Experiences with some longitudinal exploratory data mining problems

Gilbert Ritschard
Matthias Studer and Emmanuel Rousseaux

NCCR LIVES and Institute for demographic and life course studies
University of Geneva
http://mephisto.unige.ch

COMPSTAT, Geneva, August 19-22, 2014
1 Introduction

2 Frequent subsequences in TraMineR

3 Frequent Swiss life course subsequences

4 Discriminant subsequences

5 Maximal subsequences

6 Association rules

7 Conclusion
1 Introduction

- Objectives
  - The Biographical Data from the Swiss Household Panel
  - Frequent subsequences versus Frequent itemsets
Introduction

Objectives

- Data-mining-based methods (pattern mining)
  - Discovering *interesting information from sequences of life events*, i.e., on how people sequence important life events
    - What is the most *typical succession* of family or professional life events?
    - Are there *standard* ways of sequencing those events?
    - What are the most typical events that occur after a given subsequence such as after leaving home and ending education?
    - How is the sequencing of events *related to covariates*?
    - Which event sequencings do *best discriminate groups* such as men and women?
- Mining of frequent (Agrawal and Srikant, 1995; Mannila et al., 1995; Bettini et al., 1996; Mannila et al., 1997; Zaki, 2001) and discriminant event subsequences (Ritschard et al., 2013)
Introduction

Objectives

- Data-mining-based methods (pattern mining)
  - Discovering interesting information from sequences of life events, i.e., on how people sequence important life events
    - What is the most typical succession of family or professional life events?
    - Are there standard ways of sequencing those events?
    - What are the most typical events that occur after a given subsequence such as after leaving home and ending education?
    - How is the sequencing of events related to covariates?
    - Which event sequencings do best discriminate groups such as men and women?

- Mining of frequent (Agrawal and Srikant, 1995; Mannila et al., 1995; Bettini et al., 1996; Mannila et al., 1997; Zaki, 2001) and discriminant event subsequences (Ritschard et al., 2013)
Objectives

- **Data-mining-based methods (pattern mining)**
  - Discovering *interesting information from sequences of life events*, i.e., on how people sequence important life events
    - What is the most *typical succession* of family or professional life events?
    - Are there *standard* ways of sequencing those events?
    - What are the most typical events that occur after a given subsequence such as after leaving home and ending education?
    - How is the sequencing of events *related to covariates*?
    - Which event sequencings do *best discriminate groups* such as men and women?
  - Mining of frequent (Agrawal and Srikant, 1995; Mannila et al., 1995; Bettini et al., 1996; Mannila et al., 1997; Zaki, 2001) and discriminant event subsequences (Ritschard et al., 2013)
Objectives (continued)

- Recall kind of results that can be obtained by mining event subsequences
  - most frequent subsequences
  - association rules between subsequences
  (cf. Emmanuel Rousseaux, Session CS75, Friday 22, 9 am)
  - subsequences that best discriminate groups (provided covariate)

Problem How to deal with nested subsequences?
- If (LHome) \(\rightarrow\) (Marriage) \(\rightarrow\) (Childbirth) is frequent, shall we also consider people following that path when counting the frequency of subsequence (LHome) \(\rightarrow\) (Marriage)?
- Could be more interesting to know how many people with (LHome) \(\rightarrow\) (Marriage), did not have childbirth afterwards.
Objectives (continued)

- Recall kind of results that can be obtained by mining event subsequences
  - most frequent subsequences
  - association rules between subsequences
  (cf. Emmanuel Rousseaux, Session CS75, Friday 22, 9 am)
  - subsequences that best discriminate groups (provided covariate)

- Problem How to deal with nested subsequences?
  - If (LHome) → (Marriage) → (Childbirth) is frequent, shall we also consider people following that path when counting the frequency of subsequence (LHome) → (Marriage)?
  - Could be more interesting to know how many people with (LHome) → (Marriage), did not have child birth afterwards.
Event sequences versus state sequences

- **State sequence**: states last a whole interval period
  
<table>
<thead>
<tr>
<th>age</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
</tr>
</thead>
<tbody>
<tr>
<td>state</td>
<td>2P</td>
<td>2P</td>
<td>A</td>
<td>A</td>
<td>UC</td>
<td>UC</td>
<td>UC</td>
</tr>
</tbody>
</table>

- **Event sequence**: events occur at a given (time) position
  - Interest in their order, in their sequencing
  - Can be time stamped (TSE)

<table>
<thead>
<tr>
<th>id</th>
<th>Timestamp</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>22</td>
<td>Leaving Home</td>
</tr>
<tr>
<td>101</td>
<td>24</td>
<td>Start living with partner</td>
</tr>
<tr>
<td>101</td>
<td>24</td>
<td>Childbirth</td>
</tr>
</tbody>
</table>
Event sequences versus state sequences

- **State sequence**: states last a whole interval period
  
<table>
<thead>
<tr>
<th>age</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
</tr>
</thead>
<tbody>
<tr>
<td>state</td>
<td>2P</td>
<td>2P</td>
<td>A</td>
<td>A</td>
<td>UC</td>
<td>UC</td>
<td>UC</td>
</tr>
</tbody>
</table>

- **Event sequence**: events occur at a given (time) position
  - Interest in their order, in their sequencing
  - Can be time stamped (TSE)
  
<table>
<thead>
<tr>
<th>id</th>
<th>Timestamp</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>22</td>
<td>Leaving Home</td>
</tr>
<tr>
<td>101</td>
<td>24</td>
<td>Start living with partner</td>
</tr>
<tr>
<td>101</td>
<td>24</td>
<td>Childbirth</td>
</tr>
</tbody>
</table>
1 Introduction

- Objectives
- The Biographical Data from the Swiss Household Panel
- Frequent subsequences versus Frequent itemsets
Problem in Event Sequence Mining

Introduction

The Biographical Data from the Swiss Household Panel

The Biographical SHP Data

- Sequences derived from the biographical survey conducted in 2002 by the Swiss Household Panel [www.swisspanel.ch](http://www.swisspanel.ch)
- Retain the 1503 cases studied in Widmer and Ritschard (2009) with techniques for state sequences
- Two channels: Cohabitational and occupational
- Only individuals aged 45 or more at survey time
- Focus on life trajectory between 20 and 45 years
- Granularity is yearly level
Problem in Event Sequence Mining

Introduction

The Biographical Data from the Swiss Household Panel

The Cohabitational State Sequences

Cohabitational trajectories

- Biological father and mother
- One biological parent
- One biological parent with her/his partner
- Alone
- With partner
- Partner and biological child
- Partner and non-biological child
- Biological child and no partner
- Friends
- Other
The Biographical Data from the Swiss Household Panel

The Occupational State Sequences
# Problem in Event Sequence Mining

## Introduction

The Biographical Data from the Swiss Household Panel

## Short and long state labels

<table>
<thead>
<tr>
<th>Cohabitational</th>
<th>Occupational</th>
</tr>
</thead>
<tbody>
<tr>
<td>2P</td>
<td>Mi</td>
</tr>
<tr>
<td>Biological father and mother</td>
<td>Missing</td>
</tr>
<tr>
<td>1P</td>
<td>FT</td>
</tr>
<tr>
<td>One biological parent</td>
<td>Full time</td>
</tr>
<tr>
<td>PP</td>
<td>PT</td>
</tr>
<tr>
<td>One biological parent with her/his partner</td>
<td>Part time</td>
</tr>
<tr>
<td>A</td>
<td>NB</td>
</tr>
<tr>
<td>Alone</td>
<td>Neg. break</td>
</tr>
<tr>
<td>U</td>
<td>PB</td>
</tr>
<tr>
<td>With partner</td>
<td>Pos. break</td>
</tr>
<tr>
<td>UC</td>
<td>AH</td>
</tr>
<tr>
<td>Partner and biological child</td>
<td>At home</td>
</tr>
<tr>
<td>UN</td>
<td>RE</td>
</tr>
<tr>
<td>Partner and non biological child</td>
<td>Retired</td>
</tr>
<tr>
<td>C</td>
<td>ED</td>
</tr>
<tr>
<td>Biological child and no partner</td>
<td>Education</td>
</tr>
<tr>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Friends</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
</tbody>
</table>
Events associated to cohabitational state transitions

For cohabitational trajectories, we convert states to events by defining the events associated to the state transitions:

<table>
<thead>
<tr>
<th></th>
<th>2P</th>
<th>1P</th>
<th>PP</th>
<th>A</th>
<th>U</th>
<th>UC</th>
<th>UN</th>
<th>C</th>
<th>F</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>2P</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;LH,A&quot;</td>
<td>&quot;LH,U&quot;</td>
<td>&quot;LH,U,C&quot;</td>
<td>&quot;LH,U,C&quot;</td>
<td>&quot;LH,C&quot;</td>
<td>&quot;LH,A&quot;</td>
<td>&quot;LH,O&quot;</td>
</tr>
<tr>
<td>1P</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;LH,A&quot;</td>
<td>&quot;LH,U&quot;</td>
<td>&quot;LH,U,C&quot;</td>
<td>&quot;LH,U,C&quot;</td>
<td>&quot;LH,C&quot;</td>
<td>&quot;LH,A&quot;</td>
<td>&quot;LH,O&quot;</td>
</tr>
<tr>
<td>PP</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;LH,A&quot;</td>
<td>&quot;LH,U&quot;</td>
<td>&quot;LH,U,C&quot;</td>
<td>&quot;LH,U,C&quot;</td>
<td>&quot;LH,C&quot;</td>
<td>&quot;LH,A&quot;</td>
<td>&quot;LH,O&quot;</td>
</tr>
<tr>
<td>A</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;A&quot;</td>
<td>&quot;U&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;F&quot;</td>
<td>&quot;O&quot;</td>
</tr>
<tr>
<td>U</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;UE,A&quot;</td>
<td>&quot;U&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;UE,A&quot;</td>
<td>&quot;UE,O&quot;</td>
</tr>
<tr>
<td>UC</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;UE,CL,A&quot;</td>
<td>&quot;CL&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;CL,C&quot;</td>
<td>&quot;UE&quot;</td>
<td>&quot;UE,CL,A&quot;</td>
<td>&quot;UE,CL,O&quot;</td>
</tr>
<tr>
<td>UN</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;UE,CL,A&quot;</td>
<td>&quot;CL&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;UE,C&quot;</td>
<td>&quot;UE,CL,A&quot;</td>
<td>&quot;UE,CL,O&quot;</td>
</tr>
<tr>
<td>C</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;CL,A&quot;</td>
<td>&quot;CL,U&quot;</td>
<td>&quot;U&quot;</td>
<td>&quot;CL,C&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;CL,A&quot;</td>
<td>&quot;CL,O&quot;</td>
</tr>
<tr>
<td>F</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;&quot;</td>
<td>&quot;U&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;A&quot;</td>
<td>&quot;O&quot;</td>
</tr>
<tr>
<td>O</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;A&quot;</td>
<td>&quot;U&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;A&quot;</td>
<td>&quot;O&quot;</td>
</tr>
</tbody>
</table>

For occupational trajectories, we assign an event to the start of each spell in a state.
Events associated to cohabitational state transitions

- For cohabitational trajectories, we convert states to events by defining the events associated to the state transitions:

<table>
<thead>
<tr>
<th>2P</th>
<th>1P</th>
<th>PP</th>
<th>A</th>
<th>U</th>
<th>UC</th>
<th>UN</th>
<th>C</th>
<th>F</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>2P</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;LH,A&quot;</td>
<td>&quot;LH,U&quot;</td>
<td>&quot;LH,U,C&quot;</td>
<td>&quot;LH,U,C&quot;</td>
<td>&quot;LH,C&quot;</td>
<td>&quot;LH,A&quot;</td>
</tr>
<tr>
<td>1P</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;LH,A&quot;</td>
<td>&quot;LH,U&quot;</td>
<td>&quot;LH,U,C&quot;</td>
<td>&quot;LH,U,C&quot;</td>
<td>&quot;LH,C&quot;</td>
<td>&quot;LH,A&quot;</td>
</tr>
<tr>
<td>PP</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;LH,A&quot;</td>
<td>&quot;LH,U&quot;</td>
<td>&quot;LH,U,C&quot;</td>
<td>&quot;LH,U,C&quot;</td>
<td>&quot;LH,C&quot;</td>
<td>&quot;LH,A&quot;</td>
</tr>
<tr>
<td>A</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;A&quot;</td>
<td>&quot;U&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;&quot;</td>
</tr>
<tr>
<td>U</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;UE,A&quot;</td>
<td>&quot;U&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;UE,A&quot;</td>
</tr>
<tr>
<td>UC</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;UE,CL,A&quot;</td>
<td>&quot;CL&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;CL,C&quot;</td>
<td>&quot;UE&quot;</td>
<td>&quot;UE,CL,A&quot;</td>
</tr>
<tr>
<td>UN</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;UE,CL,A&quot;</td>
<td>&quot;CL&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;UE,C&quot;</td>
<td>&quot;UE,CL,A&quot;</td>
</tr>
<tr>
<td>C</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;CL,A&quot;</td>
<td>&quot;CL,U&quot;</td>
<td>&quot;U&quot;</td>
<td>&quot;CL,C&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;CL,A&quot;</td>
</tr>
<tr>
<td>F</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;&quot;</td>
<td>&quot;U&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;A&quot;</td>
</tr>
<tr>
<td>O</td>
<td>&quot;2P&quot;</td>
<td>&quot;1P&quot;</td>
<td>&quot;PP&quot;</td>
<td>&quot;A&quot;</td>
<td>&quot;U&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;U,C&quot;</td>
<td>&quot;C&quot;</td>
<td>&quot;A&quot;</td>
</tr>
</tbody>
</table>

- For occupational trajectories, we assign an event to the start of each spell in a state.
Problem in Event Sequence Mining

Introduction

The Biographical Data from the Swiss Household Panel

Rendering cohabitational event sequences
(Bürgin and Ritschard, 2014)
Problem in Event Sequence Mining

Introduction

The Biographical Data from the Swiss Household Panel

Rendering occupational event sequences
(Bürgin and Ritschard, 2014)
Introduction

- Objectives
- The Biographical Data from the Swiss Household Panel
- Frequent subsequences versus Frequent itemsets
Mining of frequent itemsets and association rules has been popularized in the 90’s with the work of Agrawal and Srikant (1994); Agrawal et al. (1995) and their Apriori algorithm.

- Find out items that customers often buy together
- Symptoms that often occur together before a failure
Interest on sequences for accounting for the time order of the buys or symptoms

Mining typical event sequences is a specialized case of the mining of frequent itemsets

- More complicated however
- Must specify a counting method: How should we count multiple occurrences of a subsequence in a same sequence?
- Which time span should be covered? Maximal gap between two events? ...

Best known algorithms by Bettini et al. (1996), Srikant and Agrawal (1996), Mannila et al. (1997) and Zaki (2001).

Algorithm in TraMineR is adaptation of the tree search described in Masseglia (2002).
Problem in Event Sequence Mining

Introduction

Frequent subsequences versus Frequent itemsets

Frequent subsequences versus Frequent itemsets - 2

- Interest on sequences for accounting for the time order of the buys or symptoms
- Mining typical event sequences is a specialized case of the mining of frequent itemsets
  - More complicated however
  - Must specify a counting method: How should we count multiple occurrences of a subsequence in a same sequence?
  - Which time span should be covered? Maximal gap between two events? ...

- Best known algorithms by Bettini et al. (1996), Srikant and Agrawal (1996), Mannila et al. (1997) and Zaki (2001).

- Algorithm in TraMineR is adaptation of the tree search described in Masseglia (2002).
1 Introduction

2 Frequent subsequences in TraMineR

3 Frequent Swiss life course subsequences

4 Discriminant subsequences

5 Maximal subsequences

6 Association rules

7 Conclusion
2 Frequent subsequences in TraMineR

- Terminology
Events and transitions

- **Event sequence**: ordered list of transitions.
- **Transition** (transaction): a set of non ordered events.

**Example**

(LHome, Union) → (Marriage) → (Childbirth)

- (LHome, Union) and (Marriage) are transitions.
- “LHome”, “Union” et “Marriage” are events.
Events and transitions

- **Event sequence**: ordered list of transitions.
- **Transition** (transaction): a set of non-ordered events.

Example

\[(\text{LHome, Union}) \rightarrow (\text{Marriage}) \rightarrow (\text{Childbirth})\]

- (LHome, Union) and (Marriage) are transitions.
- “LHome”, “Union” et “Marriage” are events.
Subsequence

- A subsequence $B$ of a sequence $A$ is an event sequence such that
  - each event of $B$ is an event of $A$,
  - events of $B$ are in same order as in $A$.

Example

$A$ (LHome, Union) $\rightarrow$ (Marriage) $\rightarrow$ (Childbirth).
$B$ (LHome, Marriage) $\rightarrow$ (Childbirth).
$C$ (LHome) $\rightarrow$ (Childbirth).

- $C$ is a subsequence of $A$ and $B$, since order of events is respected.
- $B$ is not a subsequence of $A$, since we don’t know in $B$ whether “LHome” occurs before “Marriage”.
Subsequence

- A subsequence $B$ of a sequence $A$ is an event sequence such that
  - each event of $B$ is an event of $A$,
  - events of $B$ are in same order as in $A$.

Example

$A$ (LHome, Union) $\rightarrow$ (Marriage) $\rightarrow$ (Childbirth).

$B$ (LHome, Marriage) $\rightarrow$ (Childbirth).

$C$ (LHome) $\rightarrow$ (Childbirth).

- $C$ is a subsequence of $A$ and $B$, since order of events is respected.
- $B$ is not a subsequence of $A$, since we don’t know in $B$ whether “LHome” occurs before “Marriage”.
Frequent and discriminant subsequences

- **Support of a subsequence**: number of sequences that contain the subsequence.
  - Frequent subsequence: sequence with support greater than a minimal support.
  - A subsequence is **discriminant** between groups when its support varies significantly across groups.
Frequent and discriminant subsequences

- **Support of a subsequence**: number of sequences that contain the subsequence.
  - **Frequent** subsequence: sequence with support greater than a minimal support.
  - A subsequence is **discriminant** between groups when its support varies significantly across groups.
Frequent and discriminant subsequences

- **Support of a subsequence**: number of sequences that contain the subsequence.
  - **Frequent** subsequence: sequence with support greater than a minimal support.
  - A subsequence is **discriminant** between groups when its support varies significantly across groups.
1 Introduction

2 Frequent subsequences in TraMineR

3 Frequent Swiss life course subsequences

4 Discriminant subsequences

5 Maximal subsequences

6 Association rules

7 Conclusion
Frequent cohabitational subsequences
10 most frequent subsequences, min support = 50

- With at least 2 events

Remember that we assigned the state at age 20 as start event

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Support</th>
<th>Count</th>
<th>Transitions</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ((2P) \rightarrow (LH))</td>
<td>0.621</td>
<td>934</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2 ((2P) \rightarrow (U))</td>
<td>0.582</td>
<td>874</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3 ((2P) \rightarrow (C))</td>
<td>0.477</td>
<td>717</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4 ((LH,U))</td>
<td>0.454</td>
<td>682</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5 ((U) \rightarrow (C))</td>
<td>0.429</td>
<td>645</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6 ((2P) \rightarrow (LH,U))</td>
<td>0.392</td>
<td>589</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>7 ((LH) \rightarrow (C))</td>
<td>0.382</td>
<td>574</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8 ((A) \rightarrow (U))</td>
<td>0.376</td>
<td>565</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9 ((2P) \rightarrow (LH) \rightarrow (C))</td>
<td>0.325</td>
<td>489</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>10 ((C,U))</td>
<td>0.291</td>
<td>437</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Frequent occupational subsequences
Most frequent subsequences, min support = 50

- With at least 2 events

Remember that we assigned the state at age 20 as start event

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Support</th>
<th>Count</th>
<th>#Transitions</th>
<th>#Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (ED) → (FT)</td>
<td>0.283</td>
<td>425</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2 (FT) → (AH)</td>
<td>0.265</td>
<td>398</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3 (FT) → (PT)</td>
<td>0.219</td>
<td>329</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4 (AH) → (PT)</td>
<td>0.130</td>
<td>195</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5 (ED) → (AH)</td>
<td>0.113</td>
<td>170</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6 (ED) → (PT)</td>
<td>0.112</td>
<td>168</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>7 (FT) → (FT)</td>
<td>0.112</td>
<td>168</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8 (FT) → (AH) → (PT)</td>
<td>0.105</td>
<td>158</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>9 (FT) → (ED)</td>
<td>0.073</td>
<td>109</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10 (ED) → (FT) → (PT)</td>
<td>0.071</td>
<td>107</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Here we have cohabitational and occupational trajectories

Merging the two series of time stamped events

we get mixed cohabitational/occupational event sequences
## Merged cohabitational and occupational sequences

12 most frequent subsequences, min support 150

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Support</th>
<th>Count</th>
<th>#Transitions</th>
<th>#Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (FT) → (U)</td>
<td>0.695</td>
<td>1045</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2 (2P) → (LH)</td>
<td>0.621</td>
<td>934</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3 (FT) → (C)</td>
<td>0.583</td>
<td>876</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4 (2P) → (U)</td>
<td>0.582</td>
<td>874</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5 (FT) → (LH)</td>
<td>0.555</td>
<td>834</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6 (2P) → (C)</td>
<td>0.477</td>
<td>717</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>7 (LH,U)</td>
<td>0.454</td>
<td>682</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8 (U) → (C)</td>
<td>0.429</td>
<td>645</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9 (2P) → (LH,U)</td>
<td>0.392</td>
<td>589</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>10 (LH) → (C)</td>
<td>0.382</td>
<td>574</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>11 (2P,FT)</td>
<td>0.378</td>
<td>568</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>12 (A) → (U)</td>
<td>0.376</td>
<td>565</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Problem in Event Sequence Mining
Frequent Swiss life course subsequences

Interesting frequent subsequences

- **To get interesting knowledge we need to compare**
  - most frequent subsequences
  - with longer less frequent subsequences in which they are included.

- **For example,**

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Support</th>
<th>Count</th>
<th># Transitions</th>
<th># Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (2P) → (LH)</td>
<td>0.621</td>
<td>934</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4 (2P) → (U)</td>
<td>0.582</td>
<td>874</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9 (2P) → (LH,U)</td>
<td>0.392</td>
<td>589</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

- **Here, we know that**
  - among the 62.1% who left home (LH) after living with both parents (2P) when 20 years old
  - 39.2/62.1 = 63% left home to start a union the same year
Problem in Event Sequence Mining
Frequent Swiss life course subsequences

Interesting frequent subsequences

- To get interesting knowledge we need to compare
  - most frequent subsequences
  - with longer less frequent subsequences in which they are included.

- For example,

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Support</th>
<th>Count</th>
<th>#Transitions</th>
<th>#Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (2P) → (LH)</td>
<td>0.621</td>
<td>934</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4 (2P) → (U)</td>
<td>0.582</td>
<td>874</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9 (2P) → (LH,U)</td>
<td>0.392</td>
<td>589</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

- Here, we know that
  - among the 62.1% who left home (LH) after living with both parents (2P) when 20 years old
  - 39.2/62.1 = 63% left home to start a union the same year
Interesting frequent subsequences

To get interesting knowledge we need to compare
- most frequent subsequences
- with longer less frequent subsequences in which they are included.

For example,

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Support</th>
<th>Count</th>
<th>#Transitions</th>
<th>#Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>2  (2P) → (LH)</td>
<td>0.621</td>
<td>934</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4  (2P) → (U)</td>
<td>0.582</td>
<td>874</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9  (2P) → (LH,U)</td>
<td>0.392</td>
<td>589</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Here, we know that
- among the 62.1% who left home (LH) after living with both parents (2P) when 20 years old
- 39.2/62.1 = 63% left home to start a union the same year
Interesting frequent subsequences

- To get interesting knowledge we need to compare
  - most frequent subsequences
  - with longer less frequent subsequences in which they are included.

- For example,

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Support</th>
<th>Count</th>
<th>#Transitions</th>
<th>#Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (2P) → (LH)</td>
<td>0.621</td>
<td>934</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4 (2P) → (U)</td>
<td>0.582</td>
<td>874</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9 (2P) → (LH,U)</td>
<td>0.392</td>
<td>589</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

- Here, we know that
  - among the 62.1% who left home (LH) after living with both parents (2P) when 20 years old
  - 39.2/62.1 = 63% left home to start a union the same year
Problem in Event Sequence Mining
Frequent Swiss life course subsequences

Interesting frequent subsequences

- To get interesting knowledge we need to compare
  - most frequent subsequences
  - with longer less frequent subsequences in which they are included.

For example,

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Support</th>
<th>Count</th>
<th>#Transitions</th>
<th>#Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (2P) → (LH)</td>
<td>0.621</td>
<td>934</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4 (2P) → (U)</td>
<td>0.582</td>
<td>874</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9 (2P) → (LH,U)</td>
<td>0.392</td>
<td>589</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Here, we know that
- among the 62.1% who left home (LH) after living with both parents (2P) when 20 years old
- 39.2/62.1 = 63% left home to start a union the same year
Problem in Event Sequence Mining
Frequent Swiss life course subsequences

Interesting frequent subsequences

- To get interesting knowledge we need to compare
  - most frequent subsequences
  - with longer less frequent subsequences in which they are included.

For example,

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Support</th>
<th>Count</th>
<th>#Transitions</th>
<th>#Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (2P) → (LH)</td>
<td>0.621</td>
<td>934</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4 (2P) → (U)</td>
<td>0.582</td>
<td>874</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9 (2P) → (LH,U)</td>
<td>0.392</td>
<td>589</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Here, we know that
- among the 62.1% who left home (LH) after living with both parents (2P) when 20 years old
- 39.2/62.1 = 63% left home to start a union the same year
Interesting frequent subsequences

- To get interesting knowledge we need to compare
  - most frequent subsequences
  - with longer less frequent subsequences in which they are included.

- For example,

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Support</th>
<th>Count</th>
<th>#Transitions</th>
<th>#Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (2P) → (LH)</td>
<td>0.621</td>
<td>934</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4 (2P) → (U)</td>
<td>0.582</td>
<td>874</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9 (2P) → (LH,U)</td>
<td>0.392</td>
<td>589</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

- Here, we know that
  - among the 62.1% who left home (LH) after living with both parents (2P) when 20 years old
  - $39.2/62.1 = 63\%$ left home to start a union the same year
1 Introduction

2 Frequent subsequences in TraMineR

3 Frequent Swiss life course subsequences

4 Discriminant subsequences

5 Maximal subsequences

6 Association rules

7 Conclusion
Discriminant subsequences

- Differentiating between sexes
- Differentiating among birth cohorts
Problem in Event Sequence Mining

Discriminant subsequences

Differentiating between sexes

Mixed events: Subsequences that best discriminate sex

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Chi-2</th>
<th>Support</th>
<th>Freq. Men</th>
<th>Freq. Women</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (FT) → (AH)</td>
<td>322.7</td>
<td>0.26</td>
<td>0.060</td>
<td>0.470</td>
<td>-0.410</td>
</tr>
<tr>
<td>2 (AH)</td>
<td>317.5</td>
<td>0.41</td>
<td>0.181</td>
<td>0.634</td>
<td>-0.453</td>
</tr>
<tr>
<td>3 (PT)</td>
<td>269.7</td>
<td>0.28</td>
<td>0.088</td>
<td>0.469</td>
<td>-0.381</td>
</tr>
<tr>
<td>4 (U) → (PT)</td>
<td>260.4</td>
<td>0.20</td>
<td>0.036</td>
<td>0.373</td>
<td>-0.337</td>
</tr>
<tr>
<td>5 (FT) → (PT)</td>
<td>247.5</td>
<td>0.22</td>
<td>0.051</td>
<td>0.387</td>
<td>-0.337</td>
</tr>
<tr>
<td>6 (FT) → (U) → (AH)</td>
<td>228.2</td>
<td>0.16</td>
<td>0.016</td>
<td>0.302</td>
<td>-0.286</td>
</tr>
<tr>
<td>7 (U) → (AH)</td>
<td>226.0</td>
<td>0.20</td>
<td>0.041</td>
<td>0.350</td>
<td>-0.309</td>
</tr>
<tr>
<td>8 (AH) → (PT)</td>
<td>195.5</td>
<td>0.13</td>
<td>0.008</td>
<td>0.252</td>
<td>-0.244</td>
</tr>
<tr>
<td>9 (C) → (PT)</td>
<td>193.3</td>
<td>0.15</td>
<td>0.019</td>
<td>0.273</td>
<td>-0.254</td>
</tr>
<tr>
<td>10 (FT) → (U) → (PT)</td>
<td>192.7</td>
<td>0.16</td>
<td>0.027</td>
<td>0.289</td>
<td>-0.262</td>
</tr>
</tbody>
</table>

- Mainly occupational events (FT, PT and AH)
- In conjunction with a few cohabitational ones (U and C)
Mixed events: Subsequences that best discriminate sex at the 0.1% level

Color by sign and significance of Pearson’s residual

- Red: Negative 0.01
- Orange: Negative 0.05
- Neutral
- Light Blue: Positive 0.05
- Dark Blue: Positive 0.01
Discriminant subsequences

- Differentiating between sexes
- Differentiating among birth cohorts
Problem in Event Sequence Mining

Discriminant subsequences

Differentiating among birth cohorts

Birth cohort distribution

Mixed events: Subsequences that best discriminate birth cohorts

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Chi-2</th>
<th>Support</th>
<th>1910-25</th>
<th>1926-45</th>
<th>1946-57</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (PT)</td>
<td>64.5</td>
<td>0.28</td>
<td>0.042</td>
<td>0.205</td>
<td>0.362</td>
</tr>
<tr>
<td>2 (U) → (PT)</td>
<td>63.0</td>
<td>0.20</td>
<td>0.014</td>
<td>0.135</td>
<td>0.281</td>
</tr>
<tr>
<td>3 (FT) → (PT)</td>
<td>56.1</td>
<td>0.22</td>
<td>0.014</td>
<td>0.156</td>
<td>0.291</td>
</tr>
<tr>
<td>4 (A) → (PT)</td>
<td>46.3</td>
<td>0.11</td>
<td>0.028</td>
<td>0.055</td>
<td>0.160</td>
</tr>
<tr>
<td>5 (FT) → (U) → (PT)</td>
<td>38.5</td>
<td>0.16</td>
<td>0.000</td>
<td>0.114</td>
<td>0.210</td>
</tr>
<tr>
<td>6 (ED) → (PT)</td>
<td>36.8</td>
<td>0.11</td>
<td>0.028</td>
<td>0.065</td>
<td>0.159</td>
</tr>
<tr>
<td>7 (LH) → (PT)</td>
<td>35.9</td>
<td>0.15</td>
<td>0.014</td>
<td>0.109</td>
<td>0.204</td>
</tr>
<tr>
<td>8 (U) → (C)</td>
<td>34.2</td>
<td>0.43</td>
<td>0.239</td>
<td>0.370</td>
<td>0.497</td>
</tr>
<tr>
<td>9 (C) → (PT)</td>
<td>34.0</td>
<td>0.15</td>
<td>0.014</td>
<td>0.103</td>
<td>0.194</td>
</tr>
<tr>
<td>10 (2P) → (PT)</td>
<td>32.7</td>
<td>0.17</td>
<td>0.014</td>
<td>0.126</td>
<td>0.215</td>
</tr>
</tbody>
</table>

- Mainly emergence of Part-time (PT)
Problem in Event Sequence Mining

Discriminant subsequences

Differentiating among birth cohorts

Mixed events: Subsequences that best discriminate birth cohorts
Problem in Event Sequence Mining
Maximal subsequences

1. Introduction
2. Frequent subsequences in TraMineR
3. Frequent Swiss life course subsequences
4. Discriminant subsequences
5. Maximal subsequences
6. Association rules
7. Conclusion
Too many frequent subsequences

- There are often too many frequent subsequences!
- How can we structure those subsequences?
  - Eliminate redundant subsequences: When you experience one subsequence you also experience all its subsequences.
    - Count only maximal subsequences
    - If subsequence \((FT) \rightarrow (AH) \rightarrow (PT)\) is observed,
      - we would not count the occurrence of
        \((FT) \rightarrow (AH), (FT) \rightarrow (PT)\) or \((AH) \rightarrow (PT)\)
Frequent maximal subsequence: Definition

A subsequence is frequent maximal if frequent when in each sequence we count only those subsequences that are not themselves a subsequence of another frequent subsequence present in the same sequence.

Example: The subsequence \((2P) \rightarrow (LH)\) will be considered a maximal subsequence of sequences which do not also have a frequent supersequence such as \((2P) \rightarrow (LH,U)\).
Maximal frequent sequence in pattern mining

- Our definition of a frequent maximal subsequence differs from the notion of maximal frequent sequence used in pattern mining, where a frequent sequence is said maximal if none of its supersequence is frequent.
- In pattern mining, if $s$ is a maximal frequent sequence, then none of its subsequences is a maximal frequent subsequence, even if it occurs frequently in sequences which do not include $s$.
  - e.g., if $(U) \rightarrow (C)$ is frequent, then $(U)$ would not be considered.
- This is not very useful for life trajectories where we may be interested to know that
  - It is frequent to start a union $(U)$ without having a child afterwards $(U) \rightarrow (C)$.
Maximal frequent sequence in pattern mining

- Our definition of a frequent maximal subsequence differs from the notion of maximal frequent sequence used in pattern mining, where a frequent sequence is said maximal if none of its supersequence is frequent.

- In pattern mining, if \( s \) is a maximal frequent sequence, then none of its subsequences is a maximal frequent subsequence, even if it occurs frequently in sequences which do not include \( s \).
  
  - e.g., if \( (U) \rightarrow (C) \) is frequent, then \( (U) \) would not be considered.

- This is not very useful for life trajectories where we may be interested to know that
  
  - It is frequent to start a union \( (U) \) without having a child afterwards \( (U) \rightarrow (C) \).
Our definition of a frequent maximal subsequence differs from the notion of maximal frequent sequence used in pattern mining, where a frequent sequence is said maximal if none of its supersequence is frequent.

In pattern mining, if \( s \) is a maximal frequent sequence, then none of its subsequences is a maximal frequent subsequence, even if it occurs frequently in sequences which do not include \( s \).

- e.g., if \((U) \rightarrow (C)\) is frequent, then \((U)\) would not be considered.

This is not very useful for life trajectories where we may be interested to know that

- It is frequent to start a union \((U)\) without having a child afterwards \((U) \rightarrow (C)\)
Maximal frequent sequence in pattern mining

- Our definition of a frequent maximal subsequence differs from the notion of maximal frequent sequence used in pattern mining, where a frequent sequence is said maximal if none of its supersequence is frequent.

- In pattern mining, if \( s \) is a maximal frequent sequence, then none of its subsequences is a maximal frequent subsequence, even if it occurs frequently in sequences which do not include \( s \).

  - e.g., if \((U) \rightarrow (C)\) is frequent, then \((U)\) would not be considered.

- This is not very useful for life trajectories where we may be interested to know that

  - It is frequent to start a union \((U)\) without having a child afterwards \((U) \rightarrow (C)\)
Maximal frequent sequence in pattern mining

- Our definition of a frequent maximal subsequence differs from the notion of maximal frequent sequence used in pattern mining, where a frequent sequence is said maximal if none of its supersequence is frequent.
- In pattern mining, if $s$ is a maximal frequent sequence, then none of its subsequences is a maximal frequent subsequence, even if it occurs frequently in sequences which do not include $s$.
  - e.g., if $(U) \rightarrow (C)$ is frequent, then $(U)$ would not be considered.
- This is not very useful for life trajectories where we may be interested to know that
  - It is frequent to start a union $(U)$ without having a child afterwards $(U) \rightarrow (C)$. 
Frequent maximal subsequences: algorithm

1. Find frequent subsequences for the selected support
2. Starting from the longest obtained frequent subsequence
   - Adjust the count of each of its subsequence (by reducing their counts by the number of occurrences of the considered frequent sequence).
   - Delete from the list subsequences with counts falling below the support threshold.
3. Iterate on frequent subsequences ordered in decreasing order of length (using their already adjusted counts)
## Max subsequences, cohabitational-occupational events

12 most frequent maximal subsequences, min support 150

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Support</th>
<th>Count</th>
<th>#Transitions</th>
<th>#Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1   (2P) → (C,LH,U)</td>
<td>0.160</td>
<td>241</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2   (FT) → (U) → (AH)</td>
<td>0.159</td>
<td>239</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3   (FT) → (U) → (PT)</td>
<td>0.158</td>
<td>237</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4   (FT) → (A,LH) → (U)</td>
<td>0.152</td>
<td>228</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5   (2P,ED) → (FT) → (U)</td>
<td>0.140</td>
<td>210</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>6   (FT) → (C,LH,U)</td>
<td>0.140</td>
<td>210</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>7   (AH) → (C)</td>
<td>0.137</td>
<td>206</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8   (2P) → (LH) → (AH)</td>
<td>0.133</td>
<td>200</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>9   (AH) → (U)</td>
<td>0.130</td>
<td>195</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10  (2P,FT) → (LH,U)</td>
<td>0.129</td>
<td>194</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>11  (2P) → (LH) → (PT)</td>
<td>0.128</td>
<td>193</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>12  (2P,FT) → (AH)</td>
<td>0.126</td>
<td>190</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Max subsequences, cohabitational-occupational events
12 most frequent maximal subsequences, min support 200

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Support</th>
<th>Count</th>
<th>#Transitions</th>
<th>#Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  (2P,FT) → (LH,U)</td>
<td>0.229</td>
<td>344</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2  (A) → (U) → (C)</td>
<td>0.194</td>
<td>291</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3  (2P,ED) → (LH)</td>
<td>0.189</td>
<td>284</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4  (ED) → (FT) → (C)</td>
<td>0.189</td>
<td>284</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5  (2P) → (A,LH) → (U)</td>
<td>0.181</td>
<td>272</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>6  (2P,FT) → (LH) → (C)</td>
<td>0.178</td>
<td>268</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>7  (2P) → (LH,U) → (C)</td>
<td>0.168</td>
<td>253</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>8  (2P) → (PT)</td>
<td>0.166</td>
<td>250</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9  (FT) → (LH,U) → (C)</td>
<td>0.166</td>
<td>250</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>10 (2P) → (C,LH,U)</td>
<td>0.160</td>
<td>241</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>11 (FT) → (U) → (AH)</td>
<td>0.159</td>
<td>239</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>12 (FT) → (U) → (PT)</td>
<td>0.158</td>
<td>237</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Solutions change with chosen support

- As seen, solutions vary with chosen minsupport
- For minsupport = 0, we get the set of complete event sequences.
- We are working on criteria to select an optimal minsupport
  - to minimize the number of subsequences with no representative
  - maximize the average number of representatives
  - ...

Problem in Event Sequence Mining
Maximal subsequences
Frequent max-subsequences discriminating birth cohorts
Minsupport=150

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Chi-2</th>
<th>Support</th>
<th>1910-25</th>
<th>1926-45</th>
<th>1946-57</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (A) → (PT)</td>
<td>46.3</td>
<td>0.11</td>
<td>0.028</td>
<td>0.055</td>
<td>0.160</td>
</tr>
<tr>
<td>2 (FT) → (U) → (PT)</td>
<td>38.5</td>
<td>0.16</td>
<td>0.000</td>
<td>0.114</td>
<td>0.210</td>
</tr>
<tr>
<td>3 (ED) → (PT)</td>
<td>36.8</td>
<td>0.11</td>
<td>0.028</td>
<td>0.065</td>
<td>0.159</td>
</tr>
<tr>
<td>4 (2P) → (LH) → (PT)</td>
<td>30.4</td>
<td>0.13</td>
<td>0.014</td>
<td>0.090</td>
<td>0.172</td>
</tr>
<tr>
<td>5 (2P) → (U) → (PT)</td>
<td>27.0</td>
<td>0.12</td>
<td>0.000</td>
<td>0.083</td>
<td>0.154</td>
</tr>
<tr>
<td>6 (2P) → (C,LH,U)</td>
<td>26.2</td>
<td>0.16</td>
<td>0.268</td>
<td>0.202</td>
<td>0.115</td>
</tr>
<tr>
<td>7 (U) → (UE)</td>
<td>26.1</td>
<td>0.12</td>
<td>0.028</td>
<td>0.082</td>
<td>0.159</td>
</tr>
<tr>
<td>8 (FT) → (UE)</td>
<td>22.9</td>
<td>0.10</td>
<td>0.028</td>
<td>0.068</td>
<td>0.137</td>
</tr>
<tr>
<td>9 (FT) → (C) → (PT)</td>
<td>22.8</td>
<td>0.12</td>
<td>0.000</td>
<td>0.094</td>
<td>0.155</td>
</tr>
<tr>
<td>10 (FT) → (C,LH,U)</td>
<td>21.8</td>
<td>0.14</td>
<td>0.155</td>
<td>0.185</td>
<td>0.100</td>
</tr>
</tbody>
</table>
Frequent max-subsequences discriminating between cohorts

Color by sign and significance of Pearson's residual

- Negative 0.01
- Negative 0.05
- Neutral
- Positive 0.05
- Positive 0.01
1 Introduction

2 Frequent subsequences in TraMineR

3 Frequent Swiss life course subsequences

4 Discriminant subsequences

5 Maximal subsequences

6 Association rules

7 Conclusion
Sequential association rules

A rule $\text{subseq}_1 \rightarrow \text{subseq}_2$ such that

1. Has a minimal support
2. When $\text{subseq}_1$ occurs, it is most often followed by $\text{subseq}_2$

- Extracted from frequent sequences.
- Extraction criteria:
  - Confidence: $p(\text{subseq}_2 \mid \text{subseq}_1)$
  - Lift: $\frac{p(\text{subseq}_2 \mid \text{subseq}_1)}{p(\text{subseq}_2)}$
  - ...
Sequential association rules

A rule $\text{subseq}_1 \rightarrow \text{subseq}_2$ such that

1. Has a minimal support
2. When $\text{subseq}_1$ occurs, it is most often followed by $\text{subseq}_2$

- Extracted from frequent sequences.
- Extraction criteria:
  - Confidence: $p(\text{subseq}_2 \mid \text{subseq}_1)$
  - Lift: $\frac{p(\text{subseq}_2 \mid \text{subseq}_1)}{p(\text{subseq}_2)}$
  - ...
Extracting association rules

- From the mined frequent subsequences, we can extract association rules:

<table>
<thead>
<tr>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2P,ED) =&gt; (LH)-(C)</td>
<td>167</td>
<td>0.5719</td>
<td>1.498</td>
</tr>
<tr>
<td>(FT)-(AH) =&gt; (PT)</td>
<td>158</td>
<td>0.3970</td>
<td>1.427</td>
</tr>
<tr>
<td>(2P,ED) =&gt; (LH)</td>
<td>284</td>
<td>0.9726</td>
<td>1.345</td>
</tr>
<tr>
<td>(2P) =&gt; (C,LH,U)</td>
<td>241</td>
<td>0.2349</td>
<td>1.342</td>
</tr>
<tr>
<td>(2P,FT) =&gt; (LH,U)</td>
<td>344</td>
<td>0.6056</td>
<td>1.335</td>
</tr>
<tr>
<td>(2P) =&gt; (C,LH)</td>
<td>246</td>
<td>0.2398</td>
<td>1.335</td>
</tr>
<tr>
<td>(1P) =&gt; (LH)</td>
<td>151</td>
<td>0.9557</td>
<td>1.321</td>
</tr>
<tr>
<td>(2P,FT) =&gt; (LH,U)-(C)</td>
<td>150</td>
<td>0.2641</td>
<td>1.306</td>
</tr>
<tr>
<td>(2P) =&gt; (A,LH)-(C)</td>
<td>212</td>
<td>0.2066</td>
<td>1.278</td>
</tr>
<tr>
<td>(2P,FT) =&gt; (LH)</td>
<td>523</td>
<td>0.9208</td>
<td>1.273</td>
</tr>
</tbody>
</table>
Issues with association rules

- Classical definition assume the left hand and the right hand subsequences are frequent.
- Which implication rule should be used?
  - There are over 50 interestingness criteria (Gras’ intensity of implication, ...)
- How can we get rules for rare events (or subsequences)?
  (This will be the topic of Rousseaux’s presentation)
1 Introduction

2 Frequent subsequences in TraMineR

3 Frequent Swiss life course subsequences

4 Discriminant subsequences

5 Maximal subsequences

6 Association rules

7 Conclusion
Conclusion

- Type of outcomes for event sequences
  - frequent episodes
  - discriminant episodes
  - association rules
  - cluster analysis (not addressed in this presentation)

- Complementary insights
  - most common characteristics
  - salient distinctions between groups
  - implication rules between common characteristics
  - identify types of trajectories

- Easy to extend to other types of analyses (representative sequences, discrepancy analyses, ...)

18/8/2014gr 52/57
Conclusion

- Type of outcomes for event sequences
  - frequent episodes
  - discriminant episodes
  - association rules
  - cluster analysis (not addressed in this presentation)

- Complementary insights
  - most common characteristics
  - salient distinctions between groups
  - implication rules between common characteristics
  - identify types of trajectories

- Easy to extend to other types of analyses (representative sequences, discrepancy analyses, ...)

18/8/2014gr 52/57
Conclusion

- Looking at frequent max-subsequences produces more directly interpretable results
- Issue: Solutions vary with the min-support threshold
Problem in Event Sequence Mining

Conclusion

Thank You!
References I


References II


References III


