approximate the true intensities relatively well with the exception of $q_{23}$ from time 15 - 20 and $q_{32}$ from times 10 - 20. Examining the plots further, the intensity estimates obtained from both the local EM method, using smaller bandwidths, as well as the piecewise constant method appear to estimate the intensity well, with some instances better estimated by one method versus the other. Furthermore, decreasing the grid size, $(\delta = 1/3)$, tends to flatten out the sharp peaks and valleys while also possibly underestimating these peaks and valleys.

Overall, how do we decide which method is better or which bandwidth is ideal? The next chapter will examine these issues in more detail. We will discuss methods for bandwidth selection as well as approaches for constructing confidence intervals.
Figure 3.6: Piecewise constant intensity estimates for various bandwidths ($\delta = 1$): true $\theta$ (textcolorblack—); $\theta$ PWC (- - -); BW 1 (—); BW3 (—); BW8 (—)
Figure 3.7: Piecewise constant intensity estimates for various bandwidths ($\delta = 1/3$): true $\theta$ (textcolorblack—); $\theta$ PWC (- - -); BW 1 (—); BW3 (—); BW8 (—)
Figure 3.8: Piecewise constant intensity estimates for a fixed bandwidth of 2; true $\theta$ (---); $\delta = 1$ (---); $\delta = 1/2$ (---); $\delta = 1/3$ (---);
3.2.4 Additional Simulations

In addition to the two simulated datasets above, we created other simulated datasets to examine the suitability of the local EM method in other contexts. A simulated dataset with time homogeneous intensities was studied to ascertain whether the method would perform adequately in this situation. Although it is not expected to have the exact same estimate as the time homogeneous intensities obtained from the msm package it should nonetheless be close. Figure 3.9 looks at both the local EM estimate as well as the msm estimate. We can see that both the msm estimate and the local EM estimates are close to the true intensity. Although the local EM estimates are not exactly flat they are relatively constant with respect to time with larger bandwidths being flatter than smaller ones. It is likely that decreasing the grid size would result in smoother, flatter intensity estimates. In addition, constructing a confidence interval for the intensities produced by the local EM algorithm, the method for which will be described in the following chapter, would likely contain a straight line. This indicates that the local EM can adequately capture time homogeneous intensities.

Another simulated dataset was used to determine how the local EM performs when the data within subjects was independent. In this scenario a uniform random variable was used to randomly assign subjects, in equal proportion, to 1 of 3 states. When there is no underlying temporal dependence one would expect the underlying intensity to experience extreme fluctuation with no clear trend. This is a limit of the multi-state model as transition intensities would not converge and tend to infinity. Figure 3.10 displays the estimated intensities from the simulated data, obtained using the local EM method. It is readily apparent that all 4 intensities do not follow any pattern or trend, regardless of bandwidth size, and seem to fluctuate wildly as expected. Therefore this suggests that the local EM method is able to adequately estimate time varying intensities when they exist, yet, does not find any trends when the data are independent.

3.3 Remarks

In this chapter the local EM has been introduced to address time varying intensities and simulated data sets were utilized to illustrate the method. There are several aspects of this method that are
Figure 3.9: Constant intensity estimates for various bandwidths ($\delta = 1$): true $\theta$ (---); $\theta$ constant (- - -); BW 1 (---); BW2 (---); BW3 (---); BW4 (---); BW6 (---); BW8 (---);
Figure 3.10: Intensity estimates for various bandwidths when data is independent within subject and $\delta = 1$ (there is no underlying intensity): BW 1 (--); BW2 (—); BW3 (—); BW4 (—); BW6 (—); BW8 (—);
now worth mentioning, the first of which is computational time. The computational time involved in implementing the algorithm is related to the grid size, convergence criteria and initial values. Decreasing the grid size had the most affect on computational time. It increases the number of points to $J/\delta$, of which $100 \times (1 - \delta + F\delta)\%$ or $J \cdot N \left[ \frac{1-\delta}{\delta} + F \right]$ are missing observations, where $F$ is the fraction of missing data observed when $\delta = 1$ (in our simulation $F = 3/10$). In the context of our analysis, the time it took to run the algorithm with a 2 GHz Opteron processor for a given bandwidth was:

- 3 minutes when $\delta = 1$ and it took 4 iterations to converge (approximately 45 seconds per iteration).
- 45 minutes when $\delta = 1/2$ and it took 8 iterations to converge (approximately 6 minutes per iteration).
- 2 hours when $\delta = 1/3$ and it took 8 iterations to converge (approximately 15 minutes per iteration).

With regard to the grid size $\delta$ the data may often lend itself to natural selections for $\delta$. For example, if data is collected annually, a natural choice of delta may be $1/2$, $1/4$, or $1/12$ corresponding to a semi-annual, quarterly, or monthly grid. Although we ideally would like to construct a very fine grid, the trade-off with computational time must be assessed. In the above example with annual observations selecting $\delta = 1/4$, would lead to a grid consisting of 4 times as many points, of which a minimum of 75% of the time points have no observations. Finding the expected probabilities for these missing observations will immensely increase computational time. In general, as $\delta$ decreases computational time increases. As we will show in the following chapter, the advantages of decreasing $\delta$ are quite high initially but rapidly diminish. For example, the benefit of using $\delta = 1/2$ instead of $\delta = 1$ is proportionately much greater than electing to use $\delta = 1/3$ in place of $\delta = 1/2$. We also must mention that care must be taken when interpreting intensities using various grid sizes. Since the intensity is a rate, an intensity calculated using a $\delta \neq 1$ must be multiplied by $1/\delta$ in order for the intensities to be compared on the same scale.
Clearly, changing the convergence criteria will also affect computational time. We tested two different types of convergence criteria. The first criteria was set so

\[
\left| \frac{\sum_{j=1}^{G} \sum_{u,v} \hat{q}_{juv}^{(p)}}{Gc} - \frac{\sum_{j=1}^{G} \sum_{u,v} \hat{q}_{juv}^{(p+1)}}{Gc} \right| < K
\]

where \(c\) is the length of \(\theta(t)\), \(u, v\) are states, and \(K = 0.01\) in our example.

With \(J = 20, \delta = 0.5\) and \(c = 4\) it leads to convergence when successive \(\hat{q}_{juv}\)'s are, on average, within 0.01/160 of each other. The problem with this criteria is that it examines all the entries of \(\theta(t)\) at once. If even only one of the entries has not yet converged the algorithm may stop since, on average, the \(q_{ij}\)s will have met the convergence criteria. A stricter method of achieving convergence can be found by examining the intensity estimates at each time point. We examined:

\[
\max \left| \hat{\theta}^{(p)}(g) - \hat{\theta}^{(p+1)}(g) \right| < K \quad g = 1, \ldots, G
\]

The criteria can be adjusted depending on the size of \(\delta\) and thus the computational time it would take to reach convergence. In this simulation \(K\) was set to \(10^{-5}\) for \(\delta = 1\) and increased a little bit as \(\delta\) decreased. Ultimately, it is a trade-off between the computational time and the convergence criteria.

Lastly, selection of the initial values, \(\theta^{(0)}(t)\), can affect computational time. When analyzing the simulated data we used several initial values to ensure that all led to the same estimates for \(\theta\). Choosing initial values that were an order of magnitude away from the final values led to convergence problems similar to that discussed in Kalbfleisch and Lawless [23]. This problem can be avoided by obtaining reasonable starting values by assuming time homogeneous intensities and use the resulting estimates as initial values. Starting values further away from \(\hat{\theta}(t)\) needed more iterations and thereby more time lapsed until convergence. This was clearly seen in the simulation as we had the luxury of trying starting values that were the 'true' values. Using these values enabled us to achieve convergence remarkably faster than using starting values obtained from the time homogeneous intensity estimates.
Aside from computational time, when using the local EM there are several other aspects that are important to be aware of. As previously discussed, varying the grid size has both advantages and disadvantages. The advantages to decreasing $\delta$ are the ability to smooth out the intensity estimates from rough peaks or valleys as well as to provide a more accurate estimate for $\theta(t)$. The primary disadvantages to decreasing $\delta$ is the computational time as well as underestimating the extremes such as peaks and valleys (which may lead to a biased estimate). As long as the bandwidth is selected so that there are observations in the span, a smaller $\delta$ is advantageous as long as the computational time is reasonable. If $\delta > h$ then numerical integration will be done using a single time point which is far from ideal. As a rule of thumb we would like $\delta \leq h/2$, though not too small to cause extremely lengthy computational times. This rule will permit us to use a minimum of 4 quadrature points which will yield a better approximation.

Overall, applying the local EM to the simulated datasets suggests that the local EM method provides reasonable estimates of time varying intensities. It also indicates that it performs adequately when intensities are either constant or piecewise constant with respect to time. Aside from the methods mentioned here other kernels, in particular the uniform, were tried to verify our results. A uniform kernel with various bandwidths led to similar results, (though, as expected, not identical to the Epanechnikov kernel). Appendix A.4, presents another simulated dataset with time varying intensities, again to show the local EM methodology works in other scenarios as well. The algorithm can be modified for any type of kernel (i.e. Gaussian) and a variety of support sizes. This algorithm is user intensive. That is, the user must be aware of all the choices they make and ensure that they are reasonable, as we have discussed the various problems that may arise. The results obtained from the local EM method must be examined and interpreted to ensure that they are sensible. The analysis in this chapter has suggested that with some consideration the local EM approach estimates the underlying non-homogeneous intensities well. Aside from the two simulated datasets discussed in depth here, an additional 10 simulated datasets using the time varying intensities found in Figure 3.2 with $\delta = 1$ and $1/2$ were studied and, though not discussed here, similar results were found. For these additional simulated datasets the intensity estimates approximated the underlying time varying intensities quite well. However, we have not discussed
what “well” means or how it compares to the PWC method. In particular, we must show that
the results of the local EM are significantly better than the piecewise constant method. Furthermore, how do we determine which bandwidth size is optimal for the intensity? Which bandwidth provides the “best” estimate? Last of all, as with any statistical analysis, point estimates need be accompanied with standard errors or confidence interval estimates. In the following chapter, Bandwidth Selection and Variance Estimation, we will address these issues. In particular we will illustrate these methods using the two simulated datasets discussed in this chapter. The methods will also be applied to the 10 additional simulated datasets though the results will not be discussed in depth.
Chapter 4

Bandwidth Selection and Variance Estimation

4.1 Bandwidth Selection

The local EM approach described in the previous chapter enables time varying estimates for non-homogeneous intensities. The estimates, \( \hat{\theta}(t) \), will vary depending on the bandwidth size. We require a method for bandwidth selection that will enable us to find the optimal bandwidth. With multi-state Markov models conventional bandwidth selection approaches need to be modified to enable these methods to adapt to our situation.

4.1.1 Cross Validation and Local EM

Cross validation (CV) is a valuable tool for bandwidth selection. The usual approach to \( K \)-fold cross validation, involves estimating the prediction error by dividing the data into \( K \) sections, as discussed in Efron et al. [12]. For the \( k^{th} \) section, let \( \hat{y}_i^{k(i)} \) be the fitted value for observation \( i \) using the remaining \( K - 1 \) sections of the data. Then the CV sum of squared prediction error can be found as:

\[
\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i^{k(i)})^2
\]
Choosing $k = N$ leads to the standard method of “leave one out” cross-validation.

The squared loss function is useful when the “left out” data consists of continuously-valued observations and is proportional to the log likelihood for normally distributed data. MSMs involve discretely-defined states and a loss function should consider the predicted probabilities of occupying each state. As a result, we use the likelihood of observing the “left-out” data as the loss function. To do this, the data is divided into $D$ disjoint parts. With the $d^{th}$, $d = 1, \ldots D$ piece removed, the local EM algorithm is used to find $\hat{\theta}_h^{(-d)}$, the intensity estimate for a bandwidth of size $h$ with the $d^{th}$ part removed. Using these intensity estimates, $\hat{\theta}_h^{(-d)}$, we take the data from the $d^{th}$ section that was removed and calculate the log-likelihood using the form of equation (2.9):

$$L(Y|d,h) = \prod_{i=1}^{N_d} \prod_{j=1}^{J_i} \left[ P_{S_i(t_{ij-1}),S_i(t_{ij})}(t_{ij-1},t_{ij})| \theta_h^{(-d)} \right]$$

where $N_d$ is the number of subjects in the $d^{th}$ piece of data that was removed. Notice that only observed time points are used and imputed probabilities are not included when data is missing.

Performing this over all $D$ parts yields:

$$\ell(Y|d,h) = \frac{N_d}{N} \sum_{d=1}^{D} \sum_{i=1}^{N_d} \sum_{j=1}^{J_i} \log \left[ P_{S_i(t_{ij-1}),S_i(t_{ij})}(t_{ij-1},t_{ij})| \theta_h^{(-d)} \right]$$

where $\left[ P_{S_i(t_{ij-1}),S_i(t_{ij})}(t_{ij-1},t_{ij})| \theta_h^{(-d)} \right]$ is the probability of transitioning from state $S_i(t_{ij-1})$ at time $t_{ij-1}$ to $S_i(t_{ij})$ at time $t_{ij}$ for the $N_d$ participants using $\theta_h^{(-d)}$.

The above process is performed over various bandwidth sizes and the bandwidth corresponding to the maximum log likelihood is selected. That is,

$$\hat{h} = \arg \max_h \ell(Y|d,h)$$

One must realize that this optimal bandwidth will also depend on the grid size $\delta$ chosen. The plots in chapter 3, illustrated that a smaller $\delta$ seemed to show a better approximation for $\theta(t)$. A rule of thumb was created that $\delta \leq h/2$, $h$ the bandwidth, in order for the numerical integration procedure to be a good approximation. Care must be used to make sure when using the cross
validation method that both intensity estimates and probability estimates are on the appropriate scale corresponding to the grid being used.

### 4.1.2 Bandwidth Selection with Simulated Data

We applied our cross validation technique to the simulated data from sections 3.2.2 and 3.2.3. Recall, that observations were recorded yearly, over a 20 year time period, with 30% of the data missing. The analysis was done for 3 different grid sizes, $\delta = 1, 1/2$ and $1/3$. We used 5-fold cross validation so that the data 'left out' consisted of 60 observations. Intensities for various bandwidth sizes, $\hat{\theta}_h^{(-d)}(t)$, were computed using $4/5$ of the data which were then used to maximize the likelihood using the remaining $1/5$ of the data. For smaller bandwidths, only finer grids, ($\delta < 1$), were used as results from small bandwidths and course grids were unstable as would be expected from the discussion in section 3.3.

With simulated data, aside from using the CV as a method of bandwidth selection we have an additional advantage of knowing the true underlying intensities. Therefore, we can compare the bandwidth chosen with the integrated square error (ISE) criterion in equation (4.1) to the bandwidth chosen using cross validation. Since ISE compares to the true function, it is presumably superior, but in practice we will have only cross validation to choose bandwidth. The ISE is computed as

$$ ISE = \sum_{u \neq v} \int_0^T [\hat{q}_{uv}(s) - q_{uv}(s)]^2 ds \quad \text{for states } u, v $$ (4.1)

Furthermore, the ISE of the three grid sizes was compared to the ISE obtained using the piecewise constant method to estimate the intensities. Essentially, as discussed in section 2.3.1 piecewise constant estimates can be used to approximate time varying intensities. Since there appears to be no clearly established method for determining the number of intervals with the piecewise constant method we based it loosely on Gentleman et al. [15]. Four non-overlapping intervals were chosen as follows: $[0,5], (5,10], (10,15], (15-20]$. The msm package was used to estimate constant intensities in each of the four intervals and missing observations were dropped.
by the msm function. As can be seen in the following section, when using the optimal bandwidth, the local EM appears to perform considerably better than the PWC method for time varying intensities. In the case when the underlying true intensities are PWC then, depending on the grid selected, the local EM method performs almost as well or slightly better than the PWC method.

Simulation 1: Time varying Intensity

Both the ISE and cross validation methods were applied to the time varying simulated dataset. The details of the simulated data can be found in section 3.2.2. Table 4.1 and Figure 4.1 compare the local EM method to the PWC approach as well as the ISE method with the CV method for a variety of grid and bandwidth sizes. Comparing first the local EM with the PWC method, it appears that for moderately small bandwidth sizes the local EM performs better than the existing PWC approach. In particular, using the ISE, the optimal bandwidth of 3 (2.5 for $\delta = 1/3$) has a noticeably lower ISE across all grid sizes (largest ISE = 0.0161) as compared to ISE from the PWC estimate (0.0374). This is also true using the CV method where at the optimal bandwidth (between 2 and 3) the negative log likelihood has a lower value across all grid sizes (largest is approximately 718.5) as compared to the estimate from the PWC estimate (approximately 722). Furthermore, as expected, not all bandwidth sizes in the local EM method outperform the PWC method. Examining both very small and large bandwidths, indicates the PWC method outperforms the local EM approach in some cases. More specifically, using a bandwidth of size 21, which is equivalent to a flat intensity estimate, produces both the largest ISE and highest value of the negative log-likelihood for the CV method. This can be seen in Table 4.1 though not in Figure 4.1. As expected the PWC estimate is markedly better then the estimate obtained from the constant approach. Comparing the ISE and CV approaches, we can see that both show a moderately low bandwidth is optimal. While the ISE and CV both show the optimal bandwidth is of size 3 and 2.5 for a grid of size 1 and 1/3 respectively, there is a slight discrepancy with a grid of size 1/2 with the ISE indicating a bandwidth of 3 as being optimal and the CV indicating a bandwidth of 2. Despite the slight inconsistency between the two approaches this is not a concern since examining Table 4.1 further it appears that for the bandwidths in this range the likelihood is relatively flat. The CV
was performed and the MISE, (an extension of equation (4.1) to n samples), was calculated for the additional 10 simulated datasets mentioned at the end of section 3.3. Results were consistent with those shown in Table 4.1. For a $\delta$ of 1 and 1/2 a similar range of bandwidths was found to be optimal as to those found in this example. In addition, both the MISE and CV due to the PWC method were above the MISE and CV found using the optimal bandwidth.

### Simulation 2: Piecewise Constant Intensity

In addition to the simulated data found in section 3.2.2, we also wanted to examine the performance of the local EM in a setting that favoured PWC intensities. The simulated data in section 3.2.3 reflected these intensities. Using this data we would like to address two issues. One, we would like to determine if the local EM approach can compete with the PWC method in this scenario. Two, for the simulated data with PWC intensities, which bandwidth is optimal. Unlike simulated dataset 1 with time varying intensities, very small bandwidth sizes seemed to perform well when $\delta < 1$, possibly due to the non continuous nature of the intensity. Table 4.2 and Figure 4.2 below

<table>
<thead>
<tr>
<th>BW Size</th>
<th>ISE 1</th>
<th>ISE 1/2</th>
<th>ISE 1/3</th>
<th>CV 1</th>
<th>CV 1/2</th>
<th>CV 1/3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/3</td>
<td>0.0263</td>
<td></td>
<td></td>
<td>721.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0243</td>
<td>0.0166</td>
<td></td>
<td>721.72</td>
<td>720.97</td>
<td></td>
</tr>
<tr>
<td>4/3</td>
<td>0.0164</td>
<td>0.0165</td>
<td></td>
<td>718.87</td>
<td>719.50</td>
<td>719.12</td>
</tr>
<tr>
<td>5/3</td>
<td>0.0126</td>
<td>0.0094</td>
<td></td>
<td>718.84</td>
<td>718.55</td>
<td>718.73</td>
</tr>
<tr>
<td>2</td>
<td>0.0111</td>
<td>0.0084</td>
<td></td>
<td>718.83</td>
<td></td>
<td>718.54</td>
</tr>
<tr>
<td>2.5</td>
<td>0.0123</td>
<td></td>
<td><strong>0.0071</strong></td>
<td>718.60</td>
<td>718.82</td>
<td><strong>718.32</strong></td>
</tr>
<tr>
<td>3</td>
<td><strong>0.0161</strong></td>
<td><strong>0.0105</strong></td>
<td>0.0078</td>
<td><strong>718.30</strong></td>
<td>718.56</td>
<td>718.55</td>
</tr>
<tr>
<td>4</td>
<td>0.0140</td>
<td>0.0122</td>
<td></td>
<td>718.46</td>
<td>718.92</td>
<td>719.14</td>
</tr>
<tr>
<td>6</td>
<td>0.0283</td>
<td>0.0275</td>
<td></td>
<td>720.77</td>
<td>722.15</td>
<td>722.69</td>
</tr>
<tr>
<td>8</td>
<td>0.0495</td>
<td>0.0494</td>
<td></td>
<td>726.34</td>
<td>727.99</td>
<td>728.65</td>
</tr>
<tr>
<td>10</td>
<td>0.0720</td>
<td>0.0716</td>
<td></td>
<td>732.11</td>
<td>733.49</td>
<td>734.04</td>
</tr>
<tr>
<td>21</td>
<td>0.1305</td>
<td>0.1305</td>
<td>0.1305</td>
<td>740.05</td>
<td>740.05</td>
<td>740.05</td>
</tr>
<tr>
<td>PWC</td>
<td>0.0374</td>
<td>0.0374</td>
<td>0.0374</td>
<td>722.06</td>
<td>722.06</td>
<td>722.06</td>
</tr>
</tbody>
</table>

Table 4.1: Values of the ISE and CV methods for bandwidth selection from the time varying simulated dataset. Values in bold are the optimal bandwidths for the corresponding grid size. The ISE and CV values for the PWC method are included for comparison purposes.
show that in this simulation the local EM and PWC methods are quite competitive with each other. Using the ISE we can see that the PWC method has the advantage when $\delta = 1$ (0.058 vs. 0.0702 using a bandwidth of 3) while the local EM has the advantage when $\delta < 1$ (0.0525, 0.0403 vs. 0.058). Although at first glance it appears strange for the local EM method to outperform the PWC method with underlying PWC intensities, this may due to the nature of the PWC approach which deletes missing observations. Using the EM algorithm to fill in the missing observations provides us with more useful information that may allow us to make slightly better estimates of $\theta(t)$. Examining the CV method, it is clear that for all grid and bandwidth sizes the PWC method outperforms the local EM approach, though not by much (704.9 vs. 703.49). Although this is an inconsistency between the ISE and CV approaches there is not much concern as again, the likelihood for small bandwidths is relatively flat, with a range of 704.9-707.7. Furthermore, performing bandwidth selection using the two methods reveals a few more inconsistencies. For example, the ISE indicates that a bandwidth of 3 is appropriate for $\delta = 1$, a bandwidth of 2 when $\delta = 1/2$ and a bandwidth of $5/3$ when $\delta = 1/3$. However the CV indicates that lower bandwidths
Table 4.2: Values of the ISE and CV methods for bandwidth selection from the PWC simulated dataset. Values in bold are the optimal bandwidths for the corresponding grid size. The ISE and CV values for the PWC method are included for comparison purposes.

<table>
<thead>
<tr>
<th>BW Size</th>
<th>ISE</th>
<th>CV (- log likelihood)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\delta = 1)</td>
<td>(\frac{1}{2})</td>
</tr>
<tr>
<td>2/3</td>
<td>0.0919</td>
<td>705.98</td>
</tr>
<tr>
<td>1</td>
<td>0.0636</td>
<td>0.0696</td>
</tr>
<tr>
<td>4/3</td>
<td>0.0783</td>
<td>0.0561</td>
</tr>
<tr>
<td>5/3</td>
<td>0.0784</td>
<td><strong>0.0525</strong></td>
</tr>
<tr>
<td>2</td>
<td>0.0784</td>
<td>0.0595</td>
</tr>
<tr>
<td>3</td>
<td>0.0766</td>
<td>0.0789</td>
</tr>
<tr>
<td>6</td>
<td>0.1083</td>
<td>0.1160</td>
</tr>
<tr>
<td>8</td>
<td>0.1342</td>
<td>0.1303</td>
</tr>
<tr>
<td>10</td>
<td>0.1345</td>
<td>0.1287</td>
</tr>
<tr>
<td>PWC</td>
<td>0.0580</td>
<td>0.0580</td>
</tr>
</tbody>
</table>

Effect of Finer Grid or Larger Sample Size

Using simulated dataset 1 with time varying intensities from section 4.1.2 we further examined the effect of creating a finer grid as well as increasing the sample size. Using a grid of size \(\delta = 1/5\) the ISE was computed and compared to the ISE obtained with grid sizes of \(\delta = 1, 1/2, 1/3\). Table 4.3 compares these ISE for bandwidths of sizes 2, 3, and 4. Clearly as \(\delta\) is decreased the ISE decreases as well. This indicates that finer grids allow for estimates that are closer to the true values. Needless to say, there is a clear disadvantage to the finer grid, increased computation time.

Similarly, increasing the sample size from \(N = 300\) to \(N = 1000\) also led to a smaller ISE for the intensities. Table 4.4 examines the ISE for both sample sizes using \(\delta = 1\) and various bandwidth sizes. Not only is the ISE reduced for all the bandwidth sizes, but with the increased sample size
Figure 4.2: Comparison of the ISE and CV methods for the PWC simulated dataset. Though not as consistent as the time varying simulation it is fairly comparable, especially in this extreme scenario. Overall, the PWC method is slightly better or about the same as the local EM method. $\delta = 1(- - -); 1/2 (---); 1/3 (---);$ PWC (+ + +).

<table>
<thead>
<tr>
<th>BW Size</th>
<th>$\delta = 1$</th>
<th>$1/2$</th>
<th>$1/3$</th>
<th>$1/5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.0206</td>
<td>0.0111</td>
<td>0.0084</td>
<td>0.0063</td>
</tr>
<tr>
<td>3</td>
<td><strong>0.0161</strong></td>
<td><strong>0.0105</strong></td>
<td><strong>0.0078</strong></td>
<td><strong>0.0053</strong></td>
</tr>
<tr>
<td>4</td>
<td>0.0171</td>
<td>0.0140</td>
<td>0.0122</td>
<td>0.0099</td>
</tr>
</tbody>
</table>

Table 4.3: Comparison of the ISE using 4 different values for $\delta$, and 3 different bandwidths. It is clear that as $\delta$ is decreased the ISE decreases as well.
the optimal bandwidth has decreased from 3 to 2.

<table>
<thead>
<tr>
<th>BW Size</th>
<th>N=300</th>
<th>N=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0931</td>
<td>0.0189</td>
</tr>
<tr>
<td>2</td>
<td>0.0206</td>
<td><strong>0.0093</strong></td>
</tr>
<tr>
<td>3</td>
<td><strong>0.0161</strong></td>
<td>0.0111</td>
</tr>
<tr>
<td>4</td>
<td>0.0171</td>
<td>0.0137</td>
</tr>
<tr>
<td>6</td>
<td>0.0313</td>
<td>0.0305</td>
</tr>
<tr>
<td>8</td>
<td>0.0500</td>
<td>0.0484</td>
</tr>
<tr>
<td>10</td>
<td>0.0740</td>
<td>0.0706</td>
</tr>
</tbody>
</table>

Table 4.4: This table illustrates the effect of increasing the sample size on bandwidth selection using the ISE. As expected for the larger sample, the ISE is consistently lower than the ISE from the smaller sample size. Furthermore, as one may have anticipated, the optimal bandwidth has been lowered from 3 to 2.

### 4.2 Bootstrap and Confidence Intervals

With time varying intensities, standard errors for $\hat{\theta}$ are quite difficult to find using the standard approach that employs the normal approximation. Therefore, to obtain confidence intervals for the estimated intensities we employed a non-parametric bootstrap method. The bootstrap, discussed extensively by Efron et al. [12], is a computational method that can be used to produce a variety of statistical inferences. The basis of the idea is to use the empirical distribution of the sampled data as the true population distribution. In particular we are using the bootstrap here to compute confidence intervals for $\theta$. Let $x$ be the underlying sample data of size $n$ which can be viewed as the complete data. We then draw a sample, with replacement, of size $n$ from $x$ which is called a bootstrap sample. Drawing $B$ replicates from $x$ results in $B$ bootstrap samples. Although there are many useful applications to the bootstrap we will employ it to obtain confidence intervals for our intensity estimates. We will first examine the general form of the confidence interval and then discuss how to apply the bootstrap to our scenario. More detailed information can be found in Efron et al. [12]

Suppose the parameter $\phi$ has a normally distributed estimator $\hat{\phi}$ and estimated standard de-
viation \( \hat{\sigma} \). Then we have the standard result

\[
\frac{\hat{\phi}_n - \phi}{\hat{\sigma}_n} \to_d N(0, 1)
\]

as \( n \to \infty \).

Using this result, we can then construct an interval for our estimator \( \hat{\phi} \). Let \( Z \) represent the standard normal cumulative distribution function (cdf) and \( z_\alpha \) be the 100\( \alpha \) percentile of \( Z \). To construct the interval we have:

\[
P \left( z_\alpha \leq \frac{\hat{\phi} - \phi}{\hat{\sigma}} \leq z_{1-\alpha}\right) = 1 - 2\alpha
\]

Rewritten we find:

\[
P \left( \phi \in \left[ \hat{\phi} - z_{1-\alpha} \cdot \hat{\sigma}, \hat{\phi} + z_{1-\alpha} \cdot \hat{\sigma} \right] \right) = 1 - 2\alpha
\]

In more conventional terms the interval can be written as:

\[
(\hat{\phi} - z_{1-\alpha} \cdot \hat{\sigma}, \hat{\phi} + z_{1-\alpha} \cdot \hat{\sigma}) \text{ or } \hat{\phi} \pm z_{1-\alpha} \cdot \hat{\sigma}
\]

(4.2)

To construct a 95% confidence interval we would use \( \alpha = 0.025 \) which would yield \( z_{0.025} = -1.96 \) and \( z_{0.975} = 1.96 \). In general we will refer to the interval as \((\hat{\phi}_l, \hat{\phi}_u)\). It is important to note that (4.2) should be thought of as an approximate confidence interval since the standard error is unknown and thus the above standardization is not an exact standard normal. Increasing \( n \) will lead to more accurate estimates.

### 4.2.1 Percentile Bootstrap

Though equation (4.2) could have been employed to construct confidence intervals using the non-parametric bootstrap estimates, we elected to use the percentile method. The percentile method, while using the standard normal distribution, does not require standard errors to be computed.
From the previous section we know that \((\phi_l, \phi_u)\) are the 100\(\alpha\) and 100(1–\(\alpha\)) percentiles, respectively of \(\hat{\phi}^*\). We then create \(B\) bootstrap samples of \(\hat{\phi}^*\) to find the \(\phi_l\) and \(\phi_u\) that correspond to the \(\alpha\) and 1 – \(\alpha\) percentiles. For example, suppose we replicate \(B = 200\) samples and rank them in order \(\hat{\phi}_1^* \ldots \hat{\phi}_{200}^*\). Then with \(\alpha = 0.025\) a 95% percentile confidence interval would contain \(\phi_l = \hat{\phi}^{*(\alpha)} = \hat{\phi}_5^*\) and \(\phi_u = \hat{\phi}^{*(1-\alpha)} = \hat{\phi}_{195}^*\) resulting in the confidence interval \((\hat{\phi}_5^*, \hat{\phi}_{195}^*)\).

### 4.2.2 Confidence Intervals with Simulated Data

To construct a percentile confidence interval with the simulated data we performed a non-parametric bootstrap with \(B = 200\) replications. Each replication consisted of \(n = 300\) observations, where an observation is defined as a person’s set of states over the entire time interval of interest. We analyzed the bootstrap samples using the optimal bandwidth of 2 as it was found from the cross validation method as well as using a bandwidth of 3 as it was found from the ISE method. This was done for both \(\delta = 1/2\) and \(\delta = 1/3\). However, unlike the straightforward case of percentile intervals we are dealing with multiple parameters as oppose to a single one. There are a total of 4 different intensities, each with multiple time points (40 when \(\delta = 1/2\) and 60 when \(\delta = 1/3\)). Therefore we constructed pointwise confidence intervals for each of the intensities. Recall in the simulation that \(\hat{\theta} = (\hat{\theta}_{12}, \hat{\theta}_{21}, \hat{\theta}_{23}, \hat{\theta}_{32})\) which will now be denoted as \((\theta_1, \theta_2, \theta_3, \theta_4)\). Therefore, using the analogy from the previous section, at each time point we obtain \(\hat{\theta}_i^*\) \((i = 1 \ldots, 4)\) which is a realization of the random variable \(\hat{\theta}_i\). Ranking them in order and using \(\alpha = 0.025\) we can obtain the lower and upper bounds of the interval \((\theta_{il}, \theta_{iu}) = (\hat{\theta}_{15}^*, \hat{\theta}_{195}^*)\). Figures 4.3 and 4.4 examine the pointwise confidence intervals in more detail. It appears that the estimated values fall in between the lower and upper bounds of the 95% percentile interval. Examining the true intensities, the solid black lines, it seems that the 95% interval does not cover all of the true intensity all of the time. The percentile interval underestimates the intensity in a few places though, on a whole, it seems to give a realistic interval.

In addition to putting bounds of error on our estimated intensity, the bootstrap can also be used to test whether the estimated intensity is significantly different than a constant intensity. This can be done examining each intensity individually. When intensities are non-homogeneous with
Figure 4.3: 95% Bootstrap percentile intervals for the 4 intensity estimates for a bandwidth of size 2 and $\delta = 1/3$; true intensity (—); local EM estimate (—); upper bound (—); lower bound (—).
Figure 4.4: 95% Bootstrap percentile intervals for the 4 intensity estimates for a bandwidth of size 3 and \( \delta = 1/3 \); true intensity (---); local EM estimate (---); upper bound (—); lower bound (—).
respect to time the current standard is to assume either a constant or piecewise constant intensity. However the local EM permits time varying estimates of the underlying intensity. The bootstrap can be used to assess whether the estimate obtained from the local EM is significantly different than the intensity obtained using the constant method. For example, suppose we want to construct a hypothesis test about whether the intensity is truly time varying. Under the null hypothesis we assume the intensity is constant and the alternative hypothesis is that it is time varying. Using the msm package we can find constant intensity estimates, which will be denoted as $\hat{\theta}_C$. Similarly, the local EM algorithm can be used to acquire time varying intensity estimates, denoted $\hat{\theta}_{TV}$. Under the null hypothesis we simulate $B$ bootstrap samples with intensity vector $\hat{\theta}_C$. Next, using the local EM we find intensity estimates, denoted $\theta_{TV(b)}$, for each of the $B$ samples that were simulated. Using the sum of squares (SS) as a measure of distance, we can find the SS between $\theta_{TV(b)}$ and $\hat{\theta}_C$, denoted $SS_{boot(b)} b = 1, \ldots, B$, as well as the SS between $\theta_{TV}$ and $\hat{\theta}_C$ denoted $SS_{obs}$. If the null hypothesis is correct and the true intensity is really time homogeneous we would expect $SS_{obs}$ to have a similar value to many of the $SS_{boot(b)}$. Should $SS_{obs}$ be larger than $(1 - \alpha)100\%$ of the $SS_{boot(b)}$ we would reject the null hypothesis. We can construct a bootstrap p-value, $p_B$ which, based on the above, can be found as:

$$p_B = \frac{1}{B} \sum_{b=1}^{B} I(SS_{boot(b)} > SS_{obs})$$

This p-value can be treated in the standard way. A pre-specified significance level $\alpha$ can be used to decide whether the null hypothesis is rejected or not.

Aside from utilizing the bootstrap to determine whether an underlying intensity varies with respect to time, it can be used to address a variety of other questions. For instance in the following chapter we examine the underlying intensities for two groups of subjects, those that are HIV positive and those that are HIV negative. In this case, a similar approach, used the permutation test, to ascertain whether the intensities differed between the two groups.
4.3 Remarks

In this chapter we have suggested model assessment techniques that can be used when implementing the local EM algorithm. The cross validation method for bandwidth selection enables one to locate the optimal bandwidth to use in the algorithm. Furthermore, simulation methods indicated that the CV was reasonably consistent with the ISE and can therefore be relied upon. As mentioned earlier, the additional 10 simulations not shown here provided results similar to those obtained from the two simulated datasets above. Although we used K-fold cross validation, with small samples the standard leave one out method of cross validation can be used. In addition to bandwidth selection, the bootstrap was introduced to construct percentile intervals. These intervals, though pointwise, can be used as a measure of error for our estimated intensities. Furthermore, we have shown how the bootstrap can be used to test whether the intensity estimates acquired from the local EM are significantly different than the standard time homogeneous estimates. This test can, similarly, be extended to PWC intensities. The following chapter will present an example that will be used as a platform to illustrate the local EM and the various model assessment tools.

Despite the advantages of using these model assessment tools there are several drawbacks. As discussed previously computational time is a concern. The CV method simply repeats the local EM for all $K$ parts of the data. Though the sample size is slightly smaller, the repetition considerably increases computational time. Using a small $\delta$ only exacerbates the problem. Likewise, computational time was a major factor when using the bootstrap. One replication took, on average, approximately 45 minutes to converge when using a $\delta$ of 1/2. Decreasing $\delta$ to 1/3 meant that one replication took, on average, approximately 1 3/4 hours to converge. The replications were divided and the calculations were performed using multiple processors. Nevertheless, a method with greater efficiency would be welcome. Another disadvantage to the bootstrap method is the pointwise nature of the percentile intervals. Although we tried to construct simultaneous confidence intervals, we encountered some trouble. The trouble arose when trying to rank these functions and selecting the appropriate ones to estimate the upper and lower boundaries of the percentile interval. Some of these intensity functions were smooth while others were quite rough. This method required us
to choose the more extreme intervals which were quite rough by nature. Future work that would enable the construct of simultaneous confidence intervals for non-homogeneous intensities would be greatly beneficial.
Chapter 5

Example: Polaris Study

5.1 Background

Data from The Polaris HIV Seroconversion Study, (section 2.2.2), were used to demonstrate the local EM and its ability to estimate non-homogeneous intensities. A subset of 378 of the 459 participants was selected to form a homogeneous sample. The sample excluded female and heterosexual males, as well as, HIV negative subjects who seroconverted while on the study. The remaining participants were all men who have sex with men and were between the ages of 18 and 82 at induction with the mean age being 36.2 years. The average number of follow-up observations was 7.7 with a minimum of 1 and maximum of 17. The average length of time on study was 4.1 years with a minimum of 12 days and a maximum of 9.6 years. There was no statistically significant difference between the HIV positive and HIV negative groups with regard to demographic variables or length of time on study. The starting point for the analysis was chosen as the first HIV positive test for positive participants and the most recent HIV negative test for HIV negative participants. Many other demographic and social factors were collected including a self completed survey (SCS), filled out at home by the participants, that measured psychological aspects.

The SCS examined measures of self esteem, mastery, depression, anxiety, patient well-being (MOS-SF-36), and also explored stressful life events. Depression is measured using the CESD scale, discussed by Radloff [40], which consists of 20 questions, each with an ordinal response
ranging between 0 and 3. After combining the questions one value is obtained for each individual. As discussed in Radloff [40], Riggs et al. [42], a score of 16 or greater is indicative of probable depression and a score of 23 or greater indicates severe depression. Using these reference points, three ordinal states were created for depression as: score less than 16 (no depression), score between 16 and 22 (mild depression), and score of 23 or greater (severe depression). The ordinal nature of the states is apparent due to the fact that participants must transition through the state ‘mild depression’, even if only instantaneously, when going down from the no depression state to severe depression state or when doing the reverse. It is the intensities of the changes between depression states that we were interested in estimating.

5.2 Methodology

Participants’ depression scores were gathered at the initial interview by the SCS and every six months thereafter. The scores were obtained from the CESD scale and divided into three categories: no depression (state 1), mild depression (state 2), and severe depression (state 3). The intensity vector \( \theta \) was set up as \( \theta(t) = (q_{12}, q_{21}, q_{23}, q_{32} | 1/6 \leq t \leq 5) \), where \( t \) is measured in years and more details on the time interval can be found in upcoming section 5.2.1. Thus, the intensity matrix \( Q \) has the form:

\[
Q(t) = \begin{bmatrix}
-q_{12}(t) & q_{12}(t) & 0 \\
q_{21}(t) & -q_{21}(t) - q_{23}(t) & q_{23}(t) \\
0 & q_{32}(t) & -q_{32}(t)
\end{bmatrix}
\]

5.2.1 Local EM Method

The data were unequally balanced due primarily to two factors: (1) the time interval from the start point to the first interview varied between participants and (2) participants did not always adhere exactly to the 6 month follow-up schedule. In addition, occasionally participants may have missed an interview resulting in a 12 month, or larger, period between follow-ups. This was true for both positive and negative subjects alike. The primary question of interest was whether the rates of
transition between the three depression categories differed amongst the HIV positive and negative subjects. The underlying hypothesis is that the intensities for both groups are not different than each other. Prior to examining this hypothesis, it was first necessary to estimate the underlying intensities. As described in section 3.1.2, participants without initial states can have their states estimated using the distribution found at the initial time point. At times 1 and 2 months there were only a total of 15 observations and therefore not enough information to assign suitable probabilities in place of initial states for the remaining subjects. Instead, we elected to start the analysis at an initial time of 3 months. Among the 15 observations that were available at times 1 and 2, there were 7 subjects who had no further observations and were subsequently removed from the analysis resulting in 371 remaining participants, 133 positive and 238 negative. Additionally, only the first 5 years on study were used since follow-ups after that time were rather sparse. Numerous bandwidth sizes, starting with size 3 months and increasing in 3 month increments, were tried using the local EM. The local EM method was employed to do this using a 1 month grid (i.e. $\delta = 1/6$). This allowed for flexibility with regard to the unequally spaced data and conformed to the rule of thumb that $\delta \leq h/2$. The cross validation method was used to locate the optimal bandwidth for both the HIV positive and HIV negative groups separately. The results will be discussed in section 5.3.1.

5.2.2 Permutation Test

To examine the hypothesis whether HIV positive and negative participants have different transition intensities a permutation test was employed. Under the null hypothesis it was assumed that all the intensities were the same for both groups, $q_{uv+}(t) = q_{uv-}(t)$ for states $u, v$ (where $q_{uv+}$ and $q_{uv-}$ are the intensities for the positive and negative groups respectively). The alternative hypothesis claimed that at least one intensity was different between the HIV positive and HIV negative participants or $q_{uv+}(t) \neq q_{uv-}(t)$ for at least one $u, v$. Both groups were combined and using the same proportions of positive and negative subjects as the original data, participants were randomly assigned positive or negative status, thereby randomly creating 2 new groups. A total of $B = 100$ replicates were obtained. Although 100 replicates is small with regards to hypothesis testing this example is more to illustrate the idea rather than to provide a definitive analysis of the data. For
each replicate, the local EM was applied independently to the newly formed groups. Under the null hypothesis that there is no difference in intensities between the two groups, one would expect that the sum of squares between the HIV positive and negative groups obtained from the 100 permutation samples to be similar to those obtained from the original data. The time interval of interest was divided into two disjoint intervals, up to one year and one year or greater, and the hypothesis of interest was tested separately for each of the intervals. It was thought that HIV positive individuals may exhibit certain behaviours shortly after discovering their positive status and over a longer period of time their behaviour would acclimate to the new reality. Although there is no scientific evidence to choose precisely the one year mark as the cutoff, it was chosen for convenience.

5.3 Results

There were two questions of interest: (1) Are the intensities time homogeneous? and (2) Is there a difference in the intensities between HIV positive and HIV negative participants? The results of the local EM algorithm, together with the cross validation method for bandwidth selection, will be presented first as it will enable the non-homogeneous intensities to be estimated using the optimal bandwidth. Following this, the permutation test will be used to address the second issue, whether the intensities are different between the two groups.

5.3.1 Local EM

Let \( \theta(t)_+ \) and \( \theta(t)_- \) be the time varying intensity vectors for the HIV positive and negative participants, respectively. Figures 5.1 and 5.2 examine the intensities for both groups with various bandwidths. As expected, intensities estimated by using larger bandwidths appear to be flatter than intensities estimated by using smaller bandwidths. Examining the plots further, we can see some common trends. For the HIV positive group it appears that \( q_{21}, q_{23} \) and \( q_{32} \) seem to have decreasing intensities, especially in the first few months of follow-up, while \( q_{12} \) appears to be flatter. For the HIV negative group, the intensities appear to be relatively flat with the exception of a slight
increase in $q_{21}$ and a moderate decrease in $q_{32}$. In addition, HIV positive participants appear to have larger intensities for the first 12 to 15 months of follow-up as compared to their negative counterparts. The CV method will be applied to determine the optimal bandwidth as this will enable a more informed decision to be made about the intensity estimates.

Using the 1 month grid and the CV method, we started with a bandwidth of 3 months and proceeded in 3 month increments until 21 months to find the optimal bandwidth. Table 5.1 examines the cross validation for both the HIV positive and negative participants. As is readily apparent, the log likelihood is relatively flat and aside from the bandwidth of 3 months, there

Figure 5.1: Time Varying intensity estimates for HIV positive subjects: bandwidths of 3(--), 6 (--), 9(---) and 12(---) months.
Figure 5.2: Time Varying intensities for HIV negative subjects: bandwidths of 3(—), 6 (—), 9(—) and 12(—) months.
is little variation in the log-likelihood values. This seems consistent with Figures 5.1 and 5.2 which show that these bandwidths produce relatively similar estimates for $\theta(t)_-$ and $\theta(t)_+$. The table indicates that a bandwidth of 9 months is optimal for HIV positive participants while a slightly larger bandwidth of 12 months is optimal for HIV negative participants. Due to the flat cross validation scores, numerical optimization could have been used to find a slightly better, non-integer valued, bandwidth. However, due to the flatness, the numerical optimization is unlikely to lead to considerable improvement. Figure 5.3 compares these intensities for both positive and negative participants.

It looks as though HIV positive (black line) participants generally have decreasing intensities while HIV negative (red line) subjects have relatively flat ones. In particular, the intensities are decreasing rapidly for transitions from state 2 to 3 and 3 to 2 during the first 3 or so years of follow-up for HIV positive subjects. Furthermore, for these subjects the intensity from state 2 to 1 decreases noticeably during the first 1.25 years while there appears to be a decreasing trend in the intensity going from state 1 to 2. HIV negative participants tend to have relatively flat intensities with those transitioning from state 2 to 1 having a slight increase over time and those going from state 3 to 2 having a slight decrease over the follow-up period. In general, aside from $q_{21}$, it looks as though the intensities of HIV subjects start off higher than the negative subjects while slowly decreasing and becoming roughly equivalent for the last year to two of the follow-up period. The $q_{21}$ intensity also begins with a larger value for positive subjects and decreases rapidly so that after approximately one year into the follow-up period the intensities are quite similar. After year 3 or

<table>
<thead>
<tr>
<th>BW Size (months)</th>
<th>HIV positive</th>
<th>HIV negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>137.91</td>
<td>198.21</td>
</tr>
<tr>
<td>6</td>
<td>135.89</td>
<td>196.34</td>
</tr>
<tr>
<td>9</td>
<td><strong>135.32</strong></td>
<td>195.84</td>
</tr>
<tr>
<td>12</td>
<td>135.37</td>
<td><strong>195.77</strong></td>
</tr>
<tr>
<td>15</td>
<td>135.41</td>
<td>195.77</td>
</tr>
<tr>
<td>18</td>
<td>135.43</td>
<td>195.81</td>
</tr>
<tr>
<td>21</td>
<td>135.50</td>
<td>195.96</td>
</tr>
</tbody>
</table>

Table 5.1: Cross validation for both HIV positive and negative participants using the negative log-likelihood.
Figure 5.3: Intensity estimates for both HIV positive (—) and HIV negative (—) subjects corresponding to the optimal bandwidth as found through cross validation. HIV positive subjects tend to have a decreasing trend while HIV negative subjects appear to have relatively constant intensity estimates. The time homogeneous intensities are included for comparison.
so the intensities start to diverge with the HIV negative participants having a larger intensity than the positive participants. Figure 5.3 also contains the time homogeneous intensities as estimated using the msm package. For all intensities, except $q_{21}$, the estimated intensities of the HIV positive participants are larger than the estimated intensities of the HIV negative participants as would be expected. For $q_{21}$ the time homogeneous intensities are virtually the same. In this instance, although the HIV positive group starts off with a larger intensity than the negative group, the trend for the positive subjects is downward as oppose to the negative subjects where it is an upward trend. Hence, if we use a time homogeneous estimate for the intensity the two groups will have virtually identical estimates.

Based on these results, it appears that the intensities vary with respect to time, especially for HIV positive participants. As discussed in the previous chapter, this can be confirmed using the nonparametric bootstrap method, however we will not do so here. Rather, the permutation test will be used to address the second and slightly more complex issue, whether the intensity estimates are statistically significant between the HIV positive and HIV negative groups.

5.3.2 Permutation Test

Chapter 4 introduced the general approach for implementing the bootstrap to construct intervals and find p-values. In this example we will use similar methods with the permutation test to determine whether the intensities between the HIV positive and negative subjects are significantly different from each other. Under the null hypothesis we assume that there is no difference between the two groups for all 4 intensities. The alternative hypothesis is that at least one of the 4 intensities is significantly different between the 2 groups. Again, the data was divided into two distinct time intervals: months 3 - 12 and months 13 - 60 and the hypothesis was tested separately for each interval. Using the null hypothesis we pooled all the observations together and randomly assigned a positive and negative HIV status to the subjects in proportion to the original data with a ratio of 133 positives to 238 negatives. To perform the random assignment a uniform random variable was assigned to each individual and the data was sorted in ascending order using the uniform random variable. The first 133 participants were assigned positive status and the remaining were assigned
Figure 5.4: Intensity estimates for HIV positive subjects (—) using a bandwidth of 9 months and for HIV negative subjects (—) using a bandwidth of 9 months. Comparing the bandwidth of 9 months versus 12 months for HIV negative subjects finds little difference. With the exception of $q_{32}$ that has a markedly decreasing trend from times 0.25-1.5 year and again from 3.5-5 years, the other intensity estimates are very similar to those obtained using a bandwidth of 12 months.
negative status. A total of $B = 100$ replications were run. For both HIV groups and all 100 replications, a bandwidth of size 9 months was selected. Although the CV method indicated that the optimal bandwidth for HIV negative subjects was 12 months, nevertheless as shown in table 5.1 the likelihood was generally flat (195.765 vs. 195.835 for bandwidths 9 and 12 months). Therefore, a bandwidth of 9 months can reasonably be used while having an inconsequential impact on the results as seen in Figure 5.4 where the intensities for both groups are displayed using a bandwidth of 9 months for both groups. Comparing this to Figure 5.3 that used a bandwidth of 12 months for HIV negative subjects it is clear that there are only minor differences. More specifically, the two are very similar with the exception of $q_{32}$ that has a steeper decrease at the beginning of the study.

The hypothesis test was evaluated using the two methods described in the previous chapter, the percentile interval and the p-value method. For the p-value method, the sum of squares was used to measure the distance between the intensity estimates of the two groups. It was computed as

$$SS_{uv(b)} = \sum_{t=1}^{G} (q_{uv+}(t) - q_{uv-}(t))^2 \quad b = 1, \ldots, 100 \quad u, v = 1 \ldots K \ (u \neq v)$$

for each of the replicates and similarly for the observed data, denoted $SS_{obs}$. Under the null hypothesis $SS_{obs}$ should have a similar value as $SS_{uv(b)}$ for each intensity. Under the alternative we would expect $SS_{obs}$ to have a significantly larger value than $SS_{uv(b)}$ for at least one intensity. Examining the first time interval, months 3 - 12, Figure 5.5 displays 4 histograms, one for each intensity, comparing the $SS_{uv(b)}$ for the 100 replicates to $SS_{obs}$. For $q_{12}$ and $q_{23}$ it appears that $SS_{obs}$ is in the tail of the distribution. For the other two intensities, $SS_{obs}$, does not appear to be significantly different than the $SS_{uv(b)}$ values. In terms of the p-value for $q_{23}$ it was indeed significant with a value of 0.03 while for $q_{12}$ it was not quite significant with a value of 0.12. The other two intensities had p-values of 0.14 and 0.45 for $q_{21}$ and $q_{32}$ respectively. In this setting, with relatively few replicates ($B = 100$) it would be useful to calculate the standard error of $p_B$, where $p_B$ was defined in (4.3). This can simply be done using the well-known sample proportion formula of $\sqrt{\hat{p}_B(1-\hat{p}_B)/B}$ which in the case of $p_B = 0.03$ results in $\sqrt{(0.03)(0.97)/100} = 0.017$ which using 1.96 standard deviations results in 0.03 +/- 0.033 and contains the value 0.05. Ideally,
$B$ should be increased to get a better estimate of $p_B$ to determine if it is significant.

![Histograms](image-url)

(a) $q_{12}$  
(b) $q_{21}$  
(c) $q_{23}$  
(d) $q_{32}$

Figure 5.5: This histogram illustrates the relative location of $SS_{obs}$ (vertical line) to $SS_{uv(b)}$. The p-values can be obtained from the histograms and only $q_{23}$ was significant with a p-value of 0.03.

To better visualize these results a second approach was used and we plotted the 90% percentile intervals for both the HIV positive and HIV negative subjects for each of the 4 intensities. Although it is difficult to draw inferences based on these plots, they are quite useful as a graphical tool and complement the p-value method. The percentile intervals are constructed from the pooled data and therefore under the null hypothesis it is expected that the intensities for the positive subjects would fall within the positive confidence limits and similarly the intensities of the negative subjects
Figure 5.6: 90% percentile plots for HIV positive (---) and HIV negative (----) individuals examined followed from months 3 to 12. They are to be used in conjunction with the p-value method to determine whether the intensities between the HIV positive (---) and HIV negative (----) subjects differ. Consistent with the p-value method, only $q_{23}$ is significant as it falls outside the interval.
would fall within their confidence limits. Examining Figure 5.6 it is clear the \( q_{23} \) falls outside the percentile intervals for both the HIV positive and HIV negative data resulting in the rejection of the null hypothesis at the \( p = 0.10 \) level. Intensity \( q_{12} \) appears to be inside the interval for HIV positive participants and along the lower bound for HIV negative subjects. Intensity \( q_{21} \) is initially slightly outside the interval for both positive and negative subjects for months 3 and 4 and for the remaining time points it is inside the interval, while \( q_{32} \) remains well inside the interval for the entire time period for both HIV positive and negative individuals. Thus, as they should, the p-value and percentile interval methods produce similar results.

Turning our attention to the other time interval, years 1 - 5, it looks as though the intensities between the two groups do not differ as much. Examining the histograms in Figure 5.7, only \( q_{12} \) may have significantly different HIV positive and negative intensities as \( SS_{obs} \) is in the tail of the distribution. For the other three intensities \( SS_{obs} \) is in middle of the distribution. The p-values for the 4 intensities were 0.10, 0.48, 0.38, and 0.58 for \( q_{12}, q_{21}, q_{23} \) and \( q_{32} \) respectively. Thus only, \( q_{12} \) is borderline significant.

Referring to the percentile intervals in Figure 5.8 these results are confirmed. Intensity \( q_{21} \) remains well inside the intervals for both groups over the entire time interval. Intensities \( q_{23} \) and \( q_{32} \) are generally well inside their intervals except at the beginning of the time period for \( q_{23} \) where it is slightly outside the interval and between times 21 and 23 months for \( q_{32} \) where it is borderline for the HIV negative group. Focusing on \( q_{12} \) the plot shows the intensity for both groups is outside the interval from about time 25 - 39. However, not too much credence should be given to this as the beginning of the period has the intensities a little bit inside the intervals and it is hard to judge whether this is phenomenon is correct or whether the intensity is really relatively flat in the area. Furthermore, using a small number of replicates (\( B = 100 \)), means that the percentiles may not be estimated well enough to draw concrete conclusions. Nevertheless, it is something which should probably be investigated further, especially since the p-value for \( q_{12} \) was 0.10. It is clear in this example that dividing the data into numerous intervals will produce different outcomes for our hypothesis test. However, due to multiple testing issues we have elected not to do this. In the following section we will discuss this in a little more detail along with some other issues that arose...
Figure 5.7: Histogram illustrating the p-value method for months 13-60. The vertical line, indicates the value of $SS_{obs}$ for each intensity. Aside from $q_{12}$ with a p-value of 0.10 the other intensities were not significantly different from each other.
Figure 5.8: Percentile intervals between times 13 and 60 months for HIV positive (---) and HIV negative (---) subjects confirm the results of the p-value method. Though some intensities for both the positive (—) and negative (—) groups illustrate brief periods of times where they are outside the intervals, examining the interval as a whole provides no evidence that the intensities are significantly different between the two groups.
5.4 Remarks

In section 5.3.2 the data was split into two time periods for the construction of the percentile intervals. It was also shown that between times 25 and 39 $q_{12}$ may have had significantly different intensities for HIV positive and HIV negative participants. However, we elected not to further divide the data into more intervals due to the multiple testing problem. There are already 4 intensities being tested and if we were to examine each year separately that would lead to a total of 20 tests being performed. In addition, we must realize this is for an intensity matrix with 3 ordinal states. Increasing the number of states would considerably increase the number of tests required. Figures 5.9 and 5.10 display the histograms and percentile intervals when the interval is not divided at the one year mark. Combining the two time intervals and looking at months 3 to 60 at once, the p-value method and percentile intervals lead to some interesting results. Examining the histograms it seems that none of the intensities are significantly different between HIV positive and negative subjects, though $q_{12}$ and $q_{23}$ may be close. The p-values confirm this as they are 0.11, 0.36, 0.10, and 0.64 for $q_{12}, q_{21}, q_{23}$ and $q_{32}$ respectively. However examining the percentile intervals, there are clear periods of time for which the intensities of HIV positive and negative individuals differ significantly. This warrants further examination. One idea may be to divide up the data into smaller intervals and use methods that would account for the multiple testing. We will not pursue this any further as the data was selected as an example to illustrate the local EM method rather than to give a comprehensive analysis.

This example illustrates the advantages of estimating non-homogeneous intensities. For a clinician, the ability to obtain both marginal, or population level, and conditional, or subject specific, probabilities at any given time is beneficial. Figure 5.11 examines the marginal probabilities for both a HIV positive and HIV negative randomly selected individual with no observed states. This plot can be interpreted as the average probability of occupying a particular state at a given time when no information is available. It appears that HIV negative participants have relatively constant
Figure 5.9: From the histogram, $SS_{obs}$, (vertical line), corresponding to $q_{12}$ and possibly $q_{23}$ can be seen to have values that are in the tail of the distribution. Their p-values are borderline at about 0.11 and 0.10 respectively. Construction of percentile intervals will help draw a conclusion.
Figure 5.10: The percentile intervals for months 3 - 60 for both positive (---) and negative (---) participants provides for some interesting results. For both positive (---) and negative (---) subjects only \( q_{12} \) and \( q_{23} \) appear to suggest times where the intensity clearly falls outside the intervals. Although this may be reason to reject the null hypothesis, one must be sure to account for the entire time interval.
Figure 5.11: Marginal probabilities associated with each of the HIV groups. On the left, the highest probability is associated with state 1 (both initially and ongoing) (—). There is a moderate probability associated with starting off in state 2 (—) though it slightly decreases over time. The probability of occupying state 3 (—) at any given time is low. On the right side, the probabilities associated with the HIV positive group appear to be different. There is an increasing chance of occupying state 1, a decreasing probability associated with state 3 and a constant probability of occupying state 2.

The conditional probabilities are of particular interest when trying to determine a specific individual’s transition pathway. Figure 5.12, examines a randomly chosen HIV negative and HIV positive subject. The first and most obvious difference between the two plots is the consistency of
Figure 5.12: Conditional probabilities associated with an individual from each of the HIV groups. The “x” along the horizontal axis represents the actual observed times. On the left, the HIV negative subject is relatively consistent occupying state 1 (—) the majority of the time though he started out in state 2 (—). There is very little chance of occupying state 3 (—) except possibly toward the end of the interval. The HIV positive subject on the other hand fluctuates randomly between the various states. In both cases the conditional probabilities are available for a given time point.

the HIV negative subject versus the randomness of the positive one. However, these subjects were chosen at random and not much attention should be paid to this. What is of interest, though, is the conditional probabilities at unobserved times. When a subject has two consecutive observations that are in the same state, the probability of transitioning to other states in between these two time points is slim. However, the further the two observed states are from each other the larger the chance of a transition occurring into another state. In general these time varying transition probabilities are not linear with respect to time as we would expect from non-homogeneous intensities. Conditional probabilities are especially useful when trying to make decisions regarding chronic disease. Intervening earlier with a more powerful, yet less benign, drug may be of interest if a clinician can anticipate a detrimental outcome.

Another issue that arose during this analysis, similar to the problems in the simulation analysis, was the computation time it took to perform the analysis, specifically the permutation test. Each replication took approximately 18 hours to run, or a little more than 1800 hours for 100 replicates.
Figure 5.13: Examining one individual’s observed time points in relation to the grid. The E-Step can use the observed times to calculate the probabilities while the M-Step uses the rigid 3 month grid to obtain the MLE estimates.

Using the monthly grid versus a 3 month or even 6 month grid markedly increased the computation time. Some options may have been to reduce the convergence criterion, which was set the same as the criterion for the simulation, or possibly use fewer replications, both of which were not ideal. Another alternative may have been to use a 3 or 6 month grid an idea we will now briefly outline.

In our example data was collected every 6 months, though as explained previously, it was unbalanced. One option would be to set up a 3 month grid and round people’s observations to the closest 3 months. An example of one subject’s observations can be seen in Figure 5.13. The person’s true observations are represented by the dots but they are rounded to the closest 3 months, represented by the Xs. One can then use the local EM with the 3 month grid. Clearly, this is not as good as using a monthly grid but we can improve this approach a little bit leading to a better analysis. In the E-step, parts of intervals can be used. Referring back to figure 5.13, the transition matrices in the expectation step would be:

\[
P_{ij}(t_0, t_1) = e^{2/3Q_1} \quad P_{ij}(t_1, t_2) = e^{1/3Q_1}e^{2/3Q_2} \quad P_{ij}(t_2, t_3) = e^{1/3Q_2}e^{1/3Q_3}e^{1/3Q_4}
\]

For the M-step we would use the intensities as they are set on the 3 month grid, in this case \(e^{Q_1}, e^{Q_2}, e^{Q_3}, e^{Q_4}\). Although this method will not yield the same results as using the 1 month grid, it
is nonetheless better than rigidly confining the data to a 3 month grid.

This example was presented to illustrate the local EM method and its advantages over the PWC method. The analysis shows that at least one of the intensities is different for HIV positive and HIV negative subjects over particular intervals. Further analysis is warranted and must account for the multiple testing issue.
Chapter 6

Discussion

The local EM algorithm has provided us with a unique tool to estimate non-homogeneous intensities. Comparing it to other existing methods, such as the constant and piecewise constant estimates, several advantages are clear. Unlike the existing methods that assume a time homogeneous intensity over time segments, the new approach makes no assumptions about the functional form of the intensity. Furthermore, as shown from the simulated datasets, when examining the ISE, the local EM has appears to have an advantage over the other two methods when the intensities are truly time varying. Even in the event that the true intensities are piecewise constant, we have seen the method perform more than adequately. In addition, the local EM approach allows for missing data, unlike the other two methods, which, without specialized imputation methods, would require that missing observations are simply ignored. Moreover, model assessment tools such as cross validation for bandwidth selection and percentile intervals were developed to assess the model. Lastly, an application of the local EM method was demonstrated using the depression scores of HIV subjects. The method had the advantage of allowing the intensities to change over time, reflecting a more realistic estimate of the intensities as compared to the piecewise constant method.
6.1 General Remarks

Although the local EM method provides an approach to estimating non-homogeneous intensities it is nonetheless both computationally and user intensive. The user must choose starting values, grid size, a kernel type, a bandwidth size, as well as a convergence criterion. Varying these selections will allow for better precision in the estimation of the underlying intensities, and in parallel, will also determine the computational time required for the algorithm. In our simulations numerous starting values for \( \hat{\theta}^{(0)}(t) \) were tried and all reasonable ones converged to the same \( \hat{\theta}(t) \). Using starting values that were an order of magnitude away from \( \hat{\theta} \) caused numerical instability and led to unrealistic estimates. The method discussed in Kalbfleisch and Lawless [23] did not result in any problems and was used to generate starting values when analyzing the simulated data in this thesis. Needless to say, initial values that were chosen closer to \( \hat{\theta}(t) \) converged more rapidly. Developing a way of finding initial values that are closer to \( \hat{\theta}(t) \) may be of interest to save computation time.

As was demonstrated using the simulated data, finer grids led to more accurate estimates for \( \theta \) as illustrated from the ISE. However, there is an obvious trade-off with computational time as it increases significantly with finer grid sizes. Although selecting a \( \delta < 1 \) may increase the precision of the estimate, selecting a \( \delta \) very small does significantly increase the precision of the estimate yet may not be worth the lengthy computational time involved.

The bandwidth plays a crucial role in the estimation of the time varying intensities. Both the simulation and example illustrated the various intensity estimates based on various bandwidth sizes. Though neither the choice of kernel nor bandwidth have a significant impact on computational time, performing bandwidth selection as discussed in chapter 4 is computationally intensive due to K-fold cross validation. One option may be to set \( K = 1 \) initially to get an idea of the optimal bandwidth and then perform K-fold cross validation on a range of bandwidth sizes close to the “optimal” bandwidth. In comparison to the bandwidth, the choice of kernel had an insignificant impact on the intensity estimates.

Another choice the user must make when employing the algorithm is the selection of the convergence criteria as it will also have an impact on the intensity estimate and computational
intensity. A smaller criterion will lead to slightly more accurate estimates but in our simulation we found that the increase in computational time was not warranted as the specified criterion led to 5 decimal places of accuracy. As discussed in section 3.3 convergence was set so that at each grid point the intensity estimates were within a prespecified criterion, $C$, of consecutive iterations. However, another, yet computationally slower, approach would have been to set a convergence criterion on the likelihood function. At each grid point the likelihood from the $(p + 1)^{th}$ iteration must be within the prespecified criterion, $K$, of the $p^{th}$ iteration for convergence to be achieved. This method takes a significantly longer time to achieve convergence, especially with a smaller $\delta$, as the likelihood is relatively flat. Therefore, we elected not use this approach.

One aspect that was not discussed in detail was the number of states. In both our simulation and example, 3 ordinal states were used. Extending to $K$ states is simple and straightforward, although it would presumably increase computation time. We did not consider absorbing states in any of the scenarios examined in this thesis. Absorbing states in the context of MSM models are discussed in Gentleman et al. [15], Kalbfleisch and Lawless [23, 24] and the local EM method can be extended to this scenario.

Another idea not discussed here though worthwhile mentioning, would be to use the local EM as a model selection tool. For example, one could plot an overlay of the non-parametric estimate obtained using the local EM with a parametric estimate of a given intensity to see how well the two curves agree or even to see is the parametric estimate of the intensity falls within the pointwise bootstrap intervals. Another idea may be to use the proposed local EM estimate to aid in the selection of choosing breakpoints for a piecewise constant model.

### 6.2 Further Research

Although we have taken a first step at estimating non-homogeneous intensity matrices there is much work to be done. For simplicity, we have not included any covariates in our model. It is quite reasonable to assume that intensities may depend on subjects’ demographics and therefore these should be included in the model. Furthermore, more work is needed on model assessment
tools. Currently, non-parametric bootstrap methods are used to find pointwise confidence intervals. Ideally, simultaneous confidence intervals along the lines discussed in Mandel and Betensky [34] would be advantageous. This is complicated by the fact that there are multiple intensities that are estimated simultaneously. Furthermore, throughout we have assumed the presence of some data in the smoothing window. Creating a very fine grid with a very small bandwidth may result in smoothing windows where there is no data. One possible solution would be to use a Bayesian approach. This would permit us to generate $Y_i$ from their distribution based on the parameters of the transition model. This is one focus of future research.

Another primary area of focus is to estimate the intensity as a truly continuous function of time without assuming locally constant estimates for $\theta$. The problem arises because equation (2.2) is no longer valid since assumptions on which the equation was based have been violated. For an underlying non homogeneous intensity the transition probability can be found as shown in Andersen [4]:

$$P(s, t) = \Pi(s, t)(I + dA(u))$$

As discussed earlier, this approach is difficult to implement especially in the context of unbalanced and incomplete data. To extend the method suggested in this thesis a first step may involve extending the local likelihood constant method to local linear or to any polynomial of degree $p$, possibly by using the product integral. If possible, this can be extended to the entire interval to obtain continuous estimate $\hat{\theta}(\cdot)$.

Although MSM models have been used quite extensively in the literature, the intensities are treated as time homogeneous, or when necessary, as piecewise constant. We are not aware of any methods that allow for the intensity to be truly time dependent without any prior knowledge of its functional form. Our method has shown that obtaining estimates of non-homogeneous intensities is a viable option. Further research is needed to advance these methods and find other ways of estimating $\hat{\theta}(\cdot)$.
Bibliography


Appendix A

Appendix

A.1 Eigenvalue Decomposition for $Q$

The relationship between the intensity and transition matrix involves canonical decomposition as discussed in Tuma et al. [47], Kalbfleisch and Lawless [23]. As mentioned in Chapter 2, in the time homogeneous setting, the transition matrix $P(t)$ is related to the intensity matrix $Q(t)$ by $P(t) = \exp(Q(t))$. If $Q(t)$ has $K$ distinct eigenvalues $d_1, d_2, \ldots, d_K$ then $Q(t)$ can be decomposed to $ADA^{-1}$ where $A$ is a $K \times K$ matrix with the $j^{th}$ column being the right eigenvector corresponding to $d_j$ and $D$ is $diag(d_1, d_2, \ldots, d_K)$. The relationship between $P(t)$ and $Q(t)$ is

$$P(t) = Adiag(e^{d_1 t}, e^{d_2 t}, \ldots, e^{d_K t})A^{-1}$$

For example with $K = 3$ ordinal states we have seen that:

$$Q = \begin{bmatrix} -q_{12} & q_{12} & 0 \\ q_{21} & -(q_{21} + q_{23}) & q_{23} \\ 0 & q_{32} & -q_{32} \end{bmatrix}$$

and letting:
\[ \alpha_1 = q_{12} + q_{21} \]
\[ \alpha_1 = q_{23} + q_{32} \]
\[ \sigma = \sqrt{(\alpha_1 - \alpha_2)^2 + 4q_{21}q_{23}} \]
\[ \beta_1 = -0.5(\alpha_1 + \alpha_2) - 0.5\sigma \]
\[ \beta_2 = -0.5(\alpha_1 + \alpha_2) + 0.5\sigma \]

we can use Maple software [38] to find the 3 eigenvalues and the related entries in the transition matrix:

**Eigenvalues**

\[ d_1 = 0 \quad \text{(Note: \( \text{Det}(Q) = 0 \))} \]
\[ d_2 = -\frac{1}{2}(\alpha_1 + \alpha_2) + \frac{1}{2}\sqrt{q_{12}^2 + q_{21}^2 + q_{23}^2 + q_{32}^2 + 2(q_{21}q_{12} + q_{32}q_{23} + q_{23}q_{21} - q_{32}q_{12} - q_{12}q_{23} - q_{32}q_{21})} \]
\[ d_2 = -\frac{1}{2}(\alpha_1 + \alpha_2) - \frac{1}{2}\sqrt{q_{12}^2 + q_{21}^2 + q_{23}^2 + q_{32}^2 + 2(q_{21}q_{12} + q_{32}q_{23} + q_{23}q_{21} - q_{32}q_{12} - q_{12}q_{23} - q_{32}q_{21})} \]
Transition Matrix Entries

\[
p_{11} = -\frac{2}{4(q_{12}q_{23} + q_{12}q_{32} + q_{21}q_{32})} \left[ q_{21}(q_{23} - q_{32}) + q_{12}e^{\beta_2}(\alpha_2 - \alpha_1 - \sigma) - q_{21}(q_{23} - q_{32}) - 2q_{32}q_{21}\sigma \right]
\]

\[
p_{12} = \frac{2}{4(q_{12}q_{23} + q_{12}q_{32} + q_{21}q_{32})} \left[ q_{12}e^{\beta_1}(\alpha_2 - \alpha_2 - \sigma) + q_{12}e^{\beta_2}(\alpha_2 - \alpha_2 - \sigma) - q_{21}(q_{23} - q_{32}) + 2q_{32}\sigma \right]
\]

\[
p_{13} = 1 - p_{11} - p_{12}
\]

\[
p_{21} = \frac{-2}{4(q_{12}q_{23} + q_{12}q_{32} + q_{21}q_{32})} \left[ q_{21}(q_{23} - q_{21} - \sigma) + 2q_{12}q_{23} + e^{\beta_2}(q_{32} - q_{21} - \sigma) - 2q_{32}q_{21}\sigma \right]
\]

\[
p_{22} = \frac{-2}{4(q_{12}q_{23} + q_{12}q_{32} + q_{21}q_{32})} \left[ (e^{\beta_1} - e^{\beta_2})(q_{12}q_{23} - q_{21} - \sigma) + q_{21}q_{32}(q_{32} - q_{23} - \sigma) - \sigma(e^{\beta_1} + e^{\beta_2})(q_{21}q_{32} + 1) - 2q_{12}q_{32}\sigma \right]
\]

\[
p_{23} = 1 - p_{21} - p_{22}
\]

\[
p_{31} = \frac{2}{4(q_{12}q_{23} + q_{12}q_{32} + q_{21}q_{32})} \left[ q_{12}q_{23}(\alpha_1 + \sigma) + e^{\beta_2}(\alpha_1 - \alpha_1 - \sigma) + 2\sigma \right]
\]

\[
p_{32} = \frac{2}{4(q_{12}q_{23} + q_{12}q_{32} + q_{21}q_{32})} \left[ q_{12}(\alpha_1 - \alpha_1 - \sigma) - 2q_{12}q_{32} + e^{\beta_2}(q_{12} + (\alpha_1 - \alpha_2 - \sigma) + 2q_{21}q_{32} + 2q_{12}\sigma \right]
\]

\[
p_{33} = 1 - p_{31} - p_{32}
\]

As can be seen, the transition probabilities are complex functions of the intensities. As discussed in Kalbfleisch and Lawless [23], Jennrich and Bright [22] using a similar decomposition, the first derivatives as found in equation 2.12 can be found as:

\[
\frac{\partial \mathbf{P}(t)}{\partial \Psi_w} = \mathbf{A} \mathbf{V}_w \mathbf{A}^{-1} \quad w = 1, \ldots, p
\]

where \( \mathbf{V}_w \) is a \( K \times K \) matrix with entries

\[
\begin{align*}
g_{ij}^w &= \begin{cases} 
  g_{ij}(e^{d_{ij}} - e^{d_{ij}})/(d_i - d_j) & i \neq j \\
  g_{ii}e^{d_{ii}} & i = j
\end{cases}
\end{align*}
\]

where \( g_{ij}^w \) is the \((i,j)\) entry in \( G^w = \mathbf{A}^{-1}(\partial \mathbf{Q}(t)/\partial \Psi_w)\mathbf{A} \).

This approach allows us to use the intensity matrix, eigenvalues, and eigenvectors, to find the
derivative of the transition matrix without computing the derivatives directly.

A.2 Derivation of Expectation of Second Derivative

To implement the Newton-Raphson algorithm the second derivative or Hessian matrix is required. Equation (2.13) found the second derivative of the likelihood function to be:

\[
\frac{\partial^2 \ell}{\partial \theta_w \partial \theta_z} = \sum_{i=1}^{N} \sum_{j=1}^{J_i} \sum_{u,v=1}^{K} \delta_{ijuv} \left\{ \frac{\partial^2}{\partial \theta_w \partial \theta_z} P_{uw}(t_{ij-1}, t_{ij}; \theta) \frac{P_{uv}(t_{ij-1}, t_{ij}; \theta)}{P_{uv}(t_{ij-1}, t_{ij}; \theta)} - \frac{\partial}{\partial \theta_w} P_{uv}(t_{ij-1}, t_{ij}; \theta) \frac{\partial}{\partial \theta_z} P_{uv}(t_{ij-1}, t_{ij}; \theta) \right\}
\]

As was shown in section A.1 the transition probabilities are complex functions of the intensities and finding the second derivative using a straightforward approach would only complicate matters. To avoid this problem a quasi-Newton algorithm that avoids the explicit form of the second derivative was used and uses in its place \( E \left[ -\frac{\partial^2 \ell}{\partial \theta_w \partial \theta_z} \right] \)

The expectation can be derived by first noting that the first term in the square brackets is 0 because:

\[
\sum_{v=1}^{K} p_{uv} = 1
\]
\[
\frac{\partial p_{uK}}{\partial \theta_w} = -\sum_{v=1}^{K-1} \frac{\partial p_{uv}}{\partial \theta_w}
\]
\[
\frac{\partial^2 p_{uK}}{\partial \theta_w \partial \theta_z} = -\sum_{v=1}^{K-1} \frac{\partial^2 p_{uv}}{\partial \theta_w \partial \theta_z}
\]
\[
\sum_{v=1}^{K-1} \frac{\partial^2 p_{uv}}{\partial \theta_w \partial \theta_z} + \frac{\partial^2 p_{uK}}{\partial \theta_w \partial \theta_z} = \sum_{v=1}^{K} \frac{\partial^2 p_{uv}}{\partial \theta_w \partial \theta_z} = 0
\]

Next, to find the expectation of the second term in the brackets by finding the conditional expectation:

\[
E[\delta_{ijuv}] = E\{E[\delta_{ijuv}|\delta_{iju}]\} = E[\delta_{iju}p_{uw}(t_{ij-1}, t_{ij}; \theta)] = p_{uw}(t_{ij-1}, t_{ij}; \theta)\delta_{iju}
\]
Now dividing top and bottom by $p_{uv}(t_{ij-1}, t_{ij}; \theta)$ and taking the negative expectation of the remaining terms leads to equation (2.14).

### A.3 Derivation of Equations (3.10), (3.11) and (3.12)

To derive equations (3.10), (3.11) and (3.12) we must compute the conditional probabilities. To find (3.10), we will first examine the approach using one entry of $Y_i(g_j)$. Suppose there are $K = 3$ states and let $e_1, e_2, e_3 = (1, 0, 0)', (0, 1, 0)', (0, 0, 1)'$ respectively. In addition let $S_i, i = 1, 2, 3$ be the occupied state. Looking at just the first entry $Y_i(g_j)$ we will use $e_1$ to find this probability. Then $P(S_1|Y_i(t_i^- (g_j)), Y_i(t_i^+ (g_j)))$ can be found as:

$$P(S_1|Y_i(t_i^- (g_j)), Y_i(t_i^+ (g_j))) = \frac{P(S_1, Y_i(t_i^- (g_j-1)), Y_i(t_i^+ (g_j)))}{P(Y_i(t_i^- (g_j)), Y_i(t_i^+ (g_j)))}
= \frac{P(Y_i(t_i^+ (g_j))|S_1, Y_i(t_i^- (g_j)))P(S_1|Y_i(t_i^- (g_j)))P(Y_i(t_i^- (g_j)))}{P(Y_i(t_i^+ (g_j))|Y_i(t_i^- (g_j)))P(Y_i(t_i^- (g_j)))}
= \frac{P(Y_i(t_i^+ (g_j))|S_1)P(S_1|Y_i(t_i^- (g_j)))}{P(Y_i(t_i^+ (g_j))|Y_i(t_i^- (g_j)))}	ext{ by the Markov property}
= P(t_{e_1}, t_{Y_i(t_i^- (g_j))})e_1 e_1' P(t_{Y_i(t_i^- (g_j))}, t_{e_1})
= e_1' P(t_{Y_i(t_i^- (g_j))}, t_{e_1}) e_1 P(t_{e_1}, t_{Y_i(t_i^+ (g_j))})e_1$$

where $t_{e_1}$ is the time corresponding to state $S_1$ being occupied. An identical approach can be made with $K$ being any number of states. Using $S_2$ and $S_3$ would result in a similar result. Therefore, all entries of $Y_i(g_j)$ can be found simultaneously by using equation (3.10). The same argument can be used to find (3.11).

To derive equation (3.12), the joint probability of the two consecutive unobserved states is required, conditional on the nearest observed states. As above, for one entry of $Y_i(g_j)$ and $Y_i(g_{j+1})$ this can be found as:
\[ P(Y_i(g_j)Y_i(g_{j+1})|Y_i(t_i^{-}(g_j)), Y_i(t_i^{+}(g_j))) \]
\[ = P(Y_i(g_{j+1})|Y_i(g_j), Y_i(t_i^{-}(g_j)), Y_i(t_i^{+}(g_j)))P(Y_i(g_j)|Y_i(t_i^{-}(g_j)), Y_i(t_i^{+}(g_j))) \]
\[ = \frac{P(Y_i(t_i^{+}(g_j))|Y_i(g_{j+1}), Y_i(g_j))P(Y_i(g_{j+1})|Y_i(g_j))P(Y_i(g_j)|Y_i(t_i^{-}(g_j)), Y_i(t_i^{+}(g_j)))}{P(Y_i(t_i^{+}(g_j))|Y_i(t_i^{-}(g_j)))} \]

which, reducing and using the Markov property yields:
\[ = \frac{P(Y_i(g_j)|Y_i(t_i^{-}(g_j)))P(Y_i(g_{j+1})|Y_i(g_j))P(Y_i(t_i^{+}(g_j))|Y_i(t_i^{-}(g_j)))}{P(Y_i(t_i^{+}(g_j))|Y_i(t_i^{-}(g_j)))} \]

which, when computed for all states simultaneously yields equation (3.12).

### A.4 Another Simulation

In addition to the simulations found in chapter 3 another simulation was run to further examine the intensity estimates obtained from the local EM algorithm. Two grid sizes of \( \delta = 1 \) and \( \frac{1}{3} \) were used to examine the estimates. Figures A.1 and A.2 compare the true intensity to intensity estimates obtained using the algorithm. Although the method appears to estimate the intensities well, we note that there seems to be some biased with respect to \( q_{23} \).
Figure A.1: Time Varying intensity estimates for various bandwidths with a fixed grid of $\delta = 1$: true $\theta$ (—); $\theta$ PWC (- - -); BW 1 (—); BW3 (—); BW8 (—).
Figure A.2: Time Varying intensity estimates for various bandwidths with a fixed grid of \( \delta = 1/3 \): true \( \theta \) (—); \( \theta \) PWC ( - - -); BW 1 (—); BW3 (—); BW8 (—).
A.5 R Code

Below, is some of the R functions (and code) that was used in the thesis. Unfortunately, it is not in the shape to be used by other people but it is in middle of being re-written.

A.5.1 Code for Simulating NHPP

```r
simMoveTime = function(t1, tN, lambdaList, lambdaMax) {

allProb = function(x) {
  result = 0
  for( D in 1:length(lambdaList) )
    result = result + lambdaList[[D]](x)
  result
}

meanNumberMoves = lambdaMax * (tN - t1)
Nmmove = rpois(1, meanNumberMoves)
if(Nmove == 0)
  return(Inf)

# accept/reject
moveTimesBar = runif(Nmove, t1, tN)
acceptProb = allProb(moveTimesBar) / lambdaMax
moveTimes = moveTimesBar[as.logical(rbinom(Nmove, 1, acceptProb))]
if(length(moveTimes)==0)
  return(Inf)
moveTime = min(moveTimes)
if(length(lambdaList)==1) {

```
attributes(moveTime)$state = names(lambdaList)
} else {
  # simulate state
  # NOTE: This bit only works when there are only 2 states
  stateProbSeq = NULL
  for(D in seq(1, length(lambdaList)) )
    stateProbSeq = c(stateProbSeq, lambdaList[[D]](moveTime) )

  # convert rates to probabilities conditional on a move
  stateProbSeq = stateProbSeq / sum(stateProbSeq)

  attributes(moveTime)$state =
  names(lambdaList)[as.logical(rmultinom(1,1,stateProbSeq) ) ]
  # names(lambdaList)[2-rbinom(1,1,lambdaList[[1]](moveTime) / allProb(moveTime) )]
}
return(moveTime)

### lambda is user defined

#sim 300 people 20 time points in total
lambda = list("1" = list("2"=function(x) 0.2 + 0.02*x),
       "2" = list("1" = function(x) 0.4*(x-10)^2/100 + 0.2,
       "3"=function(x)-0.4*(x-10)^2/100+ 0.6),
       "3" = list("2" = function(x) plogis(x,location = 10, scale = 2.5)/1.5 + 0.2)
)
lambdaMaxFunc <- function(lambda, currState, t1, tN) {
  lambdaMax <- numeric()
  lambdaList <- lambda[[currState]]
  x <- seq(t1, tN, by = 0.1)
  result = 0
  for( D in 1:length(lambdaList) )
    result = result + max(lambdaList[[D]](x))
  result
  lambdaMax <- result
  lambdaMax
}

simMsmInhom = function(lambda, t1, tN, s1, lambdaMax) {
  moveTime = t1
  attributes(moveTime)$state = s1
  stateSeq = timeSeq = NULL
  while(moveTime < tN) {
    stateSeq = c(stateSeq, attributes(moveTime)$state)
    timeSeq = c(timeSeq, moveTime)
    lambdaMax = lambdaMaxFunc(lambda, attributes(moveTime)$state, t1, tN)
    moveTime = simMoveTime(moveTime, tN, lambda[[attributes(moveTime)$state]],
                           lambdaMax)
  }
  data.frame(time=timeSeq, state=stateSeq)
}

numpple <- 300
### set it up

start <- ceiling(runif(numpple,0,3))

people <- list()
peopleApprox <- list()
K <- 1
t1 <- 0
tN <- 20

## set up the grid

for (p in 1:numpple) {
  people[[p]] <- cbind(p, simMsmInhom(lambda, t1, tN, start[p], 1.2))
  setup <- approx(c(people[[p]]$time, Inf),
                  as.numeric(c(people[[p]]$state, -999)), seq(t1, tN, len=(tN*K+1)),
                  method="constant")
  peopleApprox[[p]] <- data.frame(id = p, time = setup[[1]],
                                  time2 = 1:(tN*K+1), state = setup[[2]])
}

data.nonh <- as.data.frame(do.call("rbind", peopleApprox))

# create missing data - 30%

data.rand <- data.nonh
uni.time2 <- unique(data.rand$time2)
data.rand$rand <- runif(dim(data.rand)[1])
A Bibliography

A.5.2 Code for Local EM

Grid Size

## set up the grid based on the grid size

```r
fineGrid <- function(dataobs,delta) {
  delta <- 1/delta
  dataobs$numrecords <- 1:dim(dataobs)[1]
  time.new <- rep(seq(1:(delta*length(unique(dataobs$time)))),
      length(unique(dataobs$id)))
  obsnum <- 1:length(time.new)
  id.new <- rep(dataobs$id,each=delta)
  state.new <- rep(NA,length(time.new))
  numrecords.new <- 1:length(time.new)
  data.new <- data.frame(numrecords.new,obsnum,time.new,id.new,state.new)

  #### pull off the old time points where we had observed states
  data.new$state.new[(dataobs$numrecords[!is.na(dataobs$state)]*delta)-(delta-1)] <-
  dataobs$state[!is.na(dataobs$state)]

  data.new <- subset(data.new,select = -numrecords.new)
}
```

data.rand$state[data.rand$time2 != min(uni.time2) & data.rand$rand < 0.3] <- NA
data.rand$times <- data.rand$time2
names(data.new) <- c("obsnum","time","id","state")

    return(data.new)

}

E Step

## E Step

### find the expected probabilities for a missing observation

### this function keeps track of the time varying probability transition matrix

transInt.tv <- function(result,nobs.prior,nobs.post) {
  transMat.tv <- array(NA,c(Nstate,Nstate,Ntime))
  for (i in (nobs.prior+1):nobs.post) {
    transMat.tv[,nobs.prior] <- MatrixExp(result[,nobs.prior])
    transMat.tv[,i] <- transMat.tv[,i-1] %*% MatrixExp(result[,i])
  }
  return(transMat.tv)
}

#### compute the probabilities

ExpectedProbs.grid <- function(ObservedData,transInt,dataimp.arr,
data,data.supply=data.grid.train,option = "supply",Nobs=1) {
  tab <- table(data.supply$state[data.supply$time==min(unique(data.supply$time))])
\begin{verbatim}
p1 <- tab[1]/sum(tab)
p2 <- tab[2]/sum(tab)
p3 <- tab[3]/sum(tab)
if (attributes(tab)$dim < 3)
{print("Error: not sufficient data, start at later data point")}

for (i in 1:Nobs) {
  if (is.na(ObservedData[,1,1,i]))
    ObservedData[,1,i] <- dataimp.arr[,1,i] <- c(p1,p2,p3)
    loc.na <- which(is.na(ObservedData[,1,i]))
  if (length(loc.na) == 0) next #### for no missing data for an individual
  if (max(loc.na) == Ntime) {
    nobs.prior <- max(which(!is.na(ObservedData[,1,i]))
    [which(!is.na(ObservedData[,1,i]) < max(loc.na)])
    time.prior <- nobs.prior/delta
    prior <- ObservedData[,i][,nobs.prior]
    interval <- max(loc.na) - nobs.prior
    ##t(prior)%*% transInt.tv(transInt,time.prior,j)[,j-1]
    for (h in (nobs.prior+1):max(loc.na)) {
      dataimp.arr[,h,i] <- prior%*%transInt.tv(transInt,nobs.prior,h)[,h-1]
    }
    loc.na <- which(is.na(dataimp.arr[,1,i]))
  }

  for (j in loc.na) {
    nobs.prior <- max(which(!is.na(ObservedData[,1,i]))
    [which(!is.na(ObservedData[,1,i]) < j])
    time.prior <- nobs.prior/delta
  }
\end{verbatim}
nobs.post <- min(which(!is.na(ObservedData[,1,i])))

[which(!is.na(ObservedData[,1,i])) > j]

time.post <- nobs.post/delta

prior <- ObservedData[,,i][,nobs.prior]
post <- ObservedData[,,i][,nobs.post]
int.prior <- j - nobs.prior
int.post <- nobs.post - j
interval <- nobs.post-nobs.prior

numerato <- diag(transInt.tv(transInt,j,nobs.post)[,,nobs.post-1]

%*%post%*%t(prior)%*% transInt.tv(transInt,nobs.prior,j)[,,j-1])

denom <- t(prior)%*%transInt.tv(transInt,nobs.prior,nobs.post)[,,nobs.post-1]%*%post

imp.prob <- numerator/denom

dataimp.arr[1, ,j,i] <- imp.prob

}

}

return(dataimp.arr)

}

### find the expected transitions using the joint distribution

TransitProbs <- function(array.new,dataFull1,transInt,data) {

datatrans.arr <- array(NA,c(3,3,Ntime,Nobs))
datatrans.arr.f <- array(NA,c(3,3,Ntime-1,Nobs))
tab <- table(data$state[data$time==min(unique(data$time))])
p1 <- tab[1]/sum(tab)
p2 <- tab[2]/sum(tab)
p3 <- tab[3]/sum(tab)
for (i in 1:Nobs) {
loc.na <- which(is.na(array.new[,1,,i]))
first.obs <- 1

if (is.na(array.new[,1,1,i]) ) {

   nobs.post <- min(which(!is.na(array.new[,1,,i])))
   first.obs <- nobs.post
   time.post <- nobs.post/delta
   post <- array.new[,,,i][,nobs.post]
   # interval = nobs.post - min(loc.na)

   for (j in 2:nobs.post) {

      int.post <- nobs.post - j

      if (int.post == 0) {
         datatrans.arr[,,j,i] <- dataFull1[,,j-1,i]*t(dataFull1[,,j,i])
         break
      }
   }}
datatrans.arr[,j,i] <- diag(dataFull1[,j-1,i])%*%MatrixExp(transInt[,j-1])

} ### closes j loop

} ### closes if (is.na(array.new[,1,1,i]) )

last.obs <- Ntime
if (is.na(array.new[,1,Ntime,i])) {
    nobs.prior <- max(which(!is.na(array.new[,1,,i])))
    [which(!is.na(array.new[,1,,i])) < max(loc.na)])
last.obs <- nobs.prior
time.prior <- nobs.prior/delta
prior <- array.new[,i][,nobs.prior]
interval <- max(loc.na) - nobs.prior
for (h in (nobs.prior+1):max(loc.na)) {
    datatrans.arr[,h,i] <- diag(dataFull1[,h-1,i])%*%MatrixExp(transInt[,h-1])
}

} ### closes if (is.na(array.new[,1,Ntime,i])

# for ( j in 2:(Ntime-interval)) {
if (last.obs - first.obs > 0) {
    for ( j in (first.obs+1):last.obs) {
        if ( !is.na(array.new[,1,j,i]))
        { datatrans.arr[,j,i] <- dataFull1[,j-1,i]%*%t(dataFull1[,j,i]) }
    }
else{
    nobs.prior <- max(which(!is.na(array.new[,1,,i]))
[which(!is.na(array.new[,1,,i])) < j])

time.prior <- nobs.prior/delta

nobs.post <- min(which(!is.na(array.new[,1,,i]))

[which(!is.na(array.new[,1,,i])) > j])

time.post <- nobs.post/delta

prior <- array.new[,,,i][,nobs.prior]

post <- array.new[,,,i][,nobs.post]

int.prior <- j - nobs.prior

int.post <- nobs.post - j

interval <- nobs.post-nobs.prior

if (int.prior == 1) {
    datatrans.arr[,,j,i] <- dataFull1[,,j-1,i]%*%t(dataFull1[,,j,i])
}

else{
    numerator<-diag(transInt.tv(transInt,nobs.prior,j)[,,j-1-1][which(prior==1),])
    %*%MatrixExp(transInt[,,j-1])
    %*%diag(transInt.tv(transInt,j,nobs.post)[,,nobs.post-1][,which(post==1)])
    denom <- t(prior)%*%transInt.tv(transInt,nobs.prior,nobs.post)
    [,nobs.post-1]%*%post
    datatrans.arr[,,j,i] <- numerator/as.numeric(denom)
}

### closes else a few lines up

} ### closes BIG else

} ### closes j loop

} ##### closes if (last.obs - first.obs > 0)
datatrans.arr.f[,,,i] <- datatrans.arr[,,-1,i]
}
### closes i loop

return(datatrans.arr.f)

}

### M Step

### M - Step

### extract the parameters of interest
ExtractPars <- function(mat) {
para <- c(mat[4],mat[2],mat[8],mat[6])
return(para)
}

## Likelihood

likFun = function(para, data, kernel) {
    mat <- matrix(c(-para[1],para[1],0,para[2],-(para[3]+para[2]),
                   para[3],0,para[4],-para[4]),nrow=numcol,byrow=T)
    params = c(MatrixExp(mat))
    likVec = data*log(params)  ### data is the expected transitions
    #print(sum(likVec * kernel))  ### this line was inserted Sep 5
}
sum(likVec * kernel)

Monetime = function(data.p, kernel.p, start,Dindex=NULL) {

res = optim(par=ExtractPars(start), likFun, data=data.p, kernel=kernel.p ,
    method="L-BFGS-B",lower=rep(0.0000001,length(par)), control = list(fnscale = -1))

res.par <- res$par

# print(res$value)

res.mat <- matrix(c(-res.par[1],res.par[1],0, res.par[2],
    -res.par[2]-res.par[3],res.par[3], 0,res.par[4],-res.par[4]), nrow = 3, byrow = T)

# check for convergence - to do this we will see if res$convergence is 0
if (res$convergence != 0) warning("convergence problem")

return(list(res.mat=res.mat,conv=res$convergence,messagge=res$message))

}
epan.density <- function(x, min, max, h) {if (min <= x & x <= max) dens <- (3/(4*h))*(1-(x/h)^2) else {dens = 0} }
epan <- function(x, min, max, h) {sapply(x, epan.density, min, max, h)}

kernelFunP = function(x, bw, type="gaussian") {

  if(type=="gaussian") {
  res=dnorm(x, 0, bw)
  }

  if(type=="uniform") {
  res=dunif(x, -bw/2, bw/2)
  }

  if (type=="epan") {
  res = epan(x, -bw, bw, bw)
  #res = epan(x, bw)
  }

  res
### for the kernel weights

```r
getKvec = function(bw, delta, support=4*bw, type="gaussian") {
  ## if (type="epan") {bw <- bw+1}
  support = delta * round(support/delta)

  res = kernelFunP(seq(-support, support, by=delta), bw=bw, type=type)

  # get rid of zeros on either end
  notzeros = range(which(res > 0))
  res = res[seq(notzeros[1], notzeros[2])]

  res = res/sum(res)

  if(round( (length(res)-1)/2 ) != ( (length(res)-1)/2 ) )
    warning("number of elements of kernel vector is even")

  attributes(res)$N = (length(res)-1)/2

  res
}
```
### Local M-Step

Mstep.uniep <- function(dataFull, bw, delta, support=4*bw, type="gaussian", transInt) {
  Kvec <- getKvec(bw=bw, delta=delta, support=support, type=type)
  #print(Kvec)

  Ntime = (dim(dataFull)[3]) + 1
  Nkern = attributes(Kvec)$N
  result = array(NA, c(Nstate, Nstate, Ntime))
  message <- list()
  converg <- numeric()

  for (Dindex in 1:(Ntime)) {
    useSeq = seq(Dindex-Nkern, Dindex+Nkern)
    useSeq = useSeq[useSeq > 0 & useSeq < Ntime]
    if (length(useSeq)==Nkern*2+1) {
      Kvec <- getKvec(bw=bw, delta=delta, support=support, type=type)
      Kuse = Kvec[useSeq] ### fix this to give appropriate numbers
      Kuse = Kvec[1:length(useSeq)]
      Kuse = Kvec[(Nkern+1-((length(useSeq)-1)/2)):((Nkern+1+((length(useSeq)-1)/2))]
      if (Dindex < (Ntime/2)) {Kuse = Kvec[length(useSeq):1]}
      Kuse = Kuse / sum(Kuse)
    }
  }
}
if (length(useSeq) < Nkern*2+1 ) {
  Kvec <- getKvec(bw=2*bw, delta=delta, support=support, type="uniform")
  Kuse = Kvec[1:length(useSeq)]
  #### Kuse = Kvec[(Nkern+1-((length(useSeq)-1)/2)):(Nkern+1+((length(useSeq)-1)/2))]
  if (Dindex < (Ntime/2) ) {Kuse = Kvec[length(useSeq):1]}
  Kuse = Kuse / sum(Kuse)
}

dataUse = dataFull[,useSeq,]

hold = Monetime(data.p=dataUse,kernel.p=Kuse, start=transInt[,Dindex])
result[,Dindex] <- hold$res.mat
result[,Ntime] <- result[,Dindex]
converg[Dindex] <- hold[[2]]
message[[Dindex]] <- hold[[3]]
### #print(Dindex)
}

return(list(result=result,converg=converg,message=message))
}

##function to extract intensity estimates and multiply
##them by there corresponding "rate" #

extractInt <- function(estimate,delta=1,specify) {

last <- length(estimate)
    a <- (estimate[[last]]/delta)[1,2,]
    b <- (estimate[[last]]/delta)[2,1,]
    c <- (estimate[[last]]/delta)[2,3,]
    d <- (estimate[[last]]/delta)[3,2,]

ext=cbind(q12=a,q21=b,q23=c,q32=d)
colnames(ext) = c(paste("q12.",specify,sep=""),paste("q21.",specify,sep=""),
paste("q23.",specify,sep=""),paste("q32.",specify,sep=""))
return(ext)
}