Discrepancy Analysis of Complex Objects Using Dissimilarities

Matthias Studer, Gilbert Ritschard, Alexis Gabadinho, and Nicolas S. Müller

Abstract. In this article we consider objects for which we have a matrix of dissimilarities and we are interested in their links with covariates. We focus on state sequences for which pairwise dissimilarities are given for instance by edit distances. The methods discussed apply however to any kind of objects and measures of dissimilarities. We start with a generalization of the analysis of variance (ANOVA) to assess the link of complex objects (e.g. sequences) with a given categorical variable. The trick is to show that discrepancy among objects can be derived from the sole pairwise dissimilarities, which permits then to identify factors that most reduce this discrepancy. We present a general statistical test and introduce an original way of rendering the results for state sequences. We then generalize the method to the case with more than one factor and discuss its advantages and limitations especially regarding interpretation. Finally, we introduce a new tree method for analyzing discrepancy of complex objects that exploits the former test as splitting criterion. We demonstrate the scope of the methods presented through a study of the factors that most discriminate Swiss occupational trajectories. All methods presented are freely accessible in our TraMineR package for the R statistical environment.

Keywords: Distance, Dissimilarities, Analysis of Variance, Decision Tree, Tree Structured ANOVA, State Sequence, Optimal Matching.

1 Introduction

The analysis of dissimilarities is used in a wide range of areas. It includes biology with the analysis of genes and proteins (sequence alignment), ecology with the comparison of ecosystems, sociology, network analysis where similarity is a central

Matthias Studer · Gilbert Ritschard · Alexis Gabadinho · Nicolas S. Müller
Department of Econometrics and Laboratory of Demography, University of Geneva, Switzerland
e-mail: `matthias.studer@unige.ch`

F. Guillet et al. (Eds.): Advances in Knowledge Discovery and Management, SCI 292, pp. 3–19.
springerlink.com

notion or the automatic analysis of texts to name just a few. When analyzed objects are not directly measurable or complex, such as sequences or ecosystems for instance, it may be convenient to think in terms of dissimilarities between objects. Having such dissimilarities, it is customary to perform a cluster analysis to get a reduced number of groups for facilitating interpretation. Once the groups are identified, it is common practice to measure the relationship between these objects and other variables of interest by using, for instance, association test or logistic regression.

However, by focusing on clusters we loose indeed information, which may lead to unfair conclusions, particularly for borderline objects. Similarly, it is possible that some associations become less significant through this reduction of information. The latter is not controlled and grouping choices, usually made on statistical ground, may hide others alternatives that might show more interesting associations with some explanatory factors.

In this article we present a set of methods to analyze dissimilarities directly, i.e. without any prior clustering. They will allow us to measure the relationship between, on the first hand, one or more covariates and, secondly, objects described using dissimilarities. We begin by studying the link with a single variable building on the test introduced by Anderson (2001). We extend then the analysis by introducing a new test of the homogeneity of object discrepancy and propose, for the case of state sequences, a new way to display the results. As a second step, we present the method from McArdle and Anderson (2001) which enables us to include several variables at the same time. Finally, we introduce a method based on induction trees that leads to a better interpretation of the results. The method is similar to the one presented in Geurts *et al.* (2006) but is more general since it is not limited to distances that can be expressed as kernels. The criteria is also similar to the one used by Piccarreta and Billari (2007) in an unsupervised setting. Finally, we give a short overview on how to perform the presented methods in R by means of TraMineR. The scope of the discussed methods is illustrated throughout the article by applying them on occupational trajectory data.

2 The Illustrative Data Set

Let us start with a short application issue that will serve as illustration throughout this article. We consider the study of occupational trajectories and expose the problematic so that examples and their interpretations will be clearer for the reader. We are interested in the construction of professional trajectories and factors that may influence it. We focus on the study of working rates following the work of Levy *et al.* (2006). We know that, while men's trajectories are relatively homogeneous and exhibit three main phases, namely “education”, “full time work” and “retirement”, those of women are much more varied. Thus, their average curve of working rates has a camel shape with a decrease in working rate when children are very young and a recovery thereafter. This average curve results however from very distinct trajectories. Some women stop working completely or reduce their working rates and then some of them return to work while others do not. In addition, some women go back and forth between work and at home activity.

Besides the effects of sex on the trajectories, we are interested in testing the differences in trajectories between generations (2 categories), family types — number of children (4 cat.) and marital status (4 cat.) — and socio-economic situations — father social status (10 cat.), income (4 cat.) and education (3 cat.). We are also interested to test whether trajectories of younger generations are significantly more diverse than those of older ones, and thus show a pluralization of trajectories.

To answer these questions, we use the data from the biographical retrospective survey conducted by the Swiss Household Panel[1] in 2002. We know, for each individual and every year, his occupational situation distinguishing between the following states: full time work, part-time work, negative break (eg., unemployment), positive break (eg., travel), at home and training. We focus on the period between ages 25 and 40 which is the key period regarding professional career deployment. We retain all cases without missing data, that is 1560 trajectories. Since all retained individuals are aged 40 at the survey time they are all born before 1962.

3 Measuring Association Using Dissimilarities

We now present a method based on the ANOVA principle to evaluate the association between, on the one hand, objects characterized by a matrix of dissimilarities and, secondly, a categorical variable. We take as a starting point the method introduced by Anderson (2001) for analyzing ecosystems. We retain the more geometric approach of Batagelj (1988) in its generalization of the Ward criterion. Finally, we apply these methods on our example.

3.1 General Principles

Following the ANOVA principles, we seek to determine the part of the variance that is “explained” by a given partition. The ANOVA is based on the notion of “sum of squares” that is the sum of the squared Euclidian distances between each value and the mean. This sum of squares, or inertia, can also be expressed as the average of the pairwise squared Euclidian distances ($d^2_{e,ij}$). These relationships are formalized by Eq. (1).

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

[1] http://www.swisspanel.ch

The concept of sum of squares can be generalized to other dissimilarity measures in two alternative ways. Anderson (2001) proposes to replace the Euclidian distance $d_{e,ij}$ in Eq. (1) with any possibly non-Euclidian measure of dissimilarity d_{ij} yielding:

$$SS^{**} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}^2 \tag{2}$$

However, we prefer to substitute the non-Euclidean dissimilarity d_{ij} for the squared Euclidean distance $d_{e,ij}^2$ rather than for the distance itself as proposed by Batagelj (1988). We argue shortly for this choice in Sec. 3.2 below. The retained generalization of SS reads thus:

$$SS^{*} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij} \tag{3}$$

We use this expression for measuring the discrepancy of our complex objects. Indeed, using $SS = SS^*$ in the definition $s^2 = \frac{1}{n}SS$ of the sample variance we get a fairly intuitive measure of the object discrepancy. Since the variance is theoretically defined for Euclidean distances, we prefer the term "discrepancy" for this more general setting. Interestingly, the discrepancy s^2 is equal to half the average pairwise dissimilarity, that is:

$$s^2 = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} \tag{4}$$

When generalizing the notion of sum of squares to non-Euclidean measures of dissimilarity, the Huygens theorem, Eq. (5), that states that the total sum of squares (SS_T) is the between sum of squares (SS_B) plus the residual within sum of squares (SS_W) remains valid (Batagelj, 1988).

$$SS_T = SS_B + SS_W \tag{5}$$

We can thus apply the analysis of variance (ANOVA) machinery to our complex objects.

The terms in Eq. (5) can all be derived from formula (3). The total sum of squares (SS_T) and the within sum of squares (SS_W) are computed directly with formula (3), SS_W being simply the sum of the within sums of squares of each subgroup. The between sum of squares SS_B is then obtained by taking the difference between the SS_T and SS_W. Using Eq. (5) we can assess the share of discrepancy explained by a categorical or discretized continuous variable. In the spirit of ANOVA, this reduction of discrepancy is due to a difference in the positioning of the gravity centers (or centroids) of the classes. This interpretation holds for any kind of distance even though the concept of class center is not clearly defined for complex non numeric objects (Batagelj, 1988). It is likely that the gravity centers will not belong to the object space, exactly as the mean of integer values may be a real non integer value. Hence, conceptually, we look for the part of the discrepancy that is explained by differences in group positioning and we measure this part with the R^2 formula (6). Alternatively, we may consider the F that compares the explained discrepancy to the residual discrepancy. The F formula is given in Eq. (7), where n is the number of cases and m the number of parameters.

$$R^2 = \frac{SS_B}{SS_T} \tag{6}$$

$$F = \frac{SS_B/(m-1)}{SS_W/(n-m)} \tag{7}$$

The statistical significance of the association, i.e. of the explained part of discrepancy cannot be assessed with the F test as in classical ANOVA. Indeed, the F statistic (7) does not follow a Fisher distribution with our complex objects for which the normality assumption is hardly defendable. We consider therefore a permutation test (Anderson, 2001; Moore *et al.*, 2003). This test works as follows. At each step we change the complex object assigned to each case by means of a randomly chosen permutation, which is equivalent to jointly permute the content of the rows and columns of the distance matrix. We thus get a F_{perm} value for each permutation. Repeating this operation p times we end up with an empirical non parametric distribution of F that characterizes its distribution under independence, i.e. assuming the objects are assigned to the cases independently of their profile in terms of explanatory factors. From this distribution, we can assess the significance of the observed F_{obs} statistic by evaluating the proportion of F_{perm} that are higher than F_{obs}. It is generally admitted that 5000 permutations are necessary to assess a significance threshold of 1% and 1000 for a threshold of 5%.

3.2 Generalization Conditions

As mentioned above, we can generalize Eq. (1) either by substituting the dissimilarity d for the Euclidean distance d_e or for its square d_e^2. In this subsection, we justify our preference for the latter solution, i.e. equation (3). Firstly, in the Euclidian case, the second equality in Eq. (1) which links the sum of deviations to the mean to the sum of pairwise differences follows from properties of signed deviances and pairwise differences which do not hold for unsigned distances. Secondly, with this choice, the non negativity of the contribution of any object to the total discrepancy automatically results when the dissimilarity satisfies the triangle inequality.

In the Euclidian case, the equality (1) can be established by showing first the following result (Späth, 1975):

$$\sum_{i=1}^{n}(y_i - x)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - x)^2 \tag{8}$$

Indeed, we have:

$$y_i - x = (y_i - \bar{y}) + (\bar{y} - x) \tag{9}$$

$$(y_i - x)^2 = (y_i - \bar{y})^2 + 2(y_i - \bar{y})(\bar{y} - x) + (\bar{y} - x)^2$$

$$\sum_{i=1}^{n}(y_i - x)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - x)^2 + 2\sum_{i=1}^{n}(y_i - \bar{y})(\bar{y} - x) \tag{10}$$

Since, $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$, the last term in (10) vanishes which yields Eq. (8). The equality (1) results then by setting $x = y_j$ in (8) and summing over $j = 1, \cdots, n$.

Clearly, equality (9) does not hold if we replace differences $y_i - x$, $y_i - \bar{y}$ and $\bar{y} - x$ with non negative dissimilarities. Likewise, the last term in (10) would not vanish with non negative dissimilarities. Using the second solution (3), we do not have to care about the deviation between objects. We just postulate that there exists a signed deviation measure in the object space.

We now turn to our second argument regarding the contribution of an object x to the total discrepancy. This contribution $d_{x\tilde{g}}$ can be seen as the dissimilarity between x and its (possibly virtual) gravity center $\tilde{g}$. Using the same scheme (3) of generalization, it can be obtained by substituting $d_{x\tilde{g}}$ to $(\bar{y} - x)^2$ in Eq. (8) and by isolating this term, which yields (Batagelj, 1988):

$$\begin{aligned} d_{x\tilde{g}} &= \frac{1}{n}\big(\sum_{i=1}^{n} d_{xi} - SS\big) \\ &= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\big(2 \cdot d_{ix} - d_{ij}\big) \end{aligned} \tag{11}$$

This contribution to the discrepancy is non negative when the dissimilarity measure respects the triangle inequality. Indeed, according to Eq. (11), $d_{x\tilde{g}}$ is minimal when each d_{ij} is maximal. Under the triangle inequality d_{ij} cannot exceed $d_{xi} + d_{xj}$ and hence, $d_{x\tilde{g}}$ reaches its minimum when $d_{ij} = d_{xi} + d_{xj}$ for all i and j. This minimum is zero which implies $d_{x\tilde{g}} \geq 0$. The non negativity of the contribution of x cannot be deduced from the triangle inequality property of the dissimilarity if we use definition (2) of SS, i.e. if we replace the squared Euclidean distance with the squared similarity.

With the retained approach, negative contributions to the discrepancy can occur with semi-metric dissimilarities, that is when the triangle inequality does not hold. The "dissimilarity" $d_{x\tilde{g}}$ becomes negative when adding x reduces the discrepancy between the other objects. This can be the case when the distance between two objects, say y and z, becomes shorter when we can pass through x, i.e. when $d_{yz} > d_{yx} + d_{xz}$. Such situation is quite usual in social network analysis. For instance, let us consider a social network between x, y and z where the dissimilarity is equal to 1 for two people that meet often and is equal to 10 when they never meet. The dissimilarity $d_{x\tilde{g}}$ would then be negative if x often meets y and z while y never meets z. From a social network perspective, we would say that x plays a cohesive role in the network.

Though a negative contribution to the discrepancy makes sense for social networks, it is not the case for most applications. Hence, the results should be interpreted with caution when the dissimilarity measure is only semi-metric. In particular, one should be ready to admit and give sense to negative contributions to the discrepancy.

3.3 Application

We now illustrate the proposed test on our example data about the study of occupational trajectories. We use optimal matching (OM) for measuring the dissimilarities between trajectories that are indeed represented as state sequences. The OM dissimilarity, also known as the edit distance, is the minimal cost of transforming one sequence into the other using two types of transformation operations, namely indel (insert or delete) and substitution of elements. The transformation cost is determined by assigning indel and substitution costs. For our example, we computed the OM distances with an indel cost set to 1 and substitution costs at 2. Notice that the OM dissimilarity respects the triangle inequality. Indeed, dissimilarity being the minimal cost for transforming a sequence y into z, we necessarily have $d_{yz} \leq d_{yx} + d_{xz}$.

The discrepancy of the occupational trajectories of the whole data set is 0.501 which is equal to half of the average edit distance (1.02). It is 0.118 for men and 0.614 for women indicating that women's trajectories exhibit wider variety.

Table 1 summarizes the results of the discrepancy analysis for the whole population as well as for men and women separately. In each case we considered individually each of the available predictive factors. The p-values of the tests are based on 1000 permutations.

Table 1 Association test with occupational trajectories

	Total			Men			Women		
Variable	F	R^2	Sig	F	R^2	Sig	F	R^2	Sig
Sex	477.995	0.235	**0.000**						
Father soc. status	1.578	0.009	**0.029**	2.085	0.026	**0.005**	1.205	0.013	0.163
Income	1.349	0.003	0.182	3.086	0.013	**0.006**	3.553	0.013	**0.000**
Education	18.486	0.023	**0.000**	20.632	0.054	**0.000**	6.287	0.015	**0.000**
Cohort	17.037	0.011	**0.000**	6.330	0.009	**0.001**	14.911	0.018	**0.000**
Children	13.704	0.026	**0.000**	1.006	0.004	0.391	25.740	0.085	**0.000**
Marital status	9.744	0.018	**0.000**	1.783	0.007	**0.047**	18.078	0.061	**0.000**

Not surprisingly, sex explains the biggest part of the discrepancy of trajectories with a R^2 that reaches 0.235. In other words, the sex variable explains 23.5% of the discrepancy. The relationship is statistically significant since the $F_{obs} = 477.995$ was never attained amongst the thousand permutations. As for the other covariates, results show that the Father's social status and Education impact primarily male trajectories while women's trajectories are more strongly influenced by familial factors such as the number of children and the marital status. than female trajectories.

In summary, these first results show that the occupational trajectory is significantly influenced by most of the considered predictive variables. From the high significance of the significance tests, differences in the positioning of the gravity centers of groups of sequences clearly exist. Nevertheless, it is difficult to understand and interpret these differences at this stage.

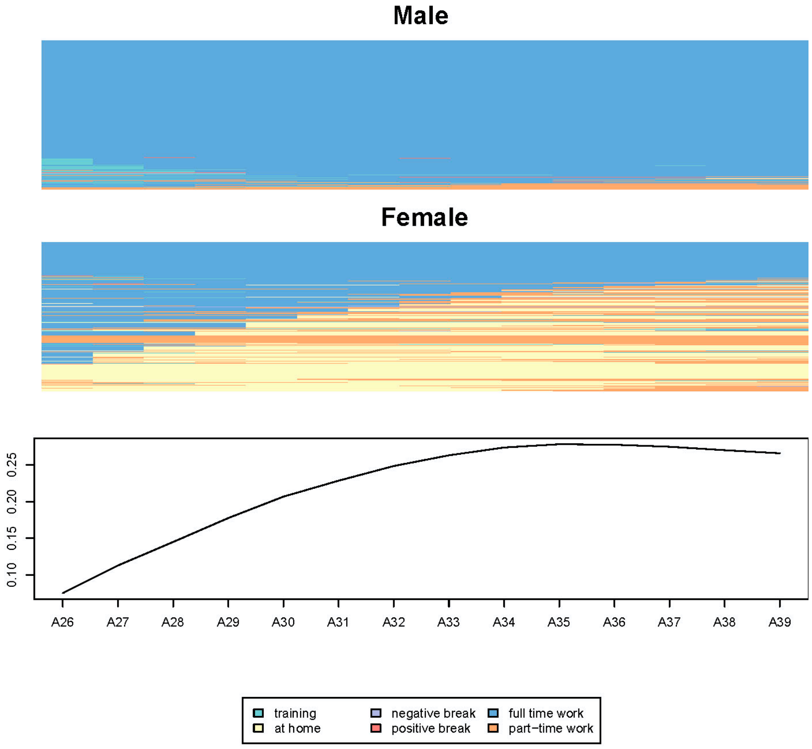

Fig. 1 Differences of trajectories according to sex

Figure 1, which presents a new way of displaying the differences between groups of sequences, should help interpretation. The first two charts show men and women trajectories using index-plots (Scherer, 2001). In these figures, each sequence is represented by a time line split into segments colored according to the corresponding occupational state.

To improve readability of the index-plots, we ordered the sequences according to the first dimension of a PCA (Principal Coordinate Analysis) (Gower, 1966). If ordering sequences by an underlying dimension facilitates the interpretation of the index-plot, the plots provide conversely useful information for interpreting the PCA axis. For instance, we observe in our case that the sequences are organized in a continuum ranging from full-time trajectories to trajectories where we stay at home during the whole sequence. The axis can thus be read as a Full-time - At home axis.

The final chart exhibits the evolution of the strength of association between the categorical covariate and a sliding two period long sub-sequence of the trajectory. For each unit of time, we extracted a sub-sequence of two consecutive states for which we calculated the distance matrix and the share of discrepancy explained by the covariate. This representation helps at identifying the periods over which the sequences are most differentiated by gender. It appears that gender differences reach their peak around 35 years old.

4 Homogeneity of Discrepancy

In some situations, it may be of interest to test whether the discrepancies within the groups differ significantly. From a geometric point of view, we are interested in measuring differences in the diameter of the distribution of sequences within each group. In classical analysis of variance, we could use a Bartlett's test (Snedecor and Cochran, 1989) that supposes equal variances under H_0 or, in other words, the homogeneity of variances. This test is based on the statistical distribution of the statistic T defined by Eq. (12), where s_i^2 stands for the discrepancy within group i. All terms in this equation can be calculated with the formulas already introduced. As for the F, it is not possible in our non-Gaussian case to assume that this statistic T has a known distribution. We use therefore again permutation tests to assess the significance of differences in discrepancy.

$$T = \frac{(n-m)\ln\left(\sum_{i=1}^{m}\frac{(n_i-1)}{(n-m)}s_i^2\right) - \sum_{i=1}^{m}(n_i-1)\ln(s_i^2)}{1+\frac{1}{3(m-1)}\left[\sum_{i=1}^{m}\frac{1}{n_i-1} - \frac{1}{n-m}\right]} \qquad (12)$$

In the previous section, we found that men's discrepancy is 0.118 against 0.614 for women. This relatively high difference is confirmed by the T_{obs} which is 460.017, a value that was attained by none of the thousand permutations. This allows us to state that the discrepancies differ significantly with the sex of the respondent. More interestingly from a sociological point of view, the discrepancy of the people born after 1945 is significantly higher than those born earlier. We thus have clear evidence that the diversity of occupational trajectories increased for younger generations.

5 Multi-factor Discrepancy Analysis

In Sec. 3.3 we examined the bivariate association between the trajectory and each of the covariates considered independently. We consider here the generalization to the multi-factor case and adopt for that the framework of the general multivariate analysis of variance. Several authors have considered such analyses from pairwise distances (Excoffier *et al.*, 1992; Gower and Krzanowski, 1999; Anderson, 2001; Zapala and Schork, 2006). We adopt the approach and formalism of McArdle and Anderson (2001) who conducted a multi-factor analysis of ecosystems on the bases the pairwise semi-metric distance of Bray-Curtis. However, as for the simple discrepancy analysis and unlike McArdle and Anderson (2001) we substitute the pairwise dissimilarity measure for the squared Euclidean distance rather than for the distance itself.

Formally, we consider the multivariate regression model: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where $\mathbf{Y}$ is the $n \times t$ matrix with n observed values of t response variables and $\mathbf{X}$ the $n \times m$ matrix with the values of m predictors including a first column of ones corresponding to the constant.

In the Euclidean case, the sum over the t response variables of their sums of squares can be derived by means of the same Gower matrix as that used in PCA (Gower, 1966). Similarly to McArdle and Anderson (2001), we generalize this analysis to any type of dissimilarities. Let $\mathbf{1}$ be a vector of ones of length n, $\mathbf{I}$ the identity matrix and $\mathbf{A}$ a matrix with generic element $a_{ij} = -\frac{1}{2}d_{ij}$, where d_{ij} is the dissimilarity between cases i and j, which we substitute for the squared Euclidean distance in the original Gower's formulation. The Gower matrix reads as follows

$$\mathbf{G} = \Big(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\Big)\mathbf{A}\Big(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\Big) \tag{13}$$

with in our case a matrix $\mathbf{A}$ that results from the available pairwise dissimilarities. The total sum of squares SS_T is equal to the trace of $\mathbf{G}$. McArdle and Anderson (2001) show that the explained sum of squares SS_B and the residual sum of squares SS_W can be written as

$$SS_B = tr\big(\mathbf{HGH}\big) \tag{14}$$

$$SS_W = tr\big[\big(\mathbf{I} - \mathbf{H}\big)\mathbf{G}\big(\mathbf{I} - \mathbf{H}\big)\big] \tag{15}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the idempotent matrix usually known as "hat" matrix in linear regression. Using these two quantities we can derive a global pseudo-R^2 and a global pseudo-F statistic by applying Eqs. (6) and (7). Formula (14) and (15), however, allow us to account of any number of covariates and specifically of categorical factors through their contrast or indicator coding.

As in the single discrepancy analysis, the F distribution is not relevant for the pseudo-F and we consider again permutations tests for assessing its significance.

We may also consider the contribution of each covariate to the total discrepancy reduction. As with multi-factor ANOVA there are different ways of looking at these individual contributions. Shaw and Mitchell-Olds (1993) distinguish for instance a Type I and a Type II method. Type I is incremental. Covariates are successively added to the model and the contribution of each covariate is measured by the SS_B increase that results when it is introduced. With this method the measured impact of each factor depends on the order in which they are introduced. With Type II, known to be robust in the absence of interaction effects, the contribution of each covariate is measured by the reduction of SS_B that occurs when we drop it out from the full model, i.e. from the model with all covariates. We retain this second method and hence compute the following F for each covariate v

$$F_v = \frac{(SS_{B_c} - SS_{B_v})/p}{SS_{W_c}/(n-m-1)} \tag{16}$$

where the SS_{B_c} and SS_{W_c} are the explained and residual sums of squares of the full model, SS_{B_v} the explained sum of squares of the model after removing variable v, and p the number of indicators or contrasts used to encode the covariate v.

Let us look at what this gives for our illustrative example. Table 2 shows the results for two models, the complete model with all variables and a model obtained after removing non significant covariates through a backward stepwise process.

Table 2 Multi-Factor Discrepancy Analysis

	Full Model			Backward Model		
Variable	F_v	ΔR^2_v	Sig	F_v	ΔR^2_v	Sig
Sex	477.196	0.218	0.000	488.627	0.224	0.000
Education	8.230	0.008	0.000	10.986	0.010	0.000
Income	0.868	0.001	0.542			
Father's soc. status	1.167	0.005	0.241			
Cohort	11.586	0.005	0.000	13.670	0.006	0.000
Children	9.887	0.014	0.000	10.313	0.014	0.000
Marital status	4.621	0.006	0.000	5.073	0.007	0.000
	F_{tot}	R^2_{tot}	Sig	F_{tot}	R^2_{tot}	Sig
Global	29.557	0.297	0.000	63.602	0.291	0.000

From the global statistics, the set of covariates provide overall significant information about the diversity of occupational trajectories.

In the full model, the sex remains the most significant covariate. If we remove this variable, the R^2 of the model ($= 0.297$) decreases by 0.218. This difference is significant since we have $F_{sex} = 477.196$, a value never attained with a thousand permutations. On the contrary, the income is for instance not significant. Removing it from the model reduces the R^2 by only 0.001 and results in a F_{income} value of 0.868, which was exceeded for $0.542 \cdot 1000 = 542$ of the thousand permutations. Likewise, the father's social status loses its significance in the multi-factor case. Indeed, it becomes non-significant as soon as we control for the education level, these two variables being strongly correlated and education being more significant.

The multi-factor approach provides information about the proper effect of the covariates on the occupational trajectory, that is the part of the its total effect that is not accounted for by already introduced factors. It is in that sense complementary to the single univariate discrepancy analysis that informs on the raw effect of each covariate. Nevertheless, while the method permits us to know which effects are significant, it does not tell us much about what the effects are, i.e. about how occupational trajectories may change with the value of the covariates. We propose for that a tree approach which can be seen as an extension of the graphical display shown in Fig. 1.

6 Tree Structured Analysis

This section introduces a new method based on the principle of induction trees for analyzing the discrepancy of objects described by a dissimilarity matrix. Induction trees work as follows (Breiman *et al.*, 1984; Kass, 1980). They start with all individuals grouped in an initial node. Then, they recursively partition each node using values of a predictor. At each node, the predictor and the split are chosen in such a way that the resulting child nodes differ as much as possible from one another or have, more or less equivalently, lowest within discrepancy. The process is repeated on each new node until some stopping criterion is reached.

Recursive partitioning is known to provide an easily comprehensible view of how each newly selected covariate nuances the effect of covariates introduced at earlier levels. This requires indeed to display suitable information about the distribution in each node. We could represent the centrotype, i.e. the observed object that minimizes the dissimilarity (11) with the group gravity center. It would be more instructive to also render the within group discrepancy. Though this is not obvious for any kind of complex objects, displaying index-plots as those used in Fig. 1 provides a good solution for state sequences.

Beside the displayed node content, the originality of our approach resides in the use of a splitting criterion derived from the pairwise dissimilarities, namely the univariate pseudo-R^2 that we described in Sec. 3. We select thus at each node the predictor and binary split for which get the highest pseudo-R^2, i.e. the split that accounts for the greatest part of the object discrepancy. An alternative would be to use the significance of the univariate pseudo-F. However, since this significance must be determined through permutation tests we would end-up with an excessive time complexity if we had to repeat it for each predictor and possible split. We consider therefore the F significance only as a stopping criteria, i.e. we stop growing a branch when we get a non-significant F for the selected split. This requires to run the permutations only once at each node, which remains tractable.

Using the pseudo-R^2 as splitting criterion condemns us to build binary trees. Indeed, the R^2 does not penalize for the number of groups and would hence always select the maximal number of groups if we allowed n-ary splits. The R^2 adjusted for the number of groups as it is used in multiple regression would not be a satisfactory solution since it is known to insufficiently penalize complexity. On the other hand, information criteria such as the BIC seem hardly derivable in our setting where we do not know the distribution of our statistics (R^2, F or SS_W) under the independence hypothesis.

It is worth mentioning that our tree building procedure resembles that proposed in Geurts *et al.* (2006). However, our formulation is more general since it works with any kind of metric and non metric dissimilarities, while Geurts *et al.* (2006)'s solution is restricted to dissimilarities that can be derived through the kernel trick. For growing a tree from semi-metric dissimilarities we should indeed be ready to accept and give sense to possible negative contributions to the variance.

Before looking at the example, let us add a few words about computational aspects. First, we can highlight that it is not necessary to recompute SS_W from scratch for each possible binary split that can be derived from a same predictor. Our algorithm makes use of partial results first collected into a symmetric $m \times m$ matrix $\mathbf{E}$, where m is the number of different observed values of the predictor. Each element $e_{k\ell}$ of $\mathbf{E}$ is defined as $e_{k\ell} = \sum_{i\in k}\sum_{j\in\ell} d_{ij}$, that is as the sum of dissimilarities between on the one hand, cases that take the k-th value of the predictor and, on the other hand, those that take the ℓ-th value. The residual sum of squares for a group of values G is then equal to $SS_{G,res} = \frac{1}{n_G}(\sum_{k\in G}\sum_{\ell\geq k,\ell\in G} e_{k\ell})$. Reusing this way the same partial sums of dissimilarities may save a great amount of computation time especially for categorical predictors with few different values.

Secondly, we may exploit the fact that the R^2 can only decrease when merging categories. From matrix $\mathbf{E}$ we can compute the R^2_{ori} that measures the part of discrepancy explained by the predictor in its original form, i.e. with all its distinct values. It then follows that this R^2_{ori} is an upper bound for the best R^2 that would result from a binary split based on the considered predictor. Hence, when the R^2_{ori} of the current predictor does not exceed the R^2 of the previously found best split, it becomes unnecessary to test the splits for the current predictor.

The global quality of the tree can be assessed through the association strength between the objects and the leaf (terminal node) membership. The global multi-factor pseudo-F gives us a way of testing the statistical significance of the obtained segmentation and the global pseudo-R^2 the part of the total discrepancy that is explained by the tree.

Figure 2 shows the dissimilarity tree grown for our example of occupational trajectories. The used stopping criteria are a p-value of 1% for the F test, a minimal leaf size of 100 and a maximal depth of 5. In each node we see the plot of the individual sequences as well as the node size and the discrepancy within the node (var). At the bottom of each parent node we indicate the retained split predictor with the associated R^2 while the definition of the binary split may be inferred from the indication at the top of the child nodes.

The overall tree R^2 is 0.302, which is higher than for the models in Table 2. The tree has thus a better explanatory power. We get this higher value by retaining only 4 predictors against 5 for the backward model. This may be explained by interaction effects that the tree automatically accounts for and that were not considered in the multi-factor discrepancy analysis. We thus can point out here that birth cohort and number of children interact in their effect on female occupational trajectories while birth cohort interacts with education in their effect on men trajectories. This automatic detection of interaction is indeed a fundamental property of all induction trees.

By looking at the displayed individual sequences, we are now able to gain knowledge about what the effect of the predictors are. Clearly, men are characterized by

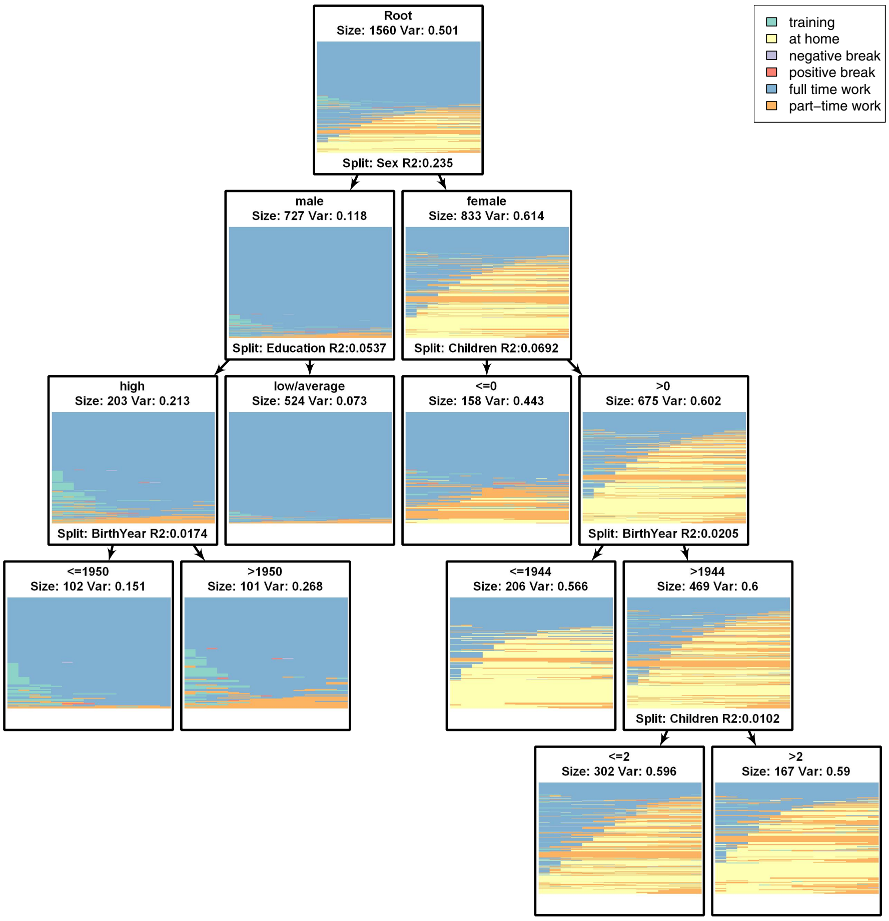

Fig. 2 Regression tree based on pairwise dissimilarities between sequences

full time trajectories while part time and at home are typically found in women's trajectories. Among men, the choice of part-time seems to be related with higher education. For women, occupational trajectories are more diversified. Those who had at least one child have higher chance to experience part time work when they were born after 1945. This birth cohort effect is, however, less pronounced among those women who had more than two children.

7 Discrepancy Analysis in R with TraMineR

The methods presented in this article are all implemented in TraMineR (Gabadinho *et al.*, 2009) our free package for the R statistical environment (R Development Core Team, 2008). We shortly show here how simple it is to use them. Assume that we have the following R objects defined in our environment: *dm* a matrix of dissimilarities between cases, *mydata* a *data.frame* with the covariates and *mysequences* an object containing the state sequences.

Univariate discrepancy analysis and test for homogeneity of discrepancy is performed by calling the *dissassoc* function. This function takes three arguments: a dissimilarity matrix, a factor and the number of permutation ($R = 1000$ by default). The results presented in Sec. 3 were obtained with the following code:

```
R> dissassoc(dm, group = mydata$sex, R = 1000)
```

Likewise, we generated the bottom part of Fig. 1 by means of function *seqdiff* with the code below.

```
R> mysequences.diff <- seqdiff(mysequences, group = mydata$sex)
R> plot(mysequences.diff)
```

The multi-factor results given in Table 2 were obtained with the *dissmfac* function. The model is specified with a classical R formula in which the left hand side is the dissimilarity matrix. The *data* argument specifies the *data.frame* containing the covariates.

```
R> dissmfac(
+  dm ~ sex + cohort + education + fathsoc + income + children + marital,
+  data = mydata, R = 1000)
```

Tree structured analysis of dissimilarities is carried out with the *disstree* function. The dissimilarity matrix and the predictors are passed to the function in the same way as in *dissmfac*. Stopping criteria can be set with the following arguments: *minSize* for the minimum node size, *maxdepth* for the maximum tree depth and *pval* for the minimum required p-value. The *R* option permits to control the number of permutations used for computing the significance.

```
R> mytree <- disstree(
+  dm ~ sex + cohort + education + fathsoc + income + children + marital,
+  data = mydata, minSize = 100, maxdepth = 5,  R = 1000, pval = 0.01)
R> print(mytree)
```

The resulting tree can then be plotted by calling the *dot* program of GraphViz[2], which is an open source graph visualization software (Gansner and North, 1999). Assuming GraphViz is on the path, we get a tree similar to that of Fig. 2 but with density plots instead of the index-plots just with the steps below. The plot is generated in file *mytree.dot.svg*.

```
R> seqtree2dot(mytree, filename = "mytree", seqs = mysequences,
+  plottype = "seqdplot")
R> shell("dot -Tsvg -O mytree.dot")
```

[2] http://www.graphviz.org/

8 Conclusion

The aim of this article was to propose tools for investigating how complex objects characterized by their pairwise dissimilarities are related to covariates or predictive attributes. The methods proposed are inspired from the classical ANOVA framework. The basic trick consists in extending results that express the classical sum of squares *SS* in terms of pairwise squared Euclidean distances to the case of any possibly non metric dissimilarity. We designate this general setting as discrepancy analysis. We proposed first a pseudo-R^2 and a pseudo-F test for the univariate case in which each covariate is examined separately. For this same univariate case we discussed also a way of testing the homogeneity of the discrepancies among groups. We then discussed the multi-factor case where we assess the impact of a covariate by controlling for the effect of the other factors. Eventually, we introduced an original tree structured method for discrepancy analysis. For both the univariate and tree structured settings we considered also the question of depicting the effect of the covariates. The difficulty is here to find a suited way of representing the distribution of the objects. We showed that index-plots prove useful when objects are of state sequences. However, more general solutions that could be used for any type of objects would here be necessary and we are presently working on that.

The work presented leaves certainly place to improvements on several aspects. For instance, we plan to further explore alternatives to the R^2 splitting criteria used in dissimilarity trees. We are looking for a way to use p-values of pseudo-F statistics and for a penalized criteria that would permit n-ary splits.

Acknowledgements. This work is part of the Swiss National Science Foundation research project FN-1000015-122230 “Mining Event Histories: Towards New Insights on Personal Swiss Life Courses”.

References

Anderson, M.J.: A new method for non-parametric multivariate analysis of variance. Austral Ecology 26, 32–46 (2001)

Batagelj, V.: Generalized Ward and related clustering problems. In: Bock, H. (ed.) Classification and related methods of data analysis, pp. 67–74. North-Holland, Amsterdam (1988)

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification And Regression Trees. Chapman and Hall, New York (1984)

Excoffier, L., Smouse, P.E., Quattro, J.M.: Analysis of Molecular Variance Inferred from Metric Distances among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. Genetics 131, 479–491 (1992)

Gabadinho, A., Ritschard, G., Studer, M., Müller, N.S.: Mining Sequence Data in R with the TraMineR package: A User’s Guide. Technical report, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva (2009), `http://mephisto.unige.ch/traminer/`

Gansner, E.R., North, S.C.: An Open Graph Visualization System and Its Applications to software engineering. Software - Practice and Experience 30, 1203–1233 (1999)

Geurts, P., Wehenkel, L., d'Alché Buc, F.: Kernelizing the output of tree-based methods. In: Cohen, W.W., Moore, A. (eds.) ICML. ACM International Conference Proceeding Series, vol. 148, pp. 345–352. ACM, New York (2006)

Gower, J.C.: Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. Biometrika 53(3/4), 325–338 (1966), http://www.jstor.org/stable/2333639

Gower, J.C., Krzanowski, W.J.: Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. Journal of the Royal Statistical Society: Series C (Applied Statistics) 48(4), 505–519 (1999)

Kass, G.V.: An exploratory technique for investigating large quantities of categorical data. Applied Statistics 29(2), 119–127 (1980)

Levy, R., Gauthier, J.-A., Widmer, E.: Entre contraintes institutionnelle et domestique : les parcours de vie masculins et féminins en Suisse. Cahiers canadiens de sociologie 31(4), 461–489 (2006)

McArdle, B.H., Anderson, M.J.: Fitting Multivariate Models to Community Data: A Comment on Distance-Based Redundancy Analysis. Ecology 82(1), 290–297 (2001), http://www.jstor.org/stable/2680104

Moore, D.S., McCabe, G., Duckworth, W., Sclove, S.: Bootstrap Methods and Permutation Tests. In: The Practice of Business Statistics: Using Data for Decisions, W. H. Freeman, New York (2003)

Piccarreta, R., Billari, F.C.: Clustering work and family trajectories by using a divisive algorithm. Journal of the Royal Statistical Society A 170(4), 1061–1078 (2007)

R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008) ISBN 3-900051-07-0, http://www.r-project.org

Scherer, S.: Early Career Patterns: A Comparison of Great Britain and West Germany. European Sociological Review 17(2), 119–144 (2001)

Shaw, R.G., Mitchell-Olds, T.: Anova for Unbalanced Data: An Overview. Ecology 74(6), 1638–1645 (1993), http://www.jstor.org/stable/1939922

Snedecor, G.W., Cochran, W.G.: Statistical methods, 8th edn. Iowa State University Press (1989)

Späth, H.: Cluster analyse algorithmen. R. Oldenbourg Verlag, München (1975)

Zapala, M.A., Schork, N.J.: Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proceedings of the National Academy of Sciences of the United States of America 103(51), 19430–19435 (2006)