Analysing the diversity of categorical sequences

Supervised and non-supervised analysis of the diversity of categorical sequences

Gilbert Ritschard
Department of Econometrics and Institute of Demography
University of Geneva
http://mephisto.unige.ch/traminer

Séminaire de recherche, Labo ERIC, Lyon, 9 décembre 2010

Outline

1. Objectives
2. Pairwise dissimilarities between sequences
3. Unsupervised clustering
4. Supervised Classification
5. Conclusion

Objectives

Aim of the presentation
Using Swiss cohabitational and occupational data from the SHP biographical survey
Demonstrate the results that can be extracted from state sequences
Unsupervised learning
Clustering
Representative sequences
Supervised learning
Discrepancy analysis (ANOVA)
Regression trees of state sequence data

State sequences: examples

- Cohabitation state sequences (from SHP)
  - 2P = with 2 parents, U = with partner, C = with child, A = alone, ...
  - Sequence
    2. 2P-2P-U-U-U-U-U-U-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC
    4. 2P-2P-2P-2P-2P-2P-U-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC-UC

- Compact representation
  - Sequence
    1. (2P,1) → (U,25)
    2. (2P,2) → (U,6) → (UC,18)
    3. (2P,1) → (A,5) → (U,2) → (UC,18)
    4. (2P,8) → (U,1) → (UC,17)
Typical questions in social sciences

- In the field of Life course analysis
  - Are there standard of lives, ideal-types?
  - What are those standards, those ideal-tyes?
  - How are those standards linked to covariates (sex, birth cohort, ...)?

- Can answer to these questions by seeking typologies and studying groupings in terms of covariates

Prerequisite: pairwise dissimilarities

- Common point between all addressed methods:
  - ... all are based on pairwise dissimilarities between sequences

Dissimilarity measure – 1

Based on count of matching attributes

- Measures Based on count of matching attributes $A(x, y)$ (proximity)

$$d(x, y) = A(x, x) + A(y, y) - 2A(x, y)$$

available in TraMineR

- LCP $A(x, y) = \text{length of longest common prefix}$
- RLCP $A(x, y) = \text{length of longest common suffix}$
- LCS $A(x, y) = \text{length of longest common subsequence}$
- HAM simple $A(x, y) = \text{half of number of matching elements}$

Dissimilarity measure – 2

Edit distances

- Edit distance: (minimal) cost of transforming $x$ into $y$
  available in R

  - OM Optimal matching of state séquences (Levenshtein, 1966)
    - indel cost (insertion/deletion)
    - pairwise substitution costs
  - Generalized HAM, Hamming = OM without indel
  - DHD, Dynamic Hamming Distance, position-varying substitution cost (Lesnard, 2006)
Discrepancy of a set of sequences

- From a dissimilarity matrix, we can define the discrepancy of a set of sequences.
- Sum of squares $SS$ can be expressed in terms of pairwise distances.

$$SC = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} (y_i - y_j)^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} d_{e,ij}^2$$

- Replacing $d^2_{e,ij}$ with the dissimilarity OM, LCP, LCS ... (or its square), we get a pseudo $SS$.

Dissimilarity based analysis

- If we know how to compute a dissimilarity, we can apply any analysis based on dissimilarities or variances.
  - Clustering (agglomerative, divisive, partitioning, ...) (Kaufman and Rousseeuw, 2005)
  - Principal coordinate analysis (PCO, MDS) (Gower, 1966)
  - Representative sequences (Gabadinho et al., 2009b)
  - ANOVA (Studer et al., 2010)
  - ...

Building a typology

- To illustrate, hierarchical clustering with Ward criterion.
- Data: Cohabitational trajectories,
  - 1503 sequences from the 2002 biographic survey of the SHP
  - alphabet of 10 states
  - yearly data, from age 20 to 45 years (length 26)

Clustering from dissimilarities

- Compute the dissimilarity matrix, for instance with `TraMineR`
  ```r
  om.coh <- seqdist(seqs.coh, method="OM", sm="TRATE", indel=1)
  om.dist.coh
  ```
  is a $1503 \times 1503$ matrix that can be passed to any clustering method that accepts a distance matrix as input.
- In R, we can use the `cluster` library (Maechler et al., 2005) which proposes among others
  - `agnes()` an agglomerative method
  - `diana()` a divisive method
  - `pam()` partitioning around medoids
- Illustration: agglomerative method with Ward
- We use the `agnes()` function
  ```r
  clw.coh <- agnes(om.coh, diss=T, method="ward")
  ```
  and retain the partition into 5 clusters
  ```r
  cutree(clw.coh, k=5)
  ```
Unsupervised clustering
Typology
Hierarchical clustering, Ward
Dendrogram

Cohabitation trajectories

Analysing the diversity of categorical sequences
Unsupervised clustering
Typology
i-plot, and order of the sequences

- When number of sequences is high, sorting sequences helps readability

Rendering the clusters

- State sequences can easily be visualized
- Example: i-plot
  Sequence 1 \((2P,1)\) - \((U,2)\)
  Sequence 2 \((2P,2)\) - \((U,6)\) - \((UC,18)\)
  Sequence 3 \((2P,1)\) - \((A,5)\) - \((U,2)\) - \((UC,18)\)
  Sequence 4 \((2P,8)\) - \((U,1)\) - \((UC,17)\)

- (Horizontal) stacking of rectangles, with color representing the state and length its duration.
- The vertical alignment informs about the distribution at each position

Typology of cohabitation state sequences
i-plot, sorted with MDS[1]
**Analyzing the diversity of categorical sequences**

**Unsupervised clustering**

**Typology**

**Typology of cohabitational state sequences**

**Discrepancies**

- Discrepancy (pseudo variance) $\sqrt{\frac{1}{n^2} \sum_i \sum_j d(i,j)}$

<table>
<thead>
<tr>
<th>Count</th>
<th>Percent</th>
<th>Discrepancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parental</td>
<td>938</td>
<td>62.4</td>
</tr>
<tr>
<td>Conjugal</td>
<td>261</td>
<td>17.4</td>
</tr>
<tr>
<td>Mixed</td>
<td>121</td>
<td>8.1</td>
</tr>
<tr>
<td>Parental Home</td>
<td>108</td>
<td>7.2</td>
</tr>
<tr>
<td>Solo</td>
<td>75</td>
<td>5.0</td>
</tr>
<tr>
<td>Total</td>
<td>1503</td>
<td>100.0</td>
</tr>
</tbody>
</table>

---

### Typology of cohabitational state sequences

**d-plot, transversal distributions**

**mt-plot, mean time in each state**

**Ht-plot, transversal entropies**

**Discrepancies**

**Conjugal**

- Type 1: Parental Trajectories (62%)
- Type 2: Conjugal Trajectories (17%)
- Type 3: Mixed Cohabitation Trajectories (8%)
- Type 4: Parental Home Trajectories (7%)
- Type 5: Solo Trajectories (5%)
Analysing the diversity of categorical sequences

Unsupervised clustering

Typology

**MDS: Cloud of points**

*Type 1: Full Time Trajectories (52%)*
*Freq. (n=797)*
*A20 A24 A28 A32 A36 A40 A44*
*0.0 0.2 0.4 0.6 0.8 1.0*

*Type 2: Mixed Occupational Trajectories (22%)*
*Freq. (n=91)*
*A20 A24 A28 A32 A36 A40 A44*
*0.0 0.2 0.4 0.6 0.8 1.0*

*Type 3: Mixed Cohabitation Trajectories (8%)*

*Type 4: At Home Trajectories (5.5%)*

*Type 5: Part Time Trajectories (5.5%)*

Aim: summarize a set of sequences

Find a small set of sequences, such that

- non redundant
- cover a minimal percentage of the set

Redundance and coverage defined in terms of neighborhood

- x and y non redundant if $d(x, y) > \delta_{tsim}$
- coverage: % of sequences that have at least one representative $r$ in their neighborhood ($d(x, r) < \delta_{tsim}$)
Sort sequences according to a representativeness criterion
- density number of sequences in its neighborhood
- centrality sum of distances to all other sequences
- others: frequency, mean of its state frequencies, likelihood, ...

Suppress redundancy
- Compute coverage of the sequence with highest score
- Then, for the next ones
  - drop out if redundant with sequences already retained
  - else, compute coverage of the new set of representatives
- Stop when the wanted coverage is reached.

Representative sequences

Heuristic

Cohabital trajectories (tsim=2, trep=6)

Occupational trajectories (tsim=1, trep=4)
Analysing the diversity of categorical sequences
Supervised Classification
ANOVA on sequential data

Association between sequences and a covariate

- For supervised clustering (according the value of a covariate)
- ... must be able to measure association between sequential data and a covariate
- Since we know how to determine the discrepancy
  \[ SS = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} d_{ij} \]
- We can compute pseudo $R^2$'s and pseudo $F$'s

\[ SS_T = SS_B + SS_W \]

**Analysis of sequence discrepancy**
(Studer et al., 2009, 2010)

- ANOVA like analysis based on pairwise dissimilarities
- We decompose the SS (Sum of squares equivalent)
  \[ SS_T = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} d_{ij} \]
  \[ SS_W = \sum_{g} \left( \frac{1}{n_g} \sum_{j=1}^{n_g} \sum_{i=j+1}^{n_g} d_{ij,g} \right) \]
  \[ SS_B = SS_T - SS_W \]

Pseudo R-square and ANOVA Table

<table>
<thead>
<tr>
<th>Discrepancy</th>
<th>df</th>
<th>Mean Discr.</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>$SS_B$</td>
<td>$df_B = m - 1$</td>
<td>$\frac{SS_B}{SS_T}$</td>
</tr>
<tr>
<td>Within</td>
<td>$SS_W$</td>
<td>$df_W = \sum n_g - m$</td>
<td>$\frac{SS_W}{SS_T}$</td>
</tr>
<tr>
<td>Total</td>
<td>$SS_T$</td>
<td>$df_T = n - 1$</td>
<td></td>
</tr>
</tbody>
</table>

Pseudo $R^2$

\[ R^2 = \frac{SS_B}{SS_T} \]

Pseudo $F$

\[ F = \frac{SS_B/(m - 1)}{SS_W/(n - m)} \]

- Normality is not defendable in this setting.
- $F$ cannot be compared with an $F$ distribution.
- The significance is assessed through a permutation test
- Permutation test: iteratively randomly reassign each covariate profile to one of the observed sequence and recompute the $F$.
- Empirical distribution of $F$ under independence.
**Supervised Classification**

**ANOVA on sequential data**

**Analysis of sequence discrepancy**

Cohabital trajectories with birth cohort

<table>
<thead>
<tr>
<th>SS</th>
<th>df</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp</td>
<td>186.63</td>
<td>2</td>
</tr>
<tr>
<td>Res</td>
<td>23149.63</td>
<td>1500</td>
</tr>
<tr>
<td>Total</td>
<td>23336.26</td>
<td>1502</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>t₀</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pseudo F</td>
<td>5.057</td>
</tr>
<tr>
<td>Pseudo Fbf</td>
<td>5.851</td>
</tr>
<tr>
<td>Pseudo R²</td>
<td>0.007</td>
</tr>
<tr>
<td>Bartlett</td>
<td>8.731</td>
</tr>
<tr>
<td>Levene</td>
<td>14.122</td>
</tr>
</tbody>
</table>

**Empirical F distribution**

Cohabital trajectories with birth cohort

Cohabital trajectories with month of birth

**Simple ANOVA, cohabital trajectories**

<table>
<thead>
<tr>
<th>categories</th>
<th>$R^2$</th>
<th>$F$</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth Cohort</td>
<td>0.008</td>
<td>6.0</td>
<td>0.001</td>
</tr>
<tr>
<td>Sex</td>
<td>0.004</td>
<td>6.6</td>
<td>0.001</td>
</tr>
<tr>
<td>Education Level</td>
<td>0.007</td>
<td>3.3</td>
<td>0.001</td>
</tr>
<tr>
<td>Birth Month</td>
<td>0.009</td>
<td>1.1</td>
<td>0.272</td>
</tr>
</tbody>
</table>

**Simple ANOVA, occupational trajectories**

<table>
<thead>
<tr>
<th>categories</th>
<th>$R^2$</th>
<th>$F$</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth Cohort</td>
<td>0.007</td>
<td>5.1</td>
<td>0.001</td>
</tr>
<tr>
<td>Sex</td>
<td>0.183</td>
<td>336.6</td>
<td>0.001</td>
</tr>
<tr>
<td>Education Level</td>
<td>0.065</td>
<td>34.8</td>
<td>0.001</td>
</tr>
</tbody>
</table>
Homogeneity of within group discrepancy

- Is discrepancy the same in all groups?
- Contribution to inertia $d_{g} = \frac{1}{n} \left( \sum_{i} d_{xi} - SS \right)$
- Letting $z_{i}$ be the dissimilarity between sequence $i$ and its group center
- Levene: $F$ test (ANOVA) on the $z_{i}$'s

Levene test, Cohabitational Trajectories

<table>
<thead>
<tr>
<th>categories</th>
<th>$L$</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth Cohort</td>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>Sex</td>
<td>2</td>
<td>9.2</td>
</tr>
<tr>
<td>Education Level</td>
<td>4</td>
<td>2.4</td>
</tr>
<tr>
<td>Birth Month</td>
<td>12</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Levene test, Occupational Trajectories

<table>
<thead>
<tr>
<th>categories</th>
<th>$L$</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth Cohort</td>
<td>3</td>
<td>0.01</td>
</tr>
<tr>
<td>Sex</td>
<td>2</td>
<td>912.8</td>
</tr>
<tr>
<td>Education Level</td>
<td>4</td>
<td>15.2</td>
</tr>
</tbody>
</table>

Multi-factor ANOVA

- Generalization to multi-factor case voir (Studer et al., 2010)
- Here, we consider Type II effects
- Measures contribution added by each factor $v$ when we control for all the others.
- The $F$ statistic is

$$ F_{v} = \frac{(SS_{Bv} - SS_{Bc})/p}{SS_{Wc}/(n - m - 1)} $$

where $SS_{Bv}$ and $SS_{Wc}$ are the explained and residual SS of the full model. $SS_{Bc}$ is the explained part of the model after deletion of $v$, and $p$ the number of indicator or contrasts used for coding variable $v$.

- Significance is again evaluated through permutation tests.
Multi-factor, derivation of the results

- Consider the linear model $Y = XB$
- Its ‘Hat’ matrix $H = X(X'X)^{-1}X'$, such that $\hat{Y} = HY$
- Gower’s matrix $G = -\frac{1}{2} \left(I - \frac{1}{n}11'\right) D \left(I - \frac{1}{n}11'\right)$
  
  with $D$ matrix of squared Euclidean distances
- We have:
  
  $SS_T = \text{tr}(G)$
  $SS_B = \text{tr}(HG)$
  $SS_W = \text{tr}((I - H)G)$

- Generalization by substituting $D$ with the matrix of dissimilarities.

Multi-factor analysis, Occupational trajectories

<table>
<thead>
<tr>
<th>Variable</th>
<th>PseudoF</th>
<th>PseudoR2</th>
<th>p_value</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>497.039</td>
<td>0.226</td>
<td>0.0000</td>
</tr>
<tr>
<td>cohort3b</td>
<td>5.281</td>
<td>0.005</td>
<td>0.0010</td>
</tr>
<tr>
<td>edu_lev</td>
<td>34.353</td>
<td>0.047</td>
<td>0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>116.800</td>
<td>0.319</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Evolution of the differences

- How do differences vary over time?
- At which age do trajectories most differ between cohorts?
- Compute $R^2$ on small sliding windows (length 2)
- \(\Rightarrow\) Series of $R^2$, and we plot their evolution
- Likewise we can plot the series of
  - the total residual discrepancies ($SS_W$)
  - the residual discrepancy of each group ($SS_C$)
**Tree structured analysis of sequence data**

- **Aim:** Find out most important predictors and their interactions.
- Iteratively segment cases using covariate values
- Form as homogeneous groups.
- at each step, select covariate and split that generates the highest $R^2$.
- Split significance tested with permutation $F$.
- Stop when the selected split is not significant.
Analysing the diversity of categorical sequences
Supervised Classification
Regression trees for categorical sequences

Rendering the tree
Occupational trajectories

ANOVA for the leaves of the tree,

<table>
<thead>
<tr>
<th></th>
<th>t0</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pseudo F</td>
<td>62.49</td>
<td>0.000</td>
</tr>
<tr>
<td>Pseudo Fbf</td>
<td>55.66</td>
<td>0.000</td>
</tr>
<tr>
<td>Pseudo R2</td>
<td>0.17</td>
<td>0.000</td>
</tr>
<tr>
<td>Bartlett</td>
<td>60.60</td>
<td>0.000</td>
</tr>
<tr>
<td>Levene</td>
<td>43.66</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Rendering the tree
Cohabitational trajectories
ANALYSING THE DIVERSITY OF CATEGORICAL SEQUENCES

CONCLUSION 1: ABOUT SEQUENCE ANALYSIS

- Analyse trajectories until 45 years => ignore recent generations
- Most recent birth year is 1957 (2002 - 45)
- Issues:
  - Granularity: year, month, day, ...
  - State definition: should we distinguish {separated, divorced, widowed} or consider a single state? works by Raffaella Piccaretta

CONCLUSION 2: MISSING DATA AND WEIGHTS

- Missing data in sequences
- TraMineR allows for differentiated handling of left, right and in-between missing values
  - consider ‘missing’ as a specific state
  - drop out (left shift of subsequent elements)
  - impute, but how?
- Weighting cases
  - Account for them in rendering of sequences (weighted transversal characteristics)
  - Implemented solutions for ANOVA and permutation test
  - Not relevant for dissimilarities and longitudinal characteristics

ANOVA FOR LEAVES OF THE TREE,

<table>
<thead>
<tr>
<th></th>
<th>t0</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pseudo F</td>
<td>5.74</td>
<td>0.000</td>
</tr>
<tr>
<td>Pseudo Fbf</td>
<td>5.62</td>
<td>0.000</td>
</tr>
<tr>
<td>Pseudo R2</td>
<td>0.01</td>
<td>0.000</td>
</tr>
<tr>
<td>Bartlett</td>
<td>0.96</td>
<td>0.049</td>
</tr>
<tr>
<td>Levene</td>
<td>2.73</td>
<td>0.041</td>
</tr>
</tbody>
</table>
Conclusion 3: Extending the analysis

- Since TraMineR is an R library, its outcome can easily be combined in a same script with any other R process
- We have seen: cluster analysis, MDS, ...
- In Widmer and Ritschard (2009).
  - Relationship between occupational and cohabitational trajectories by regressing longitudinal entropies of each of them on occupational and cohabitational types while controlling for birth cohort and sex.
  - Studied also cluster membership with logistic regressions.

Conclusion 4: Application to other kind of data

- Discrepancy based analysis
- ... applies to any data that can be characterized by their pairwise dissimilarities.
- Only aspect specific to state sequences: their visual rendering.

Conclusion 4: About TraMineR

- TraMineR is a unique toolbox for discrete sequence analysis
- Can do much more than shown in this presentation,
  - handling of sequence data
  - conversion between states and events
  - multi-channel dissimilarity for parallel sequences
  - frequent and discriminant sub-sequences
  - extracting association rules between sub-sequences
  - ...
- ... and, like R, available for free on the CRAN
  http://cran.r-project.org
- See also package’s web site
  http://mephisto.unige.ch/traminer

Thank You!
References I


References II


